A primer for postgraduate statistics

STA 501
co-ordinator:
Professor Robert Shell
Contributors: Sandra D. Matanyaire,
Noma Makepela, Rob Shell, Mike Bergh,
Denise Lievesly, Sandra Rowoldt, John C. Whitehead
3rd level, Mathematical Sciences Building,
Room 3.23.
Telephone No: 021-959-3898.
E-mail: rshell@uwc.ac.za; rshell@iafrica.com

This is the classic curve of a platykurtic distribution

THE NORMAL LAW OF ERROR STANDS OUT IN THE EXPERIENCE OF MANKIND AS ONE OF THE BROADEST GENERALIZATIONS OF NATURAL PHILOSOPHY IT SERVES AS THE GUIDING INSTRUMENT IN RESEARCHES IN THE PHYSICAL AND SOCIAL SCIENCES AND IN MEDICINE AGRICULTURE AND ENGINEERING IT IS AN INDISPENSABLE TOOL FOR THE ANALYSIS AND THE INTERPRETATION OF THE BASIC DATA OBTAINED BY OBSERVATION AND EXPERIMENT
# Contents

**Preface** ............................................................................................................. 21

**Sticky mouse syndrome and managing files** ............................................. 23
- Fighting sticky mouse syndrome ................................................................. 23
- Starting to manage your files .................................................................... 23
- The DGT subdirectory system and how to work with files .................... 25

**Quantitative methods, large datasets and the PC** .................................... 27
  *by Robert C.-H. Shell*
  - Topics covered .......................................................................................... 27
  - Intro .......................................................................................................... 27
  - Virtual Reality ........................................................................................ 29
  - Reading .................................................................................................... 31

**An introduction to spreadsheets** ................................................................. 33
  *by Robert C.-H. Shell*
  - Spreadsheet basics ................................................................................. 34
  - The third dimension of the cell .............................................................. 34
  - Cursor control ....................................................................................... 35
  - The program function keys ................................................................... 36
  - The menus ............................................................................................. 37
  - Ranges .................................................................................................... 38
  - The formula ............................................................................................ 38
  - Calculating numeric formulas .............................................................. 38
  - Documenting a formula ........................................................................ 38
  - Using formula markers ......................................................................... 39
  - Printing formulas .................................................................................. 39
<table>
<thead>
<tr>
<th>Topic</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Using @ functions</td>
<td>39</td>
</tr>
<tr>
<td>Statistical functions</td>
<td>40</td>
</tr>
<tr>
<td>Mathematical functions</td>
<td>40</td>
</tr>
<tr>
<td>Engineering functions</td>
<td>41</td>
</tr>
<tr>
<td>Logical functions:</td>
<td>41</td>
</tr>
<tr>
<td>Saving your work</td>
<td>42</td>
</tr>
<tr>
<td>Classwork</td>
<td>43</td>
</tr>
<tr>
<td>Difference between databases and spreadsheets</td>
<td>43</td>
</tr>
<tr>
<td>Fields (variables), records (cases) and values (data)</td>
<td>44</td>
</tr>
<tr>
<td><strong>A spreadsheet exercise</strong></td>
<td>45</td>
</tr>
<tr>
<td><em>by Robert C.-H. Shell</em></td>
<td>45</td>
</tr>
<tr>
<td>Aitken's iterative procedure</td>
<td>46</td>
</tr>
<tr>
<td>Dis-aggregating age groups</td>
<td>50</td>
</tr>
<tr>
<td>Insertion of procedure into a spreadsheet</td>
<td>51</td>
</tr>
<tr>
<td>List of references</td>
<td>52</td>
</tr>
<tr>
<td><strong>Basic concepts in statistics and notation</strong></td>
<td>53</td>
</tr>
<tr>
<td>Terminology</td>
<td>54</td>
</tr>
<tr>
<td><em>Population</em></td>
<td>54</td>
</tr>
<tr>
<td><em>Finite and infinite populations</em></td>
<td>54</td>
</tr>
<tr>
<td><em>Sample</em></td>
<td>54</td>
</tr>
<tr>
<td>Objectives of statistics</td>
<td>54</td>
</tr>
<tr>
<td>Statistics</td>
<td>54</td>
</tr>
<tr>
<td>Variables</td>
<td>55</td>
</tr>
<tr>
<td>Quantitative and qualitative variables</td>
<td>55</td>
</tr>
<tr>
<td><em>Quantitative variable</em></td>
<td>55</td>
</tr>
<tr>
<td><em>Qualitative variables</em></td>
<td>56</td>
</tr>
<tr>
<td>Random variables</td>
<td>59</td>
</tr>
<tr>
<td>Discrete and continuous random variables</td>
<td>59</td>
</tr>
</tbody>
</table>
Continuous ratio variables. ........................................................................................................................................ 59
Parameter and statistic ................................................................................................................................................ 60
Raw data .................................................................................................................................................................. 60
Array ...................................................................................................................................................................... 61
Range ...................................................................................................................................................................... 61
The subscript or index notation ........................................................................................................................................ 61

**Improving the quality of survey data** .................................................................................................................. 69

*by Professor Denise Lievesley* .................................................................................................................................. 69

Introduction to sampling terminology ............................................................................................................................ 69
The basic principles of sampling ....................................................................................................................................... 70
Random sampling .......................................................................................................................................................... 71
  *Simple random sampling* ........................................................................................................................................... 71
Precision and accuracy .................................................................................................................................................. 73
Non-Response is of two types: .......................................................................................................................................... 73
  *Unit non-response* ..................................................................................................................................................... 73
  *Item Non-Response* .................................................................................................................................................. 74
Ways in which surveys are used ......................................................................................................................................... 74
Alternatives to surveys .................................................................................................................................................... 75
Questions which have to be asked before every survey .................................................................................................. 75
Types of survey errors ................................................................................................................................................... 77
Sources of error ................................................................................................................................................................ 77
Errors arising at the data collection stage .......................................................................................................................... 78
  *Questionnaire induced* ............................................................................................................................................... 78
  *Interviewer induced* .................................................................................................................................................. 78
  *Respondent induced* ................................................................................................................................................ 78
  *Will we get clearance to do the study?* .......................................................................................................................... 78
  *Environment induced – or bias of the auspices* ............................................................................................................... 79
Errors arising at the data processing stage ......................................................................................................................... 79
Errors arising at the analysis stage ........................................................................................................................... 79
Fieldwork shortfalls .......................................................................................................................................................... 80
Sample members not contacted ........................................................................................................................................ 80
Sample members unwilling to participate ..................................................................................................................... 80
Sample members unable to participate ........................................................................................................................ 81
Reasons for refusal to an attitude survey .......................................................................................................................... 82
Follow-up interview with refusers (as a %) ............................................................................................................................ 83
General household survey .................................................................................................................................................. 84
Relationship of census characteristics to response rates .................................................................................................. 84
The interviewer’s tasks “on the doorstep” .......................................................................................................................... 85
Informed consent ................................................................................................................................................................. 86
The ideal respondent ............................................................................................................................................................... 87
Qualitative techniques .......................................................................................................................................................... 88
When is qualitative research appropriate? .......................................................................................................................... 88
Depth interviews ................................................................................................................................................................. 89
Group discussions or focus groups ....................................................................................................................................... 89
Advantages ............................................................................................................................................................................. 89
Disadvantages ................................................................................................................................................................. 90
Aspects of the interviewer’s role in qualitative research ....................................................................................................... 90
Additional requirements of the group leader ........................................................................................................................ 90
Barriers to communication .................................................................................................................................................. 91
Qualitative interviewing techniques ...................................................................................................................................... 91
Perceptual techniques ............................................................................................................................................................ 92
The interviewer’s task ............................................................................................................................................................ 92
Recruitment, training and quality control ............................................................................................................................. 92
Training .................................................................................................................................................................................. 93
Quality control ....................................................................................................................................................................... 93
General principles of questionnaire design .......................................................................................................................... 93
Order and format of the questionnaire .................................................................................................................................. 94
<table>
<thead>
<tr>
<th>Topic</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Respondent’s concerns</td>
<td>94</td>
</tr>
<tr>
<td>Interviewer’s concerns</td>
<td>94</td>
</tr>
<tr>
<td>Quality improvements</td>
<td>94</td>
</tr>
<tr>
<td>Types of questions</td>
<td>94</td>
</tr>
<tr>
<td>Issues with factual questions</td>
<td>95</td>
</tr>
<tr>
<td>Problems of generalizing about behaviour</td>
<td>95</td>
</tr>
<tr>
<td>Problems of memory</td>
<td>95</td>
</tr>
<tr>
<td>Dimensions of saliency</td>
<td>96</td>
</tr>
<tr>
<td>Asking about behaviour</td>
<td>97</td>
</tr>
<tr>
<td>Behaviour in a given period</td>
<td>97</td>
</tr>
<tr>
<td>Under-reporting of hospitalizations (2)</td>
<td>98</td>
</tr>
<tr>
<td>Estimated victimization rates</td>
<td>98</td>
</tr>
<tr>
<td>Comparison of publications looked at “yesterday”</td>
<td>99</td>
</tr>
<tr>
<td>Hints on asking threatening questions about behaviour</td>
<td>99</td>
</tr>
<tr>
<td>Short</td>
<td>100</td>
</tr>
<tr>
<td>Long</td>
<td>100</td>
</tr>
<tr>
<td>Instructions, contingent feedback and commitment</td>
<td>100</td>
</tr>
<tr>
<td>Issues:</td>
<td>101</td>
</tr>
<tr>
<td>Asking attitudinal questions</td>
<td>101</td>
</tr>
<tr>
<td>Advantages of open questions</td>
<td>101</td>
</tr>
<tr>
<td>Advantages of closed questions</td>
<td>102</td>
</tr>
<tr>
<td>Most important problem facing the U.S.A.</td>
<td>102</td>
</tr>
<tr>
<td>Use of ‘no opinion’ and ‘don’t know’ categories</td>
<td>103</td>
</tr>
<tr>
<td>Results</td>
<td>103</td>
</tr>
<tr>
<td>Types of order effect</td>
<td>103</td>
</tr>
<tr>
<td>General Question</td>
<td>104</td>
</tr>
<tr>
<td>Specific Question</td>
<td>104</td>
</tr>
<tr>
<td>Mode of data collection</td>
<td>104</td>
</tr>
<tr>
<td>Mail surveys</td>
<td>104</td>
</tr>
<tr>
<td>Topic</td>
<td>Page</td>
</tr>
<tr>
<td>----------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>Advantages</td>
<td>104</td>
</tr>
<tr>
<td>Disadvantages</td>
<td>105</td>
</tr>
<tr>
<td>Advantages or disadvantages?</td>
<td>105</td>
</tr>
<tr>
<td>How can response be improved?</td>
<td>105</td>
</tr>
<tr>
<td>Questionnaire appearance is critical</td>
<td>106</td>
</tr>
<tr>
<td>Can open material be collected in mail surveys?</td>
<td>107</td>
</tr>
<tr>
<td>Telephone surveys</td>
<td>108</td>
</tr>
<tr>
<td>Advantages</td>
<td>108</td>
</tr>
<tr>
<td>With centralization</td>
<td>108</td>
</tr>
<tr>
<td>Disadvantages</td>
<td>108</td>
</tr>
<tr>
<td>coverage bias</td>
<td>109</td>
</tr>
<tr>
<td>Questionnaire design</td>
<td>109</td>
</tr>
<tr>
<td>CATI</td>
<td>109</td>
</tr>
<tr>
<td>Sampling for telephone interviewing</td>
<td>110</td>
</tr>
<tr>
<td>Problems</td>
<td>110</td>
</tr>
<tr>
<td>Stages in data preparation</td>
<td>110</td>
</tr>
<tr>
<td>Steps in resolution</td>
<td>111</td>
</tr>
<tr>
<td>Flag changes</td>
<td>111</td>
</tr>
<tr>
<td>Variable generation of two types</td>
<td>112</td>
</tr>
<tr>
<td>Merging data sources</td>
<td>112</td>
</tr>
<tr>
<td>Sampling weight construction</td>
<td>112</td>
</tr>
<tr>
<td>Corrective weight adjustment</td>
<td>112</td>
</tr>
<tr>
<td>Frequency distributions</td>
<td>112</td>
</tr>
<tr>
<td>Validation</td>
<td>112</td>
</tr>
<tr>
<td>group identical answers and those with marginal differences</td>
<td>113</td>
</tr>
<tr>
<td>Measurement of errors</td>
<td>113</td>
</tr>
<tr>
<td>Problems in measuring errors</td>
<td>114</td>
</tr>
<tr>
<td>Validity checks</td>
<td>114</td>
</tr>
<tr>
<td>Section</td>
<td>Page</td>
</tr>
<tr>
<td>------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>Reliability checks</td>
<td>114</td>
</tr>
<tr>
<td>Correcting for non-response</td>
<td>114</td>
</tr>
<tr>
<td>Forms of weighting adjustments</td>
<td>115</td>
</tr>
<tr>
<td>Some imputation methods</td>
<td>115</td>
</tr>
<tr>
<td>Is missing really missing?</td>
<td>116</td>
</tr>
<tr>
<td>Preface to the first SPSS manual, by Norman H. Nie, Dale H. Bent and C Hadlai Hull</td>
<td>117</td>
</tr>
<tr>
<td>Preface</td>
<td>117</td>
</tr>
<tr>
<td>An introduction to SPSS, by Noma Makepela</td>
<td>123</td>
</tr>
<tr>
<td>Introduction</td>
<td>123</td>
</tr>
<tr>
<td>Main Menu</td>
<td>123</td>
</tr>
<tr>
<td>Status Bar</td>
<td>124</td>
</tr>
<tr>
<td>Command status</td>
<td>124</td>
</tr>
<tr>
<td>Filter status</td>
<td>124</td>
</tr>
<tr>
<td>Weight status.</td>
<td>124</td>
</tr>
<tr>
<td>Split File status</td>
<td>124</td>
</tr>
<tr>
<td>Data Editor (entering and editing data)</td>
<td>124</td>
</tr>
<tr>
<td>Creating a new data file</td>
<td>124</td>
</tr>
<tr>
<td>Inserting new variables</td>
<td>126</td>
</tr>
<tr>
<td>Statistical functions of SPSS</td>
<td>127</td>
</tr>
<tr>
<td>Statistics</td>
<td>128</td>
</tr>
<tr>
<td>Exploring data</td>
<td>129</td>
</tr>
<tr>
<td>Frequency tables</td>
<td>129</td>
</tr>
<tr>
<td>Cross tabulations</td>
<td>130</td>
</tr>
<tr>
<td>Example</td>
<td>130</td>
</tr>
<tr>
<td>Crosstabs and their statistics</td>
<td>130</td>
</tr>
<tr>
<td>Chi-square,</td>
<td>130</td>
</tr>
</tbody>
</table>
Data Analysis: Descriptive analysis and inferential analysis ................................................................. 143

Descriptive analysis 143

1. 2. Nominal/Categorical data ........................................................................................................ 144

Sandra Matanyaire 144

3. Ordinal data and Interval data .................................................................................................. 145

Table 2: Descriptive statistics appropriate for ordered variables .................................................... 145

4. Ratio-scaled data 146

Normal distribution 146

Sandra Matanyaire 146

Sandra Matanyaire 148

Graphs: 149

Inferential analysis ......................................................................................................................... 149

Summarized steps for hypothesis testing: ........................................................................................ 150

Sandra Matanyaire 150

Accelerator keys in the Data Editor ................................................................................................. 152

Sandra Matanyaire 152

In Edit mode: ................................................................................................................................... 153

Bibliography ..................................................................................................................................... 154

Life tables .......................................................................................................................................... 155

by Mike Bergh, Ph.D. 155

Introduction ......................................................................................................................................... 155

Example ........................................................................................................................................... 156

Statistics ........................................................................................................................................... 156

Fish population dynamics ................................................................................................................ 157

Human populations .......................................................................................................................... 163

Stationary populations and Lx and Tx ............................................................................................. 166

SPSS Version 11.0 ‘Advanced Models’, ‘Life Tables’ Option .......................................................... 167

Table: A dataset typical of the kind suitable for survival analysis ................................................ 168
Nuts and bolts of logistic regression

by Prof. John C. Whitehead

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Why use logistic regression?</td>
<td>173</td>
</tr>
<tr>
<td>Metadata description</td>
<td>174</td>
</tr>
<tr>
<td><strong>State Codes:</strong></td>
<td></td>
</tr>
<tr>
<td>The linear probability model</td>
<td>177</td>
</tr>
<tr>
<td>The logistic regression model</td>
<td>178</td>
</tr>
<tr>
<td>Interpreting logit coefficients</td>
<td>180</td>
</tr>
<tr>
<td>Estimation by maximum likelihood</td>
<td>181</td>
</tr>
<tr>
<td>Hypothesis testing</td>
<td>181</td>
</tr>
<tr>
<td>Evaluating the overall performance of the model</td>
<td>182</td>
</tr>
<tr>
<td>Some potential problems and solutions</td>
<td>185</td>
</tr>
<tr>
<td>Writing up results</td>
<td>186</td>
</tr>
<tr>
<td><strong>Some tips:</strong></td>
<td></td>
</tr>
<tr>
<td>Which model is preferred?</td>
<td>188</td>
</tr>
<tr>
<td>An example of an article demonstrating logistic regression</td>
<td>189</td>
</tr>
<tr>
<td><strong>Results</strong></td>
<td></td>
</tr>
<tr>
<td>A Test of Rational Choice Theory</td>
<td>190</td>
</tr>
<tr>
<td>Additional Tests of Economic Theory</td>
<td>190</td>
</tr>
<tr>
<td>Including Demographics</td>
<td>191</td>
</tr>
<tr>
<td><strong>Hands on Application:</strong></td>
<td></td>
</tr>
<tr>
<td>Here is a list of questions that need to be answered with the logistic regression model:</td>
<td>195</td>
</tr>
<tr>
<td>Hands on Application: Intermediate</td>
<td>196</td>
</tr>
<tr>
<td>Hands On Application: Advanced</td>
<td>197</td>
</tr>
<tr>
<td>Extensions to other logistic techniques</td>
<td>198</td>
</tr>
<tr>
<td>Syntax of the SPSS logistic regression statement</td>
<td>199</td>
</tr>
<tr>
<td><strong>Example</strong></td>
<td>201</td>
</tr>
<tr>
<td><strong>Overview</strong></td>
<td>201</td>
</tr>
<tr>
<td>Section</td>
<td>Page</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>Options</td>
<td>201</td>
</tr>
<tr>
<td>Basic Specification</td>
<td>202</td>
</tr>
<tr>
<td>Subcommand Order</td>
<td>203</td>
</tr>
<tr>
<td>Syntax Rules</td>
<td>203</td>
</tr>
<tr>
<td>Operations</td>
<td>204</td>
</tr>
<tr>
<td>Limitations</td>
<td>205</td>
</tr>
<tr>
<td>Example</td>
<td>205</td>
</tr>
<tr>
<td>Example</td>
<td>206</td>
</tr>
<tr>
<td>CATEGORICAL Subcommand</td>
<td>206</td>
</tr>
<tr>
<td>Example</td>
<td>207</td>
</tr>
<tr>
<td>CONTRAST Subcommand</td>
<td>207</td>
</tr>
<tr>
<td>Example</td>
<td>210</td>
</tr>
<tr>
<td>METHOD Subcommand</td>
<td>210</td>
</tr>
<tr>
<td>Example</td>
<td>212</td>
</tr>
<tr>
<td>SELECT Subcommand</td>
<td>213</td>
</tr>
<tr>
<td>Example</td>
<td>214</td>
</tr>
<tr>
<td>Example</td>
<td>215</td>
</tr>
<tr>
<td>Example</td>
<td>216</td>
</tr>
<tr>
<td>CRITERIA Subcommand</td>
<td>216</td>
</tr>
<tr>
<td>Example</td>
<td>217</td>
</tr>
</tbody>
</table>
Data-Graphics-Text ................................................................. 225

The graphical representation of quantitative data ....................... 227

by Robert C.-H. Shell ............................................................. 227

The bar graph .......................................................................... 230
The radar chart .......................................................................... 231
Frequency curves ...................................................................... 232
Chart menus in SPSS .................................................................. 232

To Edit a Chart Legend ............................................................ 232

Chart legend... ....................................................................... 232
Grouped frequency distributions and histograms ....................... 233
Pie charts ................................................................................ 233
Stem and leaf diagrams of variables ......................................... 234
Box and whisker plot of appropriate variables ......................... 235
Cross tabulations ...................................................................... 235
Understanding Pivot tables ...................................................... 235
Frequency distribution ............................................................ 236
Frequency table ....................................................................... 237

Table 2.2 .................................................................................. 237

Table 2.7 .................................................................................. 239

Histogram ................................................................................ 239
Frequency polygon ................................................................... 240
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cumulative frequencies</td>
<td>241</td>
</tr>
<tr>
<td>Cumulative frequency distribution (the ogive)</td>
<td>241</td>
</tr>
<tr>
<td>Regression analysis</td>
<td>242</td>
</tr>
<tr>
<td>Scatter diagrams</td>
<td>242</td>
</tr>
<tr>
<td>Introduction to path analysis</td>
<td>244</td>
</tr>
<tr>
<td>The population pyramid:</td>
<td>247</td>
</tr>
<tr>
<td>the demographer’s X-ray and a useful market research tool</td>
<td>247</td>
</tr>
<tr>
<td>The types of population pyramid</td>
<td>252</td>
</tr>
<tr>
<td>War</td>
<td>252</td>
</tr>
<tr>
<td>health</td>
<td>253</td>
</tr>
<tr>
<td>Immigration</td>
<td>254</td>
</tr>
<tr>
<td>Internal migration</td>
<td>255</td>
</tr>
<tr>
<td>Urban/rural</td>
<td>256</td>
</tr>
<tr>
<td>Population group</td>
<td>257</td>
</tr>
<tr>
<td>normalized comparisons</td>
<td>258</td>
</tr>
<tr>
<td>Spatial pyramids</td>
<td>259</td>
</tr>
<tr>
<td>Transition over time</td>
<td>260</td>
</tr>
<tr>
<td>The ultimate population pyramid</td>
<td>261</td>
</tr>
<tr>
<td>the world over time</td>
<td>261</td>
</tr>
<tr>
<td>How to write a research proposal for peer review</td>
<td>263</td>
</tr>
<tr>
<td>by Robert C.-H. Shell and Prof Kobus Oosthuizen</td>
<td>263</td>
</tr>
<tr>
<td>Introduction</td>
<td>263</td>
</tr>
<tr>
<td>Some general elements of expository style</td>
<td>264</td>
</tr>
<tr>
<td>The Introduction</td>
<td>267</td>
</tr>
<tr>
<td>Reliability of the data</td>
<td>267</td>
</tr>
<tr>
<td>Methodology</td>
<td>267</td>
</tr>
<tr>
<td>Technical features of demographic research proposals</td>
<td>268</td>
</tr>
</tbody>
</table>
Numbering ......................................................... 269
Tables and figures .................................................. 270
Quotations ............................................................ 270
References in the text ............................................... 271
Sources ............................................................... 272
One author: .......................................................... 273
Two authors: ....................................................... 273
Collection: ........................................................... 273
Article in collection: ............................................. 273
Journal article: ..................................................... 273
Thesis/dissertation: ............................................... 273
Unpublished manuscript: ...................................... 273
Newspaper report: ............................................... 274
Personal communication: .................................... 274
Official documents: .............................................. 274
The list of references at the end of the proposal .......... 274
Graphics .............................................................. 275
Tables ................................................................. 276
The importance of the non-technical elements ........... 277
The appendix: technical aspects of the proposal ........ 278
A final bibliographic section on "select sources" .......... 278

The social sciences can be value free ................................................. 279

by Ernest Nagel ......................................................... 279

The Internet and research ............................................. 291

by Mary Nassimbeni ...................................................... 291
Introduction .............................................................. 291
1 Types of search tools ............................................. 292
2 search engines ........................................................................................................ 292
Searching in specific fields ......................................................................................... 295
Meta-Search Engines ................................................................................................. 296
3 Directories ............................................................................................................... 297
Subject directories .................................................................................................... 297
Information Gateways ............................................................................................... 298
Universal Information Gateways ............................................................................. 298
Subject-based information gateways ....................................................................... 299
Speciality subject directories to search the “invisible web” ..................................... 301
4 finding general reference works on the www ......................................................... 302
5 Evaluating web resources ....................................................................................... 304
6 citing web resources ............................................................................................. 305

Searching for sources in the post-Caxtonian Era .................................................... 307

by Sandra Rowoldt

Searching the WWW1 ............................................................................................... 307
General reference works ......................................................................................... 308
Oxford Reference Online - Core Collection ............................................................ 308
Encyclopedia Britannica ......................................................................................... 309
wwwVirtual Library ............................................................................................... 309
Dictionaries ............................................................................................................. 309
Merck Manual of Diagnosis and Therapy ............................................................... 310
Evaluating web resources (see handout) ................................................................. 310
Some useful sites for African studies ...................................................................... 310
Africa Analysis ...................................................................................................... 310
Africa Online ......................................................................................................... 310
Africa Pulse ........................................................................................................... 311
Southern African NGO Network ............................................................................ 311
Africa Research Programme ................................................................................... 311
Africa South of the Sahara: Stanford University ..................................................... 312
Africa World Press Guide ........................................................................................................ 312
African Studies Center, University of Pennsylvania ............................................................ 312
African Studies Quarterly ................................................................................................. 312
Afrika No. .......................................................................................................................... 313
Postcolonial Web ............................................................................................................ 313
Some useful sites ............................................................................................................. 314

For demography & statistics .......................................................................................... 314
(including HIV/AIDS) ........................................................................................................ 314
AIDSLine .......................................................................................................................... 314
AIDSearch ....................................................................................................................... 314
CIA World Factbook ........................................................................................................ 314
Data Online for Population, Health and Nutrition (DOLPHN) ........................................ 315
Demographic and Health Surveys .................................................................................. 315
Demographic Association of Southern Africa ................................................................. 316
Demographic Research ................................................................................................... 316
GeoWeb ............................................................................................................................ 317
Globastat .......................................................................................................................... 318
International Data Base ................................................................................................ 318
OFFSTATS ....................................................................................................................... 319
The Pop Reporter ........................................................................................................... 320
Popline .............................................................................................................................. 320
Population Reference Bureau ......................................................................................... 320
Poverty in Africa ............................................................................................................. 321
the Social Science Information Gateway ......................................................................... 321
South African Data Archive ............................................................................................ 322
Statistical Agencies ......................................................................................................... 322
Statistical Profile of Education ....................................................................................... 323
Statistics Glossary ........................................................................................................... 323
Statistics South Africa .................................................................................................... 324
Statlib ................................................................................................................................. 325
<table>
<thead>
<tr>
<th>Source</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRAMSS</td>
<td>325</td>
</tr>
<tr>
<td>UNAIDS</td>
<td>326</td>
</tr>
<tr>
<td>UNESCO Statistics</td>
<td>326</td>
</tr>
<tr>
<td>United Nations Statistics Division</td>
<td>327</td>
</tr>
<tr>
<td>WHO (World Health Organization)</td>
<td>328</td>
</tr>
<tr>
<td>World Bank</td>
<td>328</td>
</tr>
<tr>
<td>World Population Information</td>
<td>329</td>
</tr>
<tr>
<td>WWW Virtual Library</td>
<td>330</td>
</tr>
</tbody>
</table>

### “But I found it on a website ...”

The importance of evaluating internet resources

#### QUALITY

- *Why bother to review quality?*  
- *How should you assess quality?*

<table>
<thead>
<tr>
<th>Source</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>WHO?</td>
<td>332</td>
</tr>
<tr>
<td>Tips on evaluating the creator of the resource</td>
<td>333</td>
</tr>
<tr>
<td>Topic</td>
<td>Page</td>
</tr>
<tr>
<td>----------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>Getting clues about authorship from URLs</td>
<td>333</td>
</tr>
<tr>
<td>Some useful tips about URLs</td>
<td>334</td>
</tr>
<tr>
<td>Investigating the root of an url</td>
<td>335</td>
</tr>
<tr>
<td>What?</td>
<td>336</td>
</tr>
<tr>
<td>Why?</td>
<td>336</td>
</tr>
<tr>
<td>Tips on evaluating the standpoint of a resource</td>
<td>337</td>
</tr>
<tr>
<td>where?</td>
<td>337</td>
</tr>
<tr>
<td>When?</td>
<td>337</td>
</tr>
</tbody>
</table>

**Index**

339
Preface

This e-book is designed to satisfy introductory quantitative and statistical methodology requirements in the writing of a dissertation at the honours, masters and doctoral levels. The e-book has a comprehensive table of contents which should be the basic navigational tool for the reader. There is a small index as well.

However, the authors hope the book will have far wider applications. Professional persons required to interpret data will benefit as they will be able to appreciate the techniques used to prepare and analyse aggregated output. Anyone who has to prepare a survey will be able to find how-to sections which will help them. Those who have to prepare a slide presentation will find useful sections on how to draw elegant graphics and also know something of the standard graphing conventions.

Robert C.-H. Shell

Extraordinary Professor of Historical demography

University of the Western cape
Sticky mouse syndrome and managing files

One of the big problems with labs is that people eat their food there and grime accumulates on the surface of the desk. Then, through their hands scuffing the table top the dust and dirt winds around the rolling bars and wheel inside the mouse making for a miserable computing experience. I asked my students to wash their hands before computing. Nobody seemed to mind.

Fighting sticky mouse syndrome

The three steps I take are

1. remove the cover and the mouseball: wash ball in water and detergent (cold water is fine).
2. Clean the rollers with an ear cotton swab dipped in isopyryl alcohol, moving the swab along the axle to break the ring of gunge.
3. Clean the mouse wheel in the same way using lots of light to see if the components are really clean.
4. Blow the interior of the mouse out with a hard breath
5. Replace the ball
6. Screw on the cover

Starting to manage your files

What can one say about file management? It may appear as a boring topic, but file management is the **sine qua non** of computer competency. If one does not learn file management, computers will remain a frustrating and
anxiety-ridden business. Nervous breakdowns and even severe depression may result if one does not master this skill.

On looking into some colleagues’ and student’s directory or folder structures over the years, I was appalled to see that sometimes all files were in one directory, e.g. “My Documents”. File names were confusing. Ddata, graphics and texts could not be easily recovered. I have 1,568,000 user files, some date back before the advent of the PC. All are backed up on multiple media, floppies, stiffies, hard drives, tapes, memory sticks and CDs. Yes, I have lost files in twenty years but all have been restored.

Have a system and constantly refine it.

Here is my Data Graphics Text (DGT) system. I divided up all my files into three basic types according to how I work. If you are a sound or video person, you would simply add those file types in their respective order. In my case, I have allocated a hard drive to each file category. Thus, I have a hard drive dedicated to data and another devoted to a full copy of the data. Another drive is devoted to Graphics and so on. A full copy is made every Friday evening while I am relaxing doing some reading.

Data files are my basic number crunching files which include all spread-sheet files, tables, statistics files, in my case SPSS files, Database (.DBF) files and so on. These always go into “My data” directory. In a rush they can go into the base directory of “My data” to be sorted later. We will talk about subdirectories below.

Graphics files, which chew up disk space, go into the “My Graphics” folder. These include clipart in its own folder, “My clipart”. Reset your program’s default directory structure to reflect your own organization.

Finally, text files which include all PDF files, manuals and all word processing files and so on go into “My text files” folder.

This simple division at least gives you an overview of where to put the files. It is the basis of a system.
The DGT subdirectory system and how to work with files

Here I am relying on the text of MS Windows XP help files.

To work with files, click Start, point to Programs, and then click Windows Explorer. Or you may right click on Start, point to Explorer. Windows Explorer works a lot like the old File Manager with the added benefit of displaying all your drive connections in one window.

To use the MS-DOS prompt, click Start, point to Programs, and then click MS-DOS Prompt (see DOS handout)

To run a program from the command line of any drive, click Start, and then click Run.

To copy files, use the same method as copying text: Select the files you want to copy in Windows Explorer, and then on the Edit menu, click Copy.

To paste copied files, select the folder in which you want to put the copied files, and then on the Edit menu, click Paste.

To switch between windows, click the button on the taskbar, or Alt W.

Tips: Renaming

To change the name of a file or folder

In My Computer or Windows Explorer, click the file or folder you want to rename.

On the File menu, click Rename.

Type the new name, and then press ENTER.

Notes
A Windows 98 file name can contain up to 255 characters, including spaces. It cannot contain the following characters: \ / : * ? “ < > |.

However, do not be fooled by the long filename option within Windows 98 and beyond. In the event of a crash, you might be only able to recover a brutally shortened DOS file name. For example, if you have used some long, indulgent name like “Letter to my mother-in-law on the occasion of her tongue removal operation.YAY. 1999-02-31.wpd” That crucial file may only be recovered by the cryptic “LETTER~1.wpd.” If you want to be a purist best to use strict eight character DOS file-naming conventions.

Laplink, a DOS utility for copying files from one computer to another uses DOS filenames and will reduce your long file names to 8 character filenames. Laplink does however have a Windows 98 version. But you are always safer using the system of 8 character filenames.
Quantitative methods, large datasets and the PC

by Robert C.-H. Shell

Topics covered

- the PC and its revolutionary application to quantitative methods
- Statistical concepts and notation
- Source analysis, data identification, collection and preparation
- Graphical presentation of data
- Graphing conventions and types of graphs
- An introduction to spreadsheets
- Statistical applications of spreadsheets
- Statistical programs (SPSSPC)
- Improving your survey techniques
- Preparing a proposal for peer review

Intro

The personal computer (PC) has the potential for changing most professional’s lives, yet it has made hardly any inroads in the lives of most. Why? While we may often have access to the equipment, few of us have the time, or sheer stamina to face several steep and exasperating learning curves. It takes time to become acquainted with the fledgling personal computer
industry and the slew of exciting software products on the market. Also, many students are simply unaware of the variety of new PC applications such as GIS and statistical software. Moreover, most software is far too expensive a commodity with which to experiment. This course is intended to provide the skeptic with several reasons to explore and use the PC.

This book is designed to fulfil two basic functions

(1) to introduce the student to the scope of the PC and

(2) to enable the student to obtain sufficient mastery over the Operating system, Windows 95/98 and XP and the various spreadsheet and statistical applications to complete a complex dissertation.

No computer experience is assumed.

Most computer-phobic scholars consider the computer too arcane, too difficult to learn, and therefore not worth spending time on. It is easier and more productive—they argue—to carry on with “the old ways.” There is much truth in this point of view, but only in the short run. Undoubtedly in learning all the ropes of the PC, personal productivity may go downhill. It is only now, twenty-five years after its introduction, that the personal computer has registered substantial gains in productivity in industry and commerce. No studies have appeared on academic productivity. Nevertheless, after a one year part time apprenticeship on the PC, one can expect as much as a fifty percent gain in productivity in all aspects of a scholar’s life from note-taking and composition to grading.

Moreover, some computer applications, like the spreadsheet, allow the scholar to do things that were not feasible before. However, without training time, the computer might as well stay in the closet, or remain a glorified typewriter. Because of its
versatility and speed, the personal computer can change all our lives; already three undergraduate colleges in the United States will not accept undergraduate students unless they have their own personal computers! By the year 2000 there will be many graduate students and faculty who will be completely proficient in a variety of PC applications.

**Virtual Reality**

The PC is really a simulator of jobs. The dotty letters on the screen are stand-ins for the real letters we all grew up with. The graphs on the spreadsheet screen are surrogates for the real graphs we drew by hand. Everything the PC does is ultimately the world cut down to Random Access Memory (RAM) size. Unlike the mainframe environment, the PC allows you to tailor your output much more exactly. The genius of the PC is precisely this talent for simulation. You can see what something looks like, fool around with it, and play with it until you get it right. This type of simulation is termed What-You-See-Is-What-You-Get (WYSI-WYG). Although people can get lost in attempting to create a perfect page, WYS-IWYG is a powerful tool and one in the main simply did not exist a few years ago.

There are problems—big problems—with personal computers. As they are now configured, they are probably among the most difficult and intractable machines an adult will have to face (apart from mainframes). Engineers and programmers have still not realized that human beings use them. Moreover, most manuals are written with little regard for the absolute clarity which the PC autodidact requires. Moreover, the great variety of hardware and software has already generated a tower of Babel. A certain amount of
unavoidable jargon flourishes in the PC world, for instance, after completion of this course you will know what it means “to do a warm boot on your stiffy, or why you “should read FAT into memory.” For all these reasons, it is little wonder then, that the scholarly world is reluctant to endorse the seemingly mad world of the personal computer. Yet for all its faults, the computer is probably the most useful tool since the pencil.

The philosophy underlying this course may be summed up by the concepts of “command and keystroke parsimony.” One learns only what is absolutely necessary to achieve a certain desired end. In this case, how to write and format a thesis in the shortest possible time. The emphasis is on producing useful output, not on learning all the bells and whistles of the PC. Everybody is encouraged to use the fewest commands and keystrokes to achieve results. No attempt is made to learn everything. That way lies madness, a very dangerous path. Learn according to need only. “Hacking” is positively discouraged. The big danger of the PC is its limitless possibilities which can dazzle the uninitiated and suck them into a “black hole” in which time and energy and finally, careers are squandered. Parsimony in learning is not merely a useful idea, it is an essential discipline in mastering the PC.

Therefore, in this course, no attempt is made to introduce the student to any programming skills. There is simply no way that someone unfamiliar with computers can justify attempting to write a program. To write even the simplest program, which would require several hundred hours of solving three dimension crossword puzzles, cannot be justified in any cost-benefit system of a career-oriented scholar. There are thousands of trained, skilled and talented programmers out there batting their brains out attempting to create perfect products in the Darwinian software jungle. Take advantage of their labors, do not squander your time learning programming skills.

This course will present an overview of applications for the PC. The spreadsheet is LOTUS 123 but students are encouraged to experiment with Quattropro and Excel.
Reading

The books written on the PC are mostly not worth buying: They are expensive and generally hastily prepared. The books below are exceptions. My rule of thumb is: if the computer book does not have an excellent index, put it back on the shelf.


Carolyn Jorgenson, *Mastering 1-2-3* (Sybex)
An introduction to spreadsheets

by Robert C.-H. Shell

Objectives

By the end of this chapter you should be able to start, use and save an elementary spreadsheet

An introduction to the spreadsheet

The spreadsheet is probably the strangest concept to the PC newcomer. There is simply no equivalent tool in the ordinary hardcopy world and there was no mainframe equivalent. If you can imagine the combination of a powerful pocket calculator, a sheet of paper twenty meters wide by twenty meters long, an electronic wand, a word-processor, a sorter, all in 3D and all in one, you are close. Without doubt, the spreadsheet is the most powerful, most versatile, most exciting application of the PC. The spreadsheet is also the most difficult application to master thoroughly.

These sessions will attempt nothing more than a "monkey-see, monkey-do" approach, but even after the first time you should be able to do simple operations and be in a position to know whether you will need to use the product. If you only want to balance checkbooks and grade term papers, chances are that you will not be able to justify the purchase price of a spreadsheet. If you deal with numbers in any way in your work, or need graphs derived from calculations, the spreadsheet is probably the way to go.
Lotus 1-2-3 is one of the best spreadsheet product around. Lotus 1-2-3 can even do cross-tabulations—the work horse statistic for nominal or categorical data. It can also do regression analysis. However, the size of the spreadsheet, which is resident in volatile memory, is dependent on the size of the memory of your machine. One of the most dreaded 1-2-3 messages is "memory full." Therefore, you simply cannot have unlimited databases in 1-2-3.

**Spreadsheet basics**

The spreadsheet is divided into columns and rows. Columns go up and down the spreadsheet, while rows go across the spreadsheet. Columns are referenced by letters, e.g. A, B, C, and so on, across the top. After column 26, the series starts again, thus column 27 is AA, column 28 is AB; column 29 is AC and so on. There are 256 columns. Rows are referenced by numbers, thus the first row is "1," the second is "2" and so on, until the end of the spreadsheet. There are 8192 rows. There are thus 2,097,715 cells, (using 1-2-3 come back and check this multiplication when you can).

A cell is formed where a column and a row intersect. This is known as the cell address. The cell is the basic building block of the spreadsheet organism. Knowing what can go into a cell and how it can go in, and how cells interact arithmetically and logically is what defines mastery over the spreadsheet.

**The third dimension of the cell**

All entries are made one cell at a time. Each cell can contain only one type of information at a time, but may display that information in a different form. Thus, while a formula is entered into a cell, that cell will display the result of the formula, not the formula, which only appears in the overhead status line. This is the 3D aspect of the spreadsheet and sometimes causes initial confusion. Later versions enable another spreadsheet to be placed on
top of the one in memory. Before opening sheet two, make sure you are
conversant with all the possibilities of the single sheet first.

1-2-3 automatically defaults to a column width of nine characters. If your
number or formula or label is longer than that, it will still be entered but
may not show up. Asterisks [*] in a cell always indicate that the column
width is too small. You can easily change the column width by using the /
[back-slash key [W]orksheet [C]olumn and then follow the prompts: max-
imum width is 245 characters.

The four basic types of information that may reside in a cell are:

1. A number
2. A label
3. A formula
4. A date, or a time (date and time arithmetic is one of the fortes of the
spreadsheet.)

**Cursor control**

Unlike a word-processor, the cursor actually covers the whole cell. The cell
is highlighted in reverse video. You may enter cell information only at that
highlighted spot. In a new spreadsheet, the cursor automatically falls on the
home cell, i.e. A1.

The column letter and the row number form the cell address, thus A1 is the
address of the first cell of the spreadsheet, or "home." You can move to any
cell in the spreadsheet by pressing the GO TO program function key (F5)
and the cell address.

You may also use the cursor control keys on the right hand side of the
keyboard:

- Pg Up moves the cursor 20 lines up
o Pg Dn moves the cursor 20 lines down
o Home moves you to A1
o End moves you to the end of the spreadsheet
o Tab moves you right one screen
o Shift-tab moves you left one screen

Note: the GO TO key at F5 and the cell address will move the cursor to that address. Thus F5 B230 will take the cursor from the cell known as B230. This is my favourite way to move around the spreadsheet; the exercise of having to remember where the various elements of the spreadsheet are will help you. If you cannot remember, place key cell addresses at the top left [HOME] of the spreadsheet and label them, such as...

Sta 501

Then press the HOME key whenever you are lost and you will be at the list of reminders and can use GO TO from there!

Try all of these out....

The program function keys

Each of the program function keys (on the left of the keyboard) has several functions. For the purposes of this session, I will only list the ones you are going to use, or can use. Program function keys are simply extra keys on the keyboard, which are dedicated to a particular function. Each program, like WordPerfect, Symphony &c. has different functions assigned to each program function key.

F1 HELP on-line explanation of commands
F2 EDIT allows you to edit the contents of a cell
F3 NAME view a list of range names
F4 ABSOLUTE: makes a cell constant in a formula
F5 GO TO: allows you to move to an address
F6 WINDOW: moves the cursor to or from a window
F7 QUERY: extract, find or delete records from a database
F9 CALC: recalculate a spreadsheet
F10 GRAPH: redraw a graph

Try each one of these and see the results.

The menus

1-2-3 is a menu-and icon driven program. This means instead of having to memorize commands, such as in SPSSPC or DOS (which are termed "command-driven" programs), you are confronted with menus. To access the menu system, you have to hit the slash key ("/"), which gives you entree to the dungeons and dragons of 1-2-3, a hierarchical labyrinth of commands at different levels. Familiarity with these command trees will give you mastery of the program.

Note: while you are in menu mode, you cannot enter anything into the spreadsheet.

Note: to return from lower levels of menus, you simply hit "Esc[ape]" and the program will take you to the next higher level, finally returning you to the spreadsheet itself.

Note: You may use the cursor keys to make a menu choice, but the fastest way is to type the first letter of the menu choice, which saves a few key-strokes.

Exercise:
explore the menu trees down to the bottom most level.
Ranges

One of the handiest aspects of the spreadsheet is the ability to manipulate ranges. A range is a set of cells in a column or a row, or both. A range is always in the form of a rectangle and can be copied, named, deleted, added to or manipulated at will. You can name ranges and create ranges very easily through the menu system, using the / key and the "Range" choice.

Anchoring the range is accomplished by using the "." or period. "Establishing" the range may be done by using the cursor control keys or typing in the cell addresses manually, which is the way to go if you can remember where the ranges begin and end, or if the range is very long.

The formula

There are a number of things you need to know about formulae. First, they are constructed in algebraic or logical expressions. Second, the ( ) parentheses take precedence in computations. All formulas may be preceded by an +.

Calculating numeric formulas

The number of decimal places 1-2-3 displays for a calculated value depends on the number format of the cell. 1-2-3 calculates the value to a precision of 15 digits regardless of how many it displays. Use an @function, such as @ROUND, to specify a precision of less than 15 digits.

Documenting a formula

You can annotate a formula by typing ; (semicolon) immediately after the formula and then typing the note. The note appears in the contents box only, unless you format the cell as text.
To add a comment to a cell containing a formula, use Range - Cell Comment.

**Using formula markers**

You can use formula markers to identify cells that contain formulas. Formula markers are turned off when you first start 1-2-3. To turn them on, use View - Set View Preferences.

**Printing formulas**

When you print a workbook, you can choose to print the formulas along with the cell contents. To print formulas, choose File - Preview & Page Setup, click the Include tab, and select "Formulas and cell contents" in the "Show" list.

The + indicates to 1-2-3 that you are about to start a formula. It is also the same formula for addition. The other common symbols are:

* multiplication
/ division
+ addition
- division

**Using @ functions**

An @ function in 1-2-3 jargon is a timesaving formula that has been abbreviated to a single word preceded by the symbol @, which signifies to the program that the following entry is a function and not a label. These functions operate on a range of cells or one cell, which information is specified after the open parenthesis. This is called the argument of the function.

The following is an example of a simple @function and an argument:
@SUM(A2..A10)

which expression would add all numbers in the column range of A2 through A10

Arguments must be separated by commas, and the last argument must end with the ")" parenthesis. The syntax is the name of the function plus the particular argument it requires.

There are many functions, some of the common ones are:

**Statistical functions**

@SUM
@COUNT
@AVG
@MIN
@MAX
@VAR
@STD

**Mathematical functions**

@ABS
@INT
@ROUND
Engineering functions

@ACOS; @ASIN; @ATAN; @SIN; @SQRT; @COS

Date functions:

@NOW
@DATE
@TIME

Logical functions:

@IF
@TRUE
@FALSE

This is by no means a complete list; these are the functions I use most, except for the engineering functions. Functions may be nested, i.e. a function may be performed on another function,

e.g.

@SQRT(@SUM(A1..A4))

This nested formula would add the numbers in cells A1 through A4 and then take the square root of the total. Try it and then erase the range.

Once you have read through this, we should all be ready to begin a few exercises. This is only the beginning of 1-2-3, but already you should be able to see the most incredible possibilities of this tool.

Exercises
The first thing to do with a spreadsheet is to prepare it, before you start putting in data.

Type the following labels in:

Name:.........................
Purpose of Spreadsheet........
Date of Creation:............
Date Last Worked On.........

**Saving your work**

The second thing you do with a spreadsheet is to save it. Since the spreadsheet resides in volatile memory, a power failure would result in all your work being lost. So save and name the file, call it "class."

Do this every ten minutes or so.

If you have brought your checkbooks, bring them out and we will create a simple program for balancing your checkbook. If you do not have your checkbook or are too embarrassed, make up some checks!

Type in cell A10
"No of Check:"
Type in cell B10
"Drawer:"
Type in cell C10
"Amount of Withdrawal"
Type in cell D10
"Deposits:"
Type in Cell E10
"Balance:"
Oops! We forgot to put in the date
Place the cursor on B10
Insert a new column
Type in "Date:"
These are the labels for the spreadsheet.
In the first row under the labels type in your current bank balance under "deposits." Now the formula. What is the formula that has to be written under balance?

+(E10+D10-C10) is incorrect
What is the correct formula?
Now start using some of the @ functions like @MIN, @AVG and so on until you feel comfortable with the spreadsheet environment.

**Classwork**

1. Replicate your month end statement, including all the items your bank charges you for. Improve the knowledge of how you spend your money.

**Difference between databases and spreadsheets**

Spreadsheets are not databases, but they are similar. Databases are of two kinds: one is variable field length (like Bibliography and Notebook) and the other is fixed length (like DBase IV). For notes and bibliographies one may want to have long fields, or not be bothered with limitations on record
length. However, some systems are limited in terms of sorting. Also, no numeric computations can be accomplished at all.

What is a Database management system? A database system stores, sorts and selects and performs a variety of operations based on a logical sequence of search criteria on a nearly infinite number of records on disk. A spreadsheet, on the other hand, does many of the same sort of things, except that it is mathematically more sophisticated, but it is more limited when it comes to manipulating records. A spreadsheet is more accurately conceived as a primitive and small record manager. The database system shines when it comes to combining or merging different criteria to create new information. For example, if you wanted all books on a given topic, written before 1948 and only on Africa, then a database would do this chore easily and create an output file to your exact specifications.

The main limitation of spreadsheets is the number of records it can hold. Since spreadsheets work in volatile memory, spreadsheet size is limited by volatile memory size (RAM) Databases store and work with data on disk and their total record size is limited only by your available disk size.

**Fields (variables), records (cases) and values (data)**

Fields or Variables are the units of a Record, thus "Author" and "Title" are "fields" within a "record," where the record is the publishing details of an individual book. The value is the individual datum that goes into a field.
A spreadsheet exercise
the Aitken’s procedure:

by Robert C.-H. Shell

This chapter explains how to derive interpolated approximations of age group data from irregular age groups in published censuses using the Aitken’s procedure in a revised spreadsheet form. In Aitken’s time (1930s), this used to be a tedious chore accomplished with a hand calculator, but thanks to the modern spreadsheet, regular five-year age groups or non-census year totals may be rendered rather more quickly. This chapter may be considered a cliometric note on obtaining population estimates for non-census years and dis-aggregating irregular age groups from published censuses.

Every historical demographer in South Africa knows how irritating it is when looking at some published national or colonial census displaying age group data, that the page setter has lumped age groups to fit into a table on the printed page, e.g., “5-19 years”, instead of “5-9”, “10-14”, “15-19”. To create five-year age groups for a standard population pyramid, one used to have to do tedious reiterative calculations on a hand-held calculator. Each computation was the result of numerous cross multiplications. The Aitken’s procedure however, is especially suitable for use with modern spreadsheets: one can throw away that old calculator.

The enclosed spreadsheet allows one to preserve the formulation in protected cells (red italics in the provided spreadsheet) and then simply “pump in” the irregular age groups into the input cells (green) and read out the results one age group, or one non-census year at a time. The cut-and-paste operation allows the process to be virtually automated although one has to type in the irregular age groups exactly. Using relative addresses, one could...
have the exact results display in the population pyramid data table. This is not shown here.

The same procedure may be used for interpolating population for non-census years. The census years in this country have been irregular, viz., 1865, 1875, 1891, 1904, 1911, 1921, 1936, 1946, 1950, 1960, 1970, 1980, 1985, 1991, 1996 (the Dutch East India Company did a better job with its annual *opgaafs* and *monster-rollen*, 1658-1795).

How would one obtain the number of people for a particular city in 1983 (the beginning of the AIDS epidemic in South Africa)? This procedure will do that job as well. In what follows I have simply repeated the description of Aitken’s procedure from the trusty manual of the US Bureau of the Census and then added a printout of the formulas from the spreadsheet which I prepared. This note is therefore nothing more than a modest applied procedure. Those who wish to avoid the formal description and get to work should turn to the last page.

**Aitken’s iterative procedure**

Aitken’s iterative procedure is a system of successive linear interpolations equivalent to interpolation by a polynomial of any desired degree. Aitken’s system is set up in the following basic format for interpolation between four given points for the value of $f(x)$:
Given

<table>
<thead>
<tr>
<th>Ordinates</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( f(a) )</td>
<td>(a-x)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( f(b) ) &amp; ( f(x; a, b) ) &amp; (b-x)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( f(c) ) &amp; ( f(x; a, c) ) &amp; ( f(x; a, b, c) ) &amp; (c-x)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( f(d) ) &amp; ( f(x; a, d) ) &amp; ( f(x; a, b, d) ) &amp; ( f(x; a, b, c, d) ) &amp; (d-x)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Only the first two lines would be used for two point or linear interpolation, and there would be just one computational stage. The first three lines and two computational stages would be used for three point interpolation. Additional lines and computational stages are used as required for more points. As many points as desired can be used. The first column, “given ordinates,” symbolizes the given data, i.e., the four observations. The “proportionate parts” in the extreme right-hand are differences between the given abscissa and the one for which the interpolation is wanted. The abscissa values may be transformed into simplest terms to simplify the calculations.

The entries in computational stage (1) are each calculated by computing diagonal cross-products, differencing them, and dividing by the difference between the proportionate parts as follows:

\[
f(x; a, b) = f(a)(b - x) - f(b)(a - x)
\]
Each of the expressions \( f(x; a, b), f(x; a, c), f(x; a, d) \) etc. is an estimate of \( f(x) \) obtained by linear interpolation or extrapolation of \( f(a) \) and one of the subsequent \( f(b), f(c) \) or \( f(d) \) values.

The general process of successive linear interpolations is repeated for computational stage (2), but this time we use the results of computational stage (1) and their associated diagonal multipliers. Thus,

\[
f(x; a, b, c) = f(x; a, b) - x - f(x; a, c) - x
\]

\[
( c - x ) - ( b - x )
\]

\[
f(x; a, b, d) = f(x; a, b) - x - f(x; a, d) - x
\]

\[
( d - x ) - ( b - x )
\]
Suppose, for example, we want the population in 1935 given data on populations in the census years 1920, 1930, 1940 and 1950. The calculations are summarized as follows:

<table>
<thead>
<tr>
<th>Date</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>Proportionate parts</th>
</tr>
</thead>
<tbody>
<tr>
<td>1920</td>
<td>16,321</td>
<td></td>
<td></td>
<td>1920-1935 =-15</td>
</tr>
<tr>
<td>1930</td>
<td>30,567</td>
<td>37,690</td>
<td></td>
<td>1930-1935=-5</td>
</tr>
<tr>
<td>1940</td>
<td>52,108</td>
<td>43,161</td>
<td>40,426</td>
<td>1940-1935 =+5</td>
</tr>
<tr>
<td>1950</td>
<td>87,724</td>
<td>52,022</td>
<td>41,273</td>
<td>40,002</td>
</tr>
</tbody>
</table>

The successive computations for stage (1) are:

\[
16,321(-5) - (30,567)(-15) = -81,605 + 458,505 \\
\frac{(-5)(-15)}{-5+15} = \frac{376,900}{10} = 37,690
\]
Consider the numbers of women in the age groups 14-17, 18-19, 20-24, and 25-29 for the United States in 1960. For illustrative purposes, assume that there are $K$ girls under age 14 (the number $K$ will drop out as the work progresses, so its value does not matter; it is used here simply to help clarify the exposition). The upper limit of the age range under 14 is 14.0. The number of girls aged 14 to 17 plus $K$ (the girls under 14) is then the cumulated number under 18 years old; the upper limit of that age range is 18.0. The population 18 to 19 years old, plus the population 14 to 17 years old, plus $K$ is the cumulated number under 20 years old; the upper limit of that age range is 20.0. Continuing this process, one obtains the cumulated numbers at exact ages 14.0, 18.0, 20.0, 25.0, and 30.0. The cumulated data represent the ‘ogive’ transformation of the original data for groups into data for specific points along the age scale. The transformed data are thus associated with precise points of age.

Interpolation of the transformed data can now be performed by any appropriate method but must be done twice—once for the upper limit of the subgroup for which interpolation is desired and once for the lower limit. Thus, to estimate the population in age 20 from the data for age groups aged 14-17, 18-19, 20-24, and 25-29, one estimates the population under age 20, and then estimates the population under age 21. The difference between the two estimates will be the population between the 20th and 21st birthdays. Since $K$ is common, the population under 20 and the population under 21, the subtraction causes $K$ to vanish. This means that $K$ can
be taken as zero (instead of some other arbitrary number) thereby simplifying the computing operation:

<table>
<thead>
<tr>
<th>Age group (years)</th>
<th>Upper limit of age group</th>
<th>Number of women</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aitken's procedure</td>
<td></td>
<td></td>
</tr>
<tr>
<td>In age group</td>
<td>Cumulated from youngest group</td>
<td>Computations for age 21.0 (upper limit) Proportionate parts</td>
</tr>
<tr>
<td>(1) (2) (3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14-17.....</td>
<td>18.0</td>
<td>5,516 5,516</td>
</tr>
<tr>
<td>18-19.....</td>
<td>20.0</td>
<td>2.4177,9339,142</td>
</tr>
<tr>
<td>20-24.....</td>
<td>25.0</td>
<td>5,520 13,453</td>
</tr>
<tr>
<td>25-29.....</td>
<td>30.0</td>
<td>5,537 18,990 8,918</td>
</tr>
</tbody>
</table>

The figure of 9,082 in column (3) is the interpolated estimate of the number of women cumulated to age 21.0. We also need the cumulated number to age 20.0, but that is already given as 7,933. The desired estimate of the population in age 20 is therefore the difference (1,149) between the 9,082 cumulated to age 21.0 and the 7,933 cumulated to age 20.0.

The number of women in age twenty is therefore 1,149.

### Insertion of procedure into a spreadsheet

This procedure can now be placed into a spreadsheet for storage and frequent use. Notice the first row is the column reference and the first column provides the row address. Thus the phrase “Upper limit of age group” is in cell b11 (I used the first 10 rows for identification of the spreadsheet. One could start in A1).

A:D32: (C32)
A:D33: (D32+C33)
A:D34: (D33+C34)
A:D35: (D34+C35)
A:D36: (G35)-D34
A:D40: @SUM(D39..D38)
A:E33: @ROUND(((D32)*(H33)-(D33)*H32)/((H33)-(H32)),0)
A:E34: @ROUND(((D32*(H34))-((D34)*(H32)))/((H34)-(H32)),0)
A:E35: @ROUND(((D32)*(H35))-((D35)*(H32)))/((H35)-(H32)),0)
A:F34: @ROUND(((E33)*H34)-((E34)*H33))/((H34)-(H33)),0)
A:F35: @ROUND(((E33)*(H35))-((E35)*H33))/((H35)-(H33)),0)
A:F40: @SUM(F38..F39,D34)
A:G35: @ROUND(((F34)*H35)-((F35)*H34))/((H35)-(H34)),0)

List of references

Basic concepts in statistics and notation

Objectives

By the end of this chapter you should be able to:

· have an understanding of the basic concepts of statistics
· know the sigma (Σ) notation

It seems hard enough for a graduate student to have to deal with the subject of mathematics at the same time as to be introduced to an intimidating notation style and the Greek alphabet which is

1) quite unfamiliar
2) cannot be pronounced without elocution lessons
3) which cannot be typeset on most computers
4) and which is anything but clear and intuitive
5) is not used by computer programs such as spreadsheets and statistical packages such as SPSS.

These points raise the question of whether the notation carries its weight.

The terminology of statistics is also confusing and needs some standardization. For example, many texts discuss nominal and categorical as if they were the same, but are they? Spreadsheets do not use these terms at all but talk of labels and numbers while even SPSS uses only three types of data, viz. categorical, ranked and interval
Terminology

Population

The conceptual totality of observations under consideration.

Finite and infinite populations

A population may be finite or infinite. Thus the populations of all the leaves on a tree or all the books in a library are examples of finite populations. Some finite populations have a very large number of members, e.g. the population of all the sand grains on all the beaches in the world is a very large population of sand grains, but still a finite population. An infinite population contains infinitely many members. For example, the population of all numbers is infinite.

Sample

Any subset of a population. For example, one or more members of a population is a sample.

Objectives of statistics

To make inferences about the population using the information based on the sample values. The sample should be representative of the population.

A sampling procedure that consistently over or underestimates a population characteristic is said to be biased.

Eliminate bias by choosing a simple random sample.

A simple random sample of \( n \) observations is a sample chosen in such a manner that every subset has an equal chance of being selected.

Statistics

A field of study concerned with
a) descriptive statistics (the collection, organization and summary of data)
b) inferential statistics (drawing inferences about the population using information in the sample).

The tools of statistics are employed in many fields, e.g. business, education, psychology, pharmacy, HIV/Aids, etc.

**Variables**

If in observing a characteristic, we find that it takes on different values in different persons or objects, we call the characteristic a variable, e.g. height, weight, age, blood pressure, eye colour, sex, matric mathematics result, etc.

A variable is denoted by a symbol such as \( X, Y, Z \) etc., which can assume any of a prescribed set of values. The particular values which the variables may assume are called variates, and these are normally denoted by the corresponding small letters.

Thus if \( Y \) represents tree height and the height of a particular tree, e.g. the tenth tree in the population of trees under consideration, is 3 metres, then \( y_{10} = 3 \). The little number next to the letter at the bottom right hand corner, which indicates the particular variate, is known as a subscript.

**Quantitative and qualitative variables**

**Quantitative variable**

A variable is quantitative if it is described by numbers, e.g. weight, height and age. These variables convey the concept of amount.
**Qualitative variables**

Variables that yield categorical (non-numeric) responses, e.g. eye colour, sex.

Usually these are described by words and are also called attribute variables.

Data types

1. Nominal-scaled data – mainly qualitative, assigned to a number of categories of equal importance, e.g. Surname, Town name, gender [male, female].

2. Ordinal-scaled data – mainly qualitative, assigned to coded categories; ranking implied between categories e.g. Class (1st years, 2nd years, 3rd years, Hons).

3. Interval-scaled data – quantitative; possesses both order and distance but no origin, e.g. IQ scores, and temprature.

4. Ratio-scaled data – quantitative; zero origin e.g. age, height, weight.

Nominal scale data : Gender: Male Female

Ordinal scale data : Class: 1st years 2nd years 3rd years

Interval scale data : IQ Score: 30 to 195 or more

Ratio scale data :Age: years
<table>
<thead>
<tr>
<th>ID</th>
<th>age</th>
<th>wt</th>
<th>ht</th>
<th>class</th>
<th>gender</th>
<th>study</th>
<th>HIV</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>27</td>
<td>73</td>
<td>174</td>
<td>THIRD</td>
<td>MALE</td>
<td>ALWAYS</td>
<td>neg</td>
</tr>
<tr>
<td>2</td>
<td>24</td>
<td>59</td>
<td>156</td>
<td>FIRST</td>
<td>FEMALE</td>
<td>MOST TIME</td>
<td>neg</td>
</tr>
<tr>
<td>3</td>
<td>22</td>
<td>56</td>
<td>164</td>
<td>SECOND</td>
<td>FEMALE</td>
<td>MOST TIME</td>
<td>neg</td>
</tr>
<tr>
<td>4</td>
<td>20</td>
<td>70</td>
<td>176</td>
<td>FIRST</td>
<td>MALE</td>
<td>ALWAYS</td>
<td>neg</td>
</tr>
<tr>
<td>5</td>
<td>28</td>
<td>67</td>
<td>172</td>
<td>HONS</td>
<td>MALE</td>
<td>ALWAYS</td>
<td>neg</td>
</tr>
<tr>
<td>6</td>
<td>24</td>
<td>61</td>
<td>168</td>
<td>FIRST</td>
<td>FEMALE</td>
<td>NEVER</td>
<td>pos</td>
</tr>
<tr>
<td>7</td>
<td>21</td>
<td>73</td>
<td>186</td>
<td>SECOND</td>
<td>MALE</td>
<td>SOMETIMES</td>
<td>pos</td>
</tr>
<tr>
<td>8</td>
<td>27</td>
<td>59</td>
<td>158</td>
<td>FIRST</td>
<td>FEMALE</td>
<td>NEVER</td>
<td>pos</td>
</tr>
<tr>
<td>9</td>
<td>29</td>
<td>76</td>
<td>186</td>
<td>SECOND</td>
<td>MALE</td>
<td>SOMETIMES</td>
<td>neg</td>
</tr>
<tr>
<td>10</td>
<td>28</td>
<td>61</td>
<td>170</td>
<td>THIRD</td>
<td>FEMALE</td>
<td>MOST TIME</td>
<td>neg</td>
</tr>
<tr>
<td>11</td>
<td>22</td>
<td>82</td>
<td>170</td>
<td>SECOND</td>
<td>MALE</td>
<td>ALWAYS</td>
<td>neg</td>
</tr>
<tr>
<td>12</td>
<td>24</td>
<td>63</td>
<td>168</td>
<td>FIRST</td>
<td>FEMALE</td>
<td>ALWAYS</td>
<td>neg</td>
</tr>
<tr>
<td>13</td>
<td>28</td>
<td>87</td>
<td>186</td>
<td>HONS</td>
<td>MALE</td>
<td>SOMETIMES</td>
<td>neg</td>
</tr>
<tr>
<td>14</td>
<td>25</td>
<td>76</td>
<td>176</td>
<td>FIRST</td>
<td>MALE</td>
<td>SOMETIMES</td>
<td>pos</td>
</tr>
<tr>
<td>15</td>
<td>27</td>
<td>79</td>
<td>172</td>
<td>THIRD</td>
<td>MALE</td>
<td>SOMETIMES</td>
<td>neg</td>
</tr>
<tr>
<td>16</td>
<td>24</td>
<td>57</td>
<td>166</td>
<td>SECOND</td>
<td>FEMALE</td>
<td>SOMETIMES</td>
<td>neg</td>
</tr>
<tr>
<td>17</td>
<td>29</td>
<td>81</td>
<td>180</td>
<td>THIRD</td>
<td>MALE</td>
<td>SOMETIMES</td>
<td>neg</td>
</tr>
<tr>
<td>18</td>
<td>21</td>
<td>69</td>
<td>168</td>
<td>FIRST</td>
<td>FEMALE</td>
<td>SOMETIMES</td>
<td>neg</td>
</tr>
<tr>
<td>19</td>
<td>37</td>
<td>55</td>
<td>169</td>
<td>HONS</td>
<td>FEMALE</td>
<td>ALWAYS</td>
<td>neg</td>
</tr>
<tr>
<td>20</td>
<td>28</td>
<td>91</td>
<td>182</td>
<td>FIRST</td>
<td>MALE</td>
<td>ALWAYS</td>
<td>neg</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>21</td>
<td>39</td>
<td>73</td>
<td>168</td>
<td>SECOND</td>
<td>FEMALE</td>
<td>NEVER</td>
<td>neg</td>
</tr>
<tr>
<td>22</td>
<td>24</td>
<td>89</td>
<td>178</td>
<td>THIRD</td>
<td>MALE</td>
<td>NEVER</td>
<td>pos</td>
</tr>
<tr>
<td>23</td>
<td>39</td>
<td>75</td>
<td>185</td>
<td>HONS</td>
<td>MALE</td>
<td>MOST TIME</td>
<td>neg</td>
</tr>
<tr>
<td>24</td>
<td>24</td>
<td>85</td>
<td>176</td>
<td>SECOND</td>
<td>MALE</td>
<td>MOST TIME</td>
<td>neg</td>
</tr>
<tr>
<td>25</td>
<td>22</td>
<td>68</td>
<td>166</td>
<td>SECOND</td>
<td>FEMALE</td>
<td>SOMETIMES</td>
<td>neg</td>
</tr>
<tr>
<td>26</td>
<td>20</td>
<td>89</td>
<td>178</td>
<td>SECOND</td>
<td>MALE</td>
<td>MOST TIME</td>
<td>neg</td>
</tr>
<tr>
<td>27</td>
<td>27</td>
<td>77</td>
<td>176</td>
<td>THIRD</td>
<td>MALE</td>
<td>SOMETIMES</td>
<td>neg</td>
</tr>
<tr>
<td>28</td>
<td>30</td>
<td>55</td>
<td>166</td>
<td>FIRST</td>
<td>FEMALE</td>
<td>NEVER</td>
<td>pos</td>
</tr>
<tr>
<td>29</td>
<td>23</td>
<td>81</td>
<td>170</td>
<td>SECOND</td>
<td>FEMALE</td>
<td>NEVER</td>
<td>pos</td>
</tr>
<tr>
<td>30</td>
<td>24</td>
<td>81</td>
<td>174</td>
<td>THIRD</td>
<td>MALE</td>
<td>SOMETIMES</td>
<td>pos</td>
</tr>
<tr>
<td>31</td>
<td>25</td>
<td>66</td>
<td>170</td>
<td>HONS</td>
<td>FEMALE</td>
<td>MOST TIME</td>
<td>neg</td>
</tr>
</tbody>
</table>
Random variables

When we determine the height or age, say, of a person, the result is a value of the respective variable. When the values are obtained as a result of chance factors, such as random selection, the values are values of a random variable.

Discrete and continuous random variables

A variable that can assume only certain values is called a discrete variable. Data described by a discrete variable is called discrete data. If a variable can assume only one value it is called a constant. Examples of discrete variables are, the number of children in a family, the number of goats on a farm. These can only assume values such as 0; 1; 2; ....... It is not possible for a family to have 2.6 children for example.

Other examples

Example

Number of defective parts in a sample of 5

\[ x = \{0, 1, 2, 3, 4, 5\} \]

Example

Number of correct answers in a multiple choice test paper consisting of 10 questions.

\[ x = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\} \]

Continuous ratio variables.

A variable that can assume any value between two given values is called a continuous variable. Data described by a continuous variable is called continuous data. Examples of continuous variables include sets of measurement data, such as lengths, masses, yields per hectare etc., the values re-
corded depend on the accuracy of measurement and the sensitivity of the measuring instrument. Thus the mass of a potato may be recorded as 73 g or 73.4 g or 73.4125 g etc., depending on the sensitivity and accuracy of the measuring apparatus.

In other words, a continuous variable can be defined as a random variable which can take on any value on a continuum, i.e. can assume any values within a specified interval.

**Note**

Observations which are inherently continuous are usually recorded as discrete.

**Example**

Measure height to nearest cm, weight to nearest gram.

**Parameter and statistic**

A parameter is a number that describes some property of a population, and normally Greek letters are used to denote population parameters. A number which describes some property of a sample is referred to as a sample statistic, or briefly as a statistic. Small letters from our usual (Roman) alphabet are normally used to denote sample statistics. Thus for the arithmetic mean (or average) of a population the symbol $\mu$ (mu) is used, and the symbol $\bar{x}$ (ex bar) is used to denote the arithmetic mean of a sample.

**Raw data**

Raw data are data that have not been organized numerically. Thus the numbers 17; 45; 38; 27; 6; 48; 11; 57; 34; 22; which may represent collected data, have not been organized numerically, i.e. either in ascending order (from smallest to largest), or in descending order (from largest to smallest).
Array

When raw numerical data have been arranged, e.g. in ascending or descending order of magnitude, they are said to form an array. Using the example of the raw data above, in descending order of magnitude the array is 57; 48; 45; 38; 34; 27; 22; 17; 11; 6. The reverse order would be an array in ascending order of magnitude.

Range

The range is defined as the difference between the largest and the smallest number in an array or in a raw data set. When the raw data has been arranged in ascending or descending order, it is very easy to identify the largest and the smallest number in the array and to calculate the range. For the example quoted, the range is $57 - 6 = 51$.

The subscript or index notation

Let the symbol $X$ denote a variable. Let $x_i$ denote a particular variate, i.e. the value of a particular member of $X$, in this case the first member of $X$.

The symbol $x_i$ denotes the value of $X$, when $i$ the subscript, assumes a particular value. The only values which the subscript can assume are the integers, 1; 2; 3; ...... Another word for subscript is the word, index.
The letters $i$ and $j$ are usually used as subscripts, but other letters, e.g. $k$, $p$, $q$, $s$ etc. can also be used as subscripts.

Consider the table.

Table 1.2

<table>
<thead>
<tr>
<th>col 1</th>
<th>col 2</th>
<th>col 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>row 1</td>
<td>$x_{i1}$</td>
<td>$x_{i2}$</td>
</tr>
<tr>
<td>row 2</td>
<td>$x_{i2}$</td>
<td>$x_{i2}$</td>
</tr>
<tr>
<td>row 3</td>
<td>$x_{i3}$</td>
<td>$x_{i2}$</td>
</tr>
<tr>
<td>row 4</td>
<td>$x_{i4}$</td>
<td>$x_{i2}$</td>
</tr>
</tbody>
</table>

In this table the first subscript represents the row and the second subscript the column. $x_{ip}$ represents the number in row $i$, column $p$.

**Sigma $\sum$ notation**

The symbol $\sum$ is shorthand for ‘the sum of’. Hence, the Greek capital letter $\Sigma$ is used to denote sum, which is the answer to an addition sum.

$$\sum_{i=1}^{3} x_i = x_1 + x_2 + x_3$$

In words ‘the sum of $x_i$ for $i$ going from one to three is’ $x_1 + x_2 + x_3$.

This may be written more simply as

$$\sum_{i=1}^{3} x_i = x_1 + x_2 + x_3$$ or more generally as
\[ \sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + \ldots + x_n. \]

The notation may be simplified even further, e.g. \( \sum x \) denoting the sum of all \( x \) values.

An extension to the sigma notation concept

As an extension to the concept, note

\[ \sum_{i=1}^{n} x_i y_i = x_1 y_1 + x_2 y_2 + \ldots + x_n y_n \]

which may simply be written as \( \sum xy \). Note that this is not the same as \( \sum x \sum y \), which means the sum of \( x \) multiplied by the sum of \( y \).

Furthermore consider a table with figures arranged in rows and columns, with \( r \) rows and \( c \) columns.

\[
\begin{array}{cccc}
  x_{11} & x_{12} & \cdots & x_{1c} \\
  x_{21} & x_{22} & \cdots & x_{2c} \\
  \vdots & \vdots & \ddots & \vdots \\
  x_{r1} & x_{r2} & \cdots & x_{rc} \\
\end{array}
\]

\[ \sum_{i=1}^{r} \sum_{j=1}^{c} x_{ij} = x_{11} + x_{21} + \ldots + x_{y} + \ldots + x_{rc}. \]
which represents the sum of all the figures in the table. In the case where $c$ is a constant,

$$
\sum_{i=1}^{n} c_i = c_1 + c_2 + \ldots + c_n = nc,
$$

since every $c_i$ equals the constant $c$ because the value of $c$ is constant and does not change with subscript. For this reason the subscript associated with the constant is generally omitted.

In the following where $c$ is a constant

$$
\sum_{i=1}^{n} cx_i = c(x_1 + x_2 + \ldots + x_n) = c\sum_{i=1}^{n} x_i
$$

Thus

$$
\sum_{i=1}^{n} cx_i = c\sum_{i=1}^{n} x_i
$$

Consider

$$
\sum_{i=1}^{n} (x_i + y_i) = (x_1 + y_1) + (x_2 + y_2) + \ldots + (x_n + y_n)
$$

collecting the $x$ and $y$ values separately.
\[ \sum_{i=1}^{n} (x_i + y_i) = (x_1 + x_2 + \ldots + x_n) + (y_1 + y_2 + \ldots + y_n) \]

Thus \[ \sum_{i=1}^{n} (x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i \] and

where there is no doubt as to what is meant, the above is written as
\[ \sum x + \sum y. \]

In the same way it can be shown that
\[ \sum_{i=1}^{n} (ax_i + by_i + c) = a \sum_{i=1}^{n} x_i + b \sum_{i=1}^{n} y_i + c \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i \] and

\[ \sum_{i=1}^{n} (ax_i + b) = a \sum_{i=1}^{n} x_i + nb. \]

Similarly \[ \sum_{i=1}^{n} (ax_i - b) = a \sum_{i=1}^{n} x_i - nb. \]

Example

Calculate a value for \[ \sum_{i=1}^{n} (x_i + y_i)^2 \] given that \[ \sum_{i=1}^{n} x_i^2 = 385; \]

\[ \sum_{i=1}^{n} y_i^2 = 5540 \] and \[ \sum_{i=1}^{n} x_i y_i = 770. \]
Solution
\[ \sum_{i=1}^{12} (\frac{1}{2}x_i + y_i)^2 = \sum_{i=1}^{12} (\frac{1}{2}x_i + y_i + y_i^2) \]
\[ \sum_{i=1}^{12} x_i^2 + \sum_{i=1}^{12} x_i y_i + \sum_{i=1}^{12} y_i^2 = 565 + 1540 + 1540 = 3645 \]

hence \[ \sum_{i=1}^{12} (\frac{1}{2}x_i + y_i)^2 = 3645 \]

Suppose we have observations \( Y_1, Y_2, \ldots, Y_n \)

We use the notation
\[ \sum_{i=1}^{n} Y_i \] to denote \( Y_1 + Y_2 + \ldots + Y_n \)
\[ \sum_{i=1}^{n} Y_i^2 \] to denote \( Y_1^2 + Y_2^2 + \ldots + Y_n^2 \)

If \( f \) is a function, then
\[ \sum_{i=1}^{n} f(Y_i) \] denotes \( f(Y_1) + f(Y_2) + \ldots + f(Y_n) \)
Example

\[ y_1 = 3, \quad y_2 = 1, \quad y_3 = 2 \]

\[ y_4 = 3, \quad y_5 = 5, \quad y_6 = 4 \]

Find i) \( \sum_{i=1}^{6} y_i \)  
    ii) \( \sum_{i=4}^{6} y_i \)  
    iii) \( \sum_{i=1}^{4} y_i^2 \)  
    iv) \( \left( \frac{\sum_{i=1}^{6} y_i}{6} \right)^2 \)

Solution

i) \( \sum_{i=1}^{6} y_i = 14 \)  
ii) \( \sum_{i=4}^{6} y_i = 12 \)  
iii) \( \sum_{i=1}^{4} y_i^2 = 237 \)  
iv) \( \left( \frac{\sum_{i=1}^{6} y_i}{6} \right)^2 = (14)^2 = 324 \)

\[ \left( \frac{\sum_{i=1}^{6} y_i}{6} \right)^2 = (14)^2 = 324 \]

Example

If \( f(y) = 2y - 1 \), find

\[ \sum_{i=1}^{6} f(y_i) \]  
[Solution: 30]

The subscript \( i \) is called a dummy variable. Any other subscript could be used.

\[ \sum_{i=1}^{5} y_i \]  
\[ \sum_{i=1}^{6} y_i \]  
\[ \sum_{i=1}^{7} y_i \]

The subscript is also called the index variable.
Improving the quality of survey data

Objectives

By the end of this chapter you should be able to:

· have an understanding of the basics of survey design
· know how to improve your survey instrument

by Professor Denise Lievesley

Unesco Institute of Statistics

Introduction to sampling terminology

*Population* – the group of items for which inferences are to be drawn.

*Epsem* – equal probabilities of selection method of drawing a sample.

*Sampling error* – the error which arises because a sample was drawn rather than taking a complete census.
The basic principles of sampling

A sample is a small scale representation of the population. However because it contains only a small part of the population not the whole of it, it can never be an exact replica.

How closely the sample reflects the population depends crucially on the size of the sample and the way it was selected. The securest basis for sample selection is chance although we build in constraints such as stratification.

A survey is subject to both sampling and non-sampling error whereas a census is subject only to non sampling error, yet because sampling error may be small and is usually calculated a survey may sometimes be more accurate.

Suppose we wish to estimate the average pocket money of boys in a particular school. We could draw a sample of boys and find out their pocket money. The average of the sample would be our best estimate of the population average in the absence of other information.

But another sample would not necessarily give the same result. If we continued drawing fresh samples and plotted the results we would find a pattern emerging – the distribution of the sample estimates would approximate to a normal (i.e. bell shaped) curve.

This is known as the Central Limit Theorem.

The normal distribution has certain fixed qualities which are useful to us. We can measure the standard deviation of this sampling distribution and we know that 95% of the observations lie within two standard deviations
of the mean. It follows that there are 95 chances out of a 100 that the mean of a particular sample chosen at random will be within 2 standard deviations of the true population mean. But we have only one sample. We assume that the sample value is the best estimate of the population value and we estimate the probability of being correct.

**Random sampling**

Every member of the population has a known and non-zero probability of selection.

**Simple random sampling**

Every member of the population has the same chance of selection, and every combination of ‘n’ members has the same chance of selection, (e.g. rotating urn, random number tables).

In thinking about sampling error it is conventional to define a ‘true value’ for each individual in the population. This concept was first introduced by Hansen, Hurvitz and Madow in 1953. This true value must be independent of the conditions under which the survey is carried out. The term *individual response error* is used to denote the difference between the individual true value and the information recorded for that individual. The overall response error is made up of all of the individual response errors and will comprise two main components: *random error* and *systematic error*.

Suppose one could carry out repeats of the survey under the same essential conditions (i.e., using the same sample design but drawing a fresh sample, using the same data collection procedures but doing the survey afresh etc.). Each repeat would produce a different value because of *RANDOM ERROR* (also known as *variable error*). These can be thought of as ‘random’ fluctuations about the expected value under the same essential survey conditions. If the survey estimates were averaged over a large number of repeats of the survey then the random error would disappear since it averages to zero.
Part of the random error arises because a sample was taken rather than a census and this is termed *sampling error*. It does not comprise all of the random error since if we carried out repeats of a census we would still obtain results which varied from repeat to repeat because of, for example, mistakes in answering the questions or in recording the answers.

The difference between the expected value (i.e. the average over a large number of repeats) and the true value is known as the *systematic error* or *bias*. The bias takes the same constant value in each hypothetical repeat of the survey. Biases can arise in many ways, for example, because of non-response, sampling frame deficiencies, leading questions, inadequate coding frames. Random and systematic errors differ in their implications for the survey analyst: Systematic errors are constants which cannot be measured from within the survey – it is necessary to have other data in order to make estimates of the systematic errors. On the other hand, random errors can in principle be estimated from the survey itself. Systematic errors cannot be reduced by increasing the sample size, whereas sampling errors can be reduced by increasing the sample size or by increasing the number of clusters in a multi-stage design, and other random errors can be reduced by increasing the number of other units such as coders or interviewers (this will be explained more fully later).

Since bias is not reduced by increasing the sample size, whereas random error is, bias becomes relatively more important as sample size increases. Bias and random error are not related to one another. One can have a large bias and a small random variability or vice versa.

One should never talk about a biased sample since bias is a feature of the method used rather than a particular outcome. So what do we mean by an unbiased survey method?

Having decided upon a particular sample size and a particular method if we drew a repeated number of samples and then were to take an average of all the different sample estimates of the population characteristic, we would get the right answer. This is what we mean by an *unbiased method* – the
result of a process whose long-run outcome is exactly equal to the population value being estimated.

Note that a perfectly acceptable selection method can yield an unbalanced sample – one that differs in profile in some respect from the population. This needn’t be bias – it could be the luck of the draw and it is an observable incidence of sampling variance. So the aims of sampling are to avoid bias and reduce random variability all within fixed resource constraints.

**Precision and accuracy**

The accuracy of a measurement signifies the closeness with which the measurement approaches the true value whereas precision is simply concerned with the repeatability of the measurement. In other words the precision is concerned with random errors whereas accuracy also covers bias.

- simple random samples – are ones in which every member of the population has the same chance of selection.
- proportionate stratification – uses information about the population to ensure that the sample correctly reflects the population distribution.
- disproportionate stratification – uses information about the population structure but deliberately selects the sample to have a different distribution.

Clustering – the method by which a sample is drawn in clusters in order to reduce traveling costs of interviewers.

**Non-Response is of two types:**

*Unit non-response*

When it is not possible to collect any survey data from or about a particular member of the sample.
Item Non-Response

When a co-operation member fails – or refuses – to provide some specific items of information.

Unit non-response is attributable to four causes:

- fieldwork shortfalls
- sample members not contacted
- sample members unwilling to participate
- sample members unable to participate

Ways in which surveys are used

- descriptive background data
- illumination – strategic research
- pressure for change – tactical research
- monitor or evaluate policy
- public involvement
- experimental data
- diagnostic data
- political advocacy
- delaying tactic
Alternatives to surveys

- censuses
- desk research
- administrative data
- case studies
- observation
- consultation

The choice of a research design should take into account

- Will the information be valid?
- Can we ask the relevant questions in relevant ways?
- Will the information be reliable? Will it be trustworthy, consistent and stable?
- Would we get the same results if we did the study again?
- How precise will the information be?
- How much risk due to sampling error is involved and how might this affect the use of the data for making decisions?

Questions which have to be asked before every survey

- Can the information generalize?
- Which population and conditions will the information describe?
In what ways might the data be biased?

How are human errors likely to be introduced – by faulty questions, weak procedures, etc?

How accurate will the information be?

What is the predicted total amount of error due to sampling and due to the biases introduced by interviewers and other factors in the survey?

Will we be able to extrapolate the information?

How safe will it be to use the information to describe the characteristics of populations or conditions that are different from those we have studied?

• Do we have sufficient resources?

• Are skilled staff, supplies and facilities adequate to do a worthwhile survey?

• Will we get institutional permission to do the study?

• Will we get ethical clearance clearance to do the study?

• Is permission necessary before we go into the field?

• How useful will the results be?

• Who needs the information and precisely how will they use the results to help them make decisions?
Types of survey errors

random errors
systematic errors or biases

Random errors can be thought of as random fluctuations about the expected value under the same essential survey conditions.

the survey estimates were averaged over a large number of repeats of the survey then the random error would disappear but the difference between the average and the true value is the systematic error or bias.

Part of the random error arises because a sample was taken rather than a census and this is termed sampling error.

Sampling error does not comprise all of the random error as if we conducted a census we would still obtain results which varied because of for example, mistakes in asking questions or recording answers.

Accuracy concerned with the closeness with which the measurement approaches the true value whereas precision is simply concerned with the variability of the results.

Sources of error

Errors arising at the selection stage

• sampling error

• deficiencies in the sampling frame

• selection inaccuracies and bias unit and item non-response
Errors arising at the data collection stage

Questionnaire induced

- length of questionnaire – the burn-out factor
- order of questions
- structure of questionnaire
- wording of questions
- inclusion of stressful questions

Interviewer induced

- interviewer’s characteristics
- interviewer’s expectations (role, attitude and probability expectations)
- interviewer’s way of asking questions or recording answers
- interviewer cheating

Respondent induced

- lack of knowledge
- memory difficulties
- misunderstanding the question

Will we get clearance to do the study?

- response set/boredom
- lack of motivation
• desire to mislead
• desire to please

**Environment induced – or bias of the auspices**

• mode of data collection
• place of interview
• presence of another person
• topicality of subject
• timing of the interview
• sponsorship of the survey
• confidentiality of the results
• inducement

**Errors arising at the data processing stage**

• coder errors
• inadequate coding frame
• incorrect editing
• errors in data transfer

**Errors arising at the analysis stage**

• incorrect specification of new variables
• incorrect specification of weighting or imputation adjustments
• inappropriate use of analysis techniques
incorrect inference or interpretation

This is not an exhaustive list of errors – others will arise and errors will be of different levels of importance depending on the survey. Errors will also vary in their effect within a survey.

Fieldwork shortfalls

It is usually possible to keep fieldwork shortfalls to a very low level by designing samples which take account of operation requirements. There may be a few sample members not attempted because of their inaccessibility, extreme weather conditions, unavailability or interviewers etc.

Sample members not contacted

The proportion of people who are not contacted by the end of the field period varies from survey to survey. It depends crucially upon the resources available for the survey (which determines an interviewer’s workload, and therefore the number and timing of calls upon sample members) but also depends upon the population being surveyed.

Sample members unwilling to participate

In most surveys people have a legal and moral rights not to take part, and so invariably some people refuse to co-operate. Survey organizations need to achieve a difficult balance, in minimizing refusals whilst respecting people’s right to refuse, and bearing in mind errors of reporting which might result from enforced participation and the costs involved in reducing refusals. “no one shall be subjected to arbitrary interference with his privacy….everyone has the right to the protection of the law against such interference” U.N. Declaration of Human Rights.
Sample members unable to participate

Most samples include some people too infirm, deaf or unfamiliar with the language to be interviewable. Interviewers need to be given instructions as to what should be counted as “uninterviewable” and whether proxy information or interviews via an interpreter would be acceptable. In practice the distinction between these categories is rather blurred. It may be difficult to distinguish unwillingness from inability or incapacity.

It is important to collect information on the characteristics of non-respondents for the following reasons:

- to help counter non-response and pinpoint remedial action
- to tell whether results are likely to be subject to bias and so what priority to place on reducing non-response.
- to enable adjustments to be made to compensate for non-response bias.

There is the added benefit that “knowledge of the distorting effects of non-response bias.

There is the added benefit that “knowledge of the distorting effects of non-response in probability sampling can be used in order to reveal indirectly the risk of bias in quota sampling” (Dalenius 1961), which is especially important since non-response is rarely documented in non-probability samples.

Data on non-respondents can be categorized as follows:

- aggregate population data
- individual level data – external sources
- individual level data – internal sources
- wave data
Reasons for refusal to an attitude survey
on issues of current importance

<table>
<thead>
<tr>
<th>Reason</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cannot be bothered</td>
<td>18</td>
</tr>
<tr>
<td>Disliked the subject of this survey</td>
<td>16</td>
</tr>
<tr>
<td>It was an inconvenient time</td>
<td>18</td>
</tr>
<tr>
<td>Too busy to take part in surveys</td>
<td>5</td>
</tr>
<tr>
<td>Too old/not educated enough</td>
<td>11</td>
</tr>
<tr>
<td>Surveys are a waste of time and money</td>
<td>18</td>
</tr>
<tr>
<td>Never see the results/nothing happens</td>
<td>6</td>
</tr>
<tr>
<td>Surveys are inaccurate/do not believe in sampling</td>
<td>3</td>
</tr>
<tr>
<td>Worries about selling under guise of research</td>
<td>3</td>
</tr>
<tr>
<td>Worries about confidentiality/invasion of privacy</td>
<td>6</td>
</tr>
<tr>
<td>Refused to give any information on reasons for refusal</td>
<td>10</td>
</tr>
<tr>
<td>Non-contacted by follow-up interviewer</td>
<td>8</td>
</tr>
</tbody>
</table>
**Follow-up interview with refusers (as a %)**

<table>
<thead>
<tr>
<th>Response</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes, would participate</td>
<td>22</td>
</tr>
<tr>
<td>Yes, might participate</td>
<td>21</td>
</tr>
<tr>
<td>Follow-up interview thinks they could be persuaded</td>
<td>9</td>
</tr>
<tr>
<td>Refusal was specific to this survey:</td>
<td></td>
</tr>
<tr>
<td>Might take part in another survey</td>
<td>14</td>
</tr>
<tr>
<td>No, would no participate</td>
<td>20</td>
</tr>
<tr>
<td>Insufficient information to classify</td>
<td>13</td>
</tr>
<tr>
<td>(Base: 111 interviews)</td>
<td>100</td>
</tr>
</tbody>
</table>

Conversion of refusals to an interview. Categorised by the initial interviewer’s assessment of the likelihood of doing so.
Interviewer’s assessment       %    base
Very likely/likely            72    22
Possibly                     50    43
Unlikely                    33    93
Cannot say                  23    99
Total                       28    57
% giving an interview on recall 34 314

*General household survey*

Response rates by household size

<table>
<thead>
<tr>
<th>Response Rate</th>
<th>Base</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 persons</td>
<td>82</td>
</tr>
<tr>
<td>4 persons</td>
<td>89</td>
</tr>
<tr>
<td>5 persons</td>
<td>92</td>
</tr>
<tr>
<td>All households</td>
<td>76</td>
</tr>
</tbody>
</table>

*Relationship of census characteristics to response rates*

<table>
<thead>
<tr>
<th>Correlation with response rates</th>
<th>$r^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>non-manual workers</td>
<td>-0.78</td>
</tr>
<tr>
<td>shared dwellings</td>
<td>-0.75</td>
</tr>
<tr>
<td>dwellings without a bath</td>
<td>-0.09</td>
</tr>
<tr>
<td>semi-skilled workers</td>
<td>0.41</td>
</tr>
</tbody>
</table>
furnished rented accommodation  -0.52  0.848
fertility  0.64  0.877
children 0-4 years  0.4  0.907

Little evidence of a predominant refuser category. Considerable evidence that the biases due to refusals and non-contacts are OFF-SETTING. Some evidence that particular groups are consistently under-represented in survey research:

- recruitment
- retraining
- feedback
- greater control
- re-issuing non-response

The interviewer’s tasks “on the doorstep”

- to make contact at the address
- to select and establish who is the person to interview
- to contact that person (if not the original contact)
- to persuade the selected person to take part
- to put over information which we consider members of the public approached for interview should be given.

What do we do about non-response?

Reduce it

Attempt to estimate the bias due to non-response:
by direct methods
by indirect methods
by utilizing cell/wave data
adjust the data to ‘correct’ for non-response.

Assume refusers and non-contacts who received N calls have the same characteristics as those who responded to the Nth call.

Assume people who respond at the 2nd call are a random sample of all people not contacted at the 1st call. Bartholomew.

Examination of the rate of convergence on the population characteristics over subsequent call backs. Dunkelberg and Day plus Politz Simons method.

Non response error
error is a function of the percentage not responding to the survey and the difference on the statistic between respondents and non-respondents.

Informed consent

- purpose of study
- identity of funders
- anticipated uses
- publication plans
- data storage plans
- identity of interviewer
- organization conducting survey
<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>method by which sample member has been chosen</strong></td>
<td></td>
</tr>
<tr>
<td><strong>respondent’s role in study</strong></td>
<td></td>
</tr>
<tr>
<td><strong>possible harm to respondent</strong></td>
<td></td>
</tr>
<tr>
<td><strong>degree of anonymity/confidentiality</strong></td>
<td></td>
</tr>
<tr>
<td><strong>time required</strong></td>
<td></td>
</tr>
<tr>
<td><strong>whether participation is compulsory or voluntary –</strong></td>
<td></td>
</tr>
<tr>
<td>if compulsory potential consequences of non-compliance, -if voluntary</td>
<td></td>
</tr>
<tr>
<td>entitlement to withdraw consent</td>
<td></td>
</tr>
<tr>
<td><strong>whether material facts have been withheld and</strong></td>
<td></td>
</tr>
<tr>
<td>other relevant issues.</td>
<td></td>
</tr>
</tbody>
</table>

**The ideal respondent**

- consistent
- rational
- used to examinations and to answering questions
- used to listening
- well disciplined
- answers clearly
- has minimal formal education
- is below interviewer in social status
- always does what he says he will do
- is always at home
Qualitative techniques

Qualitative techniques are so called because they can be used to identify the range of behaviours or attitudes but cannot give a basis for saying how many people behave in certain ways or hold particular views. Qualitative research tends to be small scale. Its unstructured nature allows the full exploration of attitudes and behaviours without detailed preconceived notions as to what is important.

There are two main types of qualitative research:

Depth interviews – respondents are interviewed on their own by skilled interviewers. No questionnaire is used but a guide to topics is provided. In depth interviews respondents can talk at length in their own words on individual topics giving as much or as little information as is relevant to them. They allow the interviewer to seek as much clarification or amplification as (s)he feels is necessary.

Focus group discussions – involve a group of perhaps 8-10 respondents. The group facilitator guides the discussion amongst the respondents using a list of topics. The guide will usually list the classes of information to be covered but the detail in the guide depends upon the purpose of the study and the extent to which the depth interviews/group discussions are exploratory.

When is qualitative research appropriate?

- exploratory research as a preliminary to a quantitative survey to help design the content of the questionnaire
- follow-up research to a quantitative survey to obtain more details about the attitudes of particular subgroups or on particular topics of the problem
on their own – small scale investigations, sensitive research, case-studies

**Depth interviews**

**Advantages**
- Since only one respondent is involved at a time in depth interviews they facilitate a fuller exploration of the links between individual circumstances, experiences, motivations and behaviour. Useful when the object is to develop hypotheses about individual motivations and when the psychology and circumstances of individuals need to be related.
- More appropriate for certain sorts of sensitive topics e.g. contraception, financial

**Disadvantages**
- Depth interviews are more costly and time consuming than group discussions.
- The cost and time involved in obtaining and analyzing transcripts can be very high.
- The commitment by respondents has to be substantial and the requirements for good interviewing are also high.
- Depths interviews can be threatening.

**Group discussions or focus groups**

**Advantages**
Quicker and cheaper to organize than depth interviews with the same number of respondents. Focus groups can provide a lot of general information about behaviour and attitudes and can allow researchers to see how
people interact on a topic. The chance to contrast views stimulates discussion. Can be useful to collect information on some types of sensitive topics, by discussing them with similarly placed people. Groups can be creative and can develop ideas.

**Disadvantages**

provides less detailed information than individual interviews. In particular it will not provide detailed information about an individual’s characteristics and background than can be related to his or her behaviour and attitudes.

respondents tend to be more self-selected and certain groups of individuals will be difficult to recruit, e.g. very old people, senior businessmen.

**Aspects of the interviewer’s role in qualitative research**

to be familiar with the purpose of the survey and to be able to decide on relevance or otherwise of particular issues
to get the respondent talking freely and openly on the topics with which the study is concerned
to evaluate the completeness or adequacy of respondent’s answers
to probe for amplification or clarification if necessary
to be a receptive attentive listener
to be neutral.

**Additional requirements of the group leader**

are to ensure that:
everyone has a chance to talk
don dominant respondents are controlled
the respondents talk one at a time
the group does not digress from relevant topics
differences between respondents are brought out.

**Barriers to communication**

Free and open communication may be impeded by:
- attempts to rationalize
- lack of awareness
- fear of being shown up
- over-politeness
- anxiety about what will happen to the information.

**Qualitative interviewing techniques**

**Questioning**
- asking questions which are unambiguous, non-directive and stimulating to start the respondent off on a topic or on an aspect of a topic
- using probes to amplify and to clarify the respondent’s answers
- asking about one thing at a time
- asking about both sides of an issue
- asking questions to distinguish between feelings, beliefs and evaluations.
Perceptual techniques

- using silences and expectant looks to encourage respondents to talk
- listening and appraising
- remembering what has been said
- sensitivity to the respondent.

The interviewer’s task

- Finding the person – quota filling or in random sampling checking address, named person, making selection, identifying person fulfilling role
- Persuading the person to take part and dealing with reluctance or suspicion, answering queries
- establishing ‘rapport’ and explaining the interview procedure, organizing and controlling the situation in a professional and competent way (relaxed and friendly but not too intimate)
- conducting the interview by asking the questions correctly, recording or ending the answers accurately and full, proving and explaining in but not biasing responses
- keep full and accurate records to enable the work to be checked and to ensure that all sample members have been attempted the specified number of times.

Recruitment, training and quality control

The quality of the interviewing is as vital as the design of the questionnaire and it is thus essential that the people recruited are suitable and that they are fully trained and their work monitored.
Training

This should be a combination of formal training and experience with feedback. Interviewer training practices vary considerably despite some standardization within the social and market research industry.

Quality control

- Postal and telephone checks
- Early work checks
- Progress chasing
- Monitor response rates
- Tabulate edit failures by interviewer
- Use of bonus payments and grading systems
- Interviewer training, retraining.

General principles of questionnaire design

To be an efficient data collection tool a questionnaire must:

- Help maintain respondent interest, their cooperation and involvement
- Communicate to the respondent unambiguously and in language they understand
- Help respondent work out his/her answer by indicating in what terms the response is required
- Make it easy for the interviewer to ask the correct questions and record accurate answers.
Order and format of the questionnaire

Overall principle to make it as easy and enjoyable as possible for the respondent, the interviewer is next in the list of priorities followed by the analysts. Remember that a poor layout leads to errors in asking the questions and recording the answers. Use a large clear type and leave plenty of space. Without external validation we do not know which are the correct answers and we are often in the position of simply knowing that differences occur according to the question form.

Respondent’s concerns

- willingness and motivation
- understanding of concepts and scope
- comprehension of response task
- ability to answer

Interviewer’s concerns

- easy to read and administer
- full guidance on procedures

Quality improvements

ref. Charlie Cannell’s work on contingent feedback
Eleanor Singer’s work on quality contracts with respondents

Types of questions

- classification
• factual
• behavioural
• motivation/reason
• attitudinal/opinion
• knowledge
• hypothetical
• judgemental/perceptual

Issues with factual questions
• definition (complexity and acceptability)
• communication
• accessibility (memory – recall loss and telescoping)
• social desirability and undesirability bias

Problems of generalizing about behaviour
This may require respondents to provide difficult information
  usually, most recent behaviour
  behaviour during a specific time period
  frequency information about behaviour

Problems of memory
• time contracts in memory
minor events may be overlooked or misplaced in time

- a lack of commitment on the part of the respondent to make the effort to remember
- lack of time to think and remember
- lack of awareness of the level of detail required.

Memory effects are related to saliency

**Dimensions of saliency**

- usualness
- economic/social consequences
- continuing effects
- reminders and rehearsals
- length of time exposed
- plus length of time since the event, i.e. recall period.

The event may never have passed into the long term memory, may be distorted by intervening events, and the context in which the event occurred may have changed. (memory is connective, associative). Successful recall depends upon ease of recall and the effort the respondent is prepared to make. Can we manipulate these?

Although accuracy of recall is affected by the length of time since the event with the recent events recalled best, the recall is fairly stable (i.e. it does not decline rapidly with time). Pleasant events are better recalled than unpleasant events, and both positive and negative better than neutral. Women are better than men at recalling personal events. The better educated are more accurate than others.
Recall is best in face-to-face interviews, then mail and then telephone. Intervening events can distort. Aided and bounded recall help. So do cues and prompts. Telescoping is improved by bounded recall and by the temporal frame of reference. Landmark events can help to reduce telescoping. Skilled intervention can be very helpful by the interviewer linking the event about which recollection to a salient event which occurred at around the same time.

**Asking about behaviour**

1) Make sure all reasonable alternatives are included
2) Aided recall
3) Make question specific
4) Decide time period – make it relevant if possible
5) Use bounded recall
6) Encourage reference to records
7) Use diary methods

**Behaviour in a given period**

1) Where you are going to ask about behaviour in a given period, plan preliminary questions to disarm tendency to extend the period for inclusion of such activity, for example:

   Ask about the last four weeks before asking your question about the last seven days

   Fix the furthermost limit of the seven days period before asking about behaviour in that period.
2) Similarly, plan to disarm any unwarranted tendency to answer in terms of the ‘usual’, for example by asking about usual behaviour before asking about behaviour in a specific recent period.

3) Do not combine the response distribution from open and closed question systems.

4) Make your questions as direct as possible. Where indirect questions or indices are used, see that they have been validated before putting any faith in the implied meaning of the responses they produce.

5) Preferably treat data comparatively rather than in an absolute sense. Thus applies particularly to rating scales.

**Under-reporting of hospitalizations (2)**

for the past year by no. of weeks between hospital discharge and interview

<table>
<thead>
<tr>
<th>no. of weeks, no. of hospitalizations (from records)</th>
<th>% not reported in interviews</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-10</td>
<td>114 3</td>
</tr>
<tr>
<td>11-20</td>
<td>426 6</td>
</tr>
<tr>
<td>21-30</td>
<td>459 9</td>
</tr>
<tr>
<td>31-40</td>
<td>339 11</td>
</tr>
<tr>
<td>41-52</td>
<td>364 16</td>
</tr>
</tbody>
</table>

**Estimated victimization rates**

on bounded and unbounded recall in U.S.A. national crime survey:

<table>
<thead>
<tr>
<th>Total personal victimization rates per 100 people</th>
</tr>
</thead>
<tbody>
<tr>
<td>period unbounded bounded % difference</td>
</tr>
</tbody>
</table>
Comparison of publications looked at “yesterday”

<table>
<thead>
<tr>
<th></th>
<th>% open response</th>
<th>% recall ratio (OR/CL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Daily Mail</td>
<td>16</td>
<td>12</td>
</tr>
<tr>
<td>The Times</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Daily Herald</td>
<td>9</td>
<td>5</td>
</tr>
<tr>
<td>Evening News</td>
<td>39</td>
<td>27</td>
</tr>
<tr>
<td>Daily Mirror</td>
<td>46</td>
<td>39</td>
</tr>
<tr>
<td>Daily Star</td>
<td>21</td>
<td>12</td>
</tr>
<tr>
<td>Daily Telegraph</td>
<td>17</td>
<td>13</td>
</tr>
<tr>
<td>Av. No. endorsed per person</td>
<td>1.52</td>
<td>1.12</td>
</tr>
</tbody>
</table>

Hints on asking threatening questions about behaviour

1) use long questions
2) use open questions
3) use familiar words
4) ask about others
5) deliberately load questions
6) socially undesirable to ask about the past
7) socially desirable to ask about the present
8) embed in less threatening topics
9) randomized response card sorting
10) use diaries
11) avoid reliability questions and check.

Short

Example: “What health problems have you had in the past year?”

Long

Example: “The next question asks about health problems in the last year. This is something we ask everyone – what health problems have you had in the past year?”

Instructions, contingent feedback and commitment

<table>
<thead>
<tr>
<th></th>
<th>control</th>
<th>exp. group</th>
<th>increase</th>
</tr>
</thead>
<tbody>
<tr>
<td>number of items to 16 open qs</td>
<td>64.1</td>
<td>79.3</td>
<td>24</td>
</tr>
<tr>
<td>index of precision of dates of drs visits</td>
<td>.60</td>
<td>.82</td>
<td>.37</td>
</tr>
</tbody>
</table>
checking of outside sources of info. . 70 2.29 183
no. of items reported for “pelvic region” . 63 .87 26

Pitfalls over-abstract concepts, complex tasks, double negatives, attitudes can be many faceted, intensity of views can vary substantially, measurements are very sensitive to question wording.

**Issues:**

- number of scale points
- use of a midpoint
- use of no opinion/DK
- yea saying
- question order effects

Don’t take widespread knowledge for granted. People tend to answer even when they do not have a view or have not thought about an issue, so check knowledge when asking about opinion.

**Asking attitudinal questions**

Open vs. closed questions (also field coded questions)

- allowances for DK response
- provision of middle response option
- balanced vs unbalanced questions
- order effects.

**Advantages of open questions**

- can cover all possible answers
less risk of bias from framework imposed by researcher

greater chance of identifying genuine attitudes

appropriate when respondent has not fully formulated his/her views on a subject

monitor effects of rapidly changing events avoid missing newly emergent categories

useful if set of categories very large
documenting absence of a response

**Advantages of closed questions**

provided respondents with a standard frame of reference

less risk of bias from interviewer (?)

no coding problems

quicker to administer

allows more questions per interviewer

easier for interviewer to administer

lower cost of interviewing and data preparation

less risk of bias from ‘articulate’

**Most important problem facing the U.S.A.**

<table>
<thead>
<tr>
<th></th>
<th>closed%</th>
<th>open%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crime violence</td>
<td>35</td>
<td>16</td>
</tr>
<tr>
<td>Unemployment</td>
<td>20</td>
<td>19</td>
</tr>
</tbody>
</table>
Inflation 13 13

**Use of ‘no opinion’ and ‘don’t know’ categories**

- discouraging use (no provision)
- allowing use (option)
- encouraging use (filter)

**Results**

Option 2 produces increase in use of No/DK category on general attitude questions (perhaps + 20%) Option 3 even bigger increase. Frequently repercentaging excluding No/DK gives same results – but not always. No./DK is used more frequently by elderly/less educated.

<table>
<thead>
<tr>
<th></th>
<th>Standard %</th>
<th>Filtered %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agree</td>
<td>49.9</td>
<td>39.2</td>
</tr>
<tr>
<td>Disagree</td>
<td>43.9</td>
<td>23.1</td>
</tr>
<tr>
<td>DK (volunteered)</td>
<td>15.2</td>
<td>37.6</td>
</tr>
</tbody>
</table>

**Types of order effect**

*Saliency* series of specific questions increase awareness of the issues and affects answers to later questions.

*Consistency* when questions are on related topics respondents may feel need to be consistent.
Contrast when specific question followed by general question on same topic, the specific question provides a contrast which may alter the meaning of the general question.

**General Question**

‘Do you think a woman should be allowed to have an abortion if she, and her husband want one?’

**Specific Question**

……..and there is a strong chance of a defect in the baby?’

% saying ‘yes’ to General Question when asked before Specific Question 63%

when asked after Specific Question 51%

% saying ‘yes’ to Specific Question is not affected by order – at over 80%

**Mode of data collection**

- face-to-face
- telephone
- mail
- self-completion
- computer aided interviewing

**Mail surveys**

**Advantages**

- cost (cheap way of screening)
- access (awaits respondent)
Disadvantages
response rate
limit to scope and length
no control over who completes it
no control over question order
cannot check knowledge

Advantages or disadvantages?
time to think, consult records and others
absence of an interviewer

How can response be improved?

it is not just another circular
we are prepared to nag
we are in a hurry
using incentives (what type and when should they be sent?)
covering letter important – convey importance and urgency, assure confidentiality

personalisation
get high status person to sign the letter and use relevant letterhead

Dillman’s concept of personalization “the process of creating the belief on the part of the respondent that he is receiving the researcher’s individual attention”.

How does Dillman recommend we achieve personalisation?
by the use of multiple follow-ups using mail, telephone and telegrams
by stressing the importance of his or her personal participation
by the use of names rather than for example ‘the occupier’
by individually typed letters individual salutations
hand applied signatures
first class postage

Questionnaire appearance is critical

test questionnaire
no interviewer present to help poor design
complicated routing mwithin the questionnaire ust be avoided
uncluttered design important as is space for answers
logical progression encourages systematic responses
avoid jargon and use clear unambiguous questions
make task familiar, e.g. tick boxes
questionnaire a learning procedure so keep early questions simple
- include high interest questions to encourage a return

To what extent does length matter?

**Can open material be collected in mail surveys?**

Those who respond tend to be

- favourably disposed towards the surveys
- politically and socially active
- higher socio economic group
- receptive to new ideas
- decision makers
- high achievers
- used to communicating in writing

Those who do **not** respond tend to be

- elderly
- disengaged
- urban
- feel they may be judged by their responses
- feel they are inadequate at providing the basic information
Telephone surveys

**Advantages**

- cheaper – no traveling
- higher contact rates (can call more often)
- access
- no problems of adverse weather
- immediacy
- clustering unnecessary
- better environment for interviewers

**With centralization**

- higher quality
- speedier feedback
- supervision
- can calculate interviewer effects

**Disadvantages**

- need to design specially for the telephone
- lower response
- length restrictions
- have to cope with phone failures
- selection of individual difficult
- people more likely to break
interview part way through
checking records and using show cards is difficult
pace is quicker
observation data impossible

coverage bias
- this depends in part on the frame being used

Questionnaire design
more scripting needed
cannot see whole questionnaire so more segmentation needed
authenticity hard to prove (use advance letters plus 008 numbers)
item non-response is high so minimize question dependencies

CATI
lower interviewer variability
more control – random checks
supervisor can listen in and duplicate screen
automatic filtering
can automate selection and sample control
can randomize questions, scales etc.
up to the minute results and response is possible
can build in editing checks and clarify problems with the response.
Sampling for telephone interviewing

- special frames which include telephone numbers or for which they have been collected
  - follow-up studies when respondents have given telephone numbers
  - telephone directories
  - random digit dialing and variants
  - samples from telecom

Problems

- ex-directory numbers
- mobile phones
- households with no phones
- households with 2+ phones
- malfunctioning
- unassigned numbers

Stages in data preparation

Conventionally the stages in data preparation have been:

- booking-in
- early work checks
- clerical editing
- code construction
• coding
• computer entry
• variable generation
• merging data sources
• sampling weight construction
• frequency distributions
• validation
• corrective weight construction

Steps in resolution

locate variable in error
decide change to variable
implement change
delete a value
insert a value
change a value
leave a value unchanged

Flag changes

document error rates
global v individual changes?
Variable generation of two types

- recoded variables – collapsing frames or grouping interval scale variables
- combining variables (collapsing variables particularly when filtering is used)

Merging data sources

- several questionnaires relating to one unit of analysis
- waves of a panel
- sampling information with the interview

Sampling weight construction

- to adjust for differential probabilities of selection

Corrective weight adjustment

- to adjust for differences between the sample and population distributions

Frequency distributions

- detect outliers
- become familiar with the data
- important part of documentation

Validation

- opportunity will vary considerably from survey to survey
- depends on other sources
also on resources to carry out checks

Code frame construction systematic approach required

group identical answers and those with marginal differences

make codes mutually exclusive i.e. distinct – fine judgements cause coder variability

short not necessarily better

make frames comprehensive so anticipate answers

frequency not the only criterion

So bear in mind use of the data

don’t code irrelevancies

specify level of generality/specificity

Measurement of errors

Why?

they provide a measure of the accuracy of results

they remind us of the fragility of findings

they allow us to compare the accuracy of different methods techniques and designs

they allow us to pinpoint problem staff or procedures

knowledge of how errors arise help us to reduce them

we demonstrate that we take quality seriously

we may be able to adjust the results to compensate
Problems in measuring errors

- it is complex and requires expertise
- it takes time
- it can require special designs
- it would be impossible to be exhaustive

Validity checks

- criterion validity
- construct validity
- face validity
- process validity

Reliability checks

- internal consistency
- post-replication and verification
- incorporated replication

Supervision

Correcting for non-response

Missing data occurs in surveys through:

1) Total or Unit Non-Response
2) Item non-response

- Compensate for unit non-response by weighting adjustments
- Compensate for item non-response by imputation
- Imputation assigns values for the missing responses
What should we do about partial non-responses?

**Forms of weighting adjustments**

- sample
- population – post-stratification
- raking ratio
- response probabilities
- response propensity

**Some imputation methods**

- deductive
- overall mean
- mean within classes
- cold deck
- hot deck
- flexible matching
- random
- distance function matching
- regression
Is missing really missing?

Can we assume that missingness hides a true value that is meaningful for analysis?

Example

Contrast : A: “Don’t know” for a question on income with B: “Don’t know” for a vote in an election

In A, DK is a missing value but is DK in B to be treated as missing?

If cases with missing data differ systematically from complete cases – they are not missing completely at random. So discarding incomplete cases yields major biases.

Advantages of imputation – one has a full “rectangular” dataset but missing data are not as good as having the actual data. Follow-up a subset of non-respondents more intensively – use data to impute or weight other non-respondents.
Preface to the first SPSS manual,

by Norman H. Nie, Dale H. Bent and C Hadlai Hull.

Editor’s note: Published in 1970 by McGraw-Hill, this manual went on to become a world-wide success, eventually selling 700,000 copies. The back of the manual listed 55 IBM and 7 CDC installations as of July 1970, including 3 outside North America (Cologne Germany, Edinburgh Scotland, Leiden Holland). The pictures have been added subsequently.

Preface

Statistical Package for the Social Sciences (SPSS), a system of computer programs described in the present volume, represents the culmination of several years of systems design, programming, and documentation on the part of the authors and others. In these introductory remarks, it seems appropriate to describe the reasons why this work was undertaken and to outline the philosophy which guided the development of SPSS.

The development of SPSS was evolutionary. The initial design of SPSS which was begun by Dale Bent and Norman Nie in June of 1965, while we were both graduate students at Stanford University. The impetus for its development was our continuous frustration in attempting to serve the research and teaching needs of the faculty and graduate students of the Political science department and the Institute for Political Studies. During the academic year 1964-1965 we attempted to import a library of data analysis programs which would serve the needs of the department and the institute. While this piecemeal library met some of our most elementary needs, research and teaching which required data analysis was still a very frustrating business.
Each of the programs had its own concepts and syntax, and we consequently spent an intolerable amount of time teaching users how to operate many different single-purpose programs which were so distinct that almost no transfer learning took place. The programs were written in many different languages, and consequently when a problem was encountered in one of them, it was exceedingly difficult to examine the program code to locate the bug. Further, we quickly found that social science data analysis was in practice four-fifths file editing and data handling and only in a small part actual statistical analysis. While statistical programs abounded, we were constantly forced to write many ad hoc programs to edit files, recode and transform variables, as well as accomplished a large number of other routine file liaison and housekeeping chores. Finally, we discovered that the prevailing level of documentation was poor. Many write-ups, if they existed at all, were cryptic to the point of obscurity. There appeared to be a general tendency to prepare program documentation for data-analysis programs in the fashion one encounters in the mathematical literature, and this was
simply inadequate for the average social science user. In short, after a year's experience with a piecemeal program library, we felt that we were in only slightly better circumstances than when we had begun, and in fact, it often seems to us that we had simply increased the level of frustration by announcing the availability of programs which could not be efficiently taught, easily utilised, or sufficiently supported.

The training, background, and interest of the typical social scientist aggravates the problem, for he has not thus far gained the technical grasp of computer usage to perform the data-analysis tasks required by his research. Given the number of skills that he is currently required to master, his abdication in this highly technical area may, we feel, be both justifiable and rational. We became convinced from our experience that special provisions had to be made for the social science community in order to simplify the process of data analysis so that the social science researcher himself could gain control over the day-to-day manipulation and analysis of his data.

As a result of our frustration, as well as our experience, we began to design an integrated system of programs that addressed itself to these needs. In the early period, we placed a great deal of emphasis on defining procedures that would automate the routine tasks of data processing and around which a series of statistical programs could be build. We felt that existing stand-alone programs, as well as previous statistical packages (such as widely used BMD system), had established the feasibility of implementing generalised statistical programs, but what had not been successfully accomplished was the production of an integrated system of common procedures for the management and handling of complex data files. Accordingly, a set of conventions for such a system of programs was defined and coding was initiated for the IBM 7090 computer then in use at Stanford University. The code was developed in piecemeal fashion and eventually hardened into a system, which became the forerunner of the present SPSS.

We were sufficiently encouraged by the result we had obtained with the 7090 programs to begin to make plans for the expansion and conversation
of the system to the IBM 360, which arrived at Stanford in the summer of 1967. It was at this point that Hadlai Hull joined the project and began the arduous process of converting our semioperational concepts into a working system, a task which the original two authors feel he superbly accomplished.

Conversion of the 7090 prototype system was facilitated by the exclusive use of Fortran in the 7090 program and the fact that much of the macrologic of the system could be directly transferred to the 360. In fact OS/360 provided increased speed and core, which immediately alleviated many of the problems that plagued us on the 7090. A decision was made to fix the capabilities of the system at the level documented in the present volume, and arrangements were made for the dissemination of the program and documentation to those who wished to use the package. During this period Northwestern University has proceeded with the conversion of SPSS to CDC 6000-series equipment and an operational CDC version of SPSS should be generally available by the time this volume is in print. (Appendix G is a special guide to SPSS for CDC users.)

We are aware that the current system has a great number of deficiencies, but we believe it meets a majority of the social science data-analysis needs in a way that has not been previously possible. Further, SPSS is an open-ended system, and we are hopeful that we and others will continue to improve, add to, and modify the existing version of the system. To facilitate the ability of others to add a wider variety of statistics procedures to SPSS, a technical or programmer’s guide had been included in Appendix H. Because of the design of SPSS, all future programs incorporated into the system can take complete advantage of the capabilities for file maintenance and data handling which currently exist in the package, and we fully hope that this will be an incentive to those contemplating the additional of other statistical programs.

In the rapidly changing area of computer technology, any designer of software must concern himself with the expected term of usefulness of the
programs he has created. We have tried to take a number of steps to lengthen the life expectancy of the SPSS system. First, the language selected to write the programs was Fortran IV, one of the few languages in use today with any claim to universality. Second, the system was developed on the IBM 360 and translated to CDC 6000-series equipment; these two computers represent an overwhelming proportion of all university resident machines in North America. The use of Fortran IV and a firm base in these two third-generation machines should facilitate the conversion of SPSS of forth-generation machines, which should be appearing soon.

Time sharing systems are, of course, becoming generally available throughout the United States and elsewhere. This raises the possibility of interactive or conversational programs for statistical analysis. SPSS is not a conversational system. The 7090 version was designed for conventional batch processing and the present 360 and CDC versions operate in the same fashion. At Stanford, we have had a great deal of success running SPSS from remote terminals controlled by the WYLBUR text editing and remote batch-entry system. SPSS in this environment still operates as a batch program, but the text editor and remote batch-entry and retrieval capabilities permit the user to prepare and enter all his jobs from the terminal and to print back small jobs on his terminal.

The authors have had the opportunity to examine and in some instances use true conversational statistical systems, such as APL STATPACK at the University of Alberta and IMPRESS at Dartmouth. Each of these systems is truly elegant and a great pleasure to use, but in our experience we have not encountered any conversational data-analysis system, which has the kind of data-handling capabilities, both in terms of volume and flexibility, which SPSS processes. We are convinced that such systems will emerge in the near future; they await only slightly larger and faster machines and the emergence of a more feasible remote device for the display of large amounts of output. When these conversational systems emerge, they will have a major impact on the way in which the social scientist engages in
social research. And we therefore anxiously await the pending obsolescence of our own system.

For the present, however, we feel that SPSS presents the social scientists with a useful working language for data analysis. We hope that the capabilities of SPSS will act as a minimum benchmark for the designers of new and better languages for social research. Finally, we hope that the user of SPSS will find that it provides them with a substantial increase in ease and flexibility, which they can approach, their day-to-day use of the computer.

A card deck with SPSS syntax (1979)
An introduction to SPSS

by Noma Makepela

Outcome:

at the end of this session you should have enough confidence to attempt an SPSS session, label datasets and run Frequencies and Crosstabs

Introduction

SPSS (Statistical Package for Social Sciences) is a comprehensive and flexible statistical and data management program that can take data from almost any file type to generate clearly labeled, tabulated reports, charts and plots of distributions and trends, descriptive statistics and complex statistical analyses.

SPSS is a batch driven program. Although it appears to be interactive thanks to a graphical user interface (GUI), it is actually driven by an input/output format. In this respect it is different from a spreadsheet. SPSS can handle any number of records. It is used by Government departments such as tats SA. Demonstrable competence with SPSS will enhance employment prospects.

Main Menu

Many of the tasks you want to perform with SPSS start with menu selections. Each window in SPSS has its own menu bar with menu selections appropriate for that window type. You must familiarize yourself with these menus and windows.
Status Bar

The status bar at the bottom of each SPSS window provides the following information:

**Command status**

For each procedure or command you run, a case counter indicates the number of cases processed so far. For statistical procedures that require iterative processing, the number of iterations is displayed.

**Filter status**

If you have selected a random sample or a subset of cases for analysis, the message “Filter on” indicates that some type of case filtering is currently in effect and not all cases in the data file are included in the analysis.

**Weight status.**

The message “Weight on” indicates that a weight variable is being used to weight cases for analysis.

**Split File status.**

The message Split File on indicates that the data file has been split into separate groups for analysis, based on the values of one or more grouping variables.

**Data Editor (entering and editing data)**

Data values are modified in Data view in many ways
change data values,
cut cell data
copy, and paste data values
add and delete cases
add and delete variables
change the order of variables.

You can change the data type for a variable at any time using the Define Variable Type dialog box in the Variable view, and the Data Editor will attempt to convert existing values to the new type. If no conversion is possible, the system-missing value is assigned. The conversion rules are the same as those for pasting data values to a variable with a different format type. If the change in data format may result in the loss of missing value specifications or value labels, the Data Editor displays an alert box and asks if you want to proceed with the change, or cancel it.

You can cut, copy, and paste individual cell values or groups of values in the Data Editor. You can:

Move or copy a single cell value to another cell.
Move or copy a single cell value to a group of cells.
Move or copy the values for a single case (row) to multiple cases.
Move or copy the values for a single variable (column) to multiple variables.
Move or copy a group of cell values to another group of cells.

If the defined variable types of the source and target cells are not the same, the Data Editor attempts to convert the value. If no conversion is possible, the system-missing value is inserted in the target cell.

Numeric or Date into String. Numeric (for example, numeric, dollar, dot, or comma) and date formats are converted to strings if they are pasted into a string variable cell. The string value is the numeric value as displayed in
the cell. For example, for a dollar format variable, the displayed dollar sign becomes part of the string value. Values that exceed the defined string variable width are truncated.

String into Numeric or Date. String values that contain acceptable characters for the numeric or date format of the target cell are converted to the equivalent numeric or date value. For example, a string value of 25/12/91 is converted to a valid date if the format type of the target cell is one of the day-month-year formats, but it is converted to system missing if the format type of the target cell is one of the month-day-year formats.

Date into Numeric. Date and time values are converted to a number of seconds if the target cell is one of the numeric formats (for example, numeric, dollar, dot, or comma). Since dates are stored internally as the number of seconds since October 14, 1582, converting dates to numeric values can yield some extremely large numbers. For example, the date 10/29/91 is converted to a numeric value of 12,908,073,600.

Numeric into Date or Time. Numeric values are converted to dates or times if the value represents a number of seconds that can produce a valid date or time. For dates, numeric values less than 86,400 are converted to the system-missing value.

**Creating a new data file**

You can enter data directly in the Data Editor in the Data View. You can enter data in any order. You can enter data by case or by variable, for selected areas or for individual cells. Data values are not recorded until you press Enter or select another cell. To enter anything other than simple numeric data, you must define the variable type first. If you enter a value in an empty column, the Data Editor automatically creates a new variable and assigns a variable name.
Inserting new variables

Entering data in an empty column in Data view or in an empty row in Variable view automatically creates a new variable with a default variable name (the prefix var and a sequential number) and a default data format type (numeric). The Data Editor inserts the system-missing value for all cases for the new variable. If there are any empty columns in the Data view or empty rows in the Variable view between the new variable and the existing variables, these also become new variables with the system-missing value for all cases. You can also insert new variables between existing variables.

Saving files: SPSS save formats:

(i) SPSS (*.sav). SPSS format.
(ii) SPSS 7.0 (*.sav).
(iii) SPSS/PC+ (*.sys). SPSS/PC+ format.
(iv) SPSS portable (*.por).
(v) Tab-delimited (*.dat).
(vi) Fixed ASCII (*.dat).
(vii) Excel (*.xls). Ending an SPSS session

Statistical functions of SPSS

(i) Data Tabulations

To help you uncover patterns in the data that contribute to a significant chi-square test, the Crosstabs procedure displays expected frequencies and three types of residuals (deviates) that measure the difference between observed and expected frequencies. Each cell of the table can contain any combination of counts, percentages, and residuals selected.
Counts. The number of cases actually observed and the number of cases expected if the row and column variables are independent of each other.

(ii) Percentages.

The percentages can add up across the rows or down the columns. The percentages of the total number of cases represented in the table (one layer) are also available.

(iii) Residuals.

Raw unstandardized residuals give the difference between the observed and expected values. Standardized and adjusted standardized residuals are also available.

(iv) Descriptive Statistics

Summary information about the distribution, variability, and central tendency of a variable.

Example. If each case in your data contains the daily sales totals for each member of the sales staff (for example, one entry for Bob, one for Kim, one for Brian, etc.) collected each day for several months, the Descriptive procedure can compute the average daily sales for each staff member and order the results from highest average sales to lowest.

Statistics

Sample size, mean, minimum, maximum, standard deviation, variance, range, sum, standard error of the mean, and kurtosis and skewness with their standard errors.

(v) Percentile Values.

Values of a quantitative variable that divide the ordered data into groups so that a certain percentage is above and another percentage is below. Quartiles (the 25th, 50th, and 75th percentiles) divide the observations into
four groups of equal size. If you want an equal number of groups other than four, select Cut points for n equal groups. You can also specify individual percentiles (for example, the 95th percentile, the value below which 95% of the observations fall).

(vi) Central Tendency.

Statistics that describe the location of the distribution include the mean, median, mode, and sum of all the values.

(vii) Dispersion.

Statistics that measure the amount of variation or spread in the data include the standard deviation, variance, range, minimum, maximum, and standard error of the mean.

(viii) Distribution.

Skewness and kurtosis are statistics that describe the shape and symmetry of the distribution. These statistics are displayed with their standard errors.

Values are group midpoints. If the values in your data are midpoints of groups (for example, ages of all people in their thirties are coded as 35), select this option to estimate the median and percentiles for the original, ungrouped data.

Exploring data

Frequency tables

Summary of the number of times different values of a variable occur. The Frequencies procedure provides statistics and graphical displays that are useful for describing many types of variables. For a first look at your data, the Frequencies procedure is a good place to start.
Cross tabulations

The Crosstabs procedure forms two-way and multiway tables and provides a variety of tests and measures of association for two-way tables. The structure of the table and whether categories are ordered determine what test or measure to use.

Crosstabs’ statistics and measures of association are computed for two-way tables only. If you specify a row, a column, and a layer factor (control variable), the Crosstabs procedure forms one panel of associated statistics and measures for each value of the layer factor (or a combination of values for two or more control variables). For example, if GENDER is a layer factor for a table of MARRIED (yes, no) against LIFE (is life exciting, routine, or dull), the results for a two-way table for the females are computed separately from those for the males and printed as panels following one another.

Example

Are customers from small companies more likely to be profitable in sales of services (for example, training and consulting) than those from larger companies? From a crosstabulation, you might learn that the majority of small companies (fewer than 500 employees) yield high service profits, while the majority of large companies (more than 2500 employees) yield low service profits.

Crosstabs and their statistics

Chi-square

Chi-square yields the linear-by-linear association test. For tables with two rows and two columns, select Chi-square to calculate the Pearson chi-square, the likelihood-ratio chi-square, Fisher’s exact test, and Yates’ corrected chi-square (continuity correction).
Fisher’s exact test is computed when a table that does not result from missing rows or columns in a larger table has a cell with an expected frequency of less than 5.

**Yates’ corrected chi-square**

Is computed for all other 2 by 2 tables. For tables with any number of rows and columns, select Chi-square to calculate the Pearson chi-square and the likelihood-ratio chi-square. When both table variables are quantitative,

**Correlations**

For tables in which both rows and columns contain ordered values, Correlations yields Spearman’s correlation coefficient, rho (numeric data only). Spearman’s rho is a measure of association between rank orders. When both table variables (factors) are quantitative, Correlations yields the Pearson correlation coefficient, r, a measure of linear association between the variables.

**Nominal.**

For nominal data (no intrinsic order, such as Catholic, Protestant, and Jewish), you can select Phi (coefficient) and Cramér’s V, Contingency coefficient, Lambda (symmetric and asymmetric lambdas and Goodman and Kruskal’s tau), and Uncertainty coefficient.

**Ordinal.**

For tables in which both rows and columns contain ordered values, select Gamma (zero-order for 2-way tables and conditional for 3-way to 10-way tables), Kendall’s tau-b, and Kendall’s tau-c. For predicting column categories from row categories, select Somers’ d.

**Nominal by Interval.**

When one variable is categorical and the other is quantitative, select Eta. The categorical variable must be coded numerically.
Kappa.
For tables that have the same categories in the columns as in the rows (for example, measuring agreement between two raters), select Cohen’s Kappa.

Risk.
For tables with two rows and two columns, select Risk for relative risk estimates and the odds ratio.

McNemar.
Is a nonparametric test for two related dichotomous variables.
It tests for changes in responses using the chi-square distribution.
It is useful for detecting changes in responses due to experimental intervention in “before and after” designs.

Cochran’s and Mantel-Haenszel.
Can be used to test for independence between a dichotomous factor variable and a dichotomous response variable, conditional upon covariate patterns defined by one or more layer (control) variables. The Mantel-Haenszel common odds ratio is also computed, along with Breslow-Day and Tarone’s statistics for testing the homogeneity of the common odds ratio.

ANOVA: Analysis of Variance
Analysis of variance, or ANOVA, is a method of testing the null hypothesis that several group means are equal in the population, by comparing the sample variance estimated from the group means to that estimated within the groups.

The dependent variable is quantitative. Factors are categorical. They can have numeric values or string values of up to eight characters. Covariates are quantitative variables that are related to the dependent variable.
Assumptions

The data are a random sample from a normal population; in the population, all cell variances are the same. Analysis of variance is robust to departures from normality, although the data should be symmetric. To check assumptions, you can use homogeneity of variances tests and spread-versus-level plots. You can also examine residuals and residual plots.

Measuring Linear associations

Curve Estimation

The Curve Estimation procedure produces curve estimation regression statistics and related plots for 11 different curve estimation regression models. A separate model is produced for each dependent variable. You can also save predicted values, residuals, and prediction intervals as new variables.

Example

A fire insurance company conducts a study to relate the amount of damage in serious residential fires to the distance between the closest fire station and the residence. A scatterplot reveals that the relationship between fire damage and distance to the fire station is linear. You might fit a linear model to the data and check the validity of assumptions and the goodness of fit of the model.

Data. The dependent and independent variables should be quantitative. If you select Time instead of a variable from the working data file as the
independent variable, the Curve Estimation procedure generates a time variable where the length of time between cases is uniform. If Time is selected, the dependent variable should be a time-series measure. Time-series analysis requires a data file structure in which each case (row) represents a set of observations at a different time and the length of time between cases is uniform. Assumptions. Screen your data graphically to determine how the independent and dependent variables are related (linearly, exponentially, etc.). The residuals of a good model should be randomly distributed and normal. If a linear model is used, the following assumptions should be met. For each value of the independent variable, the distribution of the dependent variable must be normal. The variance of the distribution of the dependent variable should be constant for all values of the independent variable. The relationship between the dependent variable and the independent variable should be linear, and all observations should be independent.


### Non-parametric tests

The Chi-Square Test procedure tabulates a variable into categories and computes a chi-square statistic. This goodness-of-fit test compares the observed and expected frequencies in each category to test either that all categories contain the same proportion of values or that each category contains a user-specified proportion of values. Examples. The chi-square test could be used to determine if a bag of jellybeans contains equal proportions of blue, brown, green, orange, red, and yellow candies. You could also test to see if a bag of jellybeans contains 5% blue, 30% brown, 10% green, 20% orange, 15% red, and 15% yellow candies. Statistics. Mean, standard deviation, minimum, maximum, and quartiles. The number and the
percentage of nonmissing and missing cases, the number of cases observed and expected for each category, residuals, and the chi-square statistic.

**The One-Sample Kolmogorov-Smirnov Test**

Compares the observed cumulative distribution function for a variable with a specified theoretical distribution, which may be normal, uniform, poisson, or exponential. The Kolmogorov-Smirnov Z is computed from the largest difference (in absolute value) between the observed and theoretical cumulative distribution functions. This goodness-of-fit test tests whether the observations could reasonably have come from the specified distribution. Example. Many parametric tests require normally distributed variables. The one-sample Kolmogorov-Smirnov test can be used to test that a variable, say INCOME, is normally distributed. Statistics: the mean, standard deviation, minimum, maximum, number of non-missing cases, and quartiles.

**The Runs Test procedure**

Tests whether the order of occurrence of two values of a variable is random. A run is a sequence of like observations. A sample with too many or too few runs suggests that the sample is not random. Example. Suppose that 20 people are polled to find out if they would purchase a product. The assumed randomness of the sample would be seriously questioned if all 20 people were of the same gender. The runs test can be used to determine if the sample was drawn at random. Statistics: Mean, standard deviation, minimum, maximum, number of non-missing cases, and quartiles.

**The Tests for Several Independent Samples procedure**

Compare two or more groups of cases on one variable.
**Example.**

Do three brands of 100-watt lightbulbs differ in the average time the bulbs will burn? From the Kruskal-Wallis one-way analysis of variance, you might learn that the three brands do differ in average lifetime. Statistics: Mean, standard deviation, minimum, maximum, number of nonmissing cases, and quartiles.

The Two-Independent-Samples Tests procedure compares two groups of cases on one variable. Tests: Mann-Whitney U, Moses extreme reactions, Kolmogorov-Smirnov Z, Wald-Wolfowitz runs.

**Example.**

New dental braces have been developed that are intended to be more comfortable, to look better, and to provide more rapid progress in realigning teeth. To find out if the new braces have to be worn as long as the old braces, 10 children are randomly chosen to wear the old braces, and another 10 are chosen to wear the new braces. From the Mann-Whitney U test, you might find that, on average, those with the new braces did not have to wear the braces as long as those with the old braces. Statistics: Mean, standard deviation, minimum, maximum, number of nonmissing cases, and quartiles.

Tests for Several Related Samples procedure compares the distributions of two or more variables. Tests: Friedman, Kendall’s W, and Cochran’s Q. Example. Does the public associate different amounts of prestige with a doctor, a lawyer, a police officer, and a teacher? Ten people are asked to rank these four occupations in order of prestige. Friedman’s test indicates that the public does in fact associate different amounts of prestige with these four professions. Statistics: Mean, standard deviation, minimum, maximum, number of nonmissing cases, and quartiles.
Help for SPSS learners

Sandra Matanyaire

Department of Statistics
University of the Western Cape

This introductory manual provides a step-by-step approach of data analysis in SPSS covering two areas:

- Type of analysis relevant to a data type
- How to use SPSS for data analysis

Opening SPSS

Start—Programs—SPSS for Windows—SPSS 12.0 for Windows (Click on SPSS 12.0 for Windows)

The options available in the initial window: Run the tutorial, Type in Data, Run an existing query, Create new query using Database wizard, Open an existing data source.

Select the Type in Data option and the SPSS Data Editor window will be displayed. The Data Editor allows you to enter and view data in the rectangular spreadsheet form of cases (rows) and variables (columns). Variables are the research information collected across the cases or across the subjects of study like age, sex, height, latitude, or continent. The cases contain information/data on each subject of research across the various variables of study for instance a person, household, geographic location.

The Data Editor has two interconnected windows whose access is located at the bottom left corner of the Data Editor screen:

- Data View where data is entered into the cases (rows) and
Variable View where variable information is entered.

Data Entry: Variable view and Data View

**Variable view**

The initial step in data entry is done in variable view where information on the variables is entered. The specifications of the variables that are entered into the variable view include:

- **Variable Name**: SPSS limits the variable name to 8 characters
- **Variable type**: which specifies the format of your variable options include: *Numeric, Comma, Dot, Date, Dollar, String*
- **Field width**: (number of characters entered for case entries)
- **Decimal places**: to be applied to the
- **Label option**: allows entry of up to 120 descriptive characters of the variable
- **Value labels**: allow labels to be coded for example Boys = 0 and Girls = 1, where the values of 0 and 1 are used to represent the labels boys and girls respectively.
- **The Missing option**: provides the alternatives of how missing values are treated in SPSS; either three discrete values used to denote missing values are recorded or a range of missing values plus an optional discrete value.
- **Columns option**: specifies the width of the column
- **Align entries**: either left, right or center
· Measurement Level can be chosen to either be scale (numeric data-interval or ratio), ordinal, or nominal. Nominal and ordinal data can be either string (non-numeric) or numeric. Measurement specification is important for chart procedures that identify variables either as scale or categorical. Nominal and ordinal are both treated as categorical whilst interval and ration are treated as scale.

**Data View**

In data view the data is entered for individual units of analysis across the variables specified in the first step under variable view. Data can also be imported into SPSS from excel.

File—Open—Data—Files of type: Excel (*.xls)—Select file name—Open

**Data Screening**

Before any meaningful analysis can be made to a data set, the data itself must be screened and tested for validity and appropriateness of the data values recorded. Errors can arise either in the process of data collection or data entry. The data set may contain values that stand out and apart from the other values (outliers and/or extreme values). These values can influence or distort results of the analysis hence they need to be identified and dealt with appropriately. The following procedures can be employed in SPSS during data screening:

· Generate frequencies to check for typing errors, values that deviate outside the reasonable range, check for missing values and investigate the validity of outliers.
Analyze—Descriptive Statistics—Frequencies: Select the variable(s) under investigation—OK

- Use graphs box and whisker plots and stem and leaf plots to check for outliers

Graphs—Boxplot—Simple/Clustered—Define: Select the variable under study with the category variable (x axis). If clustered select the clustering variable—OK

Or

Analyze—Descriptive Statistics—Explore: Select the dependent variable (y axis) and the factor list (x axis)—Plots—Select Boxplot and Stem-and-leaf—Continue—OK

- Analyze descriptive statistics like maximum and minimum values especially for large data sets

Analyze—Descriptive Statistics—Descriptive—OK

Or

Analyze—Descriptive Statistics—Frequencies: Select variable under study—Statistics: Select relevant statistics—Continue—OK

- List the data values using Case Summaries

Analyze—Reports—Case Summaries—OK

Crosstabulations can be useful to reveal unlikely and illogical combinations, useful mainly for categorical data. For example a deceased person providing personal details.

Analyze—Descriptive Statistics—Crosstabs—OK
Treatment of outliers/extreme values

When SPSS detects outliers one has to determine whether the outliers are recorded errors which can be corrected and if not then determine whether the data can be transformed to minimize or remove the impact of the outlying values.

SPSS provides the option of producing an outlier output.

Analyze—Descriptive Statistics—Explore: Select the dependent variable (y axis) and the factor list (x axis) —Statistics—Outliers—Continue—OK

These outliers are taken as the five smallest and the five largest values of the variable selected provided together with their case labels. Box plots and stem-and-leaf diagrams should be used to identify outliers as this output just contains the five smallest and largest values, which may not necessarily be outliers, and/or extreme values. This plot may assist in producing the case labels of possible outlying values.

Data Analysis: Descriptive analysis and inferential analysis

Descriptive analysis

The initial step in data analysis is the identification of the data type of the variable(s) under investigation to determine the most appropriate and applicable statistical analysis. The data types can be either of the following:

- Nominal/Categorical (Unordered)
- Ordinal (Ordered)
- Interval (Ordered with distance)
- Ratio (Ordered, distance plus origin)
1.2. Nominal/Categorical data

Weakest data form in terms of statistical analysis. Only frequencies or counts can be performed on numeric or non-numeric unordered variables.

Table 1: Descriptive statistics appropriate for unordered variables

<table>
<thead>
<tr>
<th>Groups</th>
<th>Counts</th>
<th>Percents</th>
<th>Valid P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequencies</td>
<td>Sample</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>Crosstabs</td>
<td>Crossed</td>
<td>✔️</td>
<td>✔️</td>
</tr>
</tbody>
</table>

Source: SPSS Base 8.0 Applications Guide 1998

The “Groups” column in the table above and those to follow indicates whether the respective statistic is computed for the entire sample (Sample), for subgroups of cases within the sample determined by a grouping variable (Groups), or for subgroups determined by combinations of grouping variables (Crossed).

Graphs appropriate for unordered variables:
· Bar Chart (Simple, Component, Multiple)

Graphs—Bar: Simple/Clustered/Stacked—Define—Category axis—OK

Or

Analyze—Descriptive Statistics—Frequencies—Charts—Bar Charts

· Pie ChartTable 2: Descriptive statistics appropriate for ordered variables

Graphs—Pie
3. **Ordinal data and Interval data**

Values that are code ordered can be described by the following statistics:

**Table 2: Descriptive statistics appropriate for ordered variables**

<table>
<thead>
<tr>
<th></th>
<th>Groups</th>
<th>Median</th>
<th>Min.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Frequencies</strong></td>
<td>Sample</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td><strong>Descriptives</strong></td>
<td>Sample</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td><strong>Explore</strong></td>
<td>Groups</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td><strong>Case Summaries</strong></td>
<td>Crossed</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td><strong>Means</strong></td>
<td>Crossed</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td><strong>One-Way ANOVA</strong></td>
<td>Groups</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td><strong>Nonparametric tests</strong></td>
<td>Sample</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

*Source: SPSS Base 8.0 Applications Guide 1998*

Graphs:
- Histogram

Graphs—Histogram

Or

Analyze—Descriptive Statistics—Frequencies—Charts—Histogram
146

Robert C.-H, Shell (ed.)

Sandra Matanyaire

- Box and Whisker plot (Boxplot)
- Stem and leaf plot

4. **Ratio-scaled data**

The strongest data type with regards to the statistical techniques that can be applied. The most common statistics like the mean, standard deviation and variance can be computed for ratio-scaled data BUT for meaningful and appropriate variable statistics and analysis the quantitative data has to follow a normal distribution or alternatively allow for a reasonable assumption of normality to be made (a symmetrical histogram can be fitting). The statistics can be misleading if the distribution of the data deviates widely from the normal distribution.

**Normal distribution**

The normal probability distribution is a continuous distribution, which is a bell shaped and symmetrical about the mean (\( \mu \)). The tails of the distribution are asymptotic (never touch the x axis) and the area under the normal curve equals to one.

The assumption of normality can be determined in the following ways using SPSS:

- Histogram with normal curve superimposed
- Generate probability plots of variables to compare to normal probability plots or other distributions

Analyze—Descriptive Statistics—Explore—Plots—Normality Plots with tests

- Compare values of the mean and median and 5% trimmed mean (obtained by removing the top 5% and bottom 5% values to remove the influence of unusual values). If the difference is marked the distribution will be skewed.
· 95% of the values should lie within two standard deviations of the mean (Mean plus or minus 2X Standard deviation)

· Skewness (measure of symmetry of the distribution) and Kurtosis (measure of peakedness) studied together with a graphical plot, by computing the ratio of each statistic to its standard error and accepting normality if the values lie between the values of –2 and 2.

In the event that the histogram is far from symmetrical and the data does not follow a normal distribution, transformations can be made to the original data to improve the quality of the statistics computed.

Transform—Compute: Insert the relevant function and the variable to be transformed—OK

Examples of transformations include: computing log to base 10 of the original variable, squaring the variable and taking the square root of the variable.

Table 3: Descriptive statistics for normally distributed ratio scaled data
<table>
<thead>
<tr>
<th>Analysis Type</th>
<th>Groups</th>
<th>Mean</th>
<th>SD</th>
<th>Std. Error of Mean</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequencies</td>
<td>Sample</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Descriptives</td>
<td>Sample</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Explore</td>
<td>Groups</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Case Summaries</td>
<td>Crossed</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Means</td>
<td>Crossed</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>One-Sample t-test</td>
<td>Sample</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>Independent t-test</td>
<td>Two</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>Paired t-test</td>
<td>Sample</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>One-Way ANOVA</td>
<td>Groups</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>GLM General Factorial</td>
<td>Crossed</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td></td>
</tr>
</tbody>
</table>
Graphs:

- Histogram
- Box and Whisker plot
- Stem and leaf plot

If the assumption of normality cannot be reasonably made at all the type of statistics that are appropriate are the same as those for interval and ordinal numeric coded data.

Inferential analysis

Inferential statistics is the part of statistics which aims to estimate true population parameters based on sample statistics. Two broad techniques of statistical inferences are available:

- Estimation of population parameters
- Hypothesis testing which involves testing a claim about the population based on sample values

Estimation of population parameters can either be a point estimate in which the sample statistic is taken to be an unbiased estimate of the population or interval estimation where a range of possible values are computed with a confidence of containing the true population value. As interval estimation provides the confidence or reliability in the estimate it is a better technique. For example a 95% confidence interval means that there is a 0.95 chance (probability) that the interval constructed will contain the true population parameter.

Analyze—Descriptive Statistics—Explore: Select the dependent variable (y axis) and the factor list (x axis) ——Statistics: Specify the Confidence Interval for the mean—Continue—OK

Or
Hypothesis tests follow in the research process from the research question(s). There are two hypotheses that are stated in the research process:

- Null hypothesis (H0) – states that the population parameter is equal to a certain hypothesized value.
- Alternative hypothesis (H1) – negates the null hypothesis by stating that the population parameter is not equal to the hypothesized value.

The equality sign should always appear in the null hypothesis either as an equal to sign (=) or greater than or equal to sign (≥) or less than or equal to sign (≤). It thus follows that the alternative hypothesis will contain the strict inequality sign, either the greater than sign (>) or the less than sign (<).

**Summarized steps for hypothesis testing:**

A hypothesis is put forward on population parameters.

The hypothesis is tested using inferential statistical analysis.

The type of inferential analysis depends on the nature of the data variables to be included in the analysis and the hypothesis put forward.

The end result of a hypothesis test is to either reject the null hypothesis or not reject the null hypothesis based on a certain significance level.

In the following sections plausible hypotheses are provided and an appropriate test is provided:

1. H0: There is no association between gender and graduating from the faculty of science.

   H1: There is an association between gender and graduating from the faculty of science.

   Test of association: chi-square test
The chi-square test measures the extent to which the observed values differ from the expected values. There are three tests for which the chi-square is used:

- Tests for independence of association
- Tests for equality of proportions when there are more than two populations
- Goodness-of-fit tests which test how well data fits a particular distribution

For the chi-square test the null hypothesis always postulates independence between the random variables. The data requirements for the chi-square test are frequency counts.

2. H0: The population mean age of male nurses is equal to the population mean age of female nurses

H1: The population mean ages are not equal

Equality of means: independent samples t-test

Analyze—Compare means—Independent samples t-test: insert the test variable and grouping variable

3. H0: The mean test score for the population is equal to eighty percent

H1: The mean test score is not equal to eighty percent

Mean equality to specified value: One sample t-test

Analyze—Compare means—One sample t-test: insert the test variable and the test value.

The t-test is used for tests on hypotheses of populations which assume normality. The population standard deviation is assumed to be not known.
and the t test is used for small samples of less than or equal to 30. The t-test can be used for the following hypothesis testing scenarios:

· Independent sample tests
· Paired sample tests (before and after scenarios)
· One sample tests

When the t test is used for independent samples test, a hypothesis of whether the samples can be assumed to have an equal variance or not has to be tested using the F-test (A null hypothesis H0 stating that the population variances are equal). If the ‘Sig’ value is more than 0.05 then the hypothesis that the variances are equal (H0) is not rejected and the t-test based on an equal variance assumed is accepted. If the ‘Sig’ value of the F test is less than 0.05 the variance equality hypothesis (H0) is rejected and the t test with equal variances not assumed is used.

This is an introductory manual. The intermediate and advanced manuals will continue with more advanced statistical methods including Analysis of Variance, regression analysis, factorial analysis, discriminant analysis and non-parametric tests.

**Accelerator keys in the Data Editor**

- Press Shift+Tab; Left Arrow To Previous variable (left)
- Enter; Down Arrow Next case (down)
- Up Arrow Previous case (up)
- Ctrl+Left Arrow First variable
- Ctrl+Right Arrow Last variable
- Ctrl+Home First variable, first case
<table>
<thead>
<tr>
<th>Keyboard Shortcut</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ctrl+End</td>
<td>Last variable, last case</td>
</tr>
<tr>
<td>F2</td>
<td>Edit cell contents (Edit mode)</td>
</tr>
<tr>
<td>Enter</td>
<td>Finish editing cell contents</td>
</tr>
<tr>
<td>Shift+F2</td>
<td>Raise the value label pop-up</td>
</tr>
<tr>
<td>Shift+Spacebar</td>
<td>Select current case (row)</td>
</tr>
<tr>
<td>Ctrl+Spacebar</td>
<td>Select current variable (column)</td>
</tr>
<tr>
<td>Shift+Left Arrow</td>
<td>Select cell to left</td>
</tr>
<tr>
<td>Shift+Right Arrow</td>
<td>Select cell to right</td>
</tr>
<tr>
<td>Tab or Right Arrow</td>
<td>Next variable (right)</td>
</tr>
<tr>
<td>Press</td>
<td>To</td>
</tr>
<tr>
<td>PgDn/PgUp</td>
<td>Scroll down/up a page</td>
</tr>
<tr>
<td>Ctrl+PgDn/Ctrl+PgUp</td>
<td>Scroll right/left a page</td>
</tr>
</tbody>
</table>

**In Edit mode:**

<table>
<thead>
<tr>
<th>Keyboard Shortcut</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Press</td>
<td>To</td>
</tr>
<tr>
<td>Right Arrow or Left Arrow</td>
<td>Next/previous character</td>
</tr>
<tr>
<td>Shift+Right Arrow</td>
<td>Extend selection</td>
</tr>
<tr>
<td>Shift+Left Arrow</td>
<td>Next cell (right)</td>
</tr>
<tr>
<td>Tab</td>
<td>Previous cell (left)</td>
</tr>
<tr>
<td>Shift+Tab</td>
<td>Move to start of value</td>
</tr>
<tr>
<td>Home</td>
<td>Move to end of value</td>
</tr>
<tr>
<td>End</td>
<td>Select to start of value</td>
</tr>
</tbody>
</table>
Shift+End   Select to end of value
Esc         Restore cell contents
Source: SPSS Help File

Bibliography
SPSS Base 8.0 1998. Applications Guide. SPSS Inc.
SPSS Help file.
Life tables
a SPSS application
(advanced models module)

by Mike Bergh, Ph.D.

SPSS-SA
Silvermine House
Steenberg Office Park
Tokai 7945

Introduction

There are many situations in which you would want to examine the distribution of times between two events, such as length of employment (time between being hired and leaving the company). However, this kind of data usually includes some cases for which the second event is not recorded (for example, people still working for the company at the end of the study). This can happen for several reasons: for some cases, the event simply does not occur before the end of the study; for other cases, we lose track of their status sometime before the end of the study; still
other cases may be unable to continue for reasons unrelated to the study (such as an employee becoming ill and taking a leave of absence). Collectively, such cases are known as **censored cases**, and they make this kind of study inappropriate for traditional techniques such as t tests or linear regression.

A statistical technique useful for this type of data is called a follow-up life table. The basic idea of the life table is to subdivide the period of observation into smaller time intervals. For each interval, all people who have been observed at least that long are used to calculate the probability of a terminal event occurring in that interval. The probabilities estimated from each of the intervals are then used to estimate the overall probability of the event occurring at different time points.

**Example**

Is a new nicotine patch therapy better than traditional patch therapy in helping people to quit smoking? You could conduct a study using two groups of smokers, one of which received the traditional therapy and the other of which received the experimental therapy. Constructing life tables from the data would allow you to compare overall abstinence rates between the two groups to determine if the experimental treatment is an improvement over the traditional therapy. You can also plot the survival or hazard functions and compare them visually for more detailed information.

**Statistics**

Number entering, number leaving, number exposed to risk, number of terminal events, proportion terminating, proportion surviving, cumulative proportion surviving (and standard error), probability density (and standard error), and hazard rate (and standard error) for each time interval for each group; median survival time for each group; and Wilcoxon (Gehan) test for comparing survival distributions between groups. Plots: function plots for survival, log survival, density, hazard rate, and one minus survival.
**Fish population dynamics**

Biological systems provide an interesting background to the consideration of life tables and the SPSS life table procedure. A subset of the general study of biological systems is concerned with analyses of fish populations and their dynamics. There are obvious links with the approaches taken in the analysis of human populations although the motivating questions have a different emphasis.

In modelling fish populations, one is very often confronted with the question of how the number of individuals in a particular cohort of fish changes over time. The very first difficulty one has to deal with is the definition of a cohort. In principle this refers to individuals which have the same birthdate, but this is obviously not a very practical definition, because no two births occur at exactly the same time. For practical purposes therefore, cohort is taken to mean individuals which are born in the same year. The kind of terminology which is typical in the fisheries literature is of the form:

\[ N_{0,y} \] : the number of individuals which are between 0 and 1 years old at the beginning of year \( y \).

If this is a pristine population, i.e. it is not being exploited or fished by humans, then after a year the number of this cohort which has survived is represented by the terminology;

\[ N_{1,y+1} \] : the number of individuals which are in age class 1 at the beginning of year \( y+1 \). These individuals will be between 1 and 2 years old at the beginning of year \( y+1 \).

Survivorship is taken to mean the proportion of individuals that survive from one year to the next. Letting \( s \) represent survivorship:

\[ N_{1,y+1} = s \cdot N_{0,y} \]
Similarly, if survivorship is the same value regardless of age or year, then a logical extension of the terminology given above leads to the following sequence of relationships:

\[
\begin{align*}
N_{2,y+2} &= s N_{1,y+1} \\
N_{3,y+3} &= s N_{2,y+2} \\
N_{4,y+4} &= s N_{3,y+3} \\
N_{5,y+5} &= s N_{4,y+4} \\
\vdots \\
N_{a,y+a} &= s N_{a-1,y+a-1} = s^a N_{0,y}
\end{align*}
\]

If one sets, say \(N_{0,y} = 1000\), and \(s = 0.87\), then the following table and figure shows how cohort numbers decline in relation to age.

<table>
<thead>
<tr>
<th>AGE</th>
<th>Survivors</th>
<th>Survivors</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1000.0</td>
<td>1000.0</td>
</tr>
<tr>
<td>1</td>
<td>1000 x 0.87</td>
<td>870.0</td>
</tr>
<tr>
<td>2</td>
<td>1000 x 0.87 x 0.87</td>
<td>756.9</td>
</tr>
<tr>
<td>3</td>
<td>1000 x 0.87 x 0.87 x 0.87</td>
<td>658.5</td>
</tr>
<tr>
<td>4</td>
<td>1000 x 0.87 x 0.87 x 0.87 x 0.87</td>
<td>572.9</td>
</tr>
<tr>
<td>5</td>
<td>1000 x 0.87 x 0.87 x 0.87 x 0.87 x 0.87</td>
<td>498.4</td>
</tr>
<tr>
<td>6</td>
<td>1000 x 0.87 x 0.87 x 0.87 x 0.87 x 0.87 x 0.87</td>
<td>433.6</td>
</tr>
<tr>
<td>7</td>
<td>1000 x 0.87 x 0.87 x 0.87 x 0.87 x 0.87 x 0.87 x 0.87</td>
<td>377.3</td>
</tr>
<tr>
<td>8</td>
<td>1000 x 0.87 x 0.87 x 0.87 x 0.87 x 0.87 x 0.87 x 0.87 x 0.87</td>
<td>328.2</td>
</tr>
<tr>
<td>9</td>
<td>1000 x 0.87 x 0.87 x 0.87 x 0.87 x 0.87 x 0.87 x 0.87 x 0.87 x 0.87</td>
<td>285.5</td>
</tr>
<tr>
<td>10</td>
<td>1000 x 0.87 x 0.87 x 0.87 x 0.87 x 0.87 x 0.87 x 0.87 x 0.87 x 0.87 x 0.87 x 0.87</td>
<td>248.4</td>
</tr>
</tbody>
</table>
If one starts with the following generic relationship describing the decline in cohort numbers with age

\[ N_{a+y+a} = s^a N_{0,y} \]

and then take the natural logarithm of each side of this relationship, the result is

\[ \ln(N_{a+y+a}) = \ln(s^a N_{0,y}) \]

This relationship simplifies to:

\[ \ln(N_{a+y+a}) = aln(s) + \ln(N_{0,y}) \]

which has the classic form, \( y = mx + c \), i.e. it is linear, where in our example \( y = \ln(N_{a+y+a}) \), \( x = a \), \( m = \ln(s) \), and \( c = \ln(N_{0,y}) \). What this tells us is that
the slope of a linear regression of \( \ln(N_{a+y}) \) versus a will be an estimate of the logarithm of the survivorship, as indicated in the following figure:

![Logarithm of cohort survivors versus age](image)

\[ y = -0.1393x + 6.9078 \]

It is easy to verify that \( \text{antilog}(-0.1393) = 0.87 \).

Partly because of the existence of this relationship, in fisheries science it is quite common to use the parameter M, where \( M = \ln(s) \), and M is called the natural mortality.

One of the nice things about M is that the instantaneous death rate of the cohort is equal to M times the instantaneous cohort survivor level. The equations shown earlier therefore change to:

\[ N_{i,y+1} = e^{-M} N_{0,y} \]
Another convenience about using natural mortality terminology rather than the survivorship terminology, is that the introduction of fishing involves the following modification to the equations:

\[
N_{x,y+1} = e^{-M-F} N_{0,y} \\
N_{2,y+2} = e^{-M-F} N_{1,y+1} \\
N_{3,y+3} = e^{-M-F} N_{2,y+2} \\
N_{4,y+4} = e^{-M-F} N_{3,y+3} \\
N_{5,y+5} = e^{-M-F} N_{4,y+4} \\
N_{a,y+a} = e^{-M-F} N_{a-1,y+a-1} = e^{-a(M+F)} N_{0,y}
\]

Where \( F \) is the fishing mortality. The following figure compares the fate of the cohort with \( M = 0.1393 \) and \( F = 0.12 \), with and without fishing:
and it is easy to verify that $0.2593 = 0.12 - \ln(0.87)$. 
The actual number of deaths that take place in any year is given by:

Deaths occurring during year $y+a-1 = \text{number present at the beginning of year } y+a-1 - \text{number present at the end of year}

$$y+a = N_{a, y+a-1} - e^{-M,F} N_{a, y+a-1} = N_{a, y+a-1} (1 - e^{-M,F}).$$

**Human populations**

A motivation behind the study of human demography is the quantification of the probability of mortality occurring. The following terminology appears in the literature:

$P_x$; the number of individuals in the population aged $x$ at the middle of the year

$d_x$; the number of deaths aged $x$ during the year

$q_x$; the probability of dying

The number of individuals present at the beginning of the year is approximately $P_x + d_x/2$, and therefore for those present at the beginning of the year the probability of dying is given approximately by the formula:

$$q_x = d_x / (P_x + d_x/2).$$

Standard values of $m_x$ are usually provided, where $m_x = d_x / P_x$, and it straightforward to verify that

$$q_x = 2m_x / (2+m_x).$$

If the values of $q_x$ are provided, as in the following table:

| Age (x) | $q_x$ |
and the radix is 100 000 individuals at birth, i.e. at age \( x = 0 \), then it is clear that the number of individuals surviving to successive ages is given by the following scheme:

<table>
<thead>
<tr>
<th>AGE</th>
<th>( l_x : \text{survivors} )</th>
<th>( l_x : \text{survivors} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>100 000</td>
<td>100 000</td>
</tr>
<tr>
<td>1</td>
<td>100 000 (1-(q_0)) 97 761</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>100 000 (1-(q_0)) (1-(q_1)) 97 584</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>100 000 (1-(q_0)) (1-(q_1)) (1-(q_2)) 97 467</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>100 000 (1-(q_0)) (1-(q_1)) (1-(q_2)) (1-(q_3)) 97 379</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>100 000 (1-(q_0)) (1-(q_1)) (1-(q_2)) (1-(q_3)) (1-(q_4)) 97 315</td>
<td></td>
</tr>
</tbody>
</table>
The actual number of deaths, \( d_x \), is given by the following:

<table>
<thead>
<tr>
<th>AGE ( d_x )</th>
<th>deaths</th>
<th>( d_x ) : deaths</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>100 000</td>
<td>( q_0 ) 2239</td>
</tr>
<tr>
<td>1</td>
<td>100 000</td>
<td>((1-q_0)) ( q_1 ) 177</td>
</tr>
<tr>
<td>2</td>
<td>100 000</td>
<td>((1-q_0)(1-q_1)) ( q_2 ) 117</td>
</tr>
<tr>
<td>3</td>
<td>100 000</td>
<td>((1-q_0)(1-q_1)(1-q_2)) ( q_3 ) 88</td>
</tr>
<tr>
<td>4</td>
<td>100 000</td>
<td>((1-q_0)(1-q_1)(1-q_2)(1-q_3)) ( q_4 ) 64</td>
</tr>
<tr>
<td>5</td>
<td>100 000</td>
<td>((1-q_0)(1-q_1)(1-q_2)(1-q_3)(1-q_4)) ( q_5 ) 56</td>
</tr>
</tbody>
</table>

The kind of question that can be answered using these life tables include for example:

- The proportion of persons surviving from one age \( a_1 \) to a higher age \( a_2 \) =

\[
(1-q_{a_1+1})(1-q_{a_1+2})\ldots(1-q_{a_2}).
\]

- The proportion of persons presently aged \( a_1 \) which will die while aged \( a_2 \) last birthday = \((1-q_{a_1+1})(1-q_{a_1+2})\ldots(1-q_{a_2})\) \( d_{a_2} \).

- The number of individuals expected to die between ages \( a_1 \) and \( a_2 \) if the radix is \( l_{a_3} \), where \( a_3 < a_1 \) and \( a_1 < a_2 \). Answer = \( l_{a_3}(l_{a_2} - l_{a_1})/l_{a_3} \).

Assume that life tables have been constructed for males and females. What is the probability that a female aged \( f_1 \) and a male aged \( m_1 \) will both die within the next \( n \) years. Answer = \( (l_{f_1+n} - l_{f_1})/l_{f_1} X (l_{m_1+n} - l_{m_1})/l_{m_1} \).
Stationary populations and $L_x$ and $T_x$

If a population experiences a constant number of births over a long period of time, and is subject to year to year invariant age specific mortality, then births will be exactly balanced by deaths and the population will attain an invariant size and age distribution. This is known as a stationary or life table population. (In biological systems this is referred to as the equilibrium population size and the age distribution is the equilibrium population vector.) This concept is useful when trying to determine quantities such as (a) the average future lifetime of a group of people, or (b) the total number of years likely to be lived in the future by a group of people.

In order to tackle these questions we recall that $l_x/l_0$ is the fraction of individuals born exactly $x$ years ago who are alive at the present time (or reference point in time). Under this definition the number present aged between $x$ and $x+1$ years is given by the integral of $l_x$ between $x$ and $x+1$. A reasonable approximation to this integral under certain circumstances is $L_x$, where

$$L_x = (l_x + l_{x+1})/2.$$  

Circumstances where such an approximation would be fair would be if births and deaths are uniformly distribute throughout the year.

The number of persons aged $x$ or greater in a life table population equals $L_x + L_{x+1} + L_{x+2} + ...$

Another useful quantity is $T_x$, where $T_x = L_x + L_{x+1} + L_{x+2} + ...$, which is the number of individuals older than age $x$ present in the population.

However, $T_x$ is useful in another context, which is apparent from the following discussion. Note that out of $l_{18}$ persons, $l_{19}$ will reach their 19th birthday and will have lived one year since their 18th birthday. The $d_{18}$ persons who died between their 18th and 19th birthday will each have lived on average 0.5 years since their 18th birthday. Collectively the number of years lived between their 18th and 19th birthdays by all people present aged
exactly 18 is \( l_{19} + 0.5d_{18} \). This quantity is also equal to \( l_{19} + 0.5(l_{18} - l_{19}) \) which is equivalent to \( 0.5(l_{18} + l_{19}) \) which in turn is exactly the same as \( L_{18} \) as defined above. Therefore \( L_{18} \) is the number of years lived by all persons between their 18th and 19th birthday, and consequently

\[
T_x = L_x + L_{x+1} + L_{x+2}
\]

is also the number of years lived by all persons after their \( x \)-th birthday. The average number of years lived beyond age \( x \) is given by \( T_x / l_x \) which is equal to a quantity known as the complete expectation of life.

**SPSS Version 11.0 ‘Advanced Models’, ‘Life Tables’ Option**

The life tables which are discussed in the previous section are linked to a broader suite of problems which fall under the general heading ‘Survival Analysis’. Survival analysis may include the quantification of the probability of mortality in human populations, however it may also include more general problems involving, for example, the time to failure of machine components. Also, medical research is likely to make extensive use of such techniques in, for example, comparing the survival rates of control and treatment groups. Typically the time interval appropriate to the latter is likely to be of the order of weeks or month rather than the annual or perhaps biannual time interval associated with actuarial life tables.

The kind of data which is encountered in this broader suite of analytical tasks is typified by the table of data given immediately below:
Table: A dataset typical of the kind suitable for survival analysis

<table>
<thead>
<tr>
<th>Group</th>
<th>Time</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>22</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>28</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>29</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>32</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>37</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>40</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>41</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>54</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>54</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>56</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>61</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>1</td>
<td>63</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>68</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>71</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>89</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>96</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>96</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>125</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>127</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>128</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>131</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>140</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>140</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>141</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>143</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>145</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>146</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>146</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>148</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>158</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>162</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>167</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>168</td>
<td>1</td>
</tr>
</tbody>
</table>
As can be seen from this table, the data provide a record of the status of individuals at particular time intervals, where the status codes are defined as follows:

1 – Died.
2 – Survived.

In this example the status “Died” is the terminal event of interest. Of course, in engineering applications, the terminal event might be “Component failed”, or a similar designation. These data are referred to as censored data since the status “Survived” refers to an individual for whom, at the time that status is recorded, the terminal event has not yet occurred. The reason that this status is recorded prior to the termination of the study is presumably that said individual has been withdrawn from the study. This obviously has an impact on the analysis of the group effect (see table) on the pattern of survivorship over time. However, “Survival Analysis” theory and software takes care of this complication by explicitly modelling the implications that these withdrawals have for the relevant statistical diagnostics.

For the purpose of this workshop, the discussion of survival analysis is contained in the separate handout entitled “Survival Analysis”. The last point to make is to note that the SPSS “Life Tables” option produces a life table analogous to an actuarial life table, using as input a dataset of the form shown above.

Life table created by SPSS Advanced Models using the km.por dataset.

<table>
<thead>
<tr>
<th>x</th>
<th>l_x</th>
<th>n/a</th>
<th>n/a</th>
<th>d_x</th>
<th>( \frac{d}{l_x} )</th>
<th>( \frac{l_x}{l_0} )</th>
<th>( \frac{l_x}{l_0} )</th>
<th>( \frac{d}{l_0} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>173</td>
<td>2</td>
<td>2</td>
<td>181</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>182</td>
<td>2</td>
<td>2</td>
<td>182</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>time interval</td>
<td>interval</td>
<td>risk events</td>
<td>terminating</td>
<td>per interval</td>
<td>at end</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------</td>
<td>----------</td>
<td>-------------</td>
<td>-------------</td>
<td>--------------</td>
<td>--------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>44</td>
<td>0</td>
<td>0</td>
<td>0.0000</td>
<td>0.0000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>44</td>
<td>0</td>
<td>0</td>
<td>0.0000</td>
<td>0.0000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>44</td>
<td>0</td>
<td>2</td>
<td>0.0455</td>
<td>0.0465</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>42</td>
<td>0</td>
<td>1</td>
<td>0.0238</td>
<td>0.0241</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>41</td>
<td>0</td>
<td>1</td>
<td>0.0244</td>
<td>0.0247</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>40</td>
<td>0</td>
<td>0</td>
<td>0.0000</td>
<td>0.0000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>40</td>
<td>0</td>
<td>1</td>
<td>0.0250</td>
<td>0.0253</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>39</td>
<td>0</td>
<td>1</td>
<td>0.0256</td>
<td>0.0260</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>38</td>
<td>0</td>
<td>0</td>
<td>0.0000</td>
<td>0.0000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>38</td>
<td>0</td>
<td>0</td>
<td>0.0000</td>
<td>0.0000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>38</td>
<td>0</td>
<td>1</td>
<td>0.0263</td>
<td>0.0267</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>37</td>
<td>0</td>
<td>0</td>
<td>0.0000</td>
<td>0.0000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>37</td>
<td>0</td>
<td>1</td>
<td>0.0270</td>
<td>0.0274</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
List of References

This chapter is partly based on course material prepared by Professor Rob Dorrington and Dr Tom Moultrie, involving readings written by Pollard, Yusuf and Pollard, and also on the 1983 UN Manual of Indirect Techniques for Demographic Estimation, as supplied to the author by Prof. Rob Shell.
Nuts and bolts of logistic regression

by Prof. John C. Whitehead

Associate Professor of Economics

Department of Economics and Finance
Cameron School of Business
University of North Carolina at Wilmington
601 South College Road
Wilmington, North Carolina 28403-5945

- Why use logistic regression?
- The linear probability model
- The logistic regression model
- Interpreting coefficient
- Estimation by maximum likelihood
- Hypothesis testing
- Evaluating the performance of the model

Why use logistic regression?

There are many important research topics for which the dependent variable is “limited” (discrete not continuous). Researchers often want to analyze whether some event occurred or not, such as voting, participation in a public program, business success or failure, morbidity, mortality, a hurricane, etc.
Binary logistic regression is a type of regression analysis where the dependent variable is a dummy variable (coded 0, 1).

A data set appropriate for logistic regression might look like this:

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Minimum</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>YES</td>
<td>122</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>BAG</td>
<td>122</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>COST</td>
<td>122</td>
<td>9.00</td>
<td></td>
</tr>
<tr>
<td>INCOME</td>
<td>122</td>
<td>5000.00</td>
<td></td>
</tr>
<tr>
<td>Valid N  (listwise)</td>
<td>122</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*This data is from a U.S. Department of the Interior survey (conducted by U.S. Bureau of the Census) which looks at a yes/no response to a question about the “willingness to pay” higher travel costs for deer hunting trips in North Carolina*

**Metadata description**

(Merged) Economic Evaluation and Screener Sections of the 1991 National Survey of Fishing, Hunting, and Wildlife-Associated Recreation (USFWS)

Selected Variables, Cleaned Data Documentation, and Downloadable Data and Programs

- BASS.DAT (n=3302) BASS.SAS
- DEER.DAT (n=6059) DEER.SAS
- TROUT.DAT (n=3238) TROUT.SAS
- WATCH.DAT (n=7213) WATCH.SAS

**Variable** **Variable Description**

- **WTNAME** The non-screener survey weighting variable: the NAME is filled in with ‘deer’, ‘bass’, ‘trout’, and ‘watch’
STATE The state tag ranges from 1-51, excluding DC. See state codes below.

URBAN Response to the question: ‘Do you consider your place of residence to be in a big city or urban area (=3), in a small city or town (=2), or in a rural area (=1)?

AGE Age of the respondent

RACE White=1, Nonwhite=0

RETIRE Has a Job/Business=0, Retired=1, Going to School=2, Keeping House=3, Something Else=4

EMPLOY Has a Job/Business=1, Not Employed=0

MARRIED Married=1, Not Married=0

EDUC Years of completed schooling

INCOME The variable is categorical and coded as the midpoint of the income category:
- $5,000—Under $10,000
- $10,000—$22,500
- $22,500—$24,900
- $24,900—$27,500
- $27,500—$29,900
- $29,900—$40,000
- $40,000—$49,900
- $49,900—$62,500
- $62,500—$74,900
- $74,900—Over $75,000

SEX Female=1, Male=0

TRIPS The total number of (deer, bass, trout, or wildlife watching) trips taken during 1991

AVGCOST The average cost of taking trips = TOTCOST/TRIPS

TOTCOST The total cost of taking trips

YES Response to the Contingent Valuation Question: ‘Would you have taken any trips during 1991 ... if the total cost of all of your trips was $A more than the amount you just reported (TOTCOST)? Yes=1, No=2

A Increase in the total cost of taking trips
AGEHUNT  Age respondent first hunted deer

BAGDEER  Response to: ‘Did you bag a deer in 1991?’ Yes=1, No=0

NUMBAG  Response to: ‘How many deer did you bad during 1991?’

BAGBUCK  Response to: ‘Did you bag a buck in 1991?’ Yes=1, No=0

CATCH  Response to: ‘About how many (bass/trout) did you catch during 1991?’ includes those caught and released

LENGTH  Response to: ‘About what was the average length in inches of the (bass/trout) you caught during 1991?’


State Codes:

01 - Alabama  02 - Alaska  03 - Arizona  04 - Arkansas  05 - California  06 - Colorado  07 - Connecticut  08 - Delaware  10 - Florida  11 - Georgia  12 - Hawaii  13 - Idaho  14 - Illinois  15 - Indiana  16 - Iowa  17 - Kansas  18 - Ken-


The linear probability model

“Why shouldn’t I just use ordinary least squares?”

Good question.

Consider the linear probability (LP) model:

\[ Y = a + BX + e \]

where

1. \( Y \) is a dummy dependent variable, =1 if event happens, =0 if event doesn’t happen,
2. \( a \) is the coefficient on the constant term,
3. \( B \) is the coefficient(s) on the independent variable(s),
4. \( X \) is the independent variable(s), and
5. \( e \) is the error term.

Use of the LP model generally gives you the correct answers in terms of the sign and significance level of the coefficients. The predicted probabilities from the model are usually where we run into trouble. There are three problems with using the LP model:

1. The error terms are heteroskedastic (heteroskedasticity occurs when the variance of the dependent variable is different with different values of
the independent variables):
\[ \text{var}(e) = p(1-p), \] where \( p \) is the probability that \( \text{EVENT}=1 \). Since \( P \) depends on \( X \) the “classical regression assumption” that the error term does not depend on the \( X \)s is violated.

2. \( e \) is not normally distributed because \( P \) takes on only two values, violating another “classical regression assumption”

3. The predicted probabilities can be greater than 1 or less than 0 which can be a problem if the predicted values are used in a subsequent analysis. Some people try to solve this problem by setting probabilities that are greater than (less than) 1 (0) to be equal to 1 (0). This amounts to an interpretation that a high probability of the Event (Nonevent) occurring is considered a sure thing.

The logistic regression model

The “logit” model solves these problems:

\[
\ln\left[\frac{p}{1-p}\right] = a + BX + e \quad \text{or} \quad \frac{p}{1-p} = \exp^a \exp^B \exp^X \exp^e
\]

where:

1. \( \ln \) is the natural logarithm, \( \log_{\exp} \), where \( \exp=2.71828\ldots \)
2. \( p \) is the probability that the event \( Y \) occurs, \( p(Y=1) \)
3. \( p/(1-p) \) is the “odds ratio”
4. \( \ln[p/(1-p)] \) is the log odds ratio, or “logit”
5. all other components of the model are the same.

The logistic regression model is simply a non-linear transformation of the linear regression. The “logistic” distribution is an S-shaped distribution function which is similar to the standard-normal distribution (which results
in a probit regression model) but easier to work with in most applications (the probabilities are easier to calculate). The logit distribution constrains the estimated probabilities to lie between 0 and 1.

For instance, the estimated probability is:

\[ p = \frac{\exp(a + BX)}{1 + \exp(a + BX)} \]

or

\[ p = \frac{1}{1 + \exp(-a - BX)} \]

With this functional form:

1. if you let \( a + BX = 0 \), then \( p = .50 \)
2. as \( a + BX \) gets really big, \( p \) approaches 1
3. as \( a + BX \) gets really small, \( p \) approaches 0.

A graphical comparison of the linear probability and logistic regression models is illustrated here.
Interpreting logit coefficients

The estimated coefficients must be interpreted with care. Instead of the slope coefficients \( B \) being the rate of change in \( Y \) (the dependent variables) as \( X \) changes (as in the LP model or OLS regression), now the slope coefficient is interpreted as the rate of change in the “log odds” as \( X \) changes. This explanation is not very intuitive. It is possible to compute the more intuitive “marginal effect” of a continuous independent variable on the probability. The marginal effect is

\[
\frac{dp}{dB} = f(BX)B
\]

where \( f(.) \) is the density function of the cumulative probability distribution function \( F(BX) \), which ranges from 0 to 1. The marginal effects depend on the values of the independent variables, so it is often useful to evaluate the marginal effects at the means of the independent variables.

SPSS doesn’t have an option for the marginal effects. If you need to compute marginal effects you can use the LIMDEP statistical package which is available on the academic mainframe.

An interpretation of the logit coefficient which is usually more intuitive (especially for dummy independent variables) is the “odds ratio”— \( \exp B \) is the effect of the independent variable on the “odds ratio” [the odds ratio is the probability of the event divided by the probability of the nonevent]. For example, if \( \exp B_3 = 2 \), then a one unit change in \( X_3 \) would make the event twice as likely \((.67/.33)\) to occur.

Odds ratios equal to 1 mean that there is a 50/50 chance that the event will occur with a small change in the independent variable. Negative coefficients lead to odds ratios less than one: if \( \exp B_2 = .67 \), then a one unit change in \( X_2 \) leads to the event being less likely \((.40/.60)\) to occur. [Odds ratios less than 1 (negative coefficients) tend to be harder to interpret than odds ratios greater than one(positive coefficients).] Note that odds ratios for continuous independent variables tend to be close to one, this does
NOT suggest that the coefficients are insignificant. Use the Wald statistic (see below) to test for statistical significance.

**Estimation by maximum likelihood**

[For those of you who just NEED to know ...] Maximum likelihood estimation (MLE) is a statistical method for estimating the coefficients of a model. MLE is usually used as an alternative to non-linear least squares for nonlinear equations.

The likelihood function (L) measures the probability of observing the particular set of dependent variable values \((p_1, p_2, ..., p_n)\) that occur in the sample. It is written as the probability of the product of the dependent variables:

\[
L = \text{Prob} \left( p_1 \times p_2 \times \cdots \times p_n \right)
\]

The higher the likelihood function, the higher the probability of observing the ps in the sample. MLE involves finding the coefficients \((a, B)\) that makes the log of the likelihood function \((LL < 0)\) as large as possible or \(-2\) times the log of the likelihood function \((-2LL)\) as small as possible. The maximum likelihood estimates solve the following condition:

\[
\{Y - p(Y=1)\}X_i = 0, \text{ summed over all observations}
\]

\{or something like that ...

**Hypothesis testing**

Testing the hypothesis that a coefficient on an independent variable is significantly different from zero is similar to OLS models. The Wald statistic for the \(B\) coefficient is:

\[
\text{Wald} = \left(\frac{B}{\text{s.e.}_B}\right)^2
\]
which is distributed chi-square with 1 degree of freedom. The Wald is simply the square of the (asymptotic) t-statistic.

The probability of a YES response from the data above was estimated with the logistic regression procedure in SPSS (click on “statistics,” “regression,” and “logistic”). The SPSS results look like this:

<table>
<thead>
<tr>
<th>Variable</th>
<th>B</th>
<th>S.E.</th>
<th>Wald</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>BAG</td>
<td>2.6390E-001</td>
<td>1.2390E-001</td>
<td>4.5347</td>
<td>1.0</td>
</tr>
<tr>
<td>INCOME</td>
<td>4.6300E-007</td>
<td>1.0700E-005</td>
<td>0.0019</td>
<td>1.0</td>
</tr>
<tr>
<td>COST</td>
<td>-0.0018</td>
<td>0.0007</td>
<td>6.5254</td>
<td>1.0</td>
</tr>
<tr>
<td>Constant</td>
<td>0.9691</td>
<td>0.5690</td>
<td>2.9005</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Notes:
[1] B is the estimated logit coefficient
[2] S.E. is the standard error of the coefficient
[4] "Sig" is the significance level of the coefficient: "the coefficient
[5] The "Partial R" = sqrt([(Wald-2)/(-2*LL(a)])]; see below for LL(a)
[6] Exp(B) is the "odds ratio" of the individual coefficient.

Evaluating the overall performance of the model

There are several statistics which can be used for comparing alternative models or evaluating the performance of a single model:

1. The model likelihood ratio (LR), or chi-square, statistic

\[
LR[i] = -2[LL(a) - LL(a.B)]
\]

or as you are reading SPSS printout:

LR[i] = [-2 Log Likelihood (of beginning model)]
- [-2 Log Likelihood (of ending model)].

where the model LR statistic is distributed chi-square with \( i \) degrees of freedom, where \( i \) is the number of independent variables. The “unconstrained model”, \( LL(a, B) \), is the log-likelihood function evaluated with all independent variables included and the “constrained model” is the log-likelihood function evaluated with only the constant included, \( LL(a) \).

Use the Model Chi-Square statistic to determine if the overall model is statistically significant.

2. The “Percent Correct Predictions” statistic assumes that if the estimated \( p \) is greater than or equal to .5 then the event is expected to occur and not occur otherwise. By assigning these probabilities 0s and 1s the following table is constructed:

![Classification Table for YES](image)

The bigger the % Correct Predictions, the better the model.

3. Most OLS researchers like the \( R^2 \) statistic. It is the proportion of the variance in the dependent variable which is explained by the variance in the independent variables. There is NO equivalent measure in logistic regression. However, there are several “Pseudo” \( R^2 \) statistics. One
psuedo $R^2$ is the McFadden’s-$R^2$ statistic (sometimes called the likelihood ratio index [LRI]):

$\text{McFadden’s-}R^2 = 1 - \frac{LL(a,B)}{LL(a)}$

$= 1 - \frac{-2LL(a,B)}{-2LL(a)}$

where the $R^2$ is a scalar measure which varies between 0 and (somewhat close to) 1 much like the $R^2$ in a LP model. Expect your Pseudo $R^2$s to be much less than what you would expect in LP model, however. Because the LRI depends on the ratio of the beginning and ending log-likelihood functions, it is very difficult to “maximize the $R^2$” in logistic regression.

The Pseudo-$R^2$ in logistic regression is best used to compare different specifications of the same model. Don’t try to compare models with different data sets with the Pseudo-$R^2$ [referees will yell at you ...].

Other Pseudo-$R^2$ statistics are printed in SPSS output but [YIKES!] I can’t figure out how these are calculated (even after consulting the manual and the SPSS discussion list)!!
Source: SPSS Output

<table>
<thead>
<tr>
<th>Source: SPSS Output</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(-2)*Initial LL</td>
<td>[1]</td>
<td>159.526</td>
</tr>
<tr>
<td>(-2)*Ending LL</td>
<td>[2]</td>
<td>147.495</td>
</tr>
<tr>
<td>Goodness of Fit</td>
<td>[3]</td>
<td>123.18</td>
</tr>
<tr>
<td>Cox &amp; Snell-R²</td>
<td></td>
<td>0.094</td>
</tr>
<tr>
<td>Nagelkerke-R²</td>
<td></td>
<td>0.129</td>
</tr>
<tr>
<td>Chi-Square [4]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>df</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>Model</td>
<td></td>
<td>12.031</td>
</tr>
</tbody>
</table>

Notes:

1. LL(a) = 159.526/(-2) = -79.763
2. LL(a,B) = 147.495/(-2) = -73.748
3. GF = [Y - P(Y=1)]²/[Y - P(Y=1)]
4. Chi-Square = -2[LL(a)-LL(a,B)] = 159.526 - 147.45
   McFadden's-R² = 1 - (147.495/159.526) = 0.075

That’s it! You are now a logistic regression expert!

Some potential problems and solutions

Logit models are subject to many of the same problems as in multiple regression:

i) **Omitted variable(s)** can result in bias in the coefficient estimates. To test for omitted variables you can conduct a likelihood ratio test:

- LR[q] = [-2LL(constrained model, i=k-q)] - [-2LL(unconstrained model, i=k)]
- where LR is distributed chi-square with q degrees of freedom, with q = 1 or more omitted variables
This test is conducted automatically by SPSS if you specify “blocks” of independent variables (look for the “block chi-square” in the SPSS output)

ii) The inclusion of irrelevant variable(s) can result in poor model fit. You can consult your Wald statistics or conduct a likelihood ratio test (see above) to search for independent variables with low explanatory power.

iii) Errors in functional form can result in biased coefficient estimates and poor model fit. You should try different functional forms and consult the Wald statistics and model chi-square statistics for overall model fit.

iv) The presence of multicollinearity will not lead to biased coefficients, but the standard errors of the coefficients will be inflated. If a variable which you think should be important (statistically significant) is not, consult the correlation coefficients. Any r(x,y) greater than .4 (.6 - .8 is usually the troublesome range) may be causing the problem.

v) You may have structural breaks in your data. Pooling the data imposes the restriction that an independent variable has the same effect on the dependent variable for different groups of data when the opposite may be true. You can conduct a likelihood ratio test:

\[ LR[i+1] = -2LL(\text{pooled model}) - [-2LL(\text{sample 1}) + -2LL(\text{sample 2})] \]

where samples 1 and 2 are pooled, and i is the number of dependent variables.

**Writing up results**

*Some tips:*

First, present descriptive statistics in a table. Make it clear that the dependent variable is discrete (0, 1) and not continuous and that you will use logistic regression. Logistic regression is a standard statistical procedure so you don’t (necessarily) need to write out the formula for it. You also (usually) don’t need to justify that you are using Logit instead of the LP model.
or Probit (similar to Logit but based on the normal distribution [the tails are less fat]).

“The dependent variable which measures the willingness to take fishing trips at different costs is YES. YES is equal to 1 if the respondent would still take fishing trips ... and 0 otherwise. Since the dependent variable is discrete, the ordinary least squares regression can be used to fit a linear probability model. However, since the linear probability model is heteroskedastic and may predict probability values beyond the (0,1) range, the logistic regression model is used to estimate the factors which influence trip-taking behavior.”

Organize your results in a table (see Table 3) stating your dependent variable (dependent variable = YES) and state that these are “logistic regression results.”

- Present coefficient estimates, t-statistics (or Wald, whichever you prefer), and (at least the) model chi-square statistic for overall model fit.
- If you are comparing several model specifications you should also present the % correct predictions and/or Pseudo-R² statistics to evaluate model performance.
- If you are comparing models with hypotheses about different blocks of coefficients or testing for structural breaks in the data, you could present the ending log-likelihood values. This will allow the reader to check your calculations.

When describing the statistics in the tables, point out the highlights for the reader. What are the significant variables? Is the overall model statistically significant?
The results from Model 1 indicate that anglers behave according to economic theory. As the costs of the trips increase, they are less likely to be willing to continue taking trips. The coefficient on the COST variable has a Wald statistic equal to 13.43 which is significant at the .01 level (99% confidence level) with a critical value of 6.635 [df=1]. The overall model is significant at the .01 level according to the Model chi-square statistic. The model predicts 61% of the responses correctly. The McFadden's $R^2$ is .053.

Which model is preferred?

Model 2 includes two additional theoretically important independent variables: INCOME and CATCH. According to the likelihood ratio test statistic, Model 2 is superior to Model 1 in terms of overall model fit. The block chi-square statistic (note: see below) is significant at the .01 level (critical value = 9.21 [df=2]), the percentage of correct predictions increases by 6%, and the McFadden's $R^2$ value is almost 100% larger. The coefficient on the CATCH and INCOME variables are statistically significant at the .05 and .10 levels.

You usually don’t need to discuss the magnitude of the coefficients—just the sign (+ or -) and statistical significance. If you are doing "risk analysis" interpreting the coefficients with the odds ratio for some other reason, you might briefly describe what it is for an unfamiliar audience.

"The 'odds ratio' for the EMPLOYED coefficient is 3.96 with a 95% confidence interval of [1.23, 12.78]. This suggests that those who are employed are almost 4 times more likely to take trips than those who are unemployed."

If your audience is unfamiliar with the extensions (beyond SPSS or SAS printouts, see below) to logistic regression, discuss the calculation of the statistics in an appendix or footnote or provide a citation. Always state the degrees of freedom for your likelihood-ratio (chi-square) tests (see above quote).
Introduction

The purpose of this paper is to estimate a model to determine the factors which influence bass fishing trips in North and South Carolina. The economic theory behind the analysis can be summarized by the following conceptual model:

Trips = f(cost, income, catch, demographics)

The willingness to take trips is expected to be inversely related with the cost of trips, positively related to the ability to pay for trips (income), and positively related to the quality of trips which is measured by the number of bass caught on all trips (catch).

Methods

The analysis is conducted with data (download SPSS data here) from the U.S. Fish and Wildlife Service of bass anglers from North and South Carolina. The variables used in the empirical analysis are described in Table 1. The descriptive statistics are presented in Table 2. A more complete description of the full data sets can be found in the appendix.
The dependent variable which measures the willingness to take fishing trips at different costs is YES. YES is equal to 1 if the respondent would still take fishing trips if the cost was $[COST]$ higher, where COST is a randomly varied amount, and 0 otherwise. Since the dependent variable is discrete, the ordinary least squares regression can be used to fit a linear probability (LP) model. However, the linear probability model is heteroskedastic and may predict probability values beyond the (0,1) range, the logistic regression model is used to estimate the factors which influence trip-taking behavior (Stynes and Peterson, 1984; Greene, 1997).

**Results**

Logistic regression results are presented in Table 3 (SPSS output). Three models are presented, the dependent variable in each is whether the Carolina bass anglers would continue to take bass fishing trips if the costs of the trips were to increase (YES=1 if they would take more trips, 0 otherwise). Each model includes different blocks of independent variables.

**A Test of Rational Choice Theory**

The most parsimonious model includes only the randomly assigned variable which specifies the higher trips costs (COST). The results from Model 1 indicate that anglers behave according to economic theory. As the costs of the trips increase, they are less likely to be willing to continue taking trips. The coefficient on the COST variable has a Wald statistic equal to 13.43 which is significant at the .01 level (99% confidence level) with a critical value of 6.635 [df=1]. The overall model is significant at the .01 level according to the model chi-square statistic. The model predicts 61% of the responses correctly. The McFadden’s $R^2$ is .053 (Amemiya, 1981).

**Additional Tests of Economic Theory**

Model 2 includes two additional theoretically important independent variables: INCOME and CATCH. According to the block chi-square statistic, Model 2 is superior to Model 1 in terms of overall model fit. The block chi-square statistic is significant at the .01 level (critical value = 9.21
[df=2]), the percentage of correct predictions increases by 6%, and the McFadden’s-$R^2$ value is almost 100% larger. The coefficient on the CATCH and INCOME variables are statistically significant at the .05 and .10 levels.

**Including Demographics**

Model 3 includes demographic variables to determine if social forces plays a role in the willingness to take trips. Males (SEX=1) and those who are EMPLOYED are more likely to take trips with higher costs. None of the other demographic variables are statistically significant according to the Wald test. The block chi-square statistic is significantly different from zero at the .05 level. The percentage correct predictions increases slightly while the McFadden’s-$R^2$ statistic increases by about 4%. According to statistical performance, Model 3 is slightly superior to Model 2.

The income coefficient becomes insignificant in Model 3. This is due to the correlation between income and the other demographic variables, especially employed. This does not suggest that ability to pay is not an important predictor of the willingness to take trips, ability to pay is now measured with the block of independent variables.

The “odds ratio” for the EMPLOYED coefficient is 3.96 with a 95% confidence interval of $[1.23, 12.78]$. This suggests that those who are employed are almost 4 times more likely to take trips than those who are unemployed (see Want, Eddy, and Fitzhugh, 1995). The “odds ratio” for the SEX coefficient is 2.67 with a 95% confidence interval of $[1.05, 6.78]$. This suggests that males are almost 2.67 times more likely to take trips than females. Since the other independent variables are either insignificantly different from zero or continuous, interpretation of there magnitude has little meaning in logistic regression.

**Additional Specification Tests**

Several other regression models were estimated to determine the sensitivity of our results to the geographic location and functional form of the regression model. First, splitting the sample into North and South Carolina resi-
dents we find no difference in the vector of coefficients for Model 3 according to the likelihood ratio test (chi-square=5.81 [8 d.f.]). We also tried two alternative functional forms. The first is a log model where the increased cost, catch, and income variables are logged. The pseudo-$R^2$ and model chi-square statistics both decrease with the log functional form indicating that the linear model is superior in terms of overall model fit. The second model includes a squared income term as an additional independent variable. The Wald statistics on both the income and squared income coefficients become insignificant indicating that this is an inferior specification.

Conclusions

The purpose of this paper was to estimate a model to determine the factors which influence bass fishing trips in North and South Carolina. The empirical results indicate economic theory is supported: anglers respond rationally to increases in fishing trip costs and their ability to pay. Also, the sex of the angler and their employment status has important effects on the willingness to take trips. Further research will be directed at estimating the economic value of these fishing trips for use in benefit-cost analysis (Loomis, 1989).
<table>
<thead>
<tr>
<th>Variable</th>
<th>Variable Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>YES</strong></td>
<td>Response to the question: ‘Would you have taken any trips during 1991 ... if the total cost of all of your trips was $[COST] more than the total cost amount you just reported?’ Yes=1, No=0</td>
</tr>
<tr>
<td><strong>COST</strong></td>
<td>Increase in the total cost of taking bass fishing trips</td>
</tr>
<tr>
<td><strong>CATCH</strong></td>
<td>Response to: ‘About how many bass did you catch during 1991?’ includes those caught and released</td>
</tr>
<tr>
<td><strong>INCOME</strong></td>
<td>The variable is categorical and coded as the midpoint of the income category and divided by 1000: $5,000--Under $10,000 $22,500--Between $10,000 and $19,900 $22,500--Between $20,000 and $24,900 $27,500--Between $25,000 and $29,900 $40,000--Between $30,000 and $49,900 $62,500--Between $50,000 and $74,900 $85,000--Over $75,000</td>
</tr>
<tr>
<td><strong>EMPLOY</strong></td>
<td>Has a Job/Business=1, Not Employed=0</td>
</tr>
</tbody>
</table>
### Table 2: Descriptive Statistics

<table>
<thead>
<tr>
<th>Variables</th>
<th>N</th>
<th>Minimum</th>
<th>Maximum</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGE</td>
<td>196.00</td>
<td>18.00</td>
<td>86.00</td>
<td></td>
</tr>
<tr>
<td>CATCH</td>
<td>196.00</td>
<td>0.00</td>
<td>600.00</td>
<td></td>
</tr>
<tr>
<td>COST</td>
<td>196.00</td>
<td>6.00</td>
<td>924.00</td>
<td></td>
</tr>
<tr>
<td>EDUCATION</td>
<td>196.00</td>
<td>4.00</td>
<td>20.00</td>
<td></td>
</tr>
<tr>
<td>EMPLOYED</td>
<td>196.00</td>
<td>0.00</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>INCOME</td>
<td>196.00</td>
<td>5.00</td>
<td>85.00</td>
<td></td>
</tr>
<tr>
<td>MARRIED</td>
<td>196.00</td>
<td>0.00</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>NC</td>
<td>196.00</td>
<td>0.00</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>SEX</td>
<td>196.00</td>
<td>0.00</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>YES</td>
<td>196.00</td>
<td>0.00</td>
<td>1.00</td>
<td></td>
</tr>
</tbody>
</table>

Valid N (listwise) = 196.00

### Table 3: Logistic Regression Results Dependent Variable = YES

<table>
<thead>
<tr>
<th>Model</th>
<th>Variable</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>Coeffi</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Constant</td>
<td>0.98*</td>
<td>13.91</td>
<td>0.1</td>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td></td>
<td>COST</td>
<td>-0.0019*</td>
<td>13.43</td>
<td>-0.0020*</td>
<td>13.33</td>
<td></td>
</tr>
<tr>
<td></td>
<td>INCOME</td>
<td>0.017*</td>
<td>3.77</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>CATCH</td>
<td>0.057*</td>
<td>6.04</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>AGE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>EDUCATION</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MARRIED</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>EMPLOYED</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SEX</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Model Chi-Square [df] = 14.488[1], 28.367[3], 31
Block Chi-Square [df] = 13.86[2], 1
% Correct Predictions = 60.71, 67.35
McFadden's-R2 = 0.053, 0.105

Note: The Wald statistics are distributed chi-square with 1 degree of freedom.
*Indicates that the coefficient is statistically significant at, at least, the .10 level.
Hands on Application:

So far, we have conducted demonstrations of dichotomous dependent variable analysis with SPSS data.

Now you will be given a chance to perform a logistic regression analysis with a similar SPSS data set (download bass.sav here). This data set from the U.S. Fish and Wildlife Service contains information on North and South Carolinians who like to fish for bass.

First, in SPSS take a look at the data (click on “Statistics,” “Summarize,” and “Descriptives”). Your results should look like this.

Then, run the bivariate logistic regression model in SPSS (click on “statistics,” “regression,” and “logistic”).

choose a dummy dependent variable from the Table above (pick YES),
specify your initial model by choosing one independent variable (try COST),
then click OK.
Here is a list of questions that need to be answered with the logistic regression model:

What is the beginning -2LL?
What is the ending -2LL?
What is the model chi-square statistic?
Is the overall model statistically significant?
What is the % correct predictions?
What is the effect of COST on YES?
Is the independent variable statistically significant?

Hands on Application: Intermediate

Now, choose two additional independent variables (try CATCH and the EMPLOYED dummy variable) and run the multivariate logistic regression model. Answer the following questions:

What is the beginning -2LL?
What is the ending -2LL?
What are the effects of the independent variables on YES?
Which of the independent variables are statistically significant?
What is the interpretation of the coefficient on the dummy variable?
Do the additional independent variables improve the overall performance of the model? [hint: compare the model chi-square and % correct predictions statistics]

Now you are ready for even more exploration of logistic regression in SPSS:
What is the 95% confidence interval for the odds ratio on your independent variables (in the Logistic Regression box, click on “options” then the “CI for exp(B)” box, then “continue”)? Does the confidence interval on the dummy variable include 1?

Save the predicted probabilities for potential use in a later analysis (in the Logistic Regression box, click on “save” and “probabilities” then “continue”).

Calculate the t-statistics on the coefficients for this model.

Calculate the McFadden’s-R² for this model.

Try to “maximize your R² by including more independent variables.

**Hands On Application: Advanced**

Next, run a logistic regression model in SPSS with the bass.sav data. Use YES as the dependent variable and include three independent variables:

Model 1: YES = f(COST, CATCH, INCOME)

Here are some advanced exercises:

Conduct hypothesis tests for groups of coefficients. Run another model adding a “block” of demographic variables: EMPLOYED, EDUCATION, MARRIED, SEX, and AGE (in the Logistic Regression box, click on “Next” then choose the demographic “covariates”). Is the block of variables statistically significant (look for the “block chi-square” statistic in the output)?

Conduct tests for structural breaks in the data. Do North and South Carolinians behave similarly? Run 3 versions of model 1: NC, SC, and pooled (in the Logistic Regression box, click on “select” then click on NC, as your “selection variable”, choose NC=1 as the “rule” and run the logit model; then do the same for NC=0). What is the likelihood ratio test statistic equal to?
Is multicollinearity a problem? Run (1) Model 1 with EMPLOYED, (2) MODEL 1 with EMPLOYED and without INCOME. What are the effects on the statistical significance of INCOME? What is the correlation between EMPLOYED and INCOME?

Conduct more tests for the appropriate model specification. In Model 1: is there a superior functional form? In the SPSS data window, select COST and “transform” and “compute” COST into a new variable: LNCOST=ln(COST). Select INCOME and “transform” and “compute” INCOME into a new variable: INCOMESQ=income^income. Run the alternative functional form:

**MODEL 2: YES = f(LNCOST, CATCH, INCOME)**

**MODEL 3: YES = f(COST, CATCH, INCOME, INCOMESQ)**

Finally, if you need to be convinced that the logistic regression model is superior to the linear probability model, here are some things to check:

Test for normality of dependent variable (choose the “skewness” option when you calculated “descriptive statistics,” if the t-stat on skewness is greater than 2 then the variable is probably non-normal).

Test for heteroskedasticity with the Park test.

Check predicted probabilities from the LP model to determine if they fall outside of the 0, 1 range (save the “unstandardized” predicted value when you run a “regression”, “linear” in SPSS).

**Extensions to other logistic techniques**

Logistic regression is also used to analyze multiple choice data. For example:
Unordered multiple \((j>2)\) choices: travel mode, treatment choice, etc., should be analyzed with the multinomial logit model

\[ p_j = \frac{\exp(B_j X_j)}{\exp(B_1 X_1 + B_2 X_2 + \ldots + B_n X_n)} \]

where \(X\) includes the constant \((1)\).

For ordered multiple \((j>2)\) choices: opinion/attitude surveys, rankings, etc., use the ordered logit model

\[ p_j = F[c_j + B_i X_i] - F[c_{j-1} + B_i X_i] \]

where \(c_j\) is a choice specific constant.

**Syntax of the SPSS logistic regression statement**

LOGISTIC REGRESSION is available in the Regression Models option.

LOGISTIC REGRESSION [VARIABLES =] dependent var

[WITH independent varlist [BY var [BY var] ... ]]

[/CATEGORICAL = var1, var2, ... ]

[/CONTRAST (categorical var) = [{INDICATOR [(refcat)] }]

{DEVIATION [(refcat)] }{SIMPLE [(refcat)] }

{DIFFERENCE }

{HELMERT }

{REPEATED }

{POLYNOMIAL[({1,2,3...})]} {metric }

{SPECIAL (matrix) }
[/METHOD = \{ENTER** } \{ALL \}]]
{BSTEP \{COND\}]} \{varlist\}
{LR }
{WALD}
{FSTEP \{COND\}]}
{LR }
{WALD}
[/SELECT = \{ALL** ]
{varname relation value}
[/\{NOORIGIN**\}]
{ORIGIN }
[/ID = [variable]]
[/PRINT = \{DEFAULT**] [SUMMARY] [CORR] [ALL] [ITER [1]] [GOODFIT]]
{n}
[CI(level)]
[/CRITERIA = [BCON ([0.001**})] [ITERATE([20**})]
[LCON([0.01**}]]
{value } {n } {value }
[PIN([0.05**})] [POUT([0.10**})] [EPS([.00000001**})]]
{value } {value } {value }
[CUT([0.5** }]]
[value ]
Temporary variables created by LOGISTIC REGRESSION are:

**Example**

LOGISTIC REGRESSION PROMOTED WITH AGE, JOBTIME, JOBRATE.

**Overview**

LOGISTIC REGRESSION regresses a dichotomous dependent variable on a set of independent variables (Aldrich and Nelson, 1984; Fox, 1984; Hosmer and Lemeshow, 1989; McCullagh and Nelder, 1989; Agresti, 1990). Categorical independent variables are replaced by sets of contrast variables, each set entering and leaving the model in a single step.

**Options**

**Processing of Independent Variables.** You can specify which independent variables are categorical in nature on the CATEGORICAL subcommand. You can control treatment of categorical independent variables by the CONTRAST subcommand. Seven methods are available for entering independent variables into the model. You can specify any one of them on
the METHOD subcommand. You can also use the keyword BY between variable names to enter interaction terms.

**Selecting Cases.** You can use the SELECT subcommand to define subsets of cases to be used in estimating a model.

**Regression through the Origin.** You can use the ORIGIN subcommand to exclude a constant term from a model.

**Specifying Termination and Model-Building Criteria.** You can further control computations when building the model by specifying criteria on the CRITERIA subcommand.

**Adding New Variables to the Working Data File.** You can save the residuals, predicted values, and diagnostics generated by LOGISTIC REGRESSION in the working data file.

**Output.** You can use the PRINT subcommand to print optional output, use the CASEWISE subcommand to request analysis of residuals, and use the ID subcommand to specify a variable whose values or value labels identify cases in output. You can request plots of the actual and predicted values for each case with the CLASSPLOT subcommand.

**Basic Specification**

· The minimum specification is the VARIABLES subcommand with one dichotomous dependent variable. You must specify a list of independent variables either following the keyword WITH on the VARIABLES subcommand or on a METHOD subcommand.

**PRED**
**PGROUP**
**RESID**
**DEV**
**LEVER**
A statistical primer for post graduates

LRESID
SRESID
ZRESID
COOK
DFBETA

· The default output includes goodness-of-fit tests for the model (–2 log-likelihood, goodness-of-fit statistic, Cox and Snell $R^2$, and NagelKerke $R^2$) and a classification table for the predicted and observed group memberships. The regression coefficient, standard error of the regression coefficient, Wald statistic and its significance level, and a multiple correlation coefficient adjusted for the number of parameters (Atkinson, 1980) are displayed for each variable in the equation.

Subcommand Order

· Subcommands can be named in any order. If the VARIABLES subcommand is not specified first, a slash (/) must precede it.

· The ordering of METHOD subcommands determines the order in which models are estimated.

Different sequences may result in different models.

Syntax Rules

· Only one dependent variable can be specified for each LOGISTIC REGRESSION.

· Any number of independent variables may be listed. The dependent variable may not appear on this list.

· The independent variable list is required if any of the METHOD subcommands are used without a variable list or if the METHOD

[Text continues on the next page]
subcommand is not used. The keyword TO cannot be used on any variable list.

- If you specify the keyword WITH on the VARIABLES subcommand, all independent variables must be listed.

- If the keyword WITH is used on the VARIABLES subcommand, interaction terms do not have to be specified on the variable list, but the individual variables that make up the interactions must be listed.

- Multiple METHOD subcommands are allowed.

- The minimum truncation for this command is LOGI REG.

**Operations**

- Independent variables specified on the CATEGORICAL subcommand are replaced by sets of contrast variables. In stepwise analyses, the set of contrast variables associated with a categorical variable is entered or removed from the model as a single step.

- Independent variables are screened to detect and eliminate redundancies.

- If the linearly dependent variable is one of a set of contrast variables, the set will be reduced by the redundant variable or variables. A warning will be issued, and the reduced set will be used.

- For the forward stepwise method, redundancy checking is done when a variable is to be entered into the model.

- When backward stepwise or direct-entry methods are requested, all variables for each METHOD subcommand are checked for redundancy before that analysis begins.
Limitations

· The dependent variable must be dichotomous for each split-file group. Specifying a dependent variable with more or less than two nonmissing values per split-file group will result in an error.

Example

LOGISTIC REGRESSION PASS WITH GPA, MAT, GRE.

· PASS is specified as the dependent variable.

· GPA, MAT, and GRE are specified as independent variables.

· LOGISTIC REGRESSION produces the default output for the logistic regression of PASS on GPA, MAT, and GRE.

VARIABLES Subcommand

VARIABLES specifies the dependent variable and, optionally, all independent variables in the model. The dependent variable appears first on the list and is separated from the independent variables by the keyword WITH.

· One VARIABLES subcommand is allowed for each Logistic Regression procedure.

· The dependent variable must be dichotomous—that is, it must have exactly two values other than system-missing and user-missing values for each split-file group.

· The dependent variable may be a string variable if its two values can be differentiated by their first eight characters.

· You can indicate an interaction term on the variable list by using the keyword BY to separate the individual variables.
If all METHOD subcommands are accompanied by independent variable lists, the keyword WITH and the list of independent variables may be omitted.

If the keyword WITH is used, all independent variables must be specified. For interaction terms, only the individual variable names that make up the interaction (for example, X1, X2) need to be specified. Specifying the actual interaction term (for example, X1 BY X2) on the VARIABLES subcommand is optional if you specify it on a METHOD subcommand.

**Example**

LOGISTIC REGRESSION PROMOTED WITH AGE, JOBTIME, JOBRATE,

AGE BY JOBTIME.

- PROMOTED is specified as the dependent variable.
- AGE, JOBTIME, JOBRATE, and the interaction AGE by JOBTIME are specified as the independent variables.
- Because no METHOD is specified, all three single independent variables and the interaction term are entered into the model.
- LOGISTIC REGRESSION produces the default output.

**CATEGORICAL Subcommand**

CATEGORICAL identifies independent variables that are nominal or ordinal. Variables that are declared to be categorical are automatically transformed to a set of contrast variables as specified on the CONTRAST subcommand. If a variable coded as is declared as categorical, its coding scheme will be changed to deviation contrasts by default.

- Independent variables not specified on CATEGORICAL are assumed to be at least interval level, except for string variables.
· Any variable specified on CATEGORICAL is ignored if it does not appear either after WITH on the VARIABLES subcommand or on any METHOD subcommand.

· Variables specified on CATEGORICAL are replaced by sets of contrast variables. If the categorical variable has \( n \) distinct values, there will be contrast variables generated.

The set of contrast variables associated with a categorical variable is entered or removed from the model as a step.

· If any one of the variables in an interaction term is specified on CATEGORICAL, the interaction term is replaced by contrast variables.

· All string variables are categorical. Only the first eight characters of each value of a string variable are used in distinguishing between values. Thus, if two values of a string variable are identical for the first eight characters, the values are treated as though they were the same.

**Example**

LOGISTIC REGRESSION PASS WITH GPA, GRE, MAT, CLASS, TEACHER

 /CATEGORICAL = CLASS,TEACHER.

· The dichotomous dependent variable PASS is regressed on the interval-level independent variables GPA, GRE, and MAT and the categorical variables CLASS and TEACHER.

**CONTRAST Subcommand**

CONTRAST specifies the type of contrast used for categorical independent variables. The interpretation of the regression coefficients for categorical variables depends on the contrasts used. The default is INDICATOR. The categorical independent variable is specified in parentheses following
CONTRAST. The closing parenthesis is followed by one of the contrast-type keywords.

- If the categorical variable has \( n \) values, there will be rows in the contrast matrix.

Each contrast matrix is treated as a set of independent variables in the analysis.

- Only one categorical independent variable can be specified per CONTRAST subcommand, but multiple CONTRAST subcommands can be specified. The following contrast types are available. See Finn (1974) and Kirk (1982) for further information on a specific type.

**INDICATOR(refcat)** *Indicator variables.* Contrasts indicate the presence or absence of category membership. By default, refcat is the last category (represented in the contrast matrix as a row of zeros). To omit a category other than the last, specify the sequence number of the omitted category (which is not necessarily the same as its value) in parentheses after the keyword INDICATOR.

**DEVIOATION(refcat)** *Deviations from the overall effect.* This is the default. The effect for each category of the independent variable except one is compared to the overall effect. Refcat is the category for which parameter estimates are not displayed (they must be calculated from the others). By default, refcat is the last category. To omit a category other than the last, specify the sequence number of the omitted category (which is not necessarily the same as its value) in parentheses after the keyword DEVIOATION.

**SIMPLE(refcat)** *Each category of the independent variable except the last is compared to the last category.* To use a category other than the last as the omitted reference category, specify its sequence number (which is not necessarily the same as its value) in parentheses following the keyword SIMPLE.
DIFFERENCE *Difference or reverse Helmert contrasts.* The effects for each category of the independent variable except the first are compared to the mean effects of the previous categories.

HELMERT *Helmert contrasts.* The effects for each category of the independent variable except the last are compared to the mean effects of subsequent categories.

POLYNOMIAL(*metric*) *Polynomial contrasts.* The first degree of freedom contains the linear effect across the categories of the independent variable, the second contains the quadratic effect, and so on. By default, the categories are assumed to be equally spaced; unequal spacing can be specified by entering a metric consisting of one integer for each category of the independent variable in parentheses after the keyword POLYNOMIAL. For example, `CONTRAST(STIMULUS)=POLYNOMIAL(1,2,4)` indicates that the three levels of STIMULUS are actually in the proportion 1:2:4. The default metric is always (1,2,...,k), where k categories are involved. Only the relative differences between the terms of the metric matter: (1,2,4) is the same metric as (2,3,5) or (20,30,50) because the difference between the second and third numbers is twice the difference between the first and second in each instance.

REPEATED *Comparison of adjacent categories.* Each category of the independent variable except the first is compared to the previous category.

SPECIAL(*matrix*) *A user-defined contrast.* After this keyword, a matrix is entered in parentheses with rows and k columns (where k is the number of categories of the independent variable). The rows of the contrast matrix contain the special contrasts indicating the desired comparisons between categories. If the special contrasts are linear combinations of each other, LOGISTIC REGRESSION reports the linear dependency and stops processing. If k rows are entered, the first row is discarded and only the last rows are used as the contrast matrix in the analysis.
**Example**

LOGISTIC REGRESSION PASS WITH GRE, CLASS

/CATEGORICAL = CLASS

/CONTRAST(CLASS)=HELMERT.

- A logistic regression analysis of the dependent variable PASS is performed on the interval independent variable GRE and the categorical independent variable CLASS.

PASS is a dichotomous variable representing course pass/fail status and CLASS identifies whether a student is in one of three classrooms. A HELMERT contrast is requested.

**Example**

LOGISTIC REGRESSION PASS WITH GRE, CLASS

/CATEGORICAL = CLASS

/CONTRAST(CLASS)=SPECIAL(2 -1 -1 0 1 -1).

- In this example, the contrasts are specified with the keyword SPECIAL.

**METHOD Subcommand**

METHOD indicates how the independent variables enter the model. The specification is the METHOD subcommand followed by a single method keyword. The keyword METHOD can be omitted. Optionally, specify the independent variables and interactions for which the method is to be used. Use the keyword BY between variable names of an interaction term.

- If no variable list is specified or if the keyword ALL is used, all of the independent variables following the keyword WITH on the VARIABLES subcommand are eligible for inclusion in the model.

- If no METHOD subcommand is specified, the default method is ENTER.
· Variables specified on CATEGORICAL are replaced by sets of contrast variables. The set of contrast variables associated with a categorical variable is entered or removed from the model as a single step.

· Any number of METHOD subcommands can appear in a Logistic Regression procedure.

METHOD subcommands are processed in the order in which they are specified. Each method starts with the results from the previous method. If BSTEP is used, all remaining eligible variables are entered at the first step. All variables are then eligible for entry and removal unless they have been excluded from the METHOD variable list.

· The beginning model for the first METHOD subcommand is either the constant variable (by default or if NOORIGIN is specified) or an empty model (if ORIGIN is specified).

The available METHOD keywords are:

**ENTER** *Forced entry*. All variables are entered in a single step. This is the default if the METHOD subcommand is omitted.

**FSTEP** *Forward stepwise*. The variables (or interaction terms) specified on FSTEP are tested for entry into the model one by one, based on the significance level of the score statistic. The variable with the smallest significance less than PIN is entered into the model. After each entry, variables that are already in the model are tested for possible removal, based on the significance of the conditional statistic, the Wald statistic, or the likelihood-ratio criterion. The variable with the largest probability greater than the specified POUT value is removed and the model is reestimated. Variables in the model are then evaluated again for removal. Once no more variables satisfy the removal criterion, covariates not in the model are evaluated for entry. Model building stops when no more variables meet entry or removal criteria, or when the current model is the same as a previous one.
**BSTEP** Backward stepwise. As a first step, the variables (or interaction terms) specified on BSTEP are entered into the model together and are tested for removal one by one.

Stepwise removal and entry then follow the same process as described for FSTEP until no more variables meet entry or removal criteria, or when the current model is the same as a previous one. The statistic used in the test for removal can be specified by an additional keyword in parentheses following FSTEP or BSTEP. If FSTEP or BSTEP is specified by itself, the default is COND.

**COND** Conditional statistic. This is the default if FSTEP or BSTEP is specified by itself.

**WALD** Wald statistic. The removal of a variable from the model is based on the significance of the Wald statistic.

**LR** Likelihood ratio. The removal of a variable from the model is based on the significance of the change in the log-likelihood. If LR is specified, the model must be reestimated without each of the variables in the model. This can substantially increase computational time. However, the likelihood-ratio statistic is the best criterion for deciding which variables are to be removed.

**Example**

```
LOGISTIC REGRESSION PROMOTED WITH AGE JOBTIME
JOBRATE RACE SEX AGENCY
/CATEGORICAL RACE SEX AGENCY
/METHOD ENTER AGE JOBTIME
/METHOD BSTEP (LR) RACE SEX JOBRATE AGENCY.
```

· AGE, JOBTIME, JOBRATE, RACE, SEX, and AGENCY are specified as independent variables.
RACE, SEX, and AGENCY are specified as categorical independent variables.

- The first METHOD subcommand enters AGE and JOBTIME into the model.
- Variables in the model at the termination of the first METHOD subcommand are included in the model at the beginning of the second METHOD subcommand.
- The second METHOD subcommand adds the variables RACE, SEX, JOBRA TE, and AGENCY to the previous model.
- Backward stepwise logistic regression analysis is then done with only the variables on the BSTEP variable list tested for removal using the LR statistic.
- The procedure continues until all variables from the BSTEP variable list have been removed or the removal of a variable will not result in a decrease in the log-likelihood with a probability larger than POUT.

**SELECT Subcommand**

By default, all cases in the working data file are considered for inclusion in LOGISTIC REGRESSION. Use the optional SELECT subcommand to include a subset of cases in the analysis.

- The specification is either a logical expression or keyword ALL. ALL is the default. Variables named on VARIABLES, CATEGORICAL, or METHOD subcommands cannot appear on SELECT.
- In the logical expression on SELECT, the relation can be EQ, NE, LT, LE, GT, or GE. The variable must be numeric and the value can be any number.
• Only cases for which the logical expression on SELECT is true are included in calculations.

All other cases, including those with missing values for the variable named on SELECT, are unselected.

• Diagnostic statistics and classification statistics are reported for both selected and unselected cases.

• Cases deleted from the working data file with the SELECT IF or SAMPLE command are not included among either the selected or unselected cases.

Example

LOGISTIC REGRESSION VARIABLES=GRADE WITH GPA, TUCE, PSI

/SELECT SEX EQ 1 /CASEWISE=RESID.

• Only cases with the value 1 for SEX are included in the logistic regression analysis.

• Residual values generated by CASEWISE are displayed for both selected and unselected cases.

ORIGIN and NOORIGIN Subcommands

ORIGIN and NOORIGIN control whether or not the constant is included. NOORIGIN (the default) includes a constant term (intercept) in all equations. ORIGIN suppresses the constant term and requests regression through the origin. (NOCONST can be used as an alias for ORIGIN.)

• The only specification is either ORIGIN or NOORIGIN.

• ORIGIN or NOORIGIN can be specified only once per Logistic Regression procedure, and it affects all METHOD subcommands.
Example

LOGISTIC REGRESSION VARIABLES=PASS WITH GPA,GRE, MAT / ORIGIN.

· ORIGIN suppresses the automatic generation of a constant term.

ID Subcommand
ID specifies a variable whose values or value labels identify the casewise listing. By default, cases are labeled by their case number.

· The only specification is the name of a single variable that exists in the working data file.

· Only the first eight characters of the variable’s value labels are used to label cases. If the variable has no value labels, the values are used.

· Only the first eight characters of a string variable are used to label cases.

PRINT Subcommand
PRINT controls the display of optional output. If PRINT is omitted, DEFAULT output (defined below) is displayed.

· The minimum specification is PRINT followed by a single keyword.

· If PRINT is used, only the requested output is displayed.

DEFAULT Goodness-of-fit tests for the model, classification tables, and statistics for the variables in and not in the equation at each step. Tables and statistics are displayed for each split file and METHOD subcommand.

SUMMARY Summary information. Same output as DEFAULT, except that the output for each step is not displayed.

CORR Correlation matrix of parameter estimates for the variables in the model.
\textbf{ITER(value)} Iterations at which parameter estimates are to be displayed. The value in parentheses controls the spacing of iteration reports. If the value is \( n \), the parameter estimates are displayed for every \( n \)th iteration starting at 0. If a value is not supplied, intermediate estimates are displayed at each iteration.

\textbf{GOODFIT} Hosmer-Lemeshow goodness-of-fit statistic (Hosmer and Lemeshow, 1989).

\textbf{CI(level)} Confidence interval for exp(B). The value in parentheses must be an integer between 1 and 99.

\textbf{ALL} All available output.

\textit{Example}

\begin{verbatim}
LOGISTIC REGRESSION VARIABLES=PASS WITH GPA,GRE,MAT 
/METHOD FSTEP 
/PRINT CORR SUMMARY ITER(2).
\end{verbatim}

· A forward stepwise logistic regression analysis of PASS on GPA, GRE, and MAT is specified.

· The PRINT subcommand requests the display of the correlation matrix of parameter estimates for the variables in the model (CORR), classification tables and statistics for the variables in and not in the equation for the final model (SUMMARY), and parameter estimates at every second iteration (ITER(2)).

\textbf{CRITERIA Subcommand}

CRITERIA controls the statistical criteria used in building the logistic regression models. The way in which these criteria are used depends on the method specified on the METHOD subLOGISTIC command. The default criteria are noted in the description of each keyword below. Iterations will stop if the criterion for BCON, LCON, or ITERATE is satisfied.
**BCON(value)** Change in parameter estimates to terminate iteration. Iteration terminates when the parameters change by less than the specified value. The default is 0.001. To eliminate this criterion, specify a value of 0.

**ITERATE** Maximum number of iterations. The default is 20.

**LCON(value)** Percentage change in the log-likelihood ratio for termination of iterations.

If the log-likelihood decreases by less than the specified value, iteration terminates. The default is 0.01. To eliminate this criterion, specify a value of 0.

**PIN(value)** Probability of score statistic for variable entry. The default is 0.05. The larger the specified probability, the easier it is for a variable to enter the model.

**POUT(value)** Probability of conditional, Wald, or LR statistic to remove a variable. The default is 0.1. The larger the specified probability, the easier it is for a variable to remain in the model.

**EPS(value)** Epsilon value used for redundancy checking. The specified value must be less than or equal to 0.05 and greater than or equal to . The default is . Larger values make it harder for variables to pass the redundancy check—that is, they are more likely to be removed from the analysis.

**CUT(value)** Cutoff value for classification. A case is assigned to a group when the predicted event probability is greater than or equal to the cutoff value. The cutoff value affects the value of the dichotomous derived variable in the classification table, the predicted group (PGROUP on CASEWISE), and the classification plot (CLASSPLOT). The default cutoff value is 0.5. You can specify a value between 0 and 1 (0 < value < 1).

**Example**

LOGISTIC REGRESSION PROMOTED WITH AGE JOBTIME RACE
A backward stepwise logistic regression analysis is performed for the
dependent variable PROMOTED and the independent variables AGE,
JOBTIME, and RACE.

- CRITERIA alters four of the statistical criteria that control the building
  of a model.

- BCON specifies that if the change in the absolute value of all of the
  parameter estimates is less than 0.01, the iterative estimation process
  should stop. Larger values lower the number of iterations required. Notice
  that the ITER and LCON criteria remain unchanged and that if either of
  them is met before BCON, iterations will terminate. (LCON can be set to 0
  if only BCON and ITER are to be used.)

- POUT requires that the probability of the statistic used to test whether a
  variable should remain in the model be smaller than 0.05. This is more
  stringent than the default value of 0.1.

- PIN requires that the probability of the score statistic used to test
  whether a variable should be included be smaller than 0.01. This makes it
  more difficult for variables to be included in the model than the default
  value of 0.05. 10 12 – 10 8 –

**CLASSPLOT Subcommand**

CLASSPLOT generates a classification plot of the actual and predicted
values of the dichotomous dependent variable at each step.

- Keyword CLASSPLOT is the only specification.

- If CLASSPLOT is not specified, plots are not generated.
Example

LOGISTIC REGRESSION PROMOTED WITH JOBTIME RACE
  /CATEGORICAL RACE
  /CLASSPLOT.

- A logistic regression model is constructed for the dichotomous dependent variable PROMOTED and the independent variables JOBTIME and RACE.
- CLASSPLOT produces a classification plot for the dependent variable PROMOTED. The vertical axis of the plot is the frequency of the variable PROMOTED. The horizontal axis is the predicted probability of membership in the second of the two levels of PROMOTED.

CASEWISE Subcommand

CASEWISE produces a casewise listing of the values of the temporary variables created by LOGISTIC REGRESSION. The following keywords are available for specifying temporary variables (see Fox, 1984). When CASEWISE is specified by itself, the default lists PRED, PGROUP, RESID, and ZRESID. If a list of variable names is given, only those named temporary variables are displayed.

**PRED** *Predicted probability*. For each case, the predicted probability of having the second of the two values of the dichotomous dependent variable.

**PGROUP** *Predicted group*. The group to which a case is assigned based on the predicted probability.

**RESID** *Difference between observed and predicted probabilities*.

**DEV** *Deviance values*. For each case, a log-likelihood-ratio statistic, which measures how well the model fits the case, is computed.

**LRESID** *Logit residual*. Residual divided by the product of PRED and 1–PRED.
**SRESID** *Studentized residual.*

**ZRESID** *Normalized residual.* Residual divided by the square root of the product of PRED and 1−\$\$ PRED.

**LEVER** *Leverage value.* A measure of the relative influence of each observation on the model’s fit.

**COOK** *Analog of Cook’s influence statistic.*

**DFBETA** *Difference in beta.* The difference in the estimated coefficients for each independent variable if the case is omitted.

The following keyword is available for restricting the cases to be displayed, based on the absolute value of SRESID:

**OUTLIER** *(value)* *Cases with absolute values of SRESID greater than or equal to the specified value are displayed.* If OUTLIER is specified with no value, the default is 2.

**Example**

```
LOGISTIC REGRESSION PROMOTED WITH JOBTIME SEX RACE
    /CATEGORYAL SEX RACE
    /METHOD ENTER
    /CASEWISE SRESID LEVER DFBETA.  CASewise produces a casewise listing of the temporary variables SRESID, LEVER, and DFBETA.

    There will be one DFBETA value for each parameter in the model. The continuous variable JOBTIME, the two-level categorical variable SEX, and the constant each require one parameter while the four-level categorical variable RACE requires three parameters. Thus, six values of DFBETA will be produced for each case.
```

MISSING Subcommand
LOGISTIC REGRESSION excludes all cases with missing values on any of the independent variables. For a case with a missing value on the dependent variable, predicted values are calculated if it has nonmissing values on all independent variables. The MISSING subcommand controls the processing of user-missing values. If the subcommand is not specified, the default is EXCLUDE.

**EXCLUDE** *Delete cases with user-missing values as well as system-missing values.* This is the default.

**INCLUDE** *Include user-missing values in the analysis.*

SAVE Subcommand

SAVE saves the temporary variables created by LOGISTIC REGRESSION. To specify variable names for the new variables, assign the new names in parentheses following each temporary variable name. If new variable names are not specified, LOGISTIC REGRESSION generates default names.

- Assigned variable names must be unique in the working data file. Scratch or system variable names (that is, names that begin with # or $) cannot be used.

- A temporary variable can be saved only once on the same SAVE subcommand.

**Example**

LOGISTIC REGRESSION PROMOTED WITH JOBTIME AGE
/SAVE PRED (PREDPRO) DFBETA (DF).

- A logistic regression analysis of PROMOTED on the independent variables JOBTIME and AGE is performed.
SAVE adds four variables to the working data file: one variable named PREDPRO, containing the predicted value from the specified model for each case, and three variables named DF0, DF1, and DF2, containing, respectively, the DFBETA values for each case of the constant, the independent variable JOBTIME, and the independent variable AGE.

**EXTERNAL Subcommand**

EXTERNAL indicates that the data for each split-file group should be held in an external scratch file during processing. This can help conserve memory resources when running complex analyses or analyses with large data sets.

- The keyword EXTERNAL is the only specification.
- Specifying EXTERNAL may result in slightly longer processing time.
- If EXTERNAL is not specified, all data are held internally and no scratch file is written.
References


1997.
Data/Graphics/Text
The graphical representation of quantitative data

*by Robert C-H. Shell*

Objectives

By the end of this chapter you should be able to:

- make an informed decision of how to chose a chart for any given set of data, and draw the chosen diagram (bar chart, pie chart, radar chart, etc.)
- plot descriptive graphics, i.e. pie charts, barcharts
- construct grouped or ungrouped frequency tables from raw numerical data and display the data by means of a histogram or ogive
- understand the principles of regression, create a scatterplot and interpret the results and their significance
<table>
<thead>
<tr>
<th>Data types</th>
<th>Definitions</th>
<th>Examples</th>
<th>Uses</th>
<th>Transformations possible</th>
<th>Graphics</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATTRIBUTE1 Nominal2 Categorical</td>
<td>NameCharacteristic</td>
<td>SHELL, RobM, FO,1 (Sex)</td>
<td>None, vertical bar if time on the X axis or Pie</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linkage Crosstabs Regression Linkage</td>
<td>Binary, Dummy, Sequential,</td>
<td></td>
<td></td>
<td>Horizontal</td>
<td></td>
</tr>
<tr>
<td>ORDINAL</td>
<td>Ranking: high, higher, highest</td>
<td>a, b, c, d, e</td>
<td>Pearson correlation</td>
<td>Any</td>
<td></td>
</tr>
<tr>
<td>Bar</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INTERVAL</td>
<td>DiscreteContinuous</td>
<td>Numeric 1, 2, 3 IQ scores</td>
<td>Regression</td>
<td>Any</td>
<td></td>
</tr>
<tr>
<td>Bar Line Scatterplot</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RATIODiscreteContinuous</td>
<td>0 intercept, numeric</td>
<td>Weight Height Space Time</td>
<td>Regression</td>
<td>Any</td>
<td></td>
</tr>
<tr>
<td>Bar Scatterplot</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The humble vertical bar chart is a powerful summary visual measure. A simple bar chart displays the count for each distinct category as a separate bar, allowing the reader or rather viewer to compare categories visually. For the simple bar graph the bar width is not important (except aestheti-
cally) only the bar length (the height of the bar) is used to indicate the magnitude of the variable.

When the data is aggregated over time a simple bar graph may also be used. Note that the X or horizontal axis is usually reserved for time.

A bar chart can plot several categorical variables in several ways: simple, clustered, stacked, bi-directional, surface chart and finally the radar chart

The radar chart

The Radar Chart shows seasonal or cyclical variation for data which could also be plotted on a bar chart. The radar chart is rarely used but is a superior convention for displaying seasonal or any cyclical fluctuation in univariate or bivariate data.
Frequency curves

Collected data can usually conveniently be considered to be a sample drawn from a large (infinite) population. The frequency polygon for a large population would closely approximate a curve. Such a curve is called a frequency curve. A smooth curve can also be drawn through the points of a cumulative frequency polygon (ogive).

Chart menus in SPSS

The Analyze and Graphs menus are available on all windows, making it easy to generate new output without having to switch windows.

To create a new chart, use the Graphs menu.

Click your right mouse button on any item in the dialog box for a description.

Gallery menu commands are used to modify an existing chart

To Edit a Chart Legend

Double-click on the chart in the Output Navigator to activate the chart for editing in a chart window.

From the menus, choose:

Chart legend...

To create a legend title, simply type one into the Legend Title box.

If labels are present, you can modify them.

To modify a label, select it in the Labels box, edit it in the Selected Label box, and click Change.
Grouped frequency distributions and histograms

A histogram also has bars, but they are plotted along an equal interval scale. The height of each bar is the count of values of a quantitative variable falling within the interval. A histogram shows the shape, center, and spread of the distribution. A normal curve superimposed on a histogram helps you judge whether the data are normally distributed.

Pie charts

A pie chart displays the contribution of parts to a whole. Each slice of a pie chart corresponds to a group defined by a single grouping variable. Select the option under Data in Chart that best describes your data.
MIGRATION AND NON-MIGRATION BY SEX

Migrants (female) 42.0%
Non-migrants (female) 58.0%

Migrants (male) 60.0%
Non-migrants (male) 40.0%

Source: UP/Brown University migration survey dataset, 2000

Stem and leaf diagrams of variables

The Descriptive group allows you to choose stem-and-leaf plots and histograms.
Box and whisker plot of appropriate variables

Summary plot based on the median, quartiles, and extreme values. The box represents the interquartile range which contains the 50% of values. The whiskers are lines that extend from the box to the highest and lowest values, excluding outliers. A line across the box indicates the median.

<table>
<thead>
<tr>
<th>Minimum</th>
<th>Q1</th>
<th>Me</th>
<th>Q3</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.0</td>
<td>10.25</td>
<td>12.5</td>
<td>18.75</td>
<td>29.0</td>
</tr>
</tbody>
</table>

Cross tabulations

(i) Producing output from cross tabulations

(ii) Interpreting cross tabulations

Understanding Pivot tables

You can change the appearance of a table either by editing table properties or by applying a TableLook. Each TableLook consists of a collection of table properties, including general appearance, footnote properties, cell properties, and borders. You can select one of the preset TableLooks or you can create and save a custom TableLook.
Frequency distribution

A frequency table or frequency distribution is an arrangement of data by classes together with the corresponding class frequencies. Note that by grouping the data like this, much of the original detail is lost in the array table. The advantage of the frequency table is that one obtains a clear “overall” picture of the data. The end numbers of the class intervals are called class limits, the lower number being the lower class limit and the larger number the upper class limit. The fourth class in Table 2.1, for example, has 16 as the lower class limit and 19 as an upper class limit. A class interval with no upper or lower class limit (e.g. for luggage data “21 kg and over”) is called an open class interval.

The data considered is continuous and a piece of luggage with weight 10.5 kg would have been measured and recorded as a 11 kg piece of luggage. Thus the “true class limits”, referred to as “class boundaries” for the fourth class in the table are 15.5 to less than 18.5. Note that the class boundaries do not coincide with actual observations, and this removes any doubt as to the placing of the data into classes.

The size or width of a class interval is the difference between the upper and lower class boundaries. The class width (size, length) used in the table is 3 units. This can be calculated by obtaining the difference between any two successive upper class limits or any two successive lower class limits.

The class mark, also called the class midpoint is obtained by adding the lower and upper class limits and dividing by two. Thus the class mark for
the fourth class in the table is 17.5 kg. The class mark is used for histograms, frequency polygons and ogives.

A useful rule of thumb when drawing up a frequency distribution is to divide the range into between 5 and 15 classes, depending on the data, and the class interval chosen should be such that the class marks coincide with actually observed data. The tally method is recommended.

**Frequency table**

How do we construct a frequency distribution?

A large mass of raw data makes little sense until it has been arranged, organized and grouped. Consider for example, the following data in Table 2.2 of monthly salaries of company employees.

**Table 2.2**

Monthly salaries of company employees (in Rands):

- 1603161516561775189515741885190615551807
- 1590148116221626150514611615159017601468
- 179117361878190815982902200211019771581
- 1660155316361666144316461469217513841654
- 1515138516471411149514311400182116731625

One way to make sense of this data would be to arrange it in the form of an array. An easy way to do this is to arrange the data into classes using a table. The entries must be done in a systematic way, for example column by column.

**Step 1**

Determine the data range
Range: difference between maximum and minimum values

**Step 2**
Decide the number of classes.

**Step 3**
**Determine the class width**
Class width: Range/number of classes (906/6 = 151 thus select 200 as class width).

**Step 4**
Determine the class limits
First class interval:
Lower limit: should be smaller or equal to minimum value which is 1384 select 1200 as lower limit
Upper limit: lower limit + class width (1200 + 200 = 1400)
    upper limit = 1400 but in order to define distinct classes the class interval is represented as 1200-<1400.

**Note**
It is sometimes useful to construct a frequency distribution with *unequal* class widths. This will be illustrated at a later stage.

Note that the class intervals are chosen in such a way that there can be no doubt about the placing of an entry.

**Step 5** Tabulate the data values
Entry table for monthly salaries of company employees (in rands)
Many questions concerning the data can now be answered with ease. For example, the highest salary is R2290 per month. The range in salary is R2290 – R1384 = R906. The three highest salaries are R2175, R2200 and R2290 respectively. Another way of handling the raw data is to draw up a tally frequency table, as shown in Table 2.6 below.

Tally table for monthly salaries of company employees (for assigning values to classes)

In the tally column (score column) each stripe represents an entry. The fifth stripe is drawn through the previous four as shown, which makes it easy to add the tally and obtain the frequency (alternatively, the stripes could be used to build squares, with the diagonal representing the fifth stripe).

Alongside the absolute frequency the relative frequency could be obtained as a (%), where the relative frequency

\[ \text{relative frequency} = \frac{\text{class frequency}}{\text{sample size}} \]

Hence relative frequency as a percentage (e.g. for the second class interval it would be.

**Table 2.7**

Frequency table for monthly salaries for company employees (in rands)

(i.e. absolute and % relative frequency distribution)

**Histogram**

A histogram, also called a frequency histogram, is a set of rectangles with bases on a horizontal (x) axis, with centers at the class marks and base lengths equal to class interval sizes. The areas of the rectangles are propor-
tional to class frequencies. If the class intervals all have equal sizes then the heights of the rectangles will also be proportional to the class frequencies.

Consider the data in Table 2.6. The class marks in order of class number are 1300; 1500; 1700; 1900; 2100; 2300. These class marks are plotted at the centers of the bases of the histogram rectangles. The width of the base of each rectangle is equal to 200 units. As there will be no gaps between the rectangles, because the data is continuous, the upper and lower limits of each rectangle base are the corresponding class boundaries, i.e. the “true class limits” see figure 2.1).

A histogram is a graphic display of a frequency distribution

x-axis: monthly salaries
y-axis: frequency

**Histogram of monthly salaries of company employees**

**Question**

What fraction of employees earn a salary between R1600 and R1800?

**Frequency polygon**

A frequency polygon is a line graph of class frequencies plotted against class marks. It can be obtained by connecting the mid points of the tops of the rectangles in the histogram.

By simply changing the vertical scale from frequency to relative frequency, and keeping exactly the same diagram, a histogram or frequency polygon can be changed into a relative frequency histogram or percentage histogram, or relative frequency polygon or percentage polygon.

In Figure 2.2, a histogram of the data and the corresponding frequency polygon are drawn to the same set of axes, to illustrate clearly the relationship between both ways of depicting the data. Note the extensions of the
A statistical primer for post graduates

Frequency polygon on either side of the histogram. Create lower interval (1000-<1200) at zero and upper interval (2400-<2600) at zero to ‘anchor’ the graph. Note also that the sum of the areas of the rectangles in the histogram equals the total area bounded by the frequency polygon and the axis.

Frequency polygon of monthly salaries of company employees superimposed on the histogram.

**Cumulative frequencies**

The total frequency of all values less than the upper class boundary of a given class interval is called the cumulative frequency ($F$). The cumulative frequency divided by the total will give the relative cumulative frequency which could be expressed as a percentage.

A table showing cumulative frequencies is called a cumulative frequency table or cumulative distribution.

A graph showing the cumulative frequencies is called a cumulative frequency polygon or ogive (pronounced oh-jive). Thus it is possible to get relative cumulative frequency distributions, percentage cumulative distributions, relative cumulative frequency polygons, percentage ogives, etc.

**Cumulative frequency distribution (the ogive)**

Cumulative frequency distributions are useful in answering the following questions: What percentage of observations fall below a specified value or above another specified value?

For each class, ask: How many observations fall below this upper class limit?

Each successive cumulative frequency is found by adding current class frequency to the immediately preceding cumulative frequency as in the AIDS ogive below.
Regression analysis

Regression is the estimation of the linear or co-linear relationship between a dependent variable and one or more independent variables or covariates.

(i) Choosing appropriate variables
(ii) Sorting data

This dialog box sorts observations or cases (rows) of the data file based on the values of one or more sorting variables. You can sort cases in ascending or descending order.

If you select multiple sort variables, cases are sorted by each variable within categories of the prior variable on the Sort list. For example, if you select GENDER as the first sorting variable and MINORITY as the second sorting variable, cases will be sorted by minority classification within each gender category.

For string variables, uppercase letters precede their lowercase counterparts in sort order. For example, the string value “Yes” comes before “yes” in sort order.

Scatter diagrams

(i) Simple Scatter plot

Plots two numeric variables against each other.

Select a variable for the Y-axis and a variable for the X-axis.

These variables must be numeric, but should not be in date format.

You can select a variable and move it into the Set Markers by box. Each value of this variable is marked by a different symbol on the scatterplot. This variable may be numeric or string.
You can select a numeric or a string variable and move it into the Label Cases by box. You can label points on the plot with this variable.

If selected, the value labels (or values if no labels are defined) of this variable are used as point labels.

If you do not select a variable to Label Cases by, case numbers can be used to label outliers and extremes.

ii) 3-D Scatterplot Option

Allow you to modify an existing chart. To create a new chart, use the Graphs menu.

Deselect Show subgroups to suppress display of subgroups in scatterplots containing a control variable. Each value of the control variable constitutes a subgroup, and is marked by a different symbol on the scatterplot. This option is available only if a variable was specified to Set Markers by when the original plot was created.

Case Labels controls the display of point labels. The current labeling status is displayed in the drop-down list.

Source of Labels allows you to control how points are labeled. In overlaid scatter-plots, if a case is represented by more than one point, selecting one point highlights all other points; turning the label on or off at one point affects all other points as well. ID variable labels the cases with the value labels (or values, if labels are not defined) of the case-identification variable. This option is not available if the ID variable was not specified for Case Labels when the plot was created. Case number labels the points with case sequence numbers in the data file.

If your data are weighted, you can deselect Use case frequency weights to chart unweighted frequencies. Deselecting this option does not restore cases that were excluded from the chart because of missing or non-positive weights.
The Spikes alternatives allow you to display None, or spikes to Floor, Centroid, or Origin. Spikes can be useful when rotating or printing 3-D plots.

You can select one of the Wireframe alternatives to display either a full wireframe (12 edges), half wireframe (9 edges), or no wireframe. The wireframe boundaries are determined by the minimum and maximum values on the scale axes. If you select no wireframe, no axis titles or axis labels are displayed.

**Introduction to path analysis**

Producing graphical output and testing for statistical significance

(i) Observed significance level

Often called the p value. The basis for deciding whether or not to reject the null hypothesis. It is the probability that a statistical result as extreme as the one observed would occur if the null hypothesis were true. If the observed significance level is small enough, usually less than 0.05 or 0.01, the null hypothesis is rejected.

(ii) Comparing curves: The confidence interval

Confidence Intervals.

By default, a 95% confidence interval for the difference in means is displayed. Enter a value between 1 and 99 to request a different confidence level.

Missing Values.

When you test several variables and data are missing for one or more variables, you can instruct the procedure which cases to include (or exclude):
Exclude missing data analysis by analysis. Each t test uses all cases that have valid data for the variables tested. Sample sizes may vary from test to test.

Exclude cases listwise. Each t test uses only cases that have valid data for all variables used in the requested t tests. The sample size is constant across tests.
The population pyramid:
the demographer’s X-ray and a useful market research tool


Population Pyramid

A very effective and quite widely used method of graphically depicting the age-sex composition of a population is called a population pyramid. A population pyramid is designed to give a detailed picture of the age-sex structure of a population, indicating either single ages, 5-year groups, or other age combinations. The basic pyramid form consists of bars, representing age groups in ascending order from the lowest to the highest, pyramided horizontally on one another (see fig. 8-8). The bars for males are given on the left of a central vertical axis and the bars for females on the right of the axis. The number of males or females in the particular age group is indicated by the length of the bars from the central axis. The age scale is usually shown straddling the central axis.

The elements of a successful population

All deaths from AIDS, 1989 to 1998 (n=454)

<table>
<thead>
<tr>
<th>Age Group</th>
<th>Males</th>
<th>Females</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-4</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>5-9</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>10-14</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>15-19</td>
<td>7.3%</td>
<td>0.9%</td>
</tr>
<tr>
<td>20-24</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>25-29</td>
<td>6.6%</td>
<td>0%</td>
</tr>
<tr>
<td>30-34</td>
<td>3.3%</td>
<td>0%</td>
</tr>
<tr>
<td>35-39</td>
<td>5.7%</td>
<td>0%</td>
</tr>
<tr>
<td>40-44</td>
<td>5.3%</td>
<td>0%</td>
</tr>
<tr>
<td>45-49</td>
<td>4.2%</td>
<td>0%</td>
</tr>
<tr>
<td>50-54</td>
<td>2%</td>
<td>0.4%</td>
</tr>
<tr>
<td>55-59</td>
<td>1.5%</td>
<td>0%</td>
</tr>
<tr>
<td>60-64</td>
<td>0.4%</td>
<td>0%</td>
</tr>
<tr>
<td>65-69</td>
<td>0%</td>
<td>0%</td>
</tr>
</tbody>
</table>

Source: Port Elizabeth: Aids Training, Instruction and Counselling Centre
axis although it may be shown at the right or left of the pyramid only, or both on the right and left, perhaps in terms of both age and year of birth (fig. 8-9). In general, the age groups in a given pyramid must have the same class interval and must be represented by bars of equal thickness. Most commonly pyramids show 5-year age groups.

A special problem is presented in the handling of the oldest age groups. If data are available for the oldest age groups in the standard class interval (e.g., 5-year age groups) until the end of the life span, the upper section of the pyramid would have an elongated needle-like form and convey little information for the space required. On the other hand, the bar for a broad terminal group is not normally used because it would not ordinarily be visually comparable with the bars for the other age groups. For this reason, pyramids are usually truncated at an age group where the data begin to run thin, e.g., 70 to 74 years, 75 to 79 years, or 80 to 84 years (fig. 8-9). An open-end interval may be employed at that point when the total number in the open interval is less than in the preceding standard age group (fig. 8-8); in this case the choice of terminal age group will depend on the particular age distribution. If the available data relate to only a broad open-end interval (e.g., 75 and over), a choice has to be made between truncating the pyramid at the next lower age (i.e., 70-74) or subdividing the data into additional standard age groups until the total number in the open-end interval is smaller than in the preceding standard age group.

Data are not usually tabulated by 5-year age groups throughout the age scale for the smaller geographic subdivisions of countries. Sometimes, for example, 10-year age groups are shown after age 25, 35, or 45, with 65 and over or 75 and over as the terminal age group. If a pyramid is to be constructed for such a population, it is neces-
sary to combine some of the figures so as to obtain totals for 10-year age groups or to subdivide the 10-year data into the component 5-year age groups. Combination creates no problem but results in the loss of some useful information. Discussion of techniques of subdividing age data into component age groups, supplementing that given earlier in this chapter, is given in chapter 22.

Pyramids may be constructed on the basis of either absolute numbers or percents. A special caution to be observed in constructing a "percent" pyramid is to be sure to calculate the percents on the basis of the grand total for the population, including both sexes and all ages (i.e., including the population in a terminal age group not shown in the pyramid). A "percent" pyramid is similar, in the geometric sense of the word, to the corresponding "absolute" pyramid. With an appropriate selection of scales, the two pyramids are identical. The choice of one or the other type of pyramid is more important when pyramids for different dates, areas, or sub-populations are to be compared. Only absolute pyramids can show the differences or changes in the overall size of the total population and in the numbers at each age. Percent pyramids show the differences or changes in the proportional size of each age-sex group. In general, absolute or percent pyramids to be compared should be drawn with the same horizontal scale and with bars of the same thickness. In any case, to minimize visual distortion, for pyramids being compared with respect to their general con-
Comparisons between pyramids for the same area at different dates and between pyramids for different areas or sub-populations may be facilitated by superimposing one pyramid on another either entirely or partly. The pyramids may be distinguished by use of different colors or cross-hatching schemes. Occasionally, in absolute pyramids and invariably in percent pyramids, the relative length of the bars in the two superimposed pyramids reverses at some ages. The drafting then becomes more complicated. For example, if one pyramid is to be drawn exactly over another and if the first pyramid is shown entirely in one color or cross-hatching scheme, then the parts of the bars in the second pyramid extending beyond the bars for the first pyramid would be shown in a second color or cross-hatching scheme, and the parts of the bars in the first pyramid extending beyond the bars for the second pyramid would be shown in a third color or cross-hatching scheme (fig. 8-10). An alternative design is to show the second pyramid wholly or partly off-set from the first one. In this design the first pyramid is presented in the conventional way except that the bars are separated from age to age. The second pyramid is drawn partially superimposed on the first, using the space between the bars wholly or in part (fig. 8-11). Any characteristic which varies by age and sex (e.g. marital status or urban-rural residence) may be added to a general population pyramid, to develop a pyramid which reflects the age-sex distribution of both the general population and the population of the categories of the characteristic (fig. 8-12). Where addi-
tional characteristics beyond age and sex are included in the pyramid, the principles of construction are essentially the same. The bar for each age is subdivided into parts representing each category of the characteristic (e.g., single, married, widowed, divorced, urban, rural). It is important that each category shown separately occupy the same position in every bar relative to the central axis and to the other category or categories shown. Again, if percents are used, all percents are calculated on a single base, the total population. Various cross-hatching schemes or coloring schemes may be used to distinguish the various categories of the characteristic represented in the pyramid. When characteristics are added to a population pyramid, the age-sex distribution is shown most clearly for the innermost category in the pyramid and for the total population covered; the distribution of the other categories is harder to interpret. Population pyramids may also be employed to depict the age-sex distribution of demographic events, such as deaths, marriages, divorces, and migration, during some period. The problem of pyramid construction is sometimes complicated by the presence in the age distribution of the category, “age not reported.” Usually, the proportion of “unknowns” is quite small, as we have seen; and a chart would hardly be affected if the “unknowns” are completely disregarded. If it is assumed that persons of unknown age have in fact the same percent distribution by age as persons of known age, it is simply necessary to compute the percents on the basis of the total number of known age. Pyramids may be analyzed and compared in terms of such characteristics as the relative magni-
The types of population pyramid

Figure 8-13. — Population of France, by Age and Sex: January 1, 1967

Populations with rather different age-sex structures are illustrated by the several pyramids shown in figure 8-13. The pyramid for Costa Rica (1963) has a very broad base and narrows very rapidly. This pyramid illustrates the case of an age-sex structure with a very large proportion of children, a very small proportion of elderly persons, and a low median age, i.e., a relatively "young" population. The pyramid for Sweden (1960) has a relatively narrow base and a middle section of nearly the same dimensions; it does not begin to converge to the vertex until after age 55. This pyramid illustrates the case of an age sex structure with a very small proportion of children, a very large...
A statistical primer for post graduates

 proportion of elderly persons, and a high median age, i.e., a relatively “old” population. The pyramids for India (1961) and the United States (1960) illustrate configurations intermediate between those for Costa Rica and Sweden. We may schematically represent these four types of pyramids by the pyramids for the stable populations which approximately correspond to them. See figure 8-14. These four pyramids were constructed on the basis of the age distributions of the four stable populations with annual growth rates ranging from 0.5 percent to 3.5 percent and with mortality levels ranging from level 23 to level 16, as defined in Coale and Demeny, op. cit. These model life tables are described in ch. 25.

Who gets sampled in the antenatal sun

Health

Source: 1004, Census for Dad, Elizabeth
Immigration

The choice of stable populations in the present discussion is partly arbitrary and should not be interpreted as reflecting a precise correspondence of actual and theoretical distributions. All four hypothetical pyramids sketched here have the same total area. Other variations in the configurations of pyramids of national populations occur. Note, for example, the pyramid for the population of France in 1967 given in figure 8-15. It reflects various irregularities associated with France’s special history and is considered further in the section below on “Analysis of Age Composition in Terms of Demographic Factors.” The pyramids of geographically very small countries and of subgroups of national populations-geographic subdivisions or socioeconomic classes—may have quite different configurations, i.e., they may vary considerably from the relatively smooth triangular and semi-elliptical shapes we have identified. For example, the pyramid for the foreign-born population of Hong Kong in 1961 has the general shape of a top (fig. 8-16). It has an extremely narrow base (i.e., a trivial percentage of children), a considerable bulge in the upper middle section (i.e., a high percentage of older adults), and a substantial asymmetry (in this case, a large excess of males). The age-sex pyramids of the married population, the labor force, heads of households, etc., have their characteristic configurations.
Internal migration

Source: See Mostert, Hofmeyr & Kok, 1990: 28-30
Urban/rural

Figure 8-12. — Percent Distribution of the Population of Ghana by Urban-Rural Residence, Age, and Sex: 1960

Population group

Figure 3-3: Age structures of Africans, Coloureds, Asians and Whites in South Africa: 1985

Sources: Van Tonder, Mostert & Hofmeyr, 1987; Mostert, Van Tonder & Hofmeyr, 1987
Percent Distribution by Age and Sex of the Populations of Sweden, United States, India, and Costa Rica: Around 1960


normalized comparisons
Spatial pyramids

Notice absence of adult males in sending region

Notice demographic pattern of industrial coastal cities

Sources: adapted from Meidany 1999a: 4, Figure 5
Transition over time
The ultimate population pyramid:

The world over time
How to write a research proposal for peer review

by Robert C.-H. Shell and Prof Kobus Oosthuizen

Outcome:

By the end of the module you should be able to write a proposal for a quantitative study which will stand a good chance of generating a bursary or grant

prepared by the late Prof Kobus Oosthuizen, Centre for Population Studies, University of Pretoria and Prof Robert C.-H. Shell, Statistics Department, UWC

Introduction

This outline presents some guidance on preparing a demographic research proposal with a South African theme and for a South African peer review audience.

The proposer must bear in mind the country's general demographic situation and population characteristics so that the analysis will be understandable and relevant to a general academic and policy-making audience who will be among your referees. For example, to write a demographic research proposal for a technical project such as "South African mortality" without mentioning the HIV/AIDS pandemic would be suicide for your proposal. NEVER underestimate the peer review of your proposal nor, on the other hand, overestimate the panel by assuming that the panel will be composed purely of demographers.

The proposal's style will depend primarily on the purpose of the proposal and the researcher's own preferences. Insofar as possible, the proposal should keep technical and non-technical aspects separated. The main text
should always be non-technical, so non-demographers can easily understand it. On any peer review panel there might well be someone from a government department who does not know what a population pyramid is. Bear that person in mind.

Some general elements of expository style

Generally, South African proposers do not write well. There are a set of stylistic shortcomings which are so universal that a list of the most egregious seems useful.

Be prepared to revise. A proposal might require five or more drafts.

If any paragraph is too long (over one page), your agenda and data may have overwhelmed you. The paragraph must be carefully constructed. The paragraph form is the same as the overall form of the paper, viz., initialization of topic, elaboration of argument, presentation and marshaling of evidence and the conclusion. In addition, the paragraph must have allusive, transitional links to the previous paragraph. Watch out for "cut-and-paste" clumsiness, one of the pitfalls of the computer age. Do not boilerplate. Miss any one of these elements and you will fail to achieve good paragraphing style.

In expository prose the paragraph is the equivalent of the sonnet. Treat the paragraph with the same respect.

Avoid long "laundry lists" of points at the beginning of your project.

Do not over use "it" constructions: "it" is usually a personal pronoun and refers to an antecedent. Too many "its" in a single paragraph, let alone a sentence, make it excessively difficult to remember the antecedent of the last "it". Excessive "its" therefore usually result in turgid prose. "It" is also used in impersonal constructions (which have no antecedent) and are also passive in construction, e.g. "It is to be avoided if it is at all possible."
Avoid the passive voice. "The passive is never used by demographers" is passive; "Demographers never use the passive" is active. Which is more direct?

In expository prose avoid all first person constructions, "I," the "royal we," the "editorial we" and so on. Such constructions are invariably pretentious and are used in conjunction with jargon (see below) to puff up a poor proposal.

Avoid all contractions e.g., "can't", "don't", "won't" "isn't".

To split or not to split infinitives? Even though Geoffrey Chaucer, the New York Times and even CBS split infinitives, avoid this practice if you can (e.g., StarTrek: "to boldly go where no man...".).

Avoid hidden gender-specific words, e.g. StarTrek: "to boldly go where no man ....").

Adverbs and adjectives leach the nouns and verbs of their power. Consider the following: "they always leach the powerful nouns and picturesque verb of their natural power." Always be very careful of the word "very". It can creep in very nearly everywhere. Moreover, the excessive overuse of unnecessary adverbs and unneeded adjectives is very juvenile and sophomoric.

Avoid crossed or mixed metaphors, e.g., "Gold was the bedrock on which this mighty empire bloomed."

Avoid elliptical logic.

Avoid tautology, verbiage, verbosity, circumlocution, rambling, elegant variation, padding and saying the same thing over and over again; in short, every form of repetition. Avoid repetition.

The Modern Language Association (MLA) in the USA has now suggested that all Latin citation forms, for example, e.g. loc. cit.(in the place cited), cf. (compare) infra (within or below); supra (above)., etc., and all Roman enumeration schemes, e.g. I, ii, iii, iv, I, II and III are not needed and
moreover, are intimidating, so use English (preferably Anglo-Saxon words), short titles and always Arabic numerals. The use of *ibid.* and *passim* in the word processing age is dangerous, since "cut-and-paste" operations, which move text and footnotes, may result in fatally erroneous and laughable *ibid.* citation antecedents.

Never rely wholly on the spell checker. Many words are homophones, but are incorrect *in situ,* e.g. "their" and "there." If a favorite word is not in your spell checker (say "autochthonous"), do not guess: use a dictionary! If English is not your home language acquire an English-speaking editor. Pay them if need be. Afrikaans-speaking, Xhosa-speaking and Zulu-speaking persons are disadvantaged in the international world of English demographic research proposals. There are few SA panels which do not include some international authority. Bear that person in mind.

Watch out for malapropisms which your spell checker might not catch, e.g. "the discovery of the Witwatersrand resulted in much revenue streaming through the imperial coiffures" (Anon., 1994); "The silent trade was characterized by the movement of bouillon across the Sahara and into Europe, from where it spread east" (Anon, 1995). "How to write a demographic research proposal for a population studies pier review" (Anon, 1999).

Master and use the little book, Strunk & Whyte, *Elements of Style.* Every time you read this short book your style will improve. This book is now online under the trade name "Grammatik," currently bundled with Corel WordPerfect. It is useful but you must learn how to use it. Properly configured, it will be an extra editor.

Avoid all jargon. The poorest proposals are invariably cloaked in excessive jargon, demography is a prime candidate for excessive jargon. Such jargon is sometimes dismissed by peer reviewers as "technicist" or "dead-ly". If you cannot make your point without jargon it is probably not worth making.
Avoid double spaces in your text. Use left justification to spot them (see spaces before the word "them")

**The Introduction**

The introduction should discuss the previous and current estimates of the particular variable (or quality of the age structure from different censuses) including graphs of historical trends which would alert your peer reviewers to the circumstance that you really have done your homework. It should follow with a brief evaluation of earlier estimates, a brief but accurately cited literature review, a discussion of the quality of the information used, and/or the techniques or procedures used. Finally, it should mention the data and techniques used for target estimates.

**Reliability of the data**

This section provides a technical discussion of the evaluation of the new and previous data revealing possible errors in collecting the information or in the areal representation of the data. If data pertain to samples, standard errors of data could be given.

**Methodology**

This section should mention the techniques that could have been applied to the available information and those that might be applied during the estimation process. It should include:

- A brief discussion about why a particular technique was selected.
- Possible data errors and their impact on the estimates.
- Biases that estimates may have because the data used did not match the assumptions required by the techniques.
An explanation of why a particular estimate was selected, if the technique offers more than one result.

The actual presentation of the demographic research proposal

Proposals must be typed. Avoid all borders and any fancy fonts. Be conservative and precise in matters of presentation.

Each page with the exception of the cover page must be numbered, top right. Bottom centre, the default in most word processors, gets mixed up with footnotes.

These points may seem pedantic, but proper form allows for consistency and shows scholarly maturity.

**Technical features of demographic research proposals**

Research proposals should be ordered systematically: It should contain headings and subheadings and pages should be numbered clearly. The pages should only be typed on the one side and the document should be bound in an inexpensive file. Aspects deserving special attention are the following:

**Title page**

The title page should contain the following information:

First name and surname of the proposer

Affiliation of the proposer

Exact title and subtitle of the proposal

Deadline and submission date

Name of the funding body and its full street address and fax number

**Table of contents**
Give a complete list of all the headings and subheadings, with the number of the page on which each begins. The same applies to tables and figures which must be listed separately under the headings "List of Tables" and "List of Figures".

Numbering

Headings and subheadings may be numbered as follows:

1. Introduction (Heading)
   1.1 Statement of the problem (first sub-heading)
   1.1.1 Theoretical perspectives (second sub-heading)

Etc.

Citation schemes, footnotes and endnotes

Citations stop us lying to each other. Do use them. An endless confusion for students is the proper form for citations. Moreover, each discipline seems to insist on a different system.

It is important that you provide proper source citations preferably at the foot of the appropriate page (called footnotes), or at the end of the proposal (called endnotes).

Proposals without citation notes at once raise suspicions. So get into the habit of footnoting. You do not, of course, need to footnote every sentence that you write. But it is desirable to use footnotes in the following circumstances:

whenever you refer to unusual information;
whenever you use an exact quotation,
whenever you use paraphrased material,
whenever you use statistical data,
whenever you refer to an idea or data which is not yours.

Some student proposal writers label their endnotes "Bibliography"! Citations have a different function from a bibliography or "select sources" (see below). Footnotes or endnotes indicate the exact sources of your information. The bibliography, on the other hand, indicates the field of literature you have consulted.

Tables and figures

Tables should be numbered chronologically, starting with table 1. Table headings should be placed above the table. If the table is not drawn up by the proposer, the source must be given below the table. Figures should be numbered in the same way and the titles of figures should appear directly below the figure. As with tables, sources must be given if the figure is not the original work of the proposer. If figures are adapted, this should also be mentioned. It is preferable to place tables and figures in the text and not in separate appendices.

Quotations

The following guidelines pertain to quotations:

Quotations should be placed in double quotation marks " ".

The spelling, grammar and punctuation of quotations must be identical to the wording of the original text.

Where part of a quotation is omitted, the omitted words should be replaced with an ellipsis - ‘dot-dot-dot’ (...).

Words not appearing in the original text should be enclosed within brackets [ ].

If the original text does not contain italics but you wish to emphasise part of the quotation by using italics, the words (emphasis added) should be placed within brackets after the italics in the quotation.
A quotation may be translated into the language in which the proposal is being written. The translation can appear as usual in the text in quotation marks, while the original is supplied in a footnote at the foot of the page. Alternatively, the original quotation may be placed in the text and the translation given in the footnote.

References to sources from which quotations are taken, should appear within brackets directly after the quotation (see below).

**References in the text**

Original sources (books, articles, reports, etc.) should be mentioned when:
- a direct quotation is taken from such a source;
- specific views or opinions of authors are given.

References to sources should be placed within brackets in the text. The following particulars must be supplied:
- the surname(s) of the author(s) if it is not supplied in the text, followed by a comma;
- the year of publication, followed by a colon;
- the page number(s).

Examples:

One author:
- Yaukey (1985: 33) noted that ...
- As one author (McDaniel, 1995: 45) said ...

Two or more authors:
- As Lucas and Meyer (1994: 95) wrote ...
- Some authors (Kok & Gelderblom, 1994: 34) hold that...
Speare, Liu and Tsay (1988: 40) argued that ...

Please note that "et al." means "and others" and it is incorrect to use this abbreviation when referring to two authors. It should be used only in the case of three or more authors, for example:

Mostert, et al. (1998: 84) describe ... (a publication by Mostert, Hofmeyr, Oost-huizen and Van Zyl).

Reference to a publication as a whole:

Zopf (1984) emphasises that ...

Series of references:


Institution:

... (Department of Labour, 1995: 45-48).

Sources

A research proposal must contain a list of all the sources referred to. The list of sources must appear at the end of the proposal after the appendix and must be compiled as follows:

All sources referred to in the text must be listed in alphabetical order according to the author's surname. Publications by the same author must be listed chronologically according to the year of publication. Publications by the same author published in the same year must be listed alphabetically according to their titles and the letters "a", "b", etc. added after the year, e.g. 1982a, 1982b. The abbreviation "anon." must be used if the author of a publication is unknown and "n.d." if the year of publication is not available. Note the use of capitals, punctuation marks and italics in the following examples.
One author:


Two authors:


Collection:


Article in collection:


Journal article:


Thesis/dissertation:


Unpublished manuscript:

Newspaper report:


Personal communication:


Official documents:


The list of references at the end of the proposal


Proper observance of citation protocols always demonstrates the maturity and professionalism of any proposal although reviewers might be reluctant to even verbalise such nitpicking details when they talk among themselves about your proposal, or refer to such shortcomings when they write their rejection report.

Text presentation of population characteristics should be supported by graphics and small tables presenting the most significant findings.

**Graphics**

The presentation of this non-technical section should include simple and elegant graphics, so the levels and trends of the estimated variables may be easily and quickly grasped. Let the values and variables do the talking. Frames and unneeded lines are graphical tautology. Make sure the lines and patterns you do use will survive xeroxing. Keep the graphic simple and always use in-line graphics, that is to say, have the graphic follow the text at the point of being mentioned, as in the next visual space after the expressed thought or adjacent to it, as in this example.
**Top twelve in Africa**

Population infected with HIV (absolute numbers)

<table>
<thead>
<tr>
<th>Country</th>
<th>Millions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Botswana</td>
<td>2</td>
</tr>
<tr>
<td>DRC</td>
<td>2</td>
</tr>
<tr>
<td>Ethiopia</td>
<td>3</td>
</tr>
<tr>
<td>Kenya</td>
<td>4</td>
</tr>
<tr>
<td>Lesotho</td>
<td>5</td>
</tr>
<tr>
<td>Malawi</td>
<td>6</td>
</tr>
<tr>
<td>Mozambique</td>
<td>7</td>
</tr>
<tr>
<td>Nigeria</td>
<td>8</td>
</tr>
<tr>
<td>South Africa</td>
<td>9</td>
</tr>
<tr>
<td>Tanzania</td>
<td>10</td>
</tr>
<tr>
<td>Zambia</td>
<td>11</td>
</tr>
<tr>
<td>Zimbabwe</td>
<td>12</td>
</tr>
</tbody>
</table>


### Tables

Follow the form of an authority like Shryock et al, *The Methods and Materials of Demography.*
The importance of the non-technical elements

The proposal should be clear and concise throughout. It should start with a summary of the main themes and specific objectives considered in the proposal (mortality, fertility, migration, etc.). It should include the impact of the target demographic characteristics on population growth and age structure, with references to social aspects such as urbanization, the potential number of students, size of the labour force, and other pertinent population issues.

This section should not include too many numbers nor any large tables; only the figures required to present a general overview should be presented.

A short section could present the degree of the estimated reliability of the information on population characteristics sought for in the proposal. For instance, one paragraph could discuss possible errors in previous estimates because they were based on data that may not have been collected properly, because they pertained to small samples, or because they were calculated indirectly. A full discussion of such problems should be included in an appendix.

Another section should compare the estimated demographic characteristics with similar estimates for other dates during the past. This section may include an analysis of trends and the meaning and consequences of such trends. It may include a comparison of the country's levels with those of other countries in the same region, or better still, with other areas of the world. The perspective within the proposal should work from the global, comparative aspect to the local or parochial micro study. Proposals which try to do too much nearly always fail. The student is advised to do a manageable topic over which they can obtain total mastery.
Finally, if the country has a policy related to population issues, a comparative analysis of the estimated levels with those targeted in the policy should be made. Some mention should be made specifically to the South African population policy document. Any demographer should be seen to be working for the people if not the state. Nevertheless, if the proposal does imply that this might result in policy being revised this should be politely (not scornfully) stated.

**The appendix: technical aspects of the proposal**

Technical discussion and methodologies used, as well as detailed data tables, should always be presented in an appendix. The technical sections of the proposal (covering population age and sex structure, mortality, etc.) may each follow a similar format.

**A final bibliographic section on "select sources"**

A list of books and articles that the proposal writer has consulted should be presented alphabetically according to author's surname at the end of your proposal. Since the word "bibliography" refers to books the writer of the proposal should use a broader term which can incorporate interviews and such non-book items such as "Reference List".
The social sciences can be value free
by Ernest Nagel

Although Max Weber was a vigorous proponent of a "value-free" social science — i.e., the achievement of rational control over the "values" involved in the various social actions that are exercised by that portion of the human species as objective agents — he appeared to dissociate two areas of social science and distinguish them as never having, or he never having realized that:

1. The concept of culture is a value-free, universalizing phenomenon to us because we consider it to be value-free in nature. In other words, it is an accurate description of the ways in which the different cultures are organized and regulated.

2. The axioms of science are value-free, as they are regarded by scientists who have made advances in the field. It is significant because it reveals relationships which are important to us because they are universal and subject to universal rules. It is also significant because it reveals relationships which are important to us because they are useful to us.

In other words, the concept of culture is a value-free, universalizing phenomenon to us because we consider it to be value-free in nature. In other words, it is an accurate description of the ways in which the different cultures are organized and regulated.

In other words, the concept of culture is a value-free, universalizing phenomenon to us because we consider it to be value-free in nature. In other words, it is an accurate description of the ways in which the different cultures are organized and regulated.
A more substantial social continuity than the personal order is embodied in the idea of a social framework which provides a basis for understanding the social world. This framework is a set of rules and norms that govern social interactions and provide a stable structure for social life. The framework is not static but is subject to change and adaptation over time. It is a product of the collective experience of a group, and it is maintained through socialization and socialization processes. The framework provides a means for individuals to understand and participate in social life.
It has therefore been noted, without saying that the study of social phenomena is as much of an imperative as a serious and well thought out. For, while undertaking this study, it is important that we keep in mind that the social sciences, particularly those that deal with human behavior, are deeply intertwined and that their understanding provides a framework for interpreting human behavior.

The social sciences, particularly those that deal with human behavior, are deeply intertwined and that their understanding provides a framework for interpreting human behavior. It is widely recognized that certain patterns of behavior are observed in all societies, whereas others are unique to specific cultures. These patterns of behavior are often the result of a combination of biological, psychological, and social factors, and they vary widely from one society to another. The study of social phenomena is therefore an important part of the social sciences, and it is important that we keep in mind that the social sciences, particularly those that deal with human behavior, are deeply intertwined and that their understanding provides a framework for interpreting human behavior.

Moreover, the social sciences are concerned with the study of human behavior, and they are concerned with the study of human behavior. The social sciences are concerned with the study of human behavior, and they are concerned with the study of human behavior. The social sciences are concerned with the study of human behavior, and they are concerned with the study of human behavior. The social sciences are concerned with the study of human behavior, and they are concerned with the study of human behavior. The social sciences are concerned with the study of human behavior, and they are concerned with the study of human behavior. The social sciences are concerned with the study of human behavior, and they are concerned with the study of human behavior.
a basic problem — to be resolved by the objective methods of scientific inquiry — concerning the adequacy of proposed means for attaining stipulated ends. Thus, economic values necessarily influence the desirability of a society, which determines, in turn, the kinds of scientific inquiry that are meaningful. For instance, if economic values are deemed more important, then the scientific inquiry should be focused on those aspects of society that directly affect economic well-being.

1. There is a more sophisticated argument for the view that the social sciences cannot be value-free. It maintains that the distinction between fact and value in a scientific inquiry is particularly problematic when the nature of human behavior is being analyzed, since the moral and ethical norms that influence human behavior cannot be separated from the scientific inquiry. Consequently, the results of scientific inquiry are necessarily influenced by the values of the society in which they are conducted.

For example, it has been argued that the social sciences cannot separate the study of human behavior from the values that guide the conduct of scientific inquiry. This is particularly true in the study of social behavior, where the values of the society in which the inquiry is conducted necessarily influence the results. For instance, if the values of a society are directed towards promoting economic growth, then the scientific inquiry will be focused on those aspects of society that contribute to economic well-being.

A statistical primer for postgraduates

By

[Text continues here...]

For the purposes of this course, we will use a simple example to illustrate the concept of statistical analysis. Consider a scenario where a researcher is interested in understanding the relationship between two variables, say, age and income. The researcher collects data from a sample of individuals and calculates the correlation coefficient, which measures the strength and direction of the relationship between the two variables. In this case, a positive correlation coefficient indicates that as age increases, income also tends to increase. This type of analysis is crucial for making informed decisions and predictions in various fields, including economics, psychology, and healthcare.
which some erroneously recognized that, more or less clearly defined, the type of anterior, judged, but which definition is included in a given instance.

It will be helpful to illustrate these two senses of "value judgment" first, and then, with the aid of a few example from biology. Animals with different their senses exhibit the condition known as "animals". An animal's visual sense is a normal number of the visual sense. This is not that, among other things, it is less able to maintain the normal internal temperature than are some other species with a "normal" blood supply and a warm brain. However, although the meaning of the term "animal" can be made quite clear, it is not in fact defined with complete precision; for example, the notion of a "natural" number of feet, which enters into the definition of the term, is not somewhat vague, since this number varies with the individual's age as well as with the side of its body and at different times (such as in an adult's age). But in any case, it is not entirely clear whether the term is natural. An investigator must judge whether the available evidence warrants the conclusion that the evidence is probable. He may perhaps think of evidence as being of several kinds, related to those in actual medical practice, or he may think of criteria as a condition that is related to those in actual medical practice, or he may think of criteria as a condition that is related to those in actual medical practice. It is not less evident that he should reasonably make an appreciating value judgment about a given instance (e.g., that it is undesirable for an animal to continue being sickly), unless he can offer a characteristic judgment about the instance independently of the preceding one (e.g., that the animal is sickly). Accordingly, although the animal's visual sense is not necessarily related by any appreciating judgment, making appreciating judgments is not a necessary condition for making interesting ones.

Let us now apply these distinctions to some of the conclusions argued in the arguments under analysis. In another fact, the term is not used to mean "animal", which is the same as the term "animals". However, it should be noted that "animals" is the same as the term "animal", which is not the same as the term "animals". However, it should be noted that "animals" is the same as the term "animals". However, it should be noted that "animals" is the same as the term "animals".

The author's intention is that it be understood that "animals" is the same as the term "animals". However, it should be noted that "animals" is the same as the term "animals". However, it should be noted that "animals" is the same as the term "animals".

The author's intention is that it be understood that "animals" is the same as the term "animals". However, it should be noted that "animals" is the same as the term "animals". However, it should be noted that "animals" is the same as the term "animals".
A statistical primer for post graduates

at the sociological of religion usual in the difference between intermediary consumer activities and those in nature in being committed to certain values. It is arising in mind that these activities are being distinguished and it can also be that a particular consumer needs an the difference between them sociological examination is in the same like that of the absence of belief, who have also acquired other certain dispositions — even the physiological structure, between, but individualism may be the to the ordinary human and in any is more prominent, in the childcare relationship and surrounding social. Indeed, because of the equation of the times, the interaction sociological is extremely difficult to decide in and the influence of some more towards the. Accounted values to be called an necessary, and if he should write, he may have his relationship unbalanced. 

In every sense, however, those kind of unproblematic grounded situation. But this may, again, who claims that a certain mandated by a given religious context, with the principle that a certain individual is a lack of understanding; a classic

value judgment. In many cases, the discussion is about what the role is necessarily something that is not valued than the values of society, and in this respect, there appears to be no difference moral and biological for that kind of logical thinking. Indeed, it would be absurd as it is undermining certain actions in a profound or emotional, sociological or psychological. The great distinction of the stability of the moral judgments, towards like empathy, honesty or helpfulness in commonly used have a widely recognized positive outcome. Accordingly, anyone who suggests such norms in characterizing human behavior can eventually be assumed to be using an equivalent of one behavior by his approach, but we test them like sociological, that they are useful, and not simply value judgments.

However, although many that certainly not all-exclusively characterizing phenomena as they by social sciences do, the exact construct to assess that always consider when a number of "social description" or "natural" values. In certain cases sometimes also have an empirical, according value association. How, one claim that a social science is lacking and jumps over judgments when the characterization becomes the sociological to sociological that are inconsistent, detailed, or individual can be realized by the equally social claim that a physical is also making such judgments which he finds a particular phenomenon as an unformal, a jump coefficient in suggesting that it is useless. For the most science in this example, the playing is demonstrating some exception in his field of research, and also like the social sciences, he is in addition expressing his discovery of the characteristics being given in these objects.

Moreover — and this is the main burden of the present conclusion — there are no clear reasons for thinking that it is necessarily appropriate to distinguish between the characterizing and the repositioning judgments implicit in these statements, whereas the judgments are assumed by scientists of human affairs as a natural assumption. To the sure, it is not always easy to make the distinction between equivalent in the social sciences. In part because much of the language employed in them is very vague, in part because agency judgments that are implicit in a statement need to be overlooked by which they are judgments to which we are necessarily committed through both expressing a sort of commitment. For it is always useful or convenient to perform this task. For many statements implicitly containing such characterizations and appealing
Robert C.-H. Shell (ed.)

The claim that the universe is orderly is based on the idea that there is a single, correct, and unchanging set of laws that govern all aspects of the universe. This idea has been supported by empirical evidence, which has shown that the laws of nature are consistent and reproducible. However, the claim that the universe is orderly is not without its limitations. For example, the laws of nature are subject to change over time and in different contexts. In addition, the claim that the universe is orderly is based on the assumption that all phenomena can be explained by the laws of nature, which may not always be the case. As a result, the claim that the universe is orderly is not a universally accepted idea, and there are many who argue that it is not possible to prove that the universe is orderly.
A statistical primer for post graduates

287

...
abroad. For example, a theoretical physicist may have to decide between two hypotheses, both of which are consistent with all the available evidence; and a theoretical sociologist may similarly have to choose between two different hypotheses concerning the relative frequency of social change under various social arrangements. But neither of these cases may have any implications for facts associated with the phenomena between which the physicist and sociologist must make their decisions. As a member of a scientific community, so does his decision require the good of all his decisions. He needs only to be aware of the implications of his decisions in the various scientific disciplines.

Moreover, reasoning on the meaning of theoretical statements depends on what particular subject matter under discussion and the theories involved in which hypotheses are to be tested. For the reasoning is entirely general; and reference to some specific subject matter becomes relevant only when a definite theoretical value is to be assigned to the risk one is prepared to take. To take such an eminently scientific decision concerning a given hypothesis, something must be said about the question whether the statistical theory is adequate to support the claim that the true application of statistical methods yields the correct answer. In the present case, the statistical theory is inadequate to support such claims for all other inquiries as well. In short, the claim we have been discussing establishes no difficulty; that supposedly exists in the search for reliable knowledge is, in the study of human behavior which is not, in the natural sciences.

A conclusion of this kind is the most radical of all. It differs from the first variety mentioned above in maintaining that there is a necessary logical connection, and not merely a

and his results of experiments, social and in consequence the influence of social change values in which he is not because of his own social involvement, determinate. This means that all empirical knowledge is determined. The knowledge of the effects of social change is determined. The knowledge of the effects of social change can be changed, its rational value is required for the further changes and every is played for the purpose is therefore not only the same particular stage in the development of human affairs. Accordingly, the substantive results of empirical science are completely the same as the interpretations of social phenomena. The logical contingency and the mutual worth of such acceptances have a "validity"; there is no contingency of social change which is not the expression of a special social phenomenon, of which it reflects the interest and values which a society of the human kind as a stage in its history. Consequently, it is not clear that a social situation can be in a normal situation between the origin of a view and these limited reality; such a situation allegedly cannot be made in inquiries and prominent representations of social relations. In this sense, the universal adequacy of the theory in general is a proposition which covers all present and relevant social situations. As an empirical program of this position possibly.

The historical and social studies of an individual can be interpreted as an attempt to understand and social engineering of the past, and to determine its present and future. Of that man, and few people see the history of mankind could only be distinguished in another for the time that in the earlier period change was not irrevocable and certain will occur which, through later amend.
Robert C.-H., Shell (ed.)
The Internet and research

by Mary Nassimbeni

Centre for Information Literacy
February 2003

Introduction

This handout is intended as a general introduction/overview of the WWW, offering a few examples of useful tools to use; it does not attempt to address individual academic disciplines. The following topics are considered.

1. Types of search tools
2. Search engines
   General search engines
   South African search engines
   Meta-search engines
3. Directories
   Subject directories
   Information gateways
   Speciality subject directories: the invisible web
4. Finding general reference works on the WWW
5. Evaluating information found on the WWW
6. Citing web resources
The WWW provides access to vast databases of information of general and specific use and value to students and researchers, and also acts as a most useful communication tool:

- E-mail discussion lists
- Newsgroups
- Research institutes and their collections
- Scholarly organisations and scholarly journals
- Museums
- Universities and their departments
- Libraries and their catalogues
- Governments and government departments
- Books, documents, reports, newspapers, proceedings of conferences, videos, sound, graphics
- E-commerce
- Search tools to find these and other sites and resources

1 Types of search tools

There is a variety of ways of finding information on the Internet. Below is a typology of search tools, with selected samples in each, and with sample queries to illustrate their capability and features.

2 search engines

Search engines are useful tools for locating information such as graphics, documents, speeches, newspapers, journals, conference proceedings, audio sounds and even video clips and e-mail addresses on the Internet. They enable us to browse through the huge amount of data available over the
Internet, quickly and efficiently. Search engines use robots, spiders, crawlers and worms to browse the Internet, find sites, index them, and add them to their vast databases. These robots and worms constantly roam the Web, looking for sites to add to their databases, and continually update their catalogues by adding either the first 100 words of the site, or the entire site’s pages to their databases. When we enter our search protocol (a string of keywords which describe the topic we want information about) into a search engine’s search box, the search engine quickly runs through its index of sites and then feeds the data back to use in the form of a list of possible sites, and/or documents. Remember that all search engines were developed independently of each other, and thus employ different methods of searching for information, and offer different choices of Internet resources for you to search, such as Newsgroups, the entire Internet, the World Wide Web, or e-mail addresses. The search engine and search methods you choose will depend on the nature of your query. You have learned about general principles of searching electronic resources, and how to construct a search statement.

Many of the search engines make provision for Boolean searching, but often this is not immediately apparent and you need to have a quick look at the ‘Advanced Search’ features in a particular search engine to make sure. The principles are the same for the WWW (e.g. Boolean logic), but conventions differ according to the search engine, or search tool that you are using. ‘OR’ frequently is the default if you do not specify that both terms must be included, for example ‘adolescents teenagers’. Common conventions for Boolean logic, phrase searching and truncation are:

- children AND violence AND television
- +children +violence +television
- Ø dogs NOT bestiality
- Ø dogs –bestiality
- v adolescents OR teenagers
§ “United Nations”, “affirmative action”: phrase searching
q child*: wild card for truncation

There is a very useful table available on a web-site constructed by librarians at Berkeley which summarises the features of their five favourite search engines, and also their search protocols. This is an excellent description of the tools used to search the Internet, and also tutorials.

Go to
http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/FindInfo.html

Examples of general search engines and their URLs are:

Google: http://www.google.com/
alta vista: http://www.altavista.com/
teoma: http://www.teoma.com/

Various facilities are offered by many search engines to improve the relevance of a set of search results. For example, Teoma allows you to refine your search by suggesting methods of limiting the results.

Example of a Simple Search on a General Search Engine Illustrating Filter Features

Go to Teoma and do a search tobacco policy in South Africa. Type in tobacco, +policy +South Africa. From there try the refinement feature.
Searching in specific fields

Some search engines like Google and Altavista have a very nice feature which allows you to limit your search to specific fields in web-sites or documents. So, for example if you wish to retrieve the URL of a web-site, you can do this by typing URL: and then the name of the organisation you are looking for. For more information on these functionalities look at the Berkeley site.

Examples of a Search Limited to Specific Fields

Go to Google, and look for a picture of the World Cup 2003 logo. Click on the image tab then type in “Cricket World Cup 2003”

There is a very useful search engine which allows you to ask for information in natural language as opposed to a search statement as shown above: find Askjeeves at http://www.askjeeves.com/.

South African search engines

If you are searching for information specific to a country, it is a good idea to use a search engine that represents this kind of information. South African examples are:

http://www.ananzi.co.za

http://www.aardvark.co.za/

Example of an advanced Boolean search on Ananzi
Do a search for government policy on HIV/AIDS on Ananzi advanced search. Type in the three search boxes: *HIV/AIDS; government, policy*, taking care to change the wording in the third drop-down box from “must not contain” to “must contain”. Ananzi will combine the three search boxes, using AND

**Meta-Search Engines**

If you are unsure about which search engine to use, you can search several search engines simultaneously by using meta-search engines. Meta-searchers conduct a search simultaneously on several search engines. They make use of popular search engines such as Lycos, Infoseek and AltaVista to conduct searches - some use up to eight search engines to ensure comprehensive searching. Results are usually combined to prevent duplication. The Berkeley Internet tutorial advises that they are a good option for simple searches or a unique term or phrase, rather than a broad topic or complex query. The Berkeley tutorial ranks the first two below as the best in this category. A recent addition to this category of tool is Vivisimo which for two years running has won the Search Engine Watch best meta-search engine award.

- [http://www.ixquick.com/](http://www.ixquick.com/)

Example of a Meta-search with Boolean Logic
Search for information on Vivisimo on the effects of television violence on children. Type in +television +violence +children. Notice the clustering of results into helpful categories.

For more information on search engines and how to use them, go to Search Engine Watch at

http://www.searchenginewatch.com/

3 DIRECTORIES

Directories are tools which offer collections of links to Web pages that are organised into hierarchical subject trees, with easily identifiable relationships between subjects. They look similar to search engines, but they differ in one crucial aspect: they are assembled by humans rather than computer programmes, and the lists of sites are therefore much shorter, but much more accurate. Instead of getting hundreds of “hits”, you will get fewer sites more relevant to what you are looking for.

Subject directories

Recommended examples of the earliest subject directories are:

YAHOO.COM

http://www.yahoo.com

Provides information on an array of subjects; allows users to suggest sites for inclusion
http://galaxy.einet.net/ Searchable subject directory of resources selected and evaluated by Internet librarians

**Information Gateways**

Information gateways guide you to sources of information by presenting subject categories, starting broadly and then moving down to finer divisions. Selecting a category takes you to the next level of detail. The information may include details of conferences, organisations, reports, documents and other forms of information. The items included have been examined by subject experts, thus guaranteeing that the quality is controlled by “peer review.”

**Universal Information Gateways**

Bubl Link (Bulletin Board for Libraries) is an information service for the UK higher education community but also of interest to researchers in other countries. Information is arranged in subject sections that can be browsed and searched.

http://bubl.ac.uk/link/
There is another excellent information gateway, also compiled by librarians, called Infomine, recommended by the Berkeley training web-site as one of their favourites. Find it at http://infomine.ucr.edu/

This gateway has searchable categories, e.g. medicine & biology, science & engineering & computing. It also has browse features.

**Subject-based information gateways**

Subject-based information gateways guide you to sources of information by presenting subject categories, which start broadly and then move down to finer divisions. Selecting a category takes you to the next level of detail, for example you may move from science to astronomy. The information may include details of conferences, organisations, reports, documents and other forms of information. The items included have been examined by subject experts, thus guaranteeing that the quality is controlled by “peer review.” In the United Kingdom there are a number of gateways which were set up for the higher education sector but which are freely available to everybody.

For another excellent SBIG, which is also searchable, go to the Social Sciences Information Gateway (SOSIG), hosted in the United Kingdom at http://www.sosig.ac.uk/. This SBIG, and others for different disciplines can be accessed via the Resource Discovery Network “a co-operative network of Internet Resource Catalogues providing access to descriptions of resources selected for their quality and accuracy by subject specialists throughout the UK academic community”. From here you can find SBIGs for health/medicine, law, architecture, engineering and many more.
Resource Discovery Network offers useful tutorials in a number of subject areas such as. Go to the RDN Virtual Training Suite at

In the USA there are also a number of good subject gateways. The WWW Virtual Library at http://www.vlib.org/Home.html is a subject gateway of selected resources in various subject categories such as business, law, science, and communications. It is one of the early and highly respected gateways compiled by subject experts. The Argus Clearinghouse styles itself as the “premier research library”. It lists selected subject guides in a number of subject categories such as engineering, the environment, and medicine at

http://www.clearinghouse.net/

Its criteria for selection and rating system are widely respected. Academic Infonet provides free access to a number of subject-based gateways, offering access to “quality educational resources”. It is worth exploring at

http://www.academicinfo.net/

There is a good South African site for research in the social sciences and humanities which has subject links which could be useful in these two areas. You can find Yenza at

http://www.nrf.ac.za/yenza
Speciality subject directories to search the “invisible web”

Because of the remarkable growth and developments on the WWW, experts have now differentiated between the “visible web” and the “invisible web”. The visible web is what you get when you search with a general search engine such as Google and Alta Vista. An article in the New York Times of 25 January 2001 reported that traditional search engines reach only a fraction of 1% of what there is on the Web, and that as many as 500 billion pieces of content are hidden from the reach of these search engines. This is the “invisible web” that may be five hundred times larger than the surface Web that the general search engines reach. Of the 500 billion documents on the WWW only one billion have been indexed by search engines. The invisible web consists of searchable information resources the content of which cannot be indexed by traditional search engines, because, for example, it is hidden behind query boxes or in databases not accessible to general search engines. Such resources include specialised databases, material in portable document format (PDF), and dictionaries and are beyond the reach of general search engines, and therefore “invisible” to them (Dahn 2000: 1). Direct Search and Invisibleweb.com are two speciality subject directories that have the ability to search the invisible web and retrieve specialised databases that are often not visible to the traditional search engines.
For more information on directories, and also guidelines on more effective searching go to the Internet Tutorials of the University of Albany Libraries at

http://library/albany.edu/internet/

It is worth subscribing to the Scout Report which offers a weekly service offering a “selection of new and newly discovered Internet resources of interest to researchers and educators”. See

http://scout.cs.wisc.edu/

which gives information about the Scout Reports for Social Sciences, Business, and Science & Engineering. The sites have been evaluated by librarians and other experts, and their quality is assured.

4 finding general reference works on the www

Encyclopaedia Britannica is freely available on the web. A search on EB provides material from the encyclopaedia, magazine articles and also websites. You can find it at
Example: Searching Encyclopaedia Britannica

Do a search for acupuncture

The WWW Virtual Library has a general reference section which gives you access to encyclopaedias, dictionaries, directories and other reference works available online. Find it at

http://home.istar.ca/~obyrne/

Dictionaries are available on a very comprehensive site: for example, Xhosa, Afrikaans, and Roget’s Thesaurus can be found at

http://www.yourdictionary.com/

A number of publishers have made their manuals and handbooks freely available on the WWW. For example the Internet edition of *Merck manual of diagnosis and therapy* is available for browsing and or searching at http://www.merck.com/pubs/mmanual/sections.htm

You would find this by going to a subject based information gateway for medicine such as Omni

http://omni.ac.uk/

or by doing a search on a search engine as Google, typing in the title of the book.
5 EVALUATING WEB RESOURCES

While there is a lot of valuable information on the WWW, there is such a vast amount that it is necessary to be able to discern the useful and significant from the irrelevant, the erroneous, the useless and the dangerous. In the same way that we evaluate what we hear, see on television, read in newspapers or books, or see advertised for authority, authenticity, accuracy or quality, we should judge or measure the value of any facts, figures, reports, results or details found via the Internet. Anyone with the know-how and publishing access rights to an Internet service provider’s database can publish pages on the Web. There are no universal quality controls for Internet publications as there are in the publishing of books or journal articles. In addition, if a user is not an expert in a topic area, it is difficult to determine differences between accurate or inaccurate; reliable or questionable, current or old, groundbreaking or well cited, and authoritative or general and indeterminate. However, this does not mean that information found via the Internet is not authoritative. For example, many print journals have been published online, and will allow complete or partial access to their articles. Similarly many scholarly organisations and learned institutions have a web presence and offer the public access to their resources.

There are a number of useful web-sites that give guidance on the particular criteria to apply to electronic resources. A whole section of Yenza deals with this issue. For example, it refers to an article by Alistair Smith, “Testing the surf: criteria for evaluating Internet resources”, in which he offers a “toolbox” of criteria that librarians can use in evaluating and selecting Internet resources, covering scope, content, design, purpose and audience, reviews, “workability” and cost. Smith’s article provides a useful introduction to the issues involved, and is available online at http://info.lib.uh.edu/pr/v8/n3/smit8n3.html.

The section on Yenza also reminds us that the sites and information found on quality information gateways and subject directories such as the examples earlier in the handout have already been filtered, evaluated and se-
lected by expert teams and can therefore be regarded as reliable. Further examples of such sites are provided, including a number of tutorials and training modules. Find this section at http://www.nrf.ac.za/yenza/internet/evaluate.htm. The RDN Virtual Training Suite also offers a section on how to evaluate materials found on the WWW. There is a useful article on how to evaluate quality on the WWW on the library site of Johns Hopkins University. Find it at: http://www.library.jhu.edu/elp/useit/evaluate/

6 CITING WEB RESOURCES

The browser allows you to cut, copy and paste sections of pages that are of interest. Remember that when you do this to make sure that you have an accurate and complete record of the title, author and url of the item, so that proper attribution may be made; plagiarism is a serious transgression. It is important to know how to reference correctly. The library site of Curtin University has a very good section on how to cite electronic sources. See it at http://lisweb.curtin.edu.au/reference/index.html. There you can choose APA, MLA or Harvard guides depending on the use in your discipline/department.
Searching for sources in the post-Caxtonian Era

A research librarian’s perspective, with special reference to the fields of Statistics, Demography and African Studies

Summary of a presentation

by Sandra Rowoldt

(University of Cape Town Libraries, South Africa)

Research Methodology
Department of Statistics
University of the Western Cape
9 March 2004

Searching the WWW¹

Types of search tools

Search engines (electronic robots)

General: google, altavista; teoma; askjeeves
(Regional e.g. SA: ananzi; aardvark)

Meta: vivisimo; iquick; metacrawler;
Searchenginewatch.com – ratings, reports of engines
Directories (compiled by humans, often librarians)

Subject directories: yahoo; galaxy.com* caveat error on page
Subject info gateways:

bubl.ac.uk;
infomine.ucr.edu
rdn.ac.uk
sosig.ac.uk

regional e.g. SA:

www.nrf.ac.za/yenza

Use Rdn and Yenza for evaluation and quality assessment as well.

Speciality subject directories: invisible web (only 1% of information reached through regular search engines) = “searchable information resources the content of which cannot be indexed by traditional search engines” including in:

Specialised databases; material in .pdf (portable document format); dictionaries, etc.

How do we reach them?

Direct Search and Invisible Web

www.freepint.com/gary/direct.htm
www.invisibleweb.com/

General reference works

Oxford Reference Online - Core Collection

http://www.oxfordreference.com/views/GLOBAL.html

The Core Collection contains about 100 dictionary, language reference, and subject reference works published by Oxford University Press. It is a
fully indexed, cross-searchable database of these books, giving subscribers unprecedented access to a comprehensive information resource. Subjects covered are:

- Art & Architecture
- Bilingual Dictionaries
- Biological Sciences
- Classics
- Computing
- Earth & Environmental Sciences
- Economics & Business
- English Dictionaries & Thesauruses
- English Language Reference
- Food & Nutrition
- General Reference
- History
- Law
- Literature
- Medicine
- Military History
- Mythology & Folklore
- Performing Arts
- Physical Sciences & Mathematics
- Politics & Social Sciences
- Quotations
- Religion & Philosophy
- Science

**Encyclopedia Britannica**

www.britannica.com

**wwwVirtual Library**

http://home.istar.ca/~obyrne/

**Dictionaries**

Afrikaans, Hindi, Xhosa, Roget’s Thesaurus (i.e. not eurocentric)

www.yourdictionary.com

and specialised:
Merck Manual of Diagnosis and Therapy

www.merck.com/pubs/mmanual/sections.htm

Evaluating web resources (see handout)

Resource Discovery Network’s Virtual Training Suite
http://www.vts.rdn.ac.uk/

Citing web resources
Title, Author, url, date accessed
See Curtin University

Some useful sites for African studies

Africa Analysis
http://217.199.168.239/
Publisher Name: Africa Analysis Ltd.
London, United Kingdom
Financial and political – subscription needed

Africa Online
http://www.africaonline.com/site/
Publisher Name: Newsread International
Includes statistical information about the economy, demographics, taxation, and import, export information on Kenya, the Ivory Coast, Ghana,
Swaziland, Tanzania, Uganda and Zimbabwe. Published by Newsread International in Nairobi.

Keywords: the Ivory Coast, Zimbabwe, agricultural exports, Kenya, Ghana, Swaziland, Tanzania, Uganda, population, imports, taxation, income

Subject Section(s): Statistics, Demography, Regional Geography

Resource Type: Books/Book Equivalents

**Africa Pulse**

http://www.africapulse.org/

Africa Pulse is an information portal for the Civil Society sector in the Southern African Development Community. It uses state-of-the-art technology to allow organisations throughout the region to publish content directly to the site. A database of website URLs (website addresses) searchable by category and country on anything from education, conflict and governance, to democracy and human rights also provides a valuable resource to the sector.

**Southern African NGO Network**

SANGONeT


Johannesburg, South Africa

Information portal for the Civil Society sector in the Southern African Development Community

**Africa Research Programme**

http://africa.gov.harvard.edu/

Harvard University
USA
Political economy of sub-Saharan Africa.
Full-text research papers downloadable from site.

**Africa South of the Sahara: Stanford University**


An invaluable and comprehensive site featuring African Studies resources and programmes, devised and maintained by Karen Fung of Stanford University.

**Africa World Press Guide**

http://worldviews.igc.org/awpguide/

Africa World Press
Trenton, NJ 08607 USA and P.O. Box 48, Asmara, Eritrea.
Annotated guide to resources from and about Africa

**African Studies Center, University of Pennsylvania**

http://www.sas.upenn.edu/African_Studies/AS.html

Online resources include ASA, country-specific pages, East African resources, African resources in Africa and other annotated selected www links.

**African Studies Quarterly**

http://web.africa.ufl.edu/asq/
ASQ is indexed in Public Affairs Information Service (PAIS) and by the Gale Group.

ASQ is available only in electronic format. There is no print version. To receive an email when a new issue of the journal is published click the “subscribe” button.

**Afrika No.**

http://www.afrika.no/newsupdate/index.html

The Norwegian Council for Africa (NCA):

The Index on Africa is a gateway to information on Africa on the Internet, with over 2,000 links sorted by country, subject and news. Africa News Update is a free of charge news and background service from the Norwegian Council for Africa

**Postcolonial Web**

http://www.postcolonialweb.org/index.html

Funded by the University Scholars Programme, National University of Singapore

Post-imperial and post-colonial literature in English on the web.

Authors of articles and their institutions are given.
Some useful sites
For demography & statistics
(including HIV/AIDS)

AIDSLine

http://www.hsl.wisc.edu/apps/admin/view_resource.cfm?resource

Coverage: 1980 to present  Update frequency: Weekly
AIDSLINE was NLM’s database containing references to literature on AIDS and HIV and no longer exists as a separate database. It can now be searched for free through the NLM Gateway. The journal literature from AIDSLINE can also be searched as a subset in PubMed by using aids [sb]. Example: tuberculosis AND aids [sb]

AIDSearch

NISC databases

http://www.nisc.com/subscribe/freeaccess_reg.htm

Free Database Registration Form is available for the databases you would like to access:

AIDSearch (MEDLINE AIDS/HIV Subset, AIDSTRIALS & AIDSDRUGS databases); Child Abuse, Child Welfare & Adoption

CIA World Factbook


The CIA World Factbook is maintained by the Central Intelligence Agency in the USA and contains information on recognised countries and regions. Each ‘country profile’ contains brief background information, and facts
data about geography, population, government and politics, economic conditions, transportation systems, conflicts and military, and trade. The information is updated and revised on a regular basis.

Keywords: world statistics, economic trends and profiles, countries, country information, economy, geography, geographical data, population, demographic data, government, transportation, transport

Subject Section(s): Statistics, Demography, Geography, Regional Geography, Comparative Geography, Economics, Tourist Industry

Resource Type: Data

Data Online for Population, Health and Nutrition (DOLPHN)

http://www.phnlp.com/dolphn/

The Data Online for Population, Health and Nutrition (DOLPHN) system is an online statistical data resource containing selected current and historical country-level demographic and health indicator data. The DOLPHN system is designed to provide users with quick and easy access to frequently used statistics, as both a reference and analytical tool. The DOLPHN database has been developed and is maintained by the Population, Health and Nutrition Information (PHNI) Project, a U.S. Agency for International Development (USAID) resource managed by Jorge Scientific Corporation.

Keywords: statistical data

Subject Section(s): International Statistics

Resource Type: Data

Demographic and Health Surveys

http://www.measuredhs.com/
The Demographic and Health Surveys (DHS+) program is a project initiated by the U.S. Agency for International Development (USAID). DHS+ seeks to “increase the use of population, health, and nutrition data for the monitoring and evaluation of programs” and to “improve and institutionalize the collection and use of data by host countries for program monitoring and evaluation of and for policy development decisions.” The site includes a searchable database of publications. + Full Record

Keywords: demographic, SPSS, SAS, raw data, survey data, less developed countries, LDCs, demographic statistics, demography, developing countries, family planning, fertility, health, housing, population policy

Subject Section(s): Social Change, Statistics, Demography, Poverty/Social Exclusion

Resource Type: Organisations/Societies

**Demographic Association of Southern Africa**


DEMSA promotes research, instruction and discussion in and of demography and population matters, stimulates collaborative relationships between persons who are involved with the study of population and the administration of population programmes.

**Demographic Research**


This is a “... free, expedited, peer-reviewed online journal of the population sciences” published by the Max Planck Institute for Demographic Research, Rostock, Germany. The full text articles are available for downloading as .pdf files and the user will require a copy of Adobe Acrobat.
Reader. The journal contents can be searched by: publication number (in reverse date order), author or subject. The site also contains details of how to register for email alerts, as well as giving full information on submission procedures, the review process, copyright notice and a statement of purpose. This site is best viewed using Internet Explorer.

Keywords: Social sciences, research, demography, statistics, population, geography

Subject Section(s): Demography  Demographic Geography

Resource Type: Journals (full text)

GeoWeb

http://geoweb.fao.org/GBR/GeoWEB.exe$ChooseCtry
http://geoweb.fao.org/FRA/GeoWEB.exe$ChooseCtry
http://geoweb.fao.org/ESP/GeoWEB.exe$ChooseCtry

Funded by the European Commission, GeoWeb is a Web based application that allows custom access to various information used by FAO’s Global Information and Early Warning System analysts to assess the crop and food supply situation for all countries in the World’. In fact, GeoWeb is both less and more than that: it concentrates on countries where food supply is problematic, so that some first world countries are less well represented; on the other hand, for those countries that are well represented the information goes well beyond crops and food supply. What is on offer via its clickable maps and pull-down boxes (in both English, French and Spanish) is a wealth of information produced by the UN on both the specific agricultural situation of each country and background political and economic factors. There are pointers, too, to reports from other sources. There is also a facility for designing one’s own maps based on information ‘collected by GIEWS and satellite images from the FAO/ARTEMIS program’. Some
other information, available only to GIEWS partners, is password-protected.

Keywords: food resources, food production, food shortages, statistical data, maps

Subject Section(s): International Statistics, Social Policy, GIS and Cartography, Agricultural and Energy Economics, Agriculture and Rural Development, Poverty and Inequality

Resource Type: Data

**Globastat**


Based on data from the CIA World Factbook, this site hosts a wealth of useful comparative data for approximately 200 countries. There are 140 categories of data, divided up into 8 sections: economy, government, geography, people, communications, transportation, military, and analysis.

Keywords: communications, military, statistics, economy, government, geography, demography, transportation

Subject Section(s): Economics, Statistics, International Statistics, National Statistics, Geography, Comparative Geography

**International Data Base**

[http://www.census.gov/ ipc/www/idbnew.html](http://www.census.gov/ ipc/www/idbnew.html)

Alternative Title: IDB

The International Data Base (IDB) is a computerised source of demographic and socio-economic statistics for all countries of the world. It is produced by the U.S. Census Bureau’s International Programs Center. The IDB combines data from country sources (especially censuses and surveys)
with IPC’s estimates and projections to provide information dating back as far as 1950 and as far ahead as 2050. The full database can be accessed online by selecting the type of data required and the countries and years of interest. As well as being displayed to screen, data can be output in a format designed for importing into a spreadsheet, or can be customised to the user’s preferences. There are also summaries provided for the demographic data available for each country, and on world population statistics.

Keywords: socio-economic indicators, demographic data, population records

Subject Section(s): International Statistics Demography Statistics Demographic Geography

Resource Type: Governmental Bodies

OFFSTATS

(Official Statistics on the Web)

http://www.auckland.ac.nz/lbr/stats/offstats/OFFSTATSmain.htm

This site has been developed by the University of Auckland Library in New Zealand. Its purpose it to provide ‘web sites offering free and easily accessible social, economic and general data from official or similar “quotable” sources, especially those that provide both current data and time series’. Towards this aim this site collects information from a range of sources including central banks and government agencies. The information contained here is searchable by country and so is a highly useful resource to anyone investigating specific countries. Some of the files shown here are available in PDF format only and so require Adobe Acrobat Reader to access them. + Full Record

Keywords: countries, official statistics, statistical information, statistics, data

Subject Section(s): Regional and Country Studies Official Statistics
Resource Type: Data

**The Pop Reporter**

http://prds.infoforhealth.org/index.php

Johns Hopkins Bloomberg School of Public Health
Information & Knowledge for Optimal Health (INFO) Project
Center for Communication Programs, Baltimore, MD 21202

Offering the opportunity to complete personalised research profiles, Pop Reporter then delivers news and research material regularly via email, customized to the expressed areas of interest.

**Popline**

http://db.jhu ccp.org/popinform/basic.html

Johns Hopkins Bloomberg School of Public Health
Information & Knowledge for Optimal Health (INFO) Project
Center for Communication Programs, Baltimore, MD 21202

**Population Reference Bureau**

http://www.prb.org/

Population Reference Bureau
Washington, DC 20009-5728

Data and analysis on international and U.S. population trends and issues such as environment, HIV/AIDS, and reproductive health.

Population dimensions of important social, economic, and political issues.
Particularly strong on information on U.S. and international population trends and their implications

**Poverty in Africa**

http://www4.worldbank.org/afr/poverty/default.cfm

This site presents the work of the World Bank Group’s Africa Poverty Monitoring. Apart from information on technical and other issues around the measurement of poverty and links to partner and related institutions, the main focus is on the extensive collection of surveys from nearly 50 African countries, covering a wide range of statistical subjects. The reports emanate mainly from a variety of sources, including the individual country’s own statistical office. Most are available in .pdf format, for which Adobe Acrobat software is required.

Keywords: africa, poverty, social policy, economic and social development, statistics

Subject Section(s): Social Policy  Official Statistics  International Statistics  Economic Development  Poverty and Inequality

Resource Type: Papers/Reports/Articles (collections)

**the Social Science Information Gateway**

(SOSIG)

http://www.sosig.ac.uk/

The Social Science Information Gateway (SOSIG) is a freely available Internet service which aims to provide a trusted source of selected, high quality Internet information for students, academics, researchers and practitioners in the social sciences, business and law. It is part of the UK Resource Discovery Network.
**SOSIG Internet Catalogue** is the main feature of the SOSIG site and is accessed via the main body of the Home Page where it can be browsed under the subject headings or searched using the main search box at the top of the page. It is an online database of high quality Internet resources. It offers users the chance to read descriptions of resources available over the Internet and to access those resources directly. The Catalogue points to thousands of resources, and each one has been selected and described by a librarian or academic. The catalogue is browsable or searchable by subject area.

**South African Data Archive**

(SADA)

http://www.nrf.ac.za/sada/index.asp

The South African Data Archive (SADA) serves as a broker between a range of data providers (e.g. statistical agencies, government departments, opinion and market research companies and academic institutions) and the research community.

The archive does not only preserve data for future use, but also adds value to the collections. It safeguards datasets and related documentation and attempts to make it as easily accessible as possible for research and educational purposes.

**Statistical Agencies**

http://www.ssb.no/english/links/

Statistics Norway features a facility to locate web-servers at statistical agencies throughout the world. Useful for locating official websites for national statistics globally listed alphabetically by country.
Statistical Profile of Education

in sub-Saharan Africa (SPESSA)


SPESSA is a database containing statistical information about education in sub-Saharan Africa. Developed by the Association for the Development of Education in Africa (ADEA) the database comprises a dataset of over 80 indicators related to education in sub-Saharan Africa using statistical information compiled from UNESCO and World Bank sources. The programme can be downloaded from this website and run on an IBM compatible computer with a minimum of 16 megabytes of random-access memory, using Microsoft Windows 95/98/NT (32-bit) operating system or later. Instructions on how to use the database are available on the website in the “Quick Start” section and there are links to the alphabetical indexes of countries and education indicators stored in the database.

Keywords: Africa, Africa south of the Sahara, educational statistics, socio-economic indicators

Subject Section(s): International Statistics Education

Resource Type: Data

Statistics Glossary

http://www.cas.lancs.ac.uk/glossary_v1.1/main.html

A list of definitions of common statistical and probability terms which can be viewed alphabetically or in subject sections such as Hypothesis Testing, Time Series Data, etc. For each entry a clear, concise definition is often followed by further practical details, relevant equations, symbols and examples. Hypertext links to related definitions are also given, but there is no search facility. Written by Valerie Easton and John McColl, the site is
hosted by the Centre for Applied Statistics at Lancaster University. + Full Record

Keywords: glossary, probability theory, statistics, statistical methods, statistical tests

Subject Section(s): Statistics Statistical Theory Psychometrics and Statistics and Methodology

Resource Type: Educational Materials

**Statistics South Africa**


The home page of Statistics South Africa (Stats SA) a government department Web site that “... aims to produce timely, accurate and accessible official statistics to help advance economic growth, development and democracy in the new South Africa.” Statistics South Africa provide a range of economic indicators which includes some census information. The site includes a current awareness section which highlights the latest statistical reports and releases (costs are made clear where relevant). Many reports, publications, discussion and technical papers are provided freely alongside information on magisterial districts; listings of cities, towns and non-urban areas; the code of conduct for public servants and contact details for branch offices. There is a search facility arranged alphabetically by topic. Selected publications and releases are made available in .pdf file format, an Adobe Acrobat Reader is required to access these documents. + Full Record

Keywords: South Africa, government, agricultural economics, national economy, land and property economics, crime, prices, transport economics, statistics, electric power, labour economics

Subject Section(s): National Statistics

Resource Type: Data
Statlib

http://lib.stat.cmu.edu/

An archive of statistics software held at the Carnegie Mellon University. In addition to a collection of statistical software it contains directory lists, archives of the s-news mailing list (s is a high level computer language for data analysis and graphics), a collection of applied statistics algorithms and various other datasets. + Full Record

Keywords: statistical software, algorithms, statistics, mathematics

Subject Section(s): Statistics, Statistical Theory

Resource Type: Resource Guides

TRAMSS

(Training Resources and Materials for Social Scientists)

http://tramss.data-archive.ac.uk/

This site is “a web-based learning and teaching resource for quantitative social science researchers, students and trainers.” Teaching materials and examples are provided to introduce users to data sources and analysis of these data through the use of “statistical software applications in event history analysis and multilevel modelling.” Two software packages are used - SABRE, for the statistical analysis of binary recurrent events, and MLwiN, for fitting multilevel models, which can only be used with the teaching datasets provided. Both can be freely downloaded. Example research questions are worked through starting with sample searches of BIRON, The Data Archive catalogue. Links are included to the Data Archive (UK) and BIRON. Notes are provided on statistical modelling, as well as a bibliography for further reading. A search facility for the site is also available. + Full Record

Keywords: Teaching resources, statistical modelling, software, statistics
UNAIDS


UNAIDS is the joint United Nations programme on HIV/AIDS. It aims to prevent the transmission of the disease and provide care and support to communities worldwide. Its website offers details of the purpose of the project and its current programmes. These cover such areas as the epidemiology of the disease, research into vaccinations, the economics of AIDS and work with women and children affected. The site also contains documents which provide details of the current state of the AIDS epidemic in nations of Asia and Africa and transcripts of recent speeches by UN officials on the topic. Users should note that some documents on this site are only offered in pdf format and therefore require access to an Adobe Acrobat Reader for use. + Full Record

Keywords: HIV/AIDS, AIDS, HIV, Africa, Asia, developing countries, sexually transmitted diseases, United Nations, prevention, UN

Subject Section(s): Health Services/National Health Service

Resource Type: Research Projects/Centres

UNESCO Statistics

http://unescostat.unesco.org/

The United Nations Educational, Scientific and Cultural Organisation (UNESCO) Statistics website gives regularly updated new information received from all countries in the world. Select the country (or region), the years and the parameters desired. Includes detailed data in the areas of...
education, culture and science, with data tables available for viewing online and for download in Excel format

Keywords: international, data, statistics, education, culture, science

Subject Section(s): International Statistics

**United Nations Statistics Division**


United Nations Statistics Division: Social Indicators

Description: This site, compiled by the United Nations Statistics Division (UNSD), provides "... social indicators covering a wide range of subject-matter fields ... from many national and international sources in the global statistical system. These indicators are issued in general and special print or machine-readable format. The indicators presented here consist mainly of the minimum list which has been proposed for followup and monitoring implementation of recent major United Nations conferences on children, population and development, social development and women. This minimum list is contained in the Report of the Expert Group on the Statistical Implications of Recent Major United Nations Conferences (E/CN.3/AC.1/1996/R.4)". + Full Record

Keywords: child-bearing, youth and elderly populations, human settlements, water supply and sanitation, income and economic activity, aging population, economic activity, UN, Social sciences, statistics, health, population, education, literacy, housing, unemployment, women, income, homelessness, social science research, wages, homeless persons, United Nations, young people, older people, demographics, united nations

Subject Section(s): International Statistics National Statistics Demography Demographic Geography Research Tools Economics

Resource Type: Data
WHO (World Health Organization)
http://www.who.int/
Website of the World Health Organization. Country-specific information can be found at http://www.who.int/country/en/ under the countries listed alphabetically.

World Bank
http://www.worldbank.org/data

World Bank Group: Development Data

Description: Includes World Bank development data (in Adobe Acrobat *.pdf format files) and related information including: development targets for the early 21st century; national statistics for countries and regions including African Development Indicators and Country at a glance tables; indicator tables, sectoral data, and links to other relevant data; assistance with working with data including methods, modelling tools, and technical assistance; and details of data publications available in print and CD-ROM formats, including World Development Indicators and World Development Report, and the opportunity to sign up for a newsletter or view the current issue + Full Record

Keywords: international, less developed countries, ldc, aid, economic co-operation, aid, data, statistics, Africa, development, trade, economic development, capital flow


Resource Type: Data
World Population Information

http://www.census.gov/ipc/www/world.html
This website leads to a range of global statistical perspectives including:

World Population Profile: 1998
http://www.census.gov/ipc/www/wp98.html
which gives the latest published compendium and analysis of data on population, fertility, mortality, contraceptive use and related demographic topics by the U.S. Census Bureau. Includes a special chapter focusing in HIV/AIDS in the Developing World

World Vital Events Per Time Unit
http://www.census.gov/cgi-bin/ipc/pcwe
which presents world births, deaths, and natural increase for the current year expressed per year, month, day, hour, minute, and second

World POPClock
http://www.census.gov/cgi-bin/ipc/popclockw/
which gives an up to the second simulation of the current world population.

World Population: 1950 to 2050
http://www.census.gov/ipc/www/worldpop.html
which has the latest estimates and projections of world population from the U.S. Census Bureau.

Historical Estimates of World
http://www.census.gov/ipc/www/worldhis.html
which presents estimates of world population from different sources for years up to 1950.

**International Data Base**

[http://www.census.gov/ipc/www/idbnew.html](http://www.census.gov/ipc/www/idbnew.html)

which is a computerized data bank containing statistical tables of demographic, and socio-economic data for all countries of the world.

**WWW Virtual Library**


The Virtual Library is the oldest catalog of the web, started by Tim Berners-Lee, the creator of html and the web itself. Unlike commercial catalogs, it is run by a loose confederation of volunteers, who compile pages of key links for particular areas in which they are expert; even though it isn’t the biggest index of the web, the VL pages are widely recognised as being amongst the highest-quality guides to particular sections of the web. Individual indexes live on hundreds of different servers around the world. A set of catalog pages linking these pages is maintained at [http://vlib.org](http://vlib.org)
The importance of evaluating internet resources

adapted from

Resource Discovery Network’s Virtual Training Suite

http://www.vts.rdn.ac.uk

Maintained by the Social Science Information Gateway (SOSIG). SOSIG aims to offer social scientists a quick and easy way of finding quality networked information that can support their work

(http://www.sosig.ac.uk)

QUALITY

WHY BOTHER TO REVIEW QUALITY?

The main reason is that, unlike printed academic books and journals, the Internet does not have a central editorial body this means that:

* Anyone can publish information on the Internet
* The best sites maintain standards of peer review but others do not
* Information can be biased
* Sites may be out of date or incomplete

With traditional books and journals, editors exercise control over content, checking its validity and excluding items which fall below their standards. However, with information on the Internet, to a large extent you will need
to do this for yourself. This is not a difficult process, in fact it forms part of
the intellectual process by which people have always conducted research. It
means that once you have gathered materials relating to your subject you
need to review them critically to select those items which are most appro-
priate to your needs.

Remember that a number of services, such as SOSIG, already employ
evaluative rules relating to these criteria, so it can save you time to go
directly there.

**HOW SHOULD YOU ASSESS QUALITY?**

There are several key questions that you should ask every time you look at
an Internet document. These can be summarised as:

Who?
What?
Why?
Where?
When?

**WHO?**

You can start by considering who created the resource.

Ask yourself whether they have the necessary knowledge to provide the
information— are they an established authority like a government, a repu-
table organisation, a university or famous scholar, or someone unknown?

You should also check whether the information is on an official or unoffi-
cial site. This is important as satirical sites are often launched during elec-
tion campaigns, but many are not labelled as such.
If the information is published on a university site, consider whether it is issued by a recognised academic department or if it is posted on the homepage of an individual, as in the latter case it may be less authoritative.

**Tips on evaluating the creator of the resource**

If your document gives an author’s name try to look for details of their position. Are they an academic scholar, a researcher or an undergraduate? You could also check whether they have published anything else by conducting searches on library catalogues (OPACs) or journal indexes such as the *International Bibliography of the Social Sciences* (IBSS).

If you are quoting information published on the Web site of an organisation, check that it is a reputable body. Look to see if it is listed in any of the directories of associations and organisations that you will find in your local library. If it is a new body, check if it quotes support or sponsorship from any other established organisations. If you are still unable to trace its provenance it may be worth contacting the appropriate professional organisations for assistance.

**Getting clues about authorship from URLs**

URL stands for Uniform Resource Locator. The URL is the Internet address or location at which you can find a particular resource. You will see it at the top of your Web browser. If you are using Netscape Navigator it will be in the Location box. If you are using Microsoft Internet Explorer it will be in the Address box.

The URL will often give you a hint as to:

* what type of organisation it is
* what type of organisation the site belongs to
* the country where the organisation is located

URLs tend to follow a pattern. They give the name of the institution, an abbreviation signifying what type of body it is and then a country code.

Let’s look at an example:

http://www.lse.ac.uk/

In this example:

LSE is an abbreviation for the London School of Economics

“ac” shows that it is an academic institution

“UK” signifies that it is based in the United Kingdom.

A similar American example would be

http://www.harvard.edu

Here Harvard is the name of the body

“edu” shows that it is an educational organisation. (US sites don’t always signify a location)

And a South African example would of course be

http://www.uwc.ac.za

Here uwc is the name of the body

“ac” shows that it is an educational organisation

“za” signifies that it is located in South Africa

**Some useful tips about URLs**

Government sites usually contain the element “gov” in their URL. For example the address of the South African Department of Health is
Companies are signified by “com” or “co” in their address. For instance Amazon Books can be found at

http://www.amazon.co.uk

in the UK

or the US site at

http://www.amazon.com

and one of our local booksellers, Clarke’s Bookshop in Long Street, Cape Town, can be found at

http://www.clarkes.co.za/

The code “org” in a URL commonly means that you are looking at a Web site maintained by an organisation. These can include, independent bodies, think tanks and charities. For example the homepage of the Independent Democrats is located at

http://www.id.org.za/

Of course many Internet addresses are far more complicated than this.

**Investigating the root of an url**

In order to trace its origin and authority a quick solution is to gradually reduce the address to its root so that you can look at the homepage of the organisation which is hosting it.

Take the example of an article by Anna Greenberg entitled “Race, Religiosity, and the Women’s Vote”. This has the URL:

The root of this URL is:

http://www.ksg.harvard.edu/prg/

In this case you could enter http://www.ksg.harvard.edu/prg/ which will take you to the homepage of the Politics Research Group John F. Kennedy School of Government, Harvard University where you will find that it is listed as a working paper published in 1998.

**What?**

Here you should consider, not only the contents of the site, but whether it meets your specific needs.

**KEY QUESTIONS TO ASK ARE:**

Is the academic level of the site appropriate? Does it treat the subject with the degree of academic rigour, which you require, or is it an introductory guide which is too basic?

Is it a commentary or the original text?

Is it published in full text, or is it an abridged format?

Is the text the right edition or date?

**Why?**

In addition to content, you will also need to examine the motivations of the author/publisher as they will influence the stance and accuracy of the information provided.

Many sites on the Internet do not present impartial facts. They exist to further specific viewpoints or to publicise particular products.
Tips on evaluating the standpoint of a resource

If you find an article or document on the Internet, trace its author and publisher using the tips in the WHO section. Look for the names of any sponsoring organisations or funding bodies and consider whether they are impartial or influenced by a particular political viewpoint. Sometimes this is not clearly marked on the site, but you can often find clues in policy, mission statements or “about us” sections on the homepage of the body.

where?

Where a document has been published may also effect its completeness and political stance.

For example, where are some areas of the world where press censorship operates. Consider whether this might be the case for your research as if so it may mean that official documents are seriously incomplete. In these circumstances, you might want to search for materials produced by non-official opposition movements or overseas pressure groups to gain a wider picture.

When?

The date when the information was published may also be important to you.

If you are looking for current news headlines or the most recent version of a treaty it is important that you are seeing the most up to date information. If you are researching a historical event, currency may be less important. However, you must always be able to quote the date of all documents and journal articles which you cite in your work.
Check the date coverage by looking for a notice of when the site was last updated. Most well designed sites readily supply this. However, beware that this may relate to the “What’s New” page rather than the section you are interested in. As a result, you may need to use your knowledge of the subject area to check if recent events are covered. If they are not mentioned it could be a sign that the resource is not regularly updated.

1 Most of the website information cited in this summary is drawn from the Social Science Information Gateway (SOSIG) q.v.
Index

A
Aitken, A.C.  52

B
bias  72
blocks of coefficients  187
Boolean searching  293

C
censored cases  156
Central Limit Theorem  70
chance
  in sampling  70
Citations  270
class limits  236
class midpoint  236
Clustering  73
confidence interval  244
Crosstabs
  procedure  130

D
Descriptive Statistics  149
Dillman’s concept of personalization  105
Dunkelberg and Day plus Politz Simons method  86

E
Epsem  69
equal probabilities of selection method  69
error
  non-response  86
effects
  random  77
estimation for structural breaks in the data  187
F
Focus groups 89
frequency distribution 236
functional form 186

H
Hansen 71
Hurvitz 71
Hypothesis tests 150

I
imputation methods 115
individual response error 71
irrelevant variable 186

M
Madow 71
multicollinearity 186

N
non-response 85
Null hypothesis 150

O
ogive 50
  mention 237
Omitted variable 185
open class interval 236

P
population parameters 149
proportionate stratification 73

Q
quotations:
  how to do them 270
random error 71
Regression 242

sample
definition 70
Sampling error 69
sampling error 72
Scatter plot 242
Shryock, Henry S. 52
Siegel, Jacob S. 52
simple random sample 73
Sorting data 242
stratification 70
structural breaks 186
systematic error 71, 72

Tables 270

unbiased method 72

variable error 71

weighting adjustments 115