UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD
Statistical Consulting Service, Department of Statistical Sciences
PD Hahn Building, Upper Campus
Tel: 021 650 5215 / 3219; Email: sta-consulting@uct.ac.za
http://www.stats.uct.ac.za/stats/consulting

Statistical Consulting Service
Database Guidelines

Data collection and capturing is the responsibility of the client.

The Statistical Consulting Service (SCS) does not retain any rights to the data submitted for analysis, and confidentiality is guaranteed.

Data is not used beyond the termination of services, except by special permission from the client, or in the case of continued joint research where the unit plays a collaborative role.

Database management is of utmost importance.

Whilst data cleaning, validation and manipulation is offered as an additional service, databases submitted for analyses are expected to conform to the specifications outlined below, within reason, and unless by pre-arrangement with the consultant.

While we aim to assist clients appropriately with their data needs, the SCS reserves the right to refuse to analyse databases not yet conforming to the specifications below.

Software

Datasets are to be submitted as Microsoft Excel workbooks or as text files, or in any equivalent formats agreed upon with the consultant(s), such as comma or tab delimited files, Microsoft Access database files, SPSS or Stata database files.

When capturing data from hard copies, it is generally useful to enter the data into Excel first, before importing it into a statistical software package.

Codebooks

Codebooks, which provide technical descriptions of the datasets, should be provided with the datasets. If submitting the data in Microsoft Excel or as a text file, a Codebook can be provided in a separate file (for example, in Microsoft Excel or Microsoft Word). When using certain statistical software, such as SPSS, the codebook is integrated into the dataset file – provided the client has appropriately prepared the dataset. In this case, the Variable View must be populated with the information indicated below.

General Dataset Guidelines

- Typically, all data should be entered into a **single table/spreadsheet**. When the data structure is more complex and a relational database has been used to record the data, there may be multiple well-defined tables (relations) in the database, with well-defined relationships amongst tables (using primary and foreign keys).
- Each **row ('observation')** should contain a unique instance of data and correspond to a single unit on which one has made observations (for example, a row may describe a single patient or company).
- Each row/unit/**observation should have a unique identifier** (which in some way corresponds to the hard copy or source of the data).
- Each **column ('field')** should correspond to a single measured variable.
- The **names of variables (columns) should only appear in the FIRST row or header** of the table, and the actual data should begin in the very next row. There should be no title rows or blank rows above the data table.
- Both the database and the variables should have **meaningful names**. Ideally, all submitted files should be dated. For example, a data file may be named 'TomJones_ICUPatients_20160128.xlsx' rather than just 'mydata.xlsx', and a variable named 'Gender' rather than 'var1'.
- **Variable names should be short, unique** and underscores are preferred to blank spaces between words.
- **Avoid punctuation marks** (such as commas, apostrophes, inverted commas and accents) in variable names and data entries, with the exception of decimal points / commas used in numeric data to separate the whole numbers from the fractional components.

Statistical Consulting Service, Department of Statistical Sciences
PD Hahn Building, Upper Campus
Tel: 021 650 5215 / 3219; Email: sta-consulting@uct.ac.za
http://www.stats.uct.ac.za/stats/consulting

- **Avoid whitespaces in variable names**, including before and after the names. For example, use "AgeOfDeath" rather than "Age of Death", and make sure that it is not accidentally "AgeOfDeath ".

- **Missing data should be represented by blank cells** in the data table (empty cells as opposed to ones containing spaces or zeroes). If a **specific missing value code** has been used during data capturing to represent missing data (for example, an Age recorded as 999 is missing), this information should be clearly indicated in the Codebook. Blank cells (missing data) and entries of 0 have distinctly different meanings.

- Any given **field should contain either only numeric data or only text data (or only dates)**. If necessary, multiple fields can be used to capture related data. For example, one field may capture the numeric Weight and another field the text Weight_Category in which the doctor indicated whether the patient is underweight or overweight.

- For **numeric fields, only the numeric component** should be recorded, and not the units or any other text. The same unit of measurement should be used for all data in a specific field. For example, a Weight entry may be 34.5, but should not be '34.5 kilograms' or '34.5 – underweight'.

- For **categorical fields, the categories may be entered as descriptive text labels or as numeric / text codes**. If entered as descriptive text, the text must be ***exactly*** the same when the category is the same. For example, for variable Gender, entries 'Male' and 'Female' (only) could be used to distinguish males from females, but one should not use a mixture of 'M' and ' M' and 'Male' and 'male' to represent males. **To reduce the possibility of discrepancies, it is preferred that codes are used**, where again the code must be ***exactly*** the same when the category is the same. For example, the number 1 could be used for males, and 0 for females (and such information provided in the Codebook). As another example, 'A', 'B' and 'C' may be used as codes, each representing some particular disease (again, to be explained in the Codebook).

- **Dates should be entered consistently**. For example, they may be entered in the form DD-MM-YYYY, YYYY-MM-DD, YYYY/MM/DD, or even as, for example, 25 January 2011, provided that ***all*** dates are treated in the same manner in the full dataset.

- **Constructed variables** can be provided within the dataset, but should ideally be created by the consultant at the time of the analysis to better maintain data integrity, provided the client specifies how these variables should be constructed. For example, the consultant may create a body mass index variable, to be used in the analysis, from variables Height and Weight, or an overall domain score by summing a number of individual Likert scale responses captured in the dataset.

- **Confidential identifying data,** such as patient names, should be removed from the dataset by the client before providing the data to the SCS.

- **Remove all hidden rows and columns** from the sheet.

- **Blank rows and columns in the worksheet are not necessary.**

- If you have already done some analyses or produced some graphs in Excel, remove them from the spreadsheet you provide to the consulting service. The data file you provide to the consulting service should include the data and variable names only.

Statistical Consulting Service, Department of Statistical Sciences
PD Hahn Building, Upper Campus
Tel: 021 650 5215 / 3219; Email: sta-consulting@uct.ac.za
http://www.stats.uct.ac.za/stats/consulting

Codebook Guidelines

Codebooks can contain various details about the dataset. Typically, for a SCS project, the Codebook should be a neat table containing the following information:

- **Variable name**
- Variable **description** or label
- **Data type**, for example, text, numeric (count), numeric, date
- **Units/Range/Format**
- **Codes** for categorical variables, providing the codes used and their meaning; for example, for a Likert scale variable one could specify 0-Strongly Disagree, 1-Disagree, 2-Neutral, 3-Agree, 4-Strongly Agree
- **Calculations** used to derive 'constructed variables' contained in the dataset, or notes on the relationship between variables

Statistical Consulting Service, Department of Statistical Sciences
PD Hahn Building, Upper Campus
Tel: 021 650 5215 / 3219; Email: sta-consulting@uct.ac.za
http://www.stats.uct.ac.za/stats/consulting

Example of Dataset (Screenshot of Microsoft Excel Dataset)

This hypothetical (unusually small) dataset is for the study of Outcome two weeks after treatment initiation, for youth, in a hospital ward, with a particular disease. The role of certain variables (such as BMI and Gender) is also to be considered, and the symptoms of the patients one week after treatment are to be described.

Patient_ID	Doctor_ID	Start	Gender	Age	Weight	Height	BMI	Symptoms	Fever	Weightloss	ShortBreath	Other	OtherType	Outcome
AS002	T	2015/05/26	M	18	45	150	20.00	1	1	0	1	0		1
AT015	S	2014/07/23	F	12	37	98	38.53	1	1	0	0	1	2	1
BR002	S	2014/12/10	F	14	42	104	38.83	1	0	0	1	0		2
BR005	T	2015/03/14	M	10	28	105	25.40	1	0	0	1	0		2
SX014	V	2015/04/01	M	14				0	0	0	0	0		1
RT058	S	2015/08/12		8	12	75	21.33	0	0	0	0	0		1
AM005	Th	2014/11/08	F	9	16	80	25.00	0	0	0	0	0		1
OL058	Th	2014/09/05	F	14	25	109	21.04	0	0	0	0	0		3
OL035	V	2014/11/23	M	16	34	152	14.72	1						3
OL057	S	2015/01/01	M	21	55	142	27.28	1	0	0	0	1	1	1
RT012	V	2015/08/07	M	18	70	120	48.61	0	0	0	0	0		1
AT014	V	2015/12/12	M	17	54	160	21.09	1	0	0	1	0		1
AR099	Th	2014/04/16	M		81	139	41.92	1	1	1	1	0		2

Example of Corresponding Codebook (Screenshot of Microsoft Excel Codebook)

Name	Label	Type of data	Units/range/format	Codes	Notes
Patient_ID	Patient identifier (folder number)	Text			Each row has a unique patient
Doctor_ID	Primary doctor identifier	Text		(the IDs are used to protect doctor identies)	(Will not be used for analysis)
Start	Date of treatment initiation	Date	YYYY/MM/DD		(Will not be used for analysis)
Gender	Gender	Text category		M = Male F = Female	
Age	Age in years	Numeric	Whole number, 8-21 inclusive (youth ward only allows these ages)		
Weight	Weight in kilograms (kg)	Numeric	Positive, expected range 10-100 inclusive		
Height	Height in centimeters (cm)	Numeric	Positive, expected range 50-200 inclusive		
BMI	Body mass index (kg / m^2)	Numeric			Calculated field: Weight / (Height/100)^2
Symptoms	Symptoms one week after treatment initiation	Numeric category (binary)	0/1	0 = No 1 = Yes	If this is 0, Fever-Other should all be 0. If this is 1, then Fever-Other are used to record the particular types of symptoms.
Fever	Fever one week after treatment intitation	Numeric category (binary)	0/1	0 = No 1 = Yes	
Weightloss	Weightloss one week after treatment initiation	Numeric category (binary)	0/1	0 = No 1 = Yes	
ShortBreath	Shortness of breath one week after treatment initation	Numeric category (binary)	0/1	0 = No 1 = Yes	
Other	Any other symptoms one week after treatment intitation	Numeric category (binary)	0/1	0 = No 1 = Yes	
OtherType	Type of any other symptoms one week after treatment intitation	Numeric category	0/1/2	0 = Likely related to disease 1 = Likely related to medication side-effects 2 = Unrelated symptoms	If Other is 1, this will be populated if known.
Outcome	Treatment outcome two weeks after treatment initiation	Numeric category (ordinal)	0/1/2/3	0 = Death 1 = Deterioration / no change 2 = Some recovery 3 = Full recovery	

Statistical Consulting Service, Department of Statistical Sciences
PD Hahn Building, Upper Campus
Tel: 021 650 5215 / 3219; Email: sta-consulting@uct.ac.za
http://www.stats.uct.ac.za/stats/consulting

Microsoft Excel Dataset Guidelines

In addition to the guidelines above:

- Typically, all data should be in a single table on a single worksheet, starting in cell A1. For example, if there are individual datasets related to different countries, each country's data should not be on its own worksheet but rather all data merged into a single table provided on one worksheet and Country included as a variable in this complete table of data.
- Additional information (such as graphs and summary statistics such as averages of columns) should be removed from this sheet.
- Formatting of text and cells, such as using boldface font and coloured cells, should not be used to provide information about the data. *All of these details will be lost when data is imported into a statistical package.*
- Do not merge cells, even those containing variable names.

SPSS Dataset Guidelines

In addition to the guidelines above:

- During the preparation of data in SPSS, the client should ensure that all variables' types and measures are well specified, variables are labelled and value labels are provided. This may eliminate the need for a separate Codebook.

'Repeated Measures' Datasets

A common type of data that provides some exceptions to the guidelines above are those containing 'repeated measures', which occur when multiple measurements are taken on each unit of interest. Such data is recorded in either 'wide' or 'long' format, illustrated through example below. Conversion between the two formats is straightforward to implement by the consultant using standard software.

This example hypothetical dataset captures the 2016 monthly return for each of a number of funds.

Long data

Fund	Platform	Month	Return
A	Coronation	1	-0.01
A	Coronation	2	0.02
A	Coronation	3	0.02
B	Allan Gray	1	0.04
B	Allan Gray	2	0.05
B	Allan Gray	3	0.06
C	Allan Gray	1	-0.02
C	Allan Gray	2	-0.06
C	Allan Gray	3	0.00

Wide data

Fund	Platform	Return_Jan	Return_Feb	Return_Mar
A	Coronation	-0.01	0.02	0.02
B	Allan Gray	0.04	0.05	0.06
C	Allan Gray	-0.02	-0.06	0.00

Database Integrity

Regardless of the software used in the analysis of data, database integrity is maintained through the use of a scripted program. Databases submitted for analysis are thus not altered. In this way, analyses can be re-run/checked at a later stage, either at the request of reviewers, or should new information have come to light.

Copies of the original, unaltered database are preserved for the duration of the SCS involvement; however, it is recommended that the client maintains his/her own backup of the provided database as well.