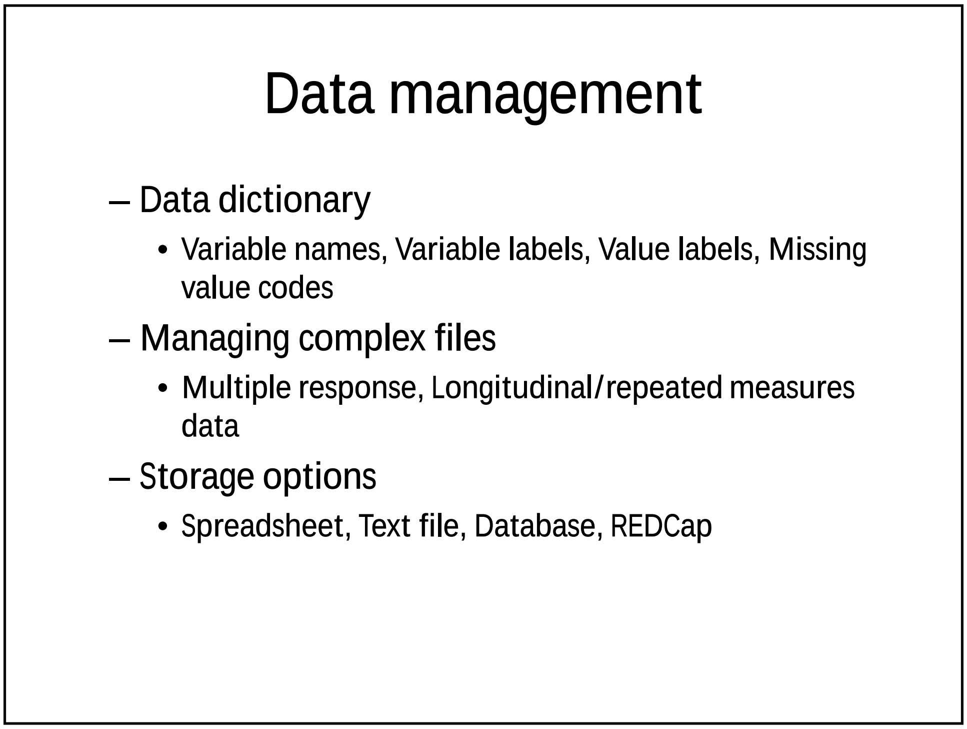
Video 11 - Data management

Steve Simon



Data management

– Data dictionary
  - Variable names, Variable labels, Value labels, Missing value codes

– Managing complex files
  - Multiple response, Longitudinal/repeated measures data

– Storage options
  - Spreadsheet, Text file, Database, REDCap

Data dictionary

- – Also called a code book
- – Start before collecting data
- – Revise as needed

Variable names

- – Brief, but descriptive explanation
- – Roughly 4 to 16 characters
- – No blanks and (almost) no symbols
- – One to three words

Good and bad variable names

- Names to avoid (www.writersexchange.com)
  - systolic blood pressure
  - systolic-blood-pressure
- Names that work
  - systolic_blood_pressure
  - systolic.blood.pressure
  - SystolicBloodPressure
- NEVER USE ALL CAPS FOR VARIABLE NAMES
  - Lower case ascenders (e.g., f and l)
  - Lower case descenders (e.g., g and y)

Variable labels

- Longer descriptions
  - Can include spaces and punctuation
  - Ideal length is 20-40 characters
  - Mention units of measurement, special qualifiers

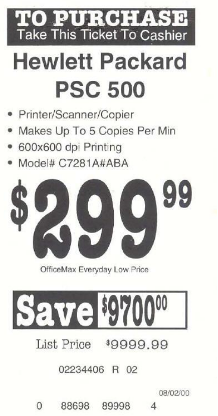
Missing value codes

- Explain WHY the value is missing
- For a survey
  - Did not answer
  - Not applicable
- For a lab result
  - Below the limit of detection
  - Insufficient volume for testing
  - Dropped the test tube and it shattered making a huge mess

Example of missing value codes

- Use extreme number code
  - 9, 99, 999
  - -1
- Use symbols
  - NA
  - (asterisk)
  - (dot)
- Never use blanks to designate missing
- Note missing value code on data dictionary

Missing value example



TO PURCHASE
Take This Ticket To Cashier

Hewlett Packard
PSC 500

- Printer/Scanner/Copier
- Makes Up To 5 Copies Per Min
- 600x600 dpi Printing
- Model# C7281A#ABA

$299 99

OfficeMax Everyday Low Price

Save $9700 00

List Price $9999.99

02234406 R 02

08/02/00

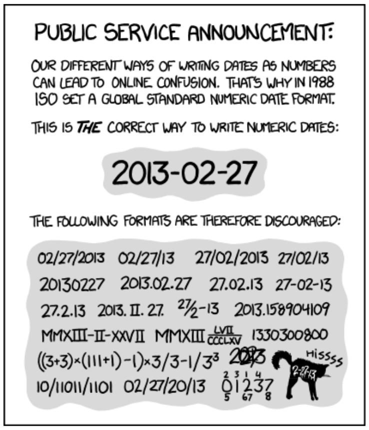
0 88698 89998 4

Price tage from computer store

First break

- What have you learned
  - variable names
  - Variable labels
  - Missing value codes
- What’s coming up
  - Date formats
  - Categorical values

Date formats


PUBLIC SERVICE ANNOUNCEMENT:
OUR DIFFERENT WAYS OF WRITING DATES AS NUMBERS CAN LEAD TO ONLINE CONFUSION. THAT'S WHY IN 1988 ISO SET A GLOBAL STANDARD NUMERIC DATE FORMAT.
THIS IS THE CORRECT WAY TO WRITE NUMERIC DATES:
2013-02-27
THE FOLLOWING FORMATS ARE THEREFORE DISCOURAGED:
02/27/2013 02/27/13 27/02/2013 27/02/13
20130227 2013.02.27 27.02.13 27-02-13
27.2.13 2013. II. 27. 27/2-13 2013.158904109
MMXIII-II-XXVII MMXIII LVII/CCCLXV 1330300800
((3+3)×(111+1)-1)×3/3-1/3³ 2013
10/11011/1101 02/27/20/13 01237
HISSSSS
2-27-13

Cartoon showing variety of data formats

Internal storage formats

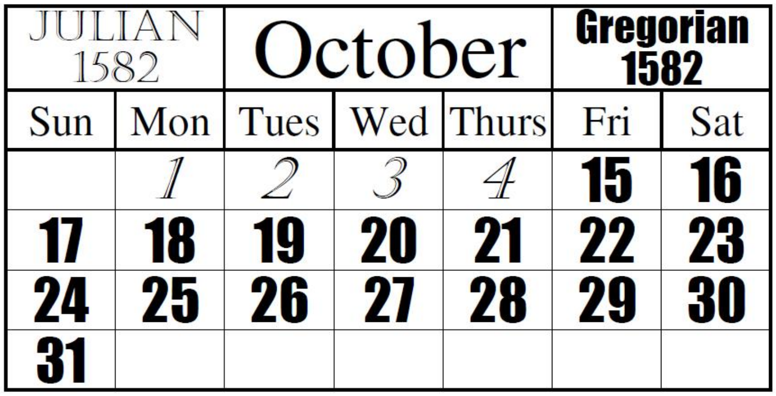
– Excel - number of days since 1899-12-31 (1900-01-00)

– R - number of days since January 1, 1970

– SAS - number of days since January 1, 1960

– SPSS - number of seconds since October 14, 1582

Gregorian calendar



JULIAN 1582		October			Gregorian 1582	
Sun	Mon	Tues	Wed	Thurs	Fri	Sat
	1	2	3	4	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

Transition to the Gregorian calendar

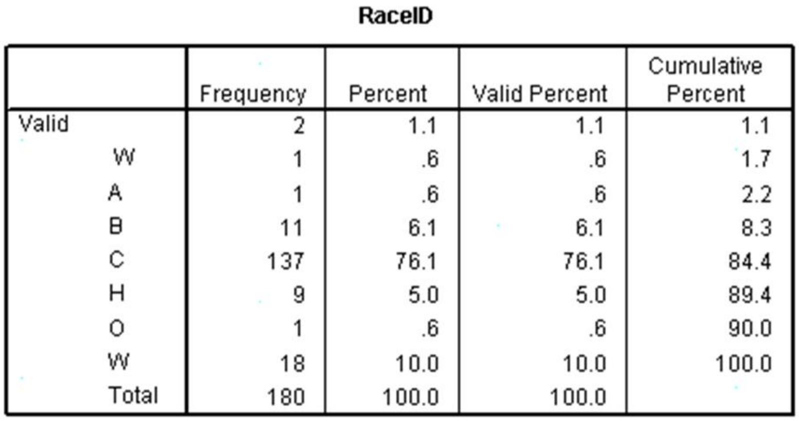
Gregorian calendar

Painting of Pope Gergory XIII

Categorical values

- – Definition: small number of possible values
- – Beware of ambiguities
  - YES, yes, and Yes are three distinct levels.
- – Use number codes
  - 0, 1, 9 for binary variables
- – Single letter codes
  - M, F, and U for gender
  - Potentially ambiguous
  - Consistent case is important.

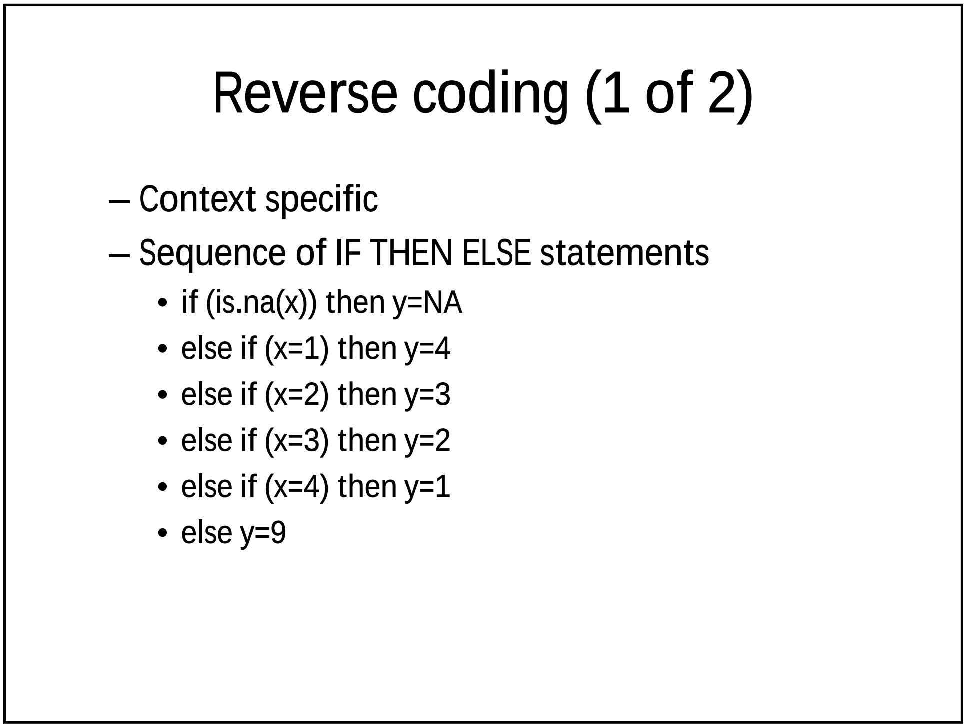
Example of ambiguous coding



RaceID

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid		2	1.1	1.1	1.1
	W	1	.6	.6	1.7
	A	1	.6	.6	2.2
	B	11	6.1	6.1	8.3
	C	137	76.1	76.1	84.4
	H	9	5.0	5.0	89.4
	O	1	.6	.6	90.0
	W	18	10.0	10.0	100.0
	Total	180	100.0	100.0	

SPSS frequencies table for RaceID

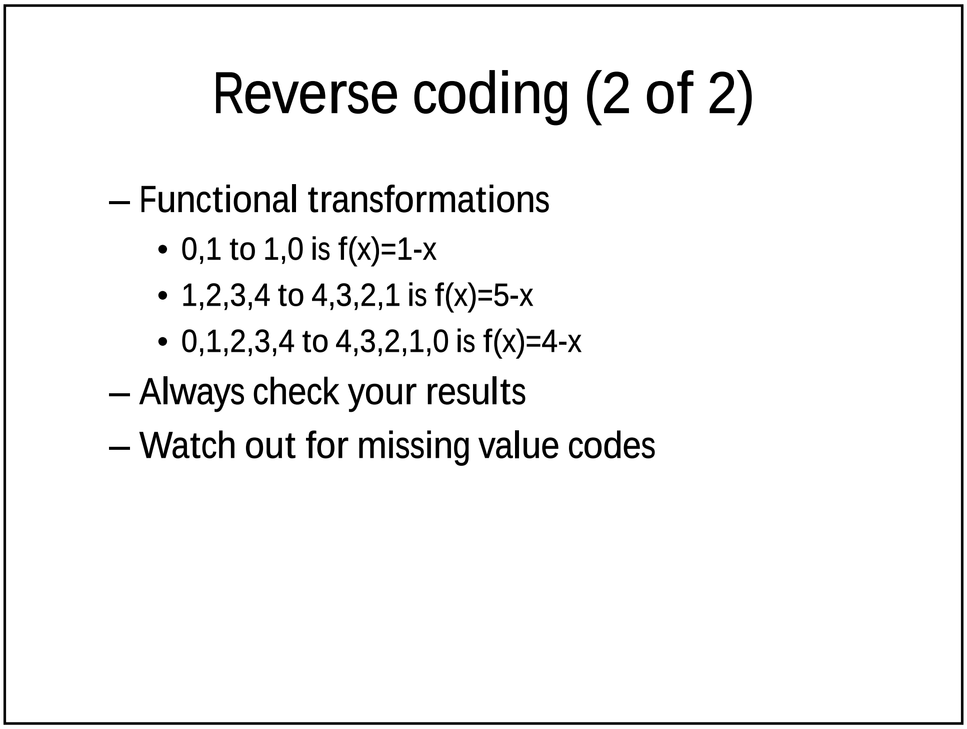


Reverse coding (1 of 2)

– Context specific

– Sequence of IF THEN ELSE statements

- if (is.na(x)) then y=NA
- else if (x=1) then y=4
- else if (x=2) then y=3
- else if (x=3) then y=2
- else if (x=4) then y=1
- else y=9



Reverse coding (2 of 2)

– Functional transformations

- 0,1 to 1,0 is f(x)=1-x
- 1,2,3,4 to 4,3,2,1 is f(x)=5-x
- 0,1,2,3,4 to 4,3,2,1,0 is f(x)=4-x

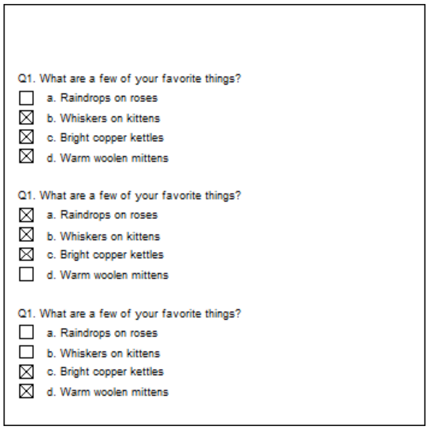
– Always check your results

– Watch out for missing value codes

Second break

- What have you learned
  - Dates
  - Value labels
  - Reverse coding
- What's coming next
  - Multiple response
  - Longitudinal/repeated measures data

A multiple response example



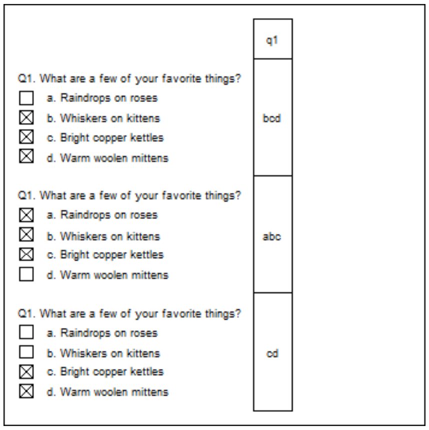
Q1. What are a few of your favorite things?
☐ a. Raindrops on roses
☒ b. Whiskers on kittens
☒ c. Bright copper kettles
☒ d. Warm woolen mittens

Q1. What are a few of your favorite things?
☒ a. Raindrops on roses
☒ b. Whiskers on kittens
☒ c. Bright copper kettles
☐ d. Warm woolen mittens

Q1. What are a few of your favorite things?
☐ a. Raindrops on roses
☐ b. Whiskers on kittens
☒ c. Bright copper kettles
☒ d. Warm woolen mittens

Questionnaire with a multiple response question

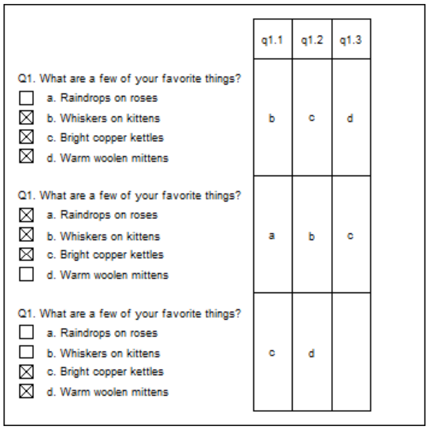
Coding multiple response with a single column



	q1
Q1. What are a few of your favorite things? ☐ a. Raindrops on roses ☒ b. Whiskers on kittens ☒ c. Bright copper kettles ☒ d. Warm woolen mittens	bcd
Q1. What are a few of your favorite things? ☒ a. Raindrops on roses ☒ b. Whiskers on kittens ☒ c. Bright copper kettles ☐ d. Warm woolen mittens	abc
Q1. What are a few of your favorite things? ☐ a. Raindrops on roses ☐ b. Whiskers on kittens ☒ c. Bright copper kettles ☒ d. Warm woolen mittens	cd

Multiple response coded into a single column

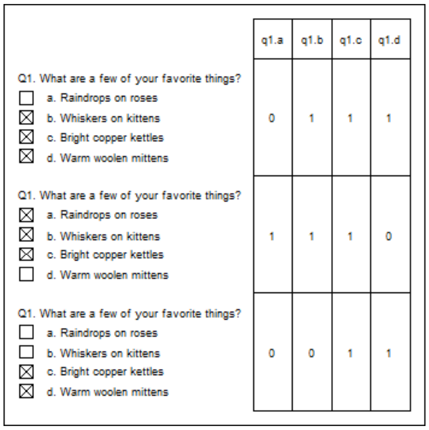
A different way to code multiple response



	q1.1	q1.2	q1.3
Q1. What are a few of your favorite things? ☐ a. Raindrops on roses ☒ b. Whiskers on kittens ☒ c. Bright copper kettles ☒ d. Warm woolen mittens	b	c	d
Q1. What are a few of your favorite things? ☒ a. Raindrops on roses ☒ b. Whiskers on kittens ☒ c. Bright copper kettles ☐ d. Warm woolen mittens	a	b	c
Q1. What are a few of your favorite things? ☐ a. Raindrops on roses ☐ b. Whiskers on kittens ☒ c. Bright copper kettles ☒ d. Warm woolen mittens	c	d	

A multiple response question coded into three columns

The recommended way to code multiple response



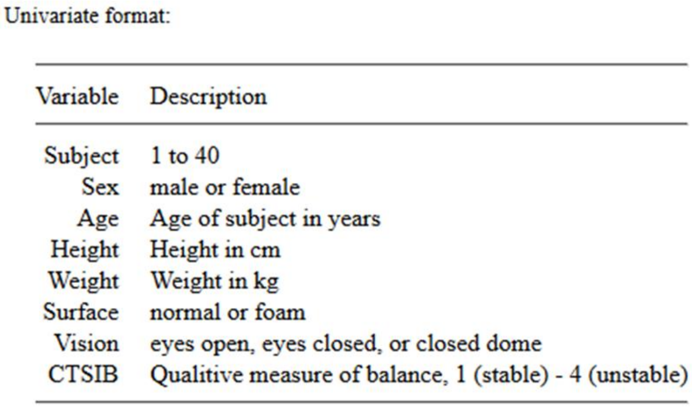
	q1.a	q1.b	q1.c	q1.d
Q1. What are a few of your favorite things? ☐ a. Raindrops on roses ☒ b. Whiskers on kittens ☒ c. Bright copper kettles ☒ d. Warm woolen mittens	0	1	1	1
Q1. What are a few of your favorite things? ☒ a. Raindrops on roses ☒ b. Whiskers on kittens ☒ c. Bright copper kettles ☐ d. Warm woolen mittens	1	1	1	0
Q1. What are a few of your favorite things? ☐ a. Raindrops on roses ☐ b. Whiskers on kittens ☒ c. Bright copper kettles ☒ d. Warm woolen mittens	0	0	1	1

Multiple response coded with individual item indicators

Longitudinal data, Repeated measures data

- – Longitunal
  - Multiple time points per patient
- – Repeated measurements
  - Measuring patient repeatedly under different conditions
- – Tall and thin format
  - One line per visit/measurement
- – Short and fat format
  - One line per patient

Example of tall/thin, dictionary

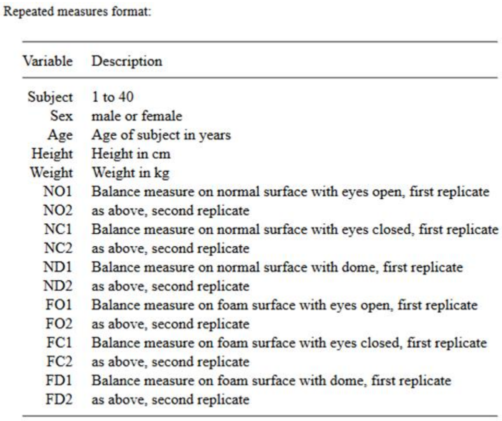


Univariate format:

Variable	Description
Subject	1 to 40
Sex	male or female
Age	Age of subject in years
Height	Height in cm
Weight	Weight in kg
Surface	normal or foam
Vision	eyes open, eyes closed, or closed dome
CTSIB	Qualitive measure of balance, 1 (stable) - 4 (unstable)

Data dictionary for tall and thin format

Example of short/fat, dictionary

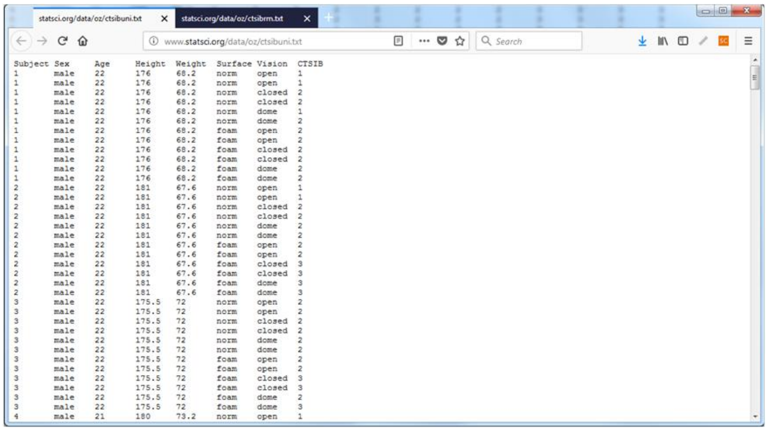


Repeated measures format:

Variable	Description
Subject	1 to 40
Sex	male or female
Age	Age of subject in years
Height	Height in cm
Weight	Weight in kg
NO1	Balance measure on normal surface with eyes open, first replicate
NO2	as above, second replicate
NC1	Balance measure on normal surface with eyes closed, first replicate
NC2	as above, second replicate
ND1	Balance measure on normal surface with dome, first replicate
ND2	as above, second replicate
FO1	Balance measure on foam surface with eyes open, first replicate
FO2	as above, second replicate
FC1	Balance measure on foam surface with eyes closed, first replicate
FC2	as above, second replicate
FD1	Balance measure on foam surface with dome, first replicate
FD2	as above, second replicate

Repeated measures example in short and fat format

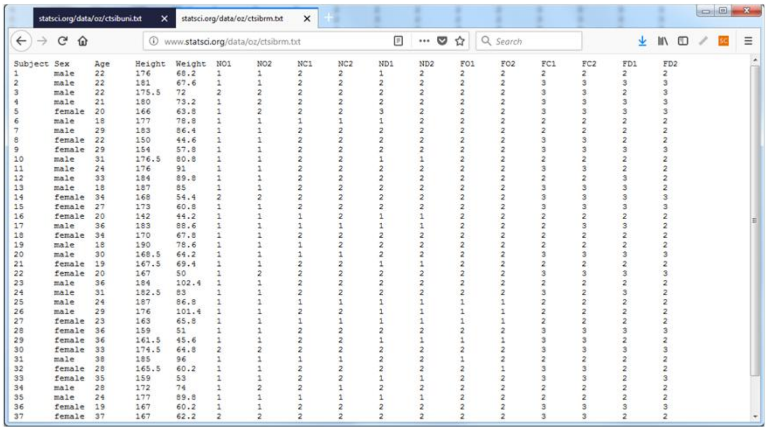
Example of tall/thin, data


statsci.org/data/oz/ctsibuni.txt
statsci.org/data/oz/ctsibrm.txt
www.statsci.org/data/oz/ctsibuni.txt
Search

```
Subject Sex     Age     Height  Weight  Surface Vision  CTSIB
1       male    22      176     68.2    norm    open    1
1       male    22      176     68.2    norm    open    1
1       male    22      176     68.2    norm    closed  2
1       male    22      176     68.2    norm    closed  2
1       male    22      176     68.2    norm    dome    1
1       male    22      176     68.2    norm    dome    2
1       male    22      176     68.2    foam    open    2
1       male    22      176     68.2    foam    open    2
1       male    22      176     68.2    foam    closed  2
1       male    22      176     68.2    foam    closed  2
1       male    22      176     68.2    foam    dome    2
1       male    22      176     68.2    foam    dome    2
2       male    22      181     67.6    norm    open    1
2       male    22      181     67.6    norm    open    1
2       male    22      181     67.6    norm    closed  2
2       male    22      181     67.6    norm    closed  2
2       male    22      181     67.6    norm    dome    2
2       male    22      181     67.6    norm    dome    2
2       male    22      181     67.6    foam    open    2
2       male    22      181     67.6    foam    open    2
2       male    22      181     67.6    foam    closed  3
2       male    22      181     67.6    foam    closed  3
2       male    22      181     67.6    foam    dome    3
2       male    22      181     67.6    foam    dome    3
3       male    22      175.5   72      norm    open    2
3       male    22      175.5   72      norm    open    2
3       male    22      175.5   72      norm    closed  2
3       male    22      175.5   72      norm    closed  2
3       male    22      175.5   72      norm    dome    2
3       male    22      175.5   72      norm    dome    2
3       male    22      175.5   72      foam    open    2
3       male    22      175.5   72      foam    open    2
3       male    22      175.5   72      foam    closed  3
3       male    22      175.5   72      foam    closed  3
3       male    22      175.5   72      foam    dome    2
3       male    22      175.5   72      foam    dome    3
4       male    21      180     73.2    norm    open    1
```

Repeated measures data in tall and thin format

Example of short/fat, data


statsci.org/data/oz/ctsibuni.txt
statsci.org/data/oz/ctsibrm.txt
www.statsci.org/data/oz/ctsibrm.txt
Search

```
Subject Sex     Age     Height  Weight  NO1     NO2     NC1     NC2     ND1     ND2     FO1     FO2     FC1     FC2     FD1     FD2
1       male    22      176     68.2    1       1       2       2       1       2       2       2       2       2       2       2
2       male    22      181     67.6    1       1       2       2       2       2       2       2       3       3       3       3
3       male    22      175.5   72      2       2       2       2       2       2       2       2       3       3       2       3
4       male    21      180     73.2    1       2       2       2       2       2       2       2       3       3       3       3
5       female  20      166     63.8    1       2       2       2       3       2       2       2       3       3       3       3
6       male    18      177     78.8    1       1       1       1       1       2       2       2       2       2       2       2
7       male    29      183     86.4    1       1       2       2       2       2       2       2       2       2       2       2
8       female  22      150     44.6    1       1       2       2       2       2       2       2       3       3       2       2
9       female  29      154     57.8    1       1       2       2       2       2       2       2       3       3       3       3
10      male    31      176.5   80.8    1       1       2       2       1       1       2       2       2       2       2       2
11      male    24      176     91      1       1       2       2       2       2       2       2       3       3       2       2
12      male    33      184     89.8    1       1       2       2       2       2       2       2       2       2       3       2
13      male    18      187     85      1       1       2       2       2       2       2       2       3       3       3       2
14      female  34      168     54.4    2       2       2       2       2       2       2       2       3       3       3       3
15      female  27      173     60.8    1       1       2       2       2       2       2       2       3       3       3       3
16      female  20      142     44.2    1       1       1       2       1       1       2       2       2       2       2       2
17      male    36      183     88.6    1       1       1       1       1       1       2       2       2       3       3       2
18      female  34      170     67.8    1       1       2       2       2       2       2       2       2       2       2       2
19      male    18      190     78.6    1       1       1       2       2       2       2       2       2       2       2       2
20      male    30      168.5   64.2    1       1       1       1       2       2       2       2       3       3       3       3
21      female  19      167.5   69.4    1       1       2       2       1       1       2       2       2       2       2       2
22      female  20      167     50      1       2       2       2       2       2       2       2       3       3       3       3
23      male    36      184     102.4   1       1       2       2       2       2       2       2       2       2       2       2
24      male    31      182.5   83      1       1       2       2       2       2       2       2       3       2       3       2
25      male    24      187     86.8    1       1       1       1       1       1       1       1       2       2       2       2
26      male    29      176     101.4   1       1       2       2       1       1       1       1       2       2       2       2
27      female  23      163     65.8    1       1       1       1       1       1       1       1       2       2       2       2
28      female  36      159     51      1       1       2       2       2       2       2       2       3       3       3       3
29      female  36      161.5   45.6    1       1       2       2       1       1       1       1       3       3       2       2
30      female  33      174.5   64.8    2       2       2       2       2       2       2       2       3       3       3       3
31      male    38      185     96      1       1       1       1       2       2       1       2       2       2       2       2
32      female  28      165.5   60.2    1       1       2       2       2       2       2       1       3       3       2       2
33      female  35      159     53      1       1       2       2       1       1       2       2       3       3       2       3
34      male    28      172     74      1       2       2       1       2       2       2       2       2       2       2       2
35      male    24      177     89.8    1       1       1       1       1       1       2       2       2       2       2       2
36      female  19      167     60.2    1       1       2       2       2       2       2       2       3       3       3       3
37      female  37      167     62.2    2       2       2       2       2       2       2       2       3       3       2       2
```

Repeated measures data in short and fat format

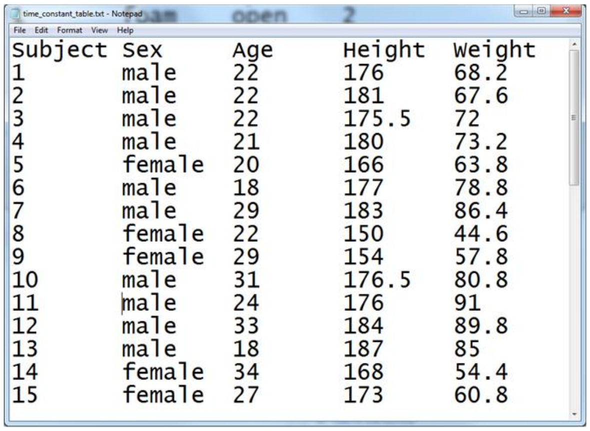
A simple alternative to both tall/thin and short/fat

– Disadvantages of tall/thin
  - Too much repetition

– Disadvantages of short/fat

– Database format
  - Time constant table
  - Time varying table

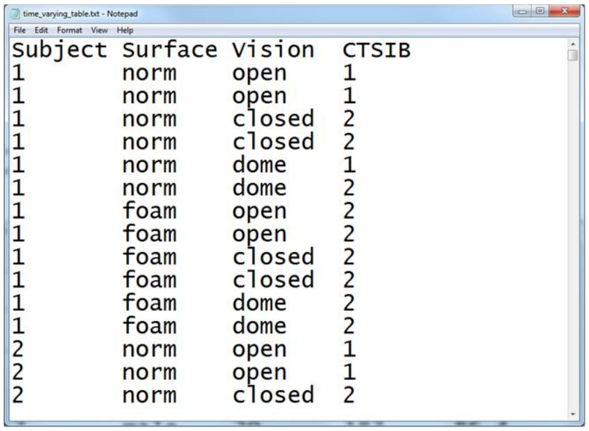
Time constant data



Subject	Sex	Age	Height	Weight
1	male	22	176	68.2
2	male	22	181	67.6
3	male	22	175.5	72
4	male	21	180	73.2
5	female	20	166	63.8
6	male	18	177	78.8
7	male	29	183	86.4
8	female	22	150	44.6
9	female	29	154	57.8
10	male	31	176.5	80.8
11	male	24	176	91
12	male	33	184	89.8
13	male	18	187	85
14	female	34	168	54.4
15	female	27	173	60.8

Table listing time constant data only

Time varying data



```
Subject Surface Vision  CTSIB
1       norm    open    1
1       norm    open    1
1       norm    closed  2
1       norm    closed  2
1       norm    dome    1
1       norm    dome    2
1       foam    open    2
1       foam    open    2
1       foam    closed  2
1       foam    closed  2
1       foam    dome    2
1       foam    dome    2
2       norm    open    1
2       norm    open    1
2       norm    closed  2
```

Table listing time varying data only

Contents of a data dictionary

– Variable names
– Variable labels
– Units of measurement
– Permissible/impermissible values
– Value labels
– Missing value codes
– Source
– License

Third break

- What have you learned
  - Multiple response variables
  - Longitudinal/repeated measures data
- What's next
  - Double entry coding
  - Excel files

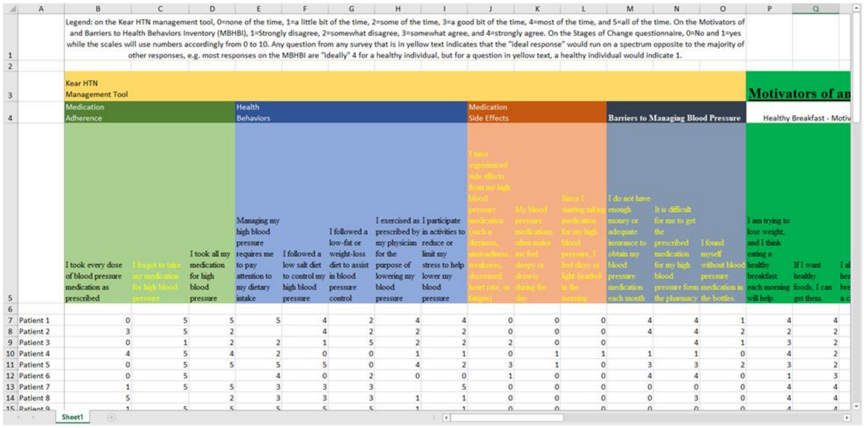
Double entry coding

- Great quality check
  - If you can afford it
- Prepare a code book first
  - Count the proportion of discrepancies
- If too many discrepancies
  - Revise the code book and re-do the data entry.
- If discrepancies small enough
  - Report this number in your publication

If you enter data into Excel

- – Do not use colors
- – Do not include summary statistics
- – Rectangular grid
- – Don’t squeeze two data values into one cell
    - Systolic/diastolic blood pressures
    - 44M for a 44 year old male
- – Variable names in first row
- – No blank cells
    - Contradicts your book

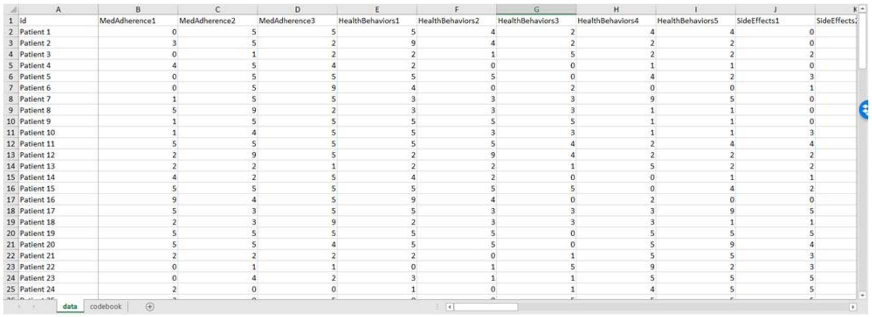
A poorly structured spreadsheet



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	
1		Legend: on the Kear HTN management tool, 0=none of the time, 1=a little bit of the time, 2=some of the time, 3=a good bit of the time, 4=most of the time, and 5=all of the time. On the Motivators of and Barriers to Health Behaviors Inventory (MBHBI), 1=Strongly disagree, 2=somewhat disagree, 3=somewhat agree, and 4=strongly agree. On the Stages of Change questionnaire, 0=No and 1=yes while the scales will use numbers accordingly from 0 to 10. Any question from any survey that is in yellow text indicates that the "ideal response" would run on a spectrum opposite to the majority of other responses, e.g. most responses on the MBHBI are "ideally" 4 for a healthy individual, but for a question in yellow text, a healthy individual would indicate 1.																
2																		
3		Kear HTN Management Tool														Motivators of an		
4		Medication Adherence			Health Behaviors					Medication Side Effects			Barriers to Managing Blood Pressure			Healthy Breakfast - Motiv		
5		I took every dose of blood pressure medication as prescribed	I forgot to take my medication for high blood pressure	I took all my medication for high blood pressure	Managing my high blood pressure requires me to pay attention to my dietary intake	I followed a low salt diet to control my high blood pressure	I followed a low-fat or weight-loss diet to assist in blood pressure control	I exercised as prescribed by my physician for the purpose of lowering my blood pressure	I participate in activities to reduce or limit my stress to help lower my blood pressure	I have experienced side effects from my high blood pressure medication (such a dizziness, unsteadiness, weakness, decreased heart rate, or fatigue)	My blood pressure medications often make me feel sleepy or drowsy during the day	Since I starting taking medication for my high blood pressure, I feel dizzy or light-headed in the morning	I do not have enough money or adequate insurance to obtain my blood pressure medication each month	It is difficult for me to get the prescribed medication for my high blood pressure form the pharmacy	I found myself without blood pressure medication in the bottles.	I am trying to lose weight, and I think eating a healthy breakfast each morning will help.	If I want healthy foods, I can get them.	I al hea bre a c
6																		
7	Patient 1	0	5	5	5	4	2	4	4	0	0	0	4	4	1	4	4	
8	Patient 2	3	5	2		4	2	2	2	0	0	0	4	4	2	2	2	
9	Patient 3	0	1	2	2	1	5	2	2	2	0	0		4	1	3	2	
10	Patient 4	4	5	4	2	0	0	1	1	0	1	1	1	1	0	4	2	
11	Patient 5	0	5	5	5	5	0	4	2	3	1	0	3	3	2	3	2	
12	Patient 6	0	5		4	0	2	0	0	1	0	0	4	4	0	1	3	
13	Patient 7	1	5	5	3	3	3		5	0	0	0	0	0	0	4	4	
14	Patient 8	5		2	3	3	3	1	1	0	0	0	0	3	0	4	4	
15	Patient 9	1	5	5	5	5	5	1	1	0	0	0	0	0	0	4	4	

A spreadsheet with data

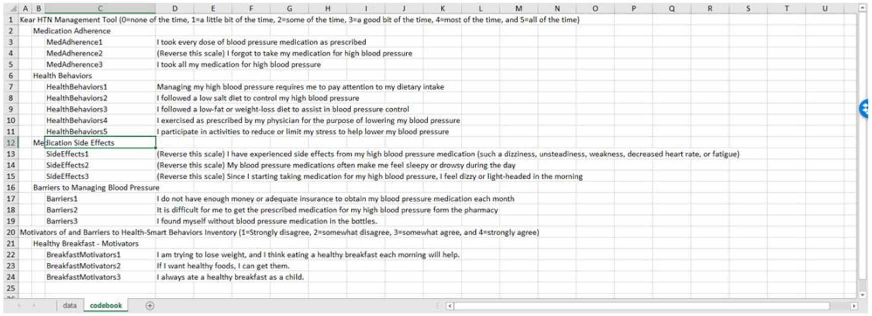
Revisions to this spreadsheet



id	MedAdherence1	MedAdherence2	MedAdherence3	HealthBehaviors1	HealthBehaviors2	HealthBehaviors3	HealthBehaviors4	HealthBehaviors5	SideEffects1
Patient 1	0	5	5	5	4	2	4	4	0
Patient 2	3	5	2	9	4	2	2	2	0
Patient 3	0	1	2	2	1	5	2	2	2
Patient 4	4	5	4	2	0	0	1	1	0
Patient 5	0	5	5	5	5	0	4	2	3
Patient 6	0	5	9	4	0	2	0	0	1
Patient 7	1	5	5	3	3	3	9	5	0
Patient 8	5	9	2	3	3	3	1	1	0
Patient 9	1	5	5	5	5	5	1	1	0
Patient 10	1	4	5	5	3	3	1	1	3
Patient 11	5	5	5	5	5	4	2	4	4
Patient 12	2	9	5	2	9	4	2	2	2
Patient 13	2	2	1	2	2	1	5	2	2
Patient 14	4	2	5	4	2	0	0	1	1
Patient 15	5	5	5	5	5	5	0	4	2
Patient 16	9	4	5	9	4	0	2	0	0
Patient 17	5	3	5	5	3	3	3	9	5
Patient 18	2	3	9	2	3	3	3	1	1
Patient 19	5	5	5	5	5	0	5	5	5
Patient 20	5	5	4	5	5	0	5	9	4
Patient 21	2	2	2	2	0	1	5	5	3
Patient 22	0	1	1	0	1	5	9	2	3
Patient 23	0	4	2	3	1	1	5	5	5
Patient 24	2	0	0	1	0	1	4	5	5

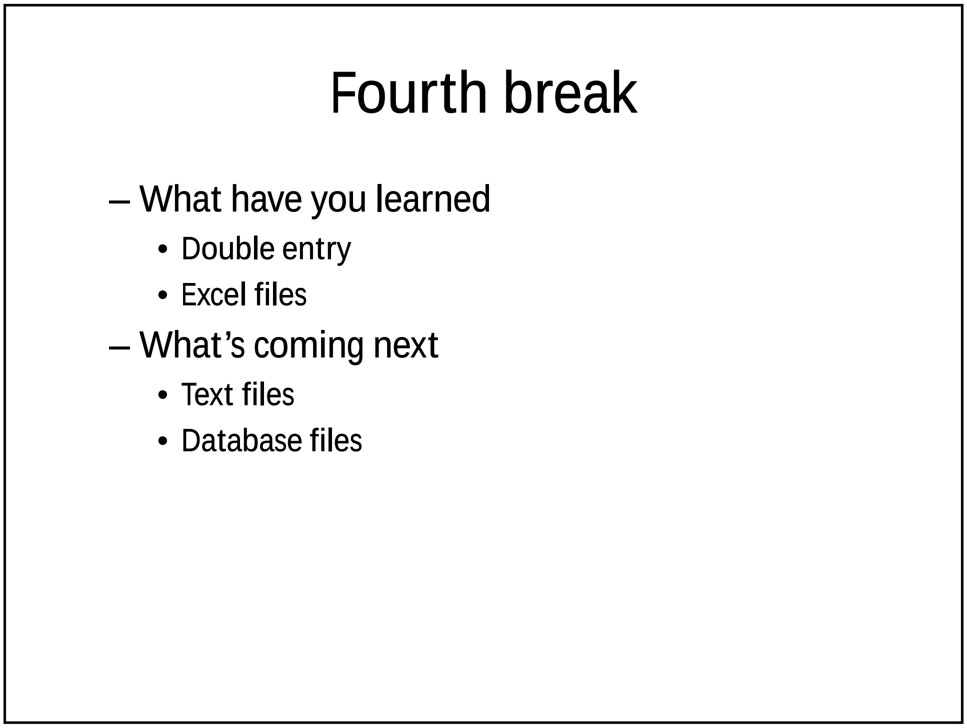
A revised and better organized spreadsheet

The codebook from this spreadsheet



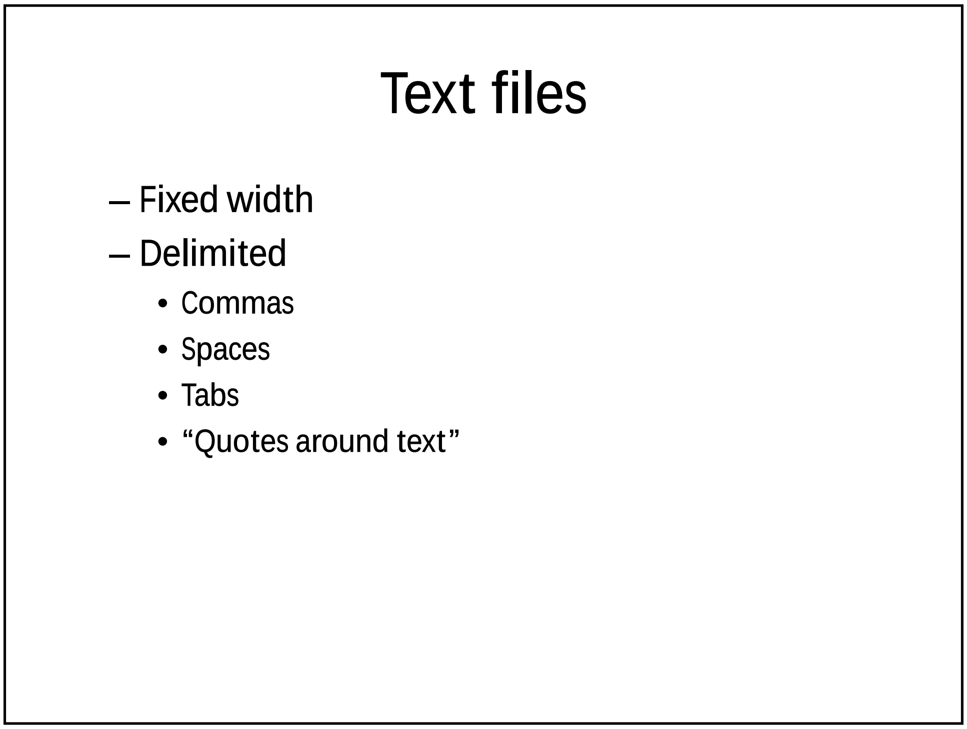
Kear HTN Management Tool (0=none of the time, 1=a little bit of the time, 2=some of the time, 3=a good bit of the time, 4=most of the time, and 5=all of the time)	
Medication Adherence	
MedAdherence1	I took every dose of blood pressure medication as prescribed
MedAdherence2	(Reverse this scale) I forgot to take my medication for high blood pressure
MedAdherence3	I took all my medication for high blood pressure
Health Behaviors	
HealthBehaviors1	Managing my high blood pressure requires me to pay attention to my dietary intake
HealthBehaviors2	I followed a low salt diet to control my high blood pressure
HealthBehaviors3	I followed a low-fat or weight-loss diet to assist in blood pressure control
HealthBehaviors4	I exercised as prescribed by my physician for the purpose of lowering my blood pressure
HealthBehaviors5	I participate in activities to reduce or limit my stress to help lower my blood pressure
Medication Side Effects	
SideEffects1	(Reverse this scale) I have experienced side effects from my high blood pressure medication (such a dizziness, unsteadiness, weakness, decreased heart rate, or fatigue)
SideEffects2	(Reverse this scale) My blood pressure medications often make me feel sleepy or drowsy during the day
SideEffects3	(Reverse this scale) Since I starting taking medication for my high blood pressure, I feel dizzy or light-headed in the morning
Barriers to Managing Blood Pressure	
Barriers1	I do not have enough money or adequate insurance to obtain my blood pressure medication each month
Barriers2	It is difficult for me to get the prescribed medication for my high blood pressure form the pharmacy
Barriers3	I found myself without blood pressure medication in the bottles.
Motivators of and Barriers to Health-Smart Behaviors Inventory (1=Strongly disagree, 2=somewhat disagree, 3=somewhat agree, and 4=strongly agree)	
Healthy Breakfast - Motivators	
BreakfastMotivators1	I am trying to lose weight, and I think eating a healthy breakfast each morning will help.
BreakfastMotivators2	If I want healthy foods, I can get them.
BreakfastMotivators3	I always ate a healthy breakfast as a child.

A codebook associated with revised spreadsheet



Fourth break

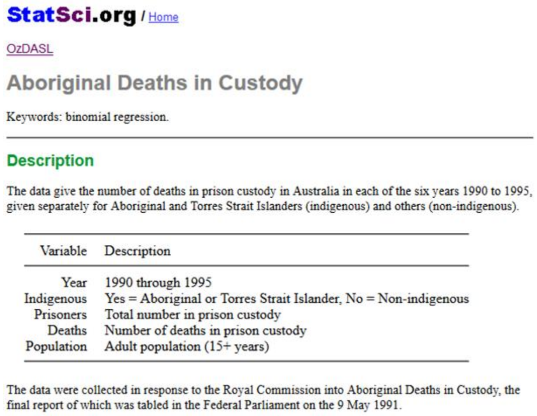
- What have you learned
  - Double entry
  - Excel files
- What’s coming next
  - Text files
  - Database files



Text files

- Fixed width
- Delimited
  - Commas
  - Spaces
  - Tabs
  - “Quotes around text”

Data dictionary for aboriginal prison death study



StatSci.org / Home

OzDASL

Aboriginal Deaths in Custody

Keywords: binomial regression.

Description

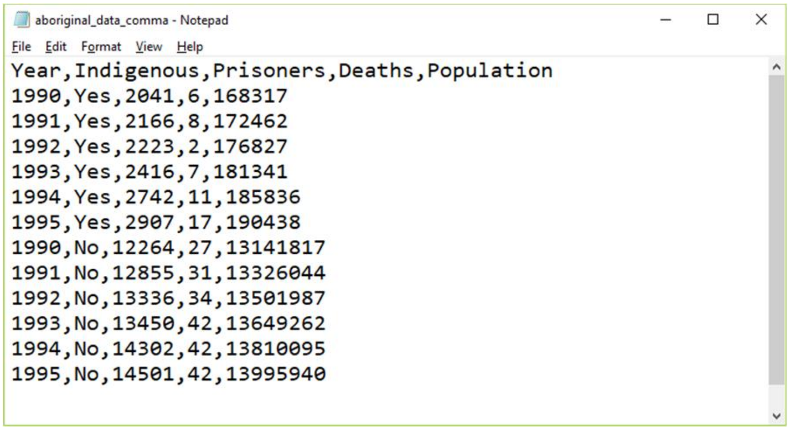
The data give the number of deaths in prison custody in Australia in each of the six years 1990 to 1995, given separately for Aboriginal and Torres Strait Islanders (indigenous) and others (non-indigenous).

Variable	Description
Year	1990 through 1995
Indigenous	Yes = Aboriginal or Torres Strait Islander, No = Non-indigenous
Prisoners	Total number in prison custody
Deaths	Number of deaths in prison custody
Population	Adult population (15+ years)

The data were collected in response to the Royal Commission into Aboriginal Deaths in Custody, the final report of which was tabled in the Federal Parliament on the 9 May 1991.

Data dictionary

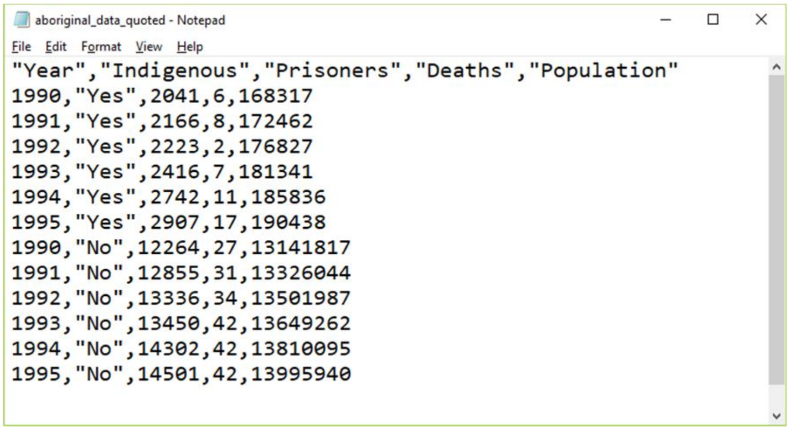
Comma separated values (csv)



```
Year,Indigenous,Prisoners,Deaths,Population
1990,Yes,2041,6,168317
1991,Yes,2166,8,172462
1992,Yes,2223,2,176827
1993,Yes,2416,7,181341
1994,Yes,2742,11,185836
1995,Yes,2907,17,190438
1990,No,12264,27,13141817
1991,No,12855,31,13326044
1992,No,13336,34,13501987
1993,No,13450,42,13649262
1994,No,14302,42,13810095
1995,No,14501,42,13995940
```

Data set using a comma separated value format

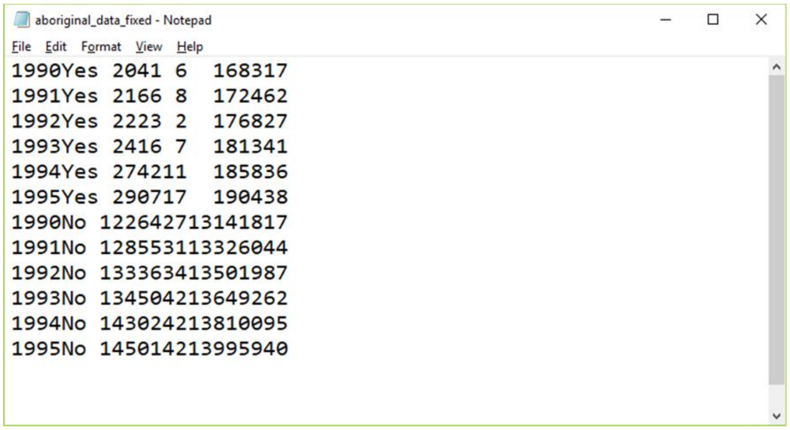
Comma separated values with quotes



```
"Year","Indigenous","Prisoners","Deaths","Population"
1990,"Yes",2041,6,168317
1991,"Yes",2166,8,172462
1992,"Yes",2223,2,176827
1993,"Yes",2416,7,181341
1994,"Yes",2742,11,185836
1995,"Yes",2907,17,190438
1990,"No",12264,27,13141817
1991,"No",12855,31,13326044
1992,"No",13336,34,13501987
1993,"No",13450,42,13649262
1994,"No",14302,42,13810095
1995,"No",14501,42,13995940
```

Data set using a quoted format

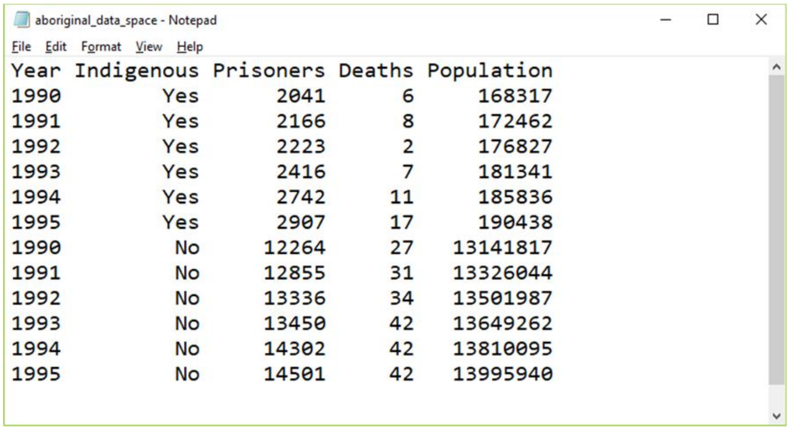
Fixed width format



```
1990Yes 2041 6  168317
1991Yes 2166 8  172462
1992Yes 2223 2  176827
1993Yes 2416 7  181341
1994Yes 274211  185836
1995Yes 290717  190438
1990No 122642713141817
1991No 128553113326044
1992No 133363413501987
1993No 134504213649262
1994No 143024213810095
1995No 145014213995940
```

Data set using a fixed width format

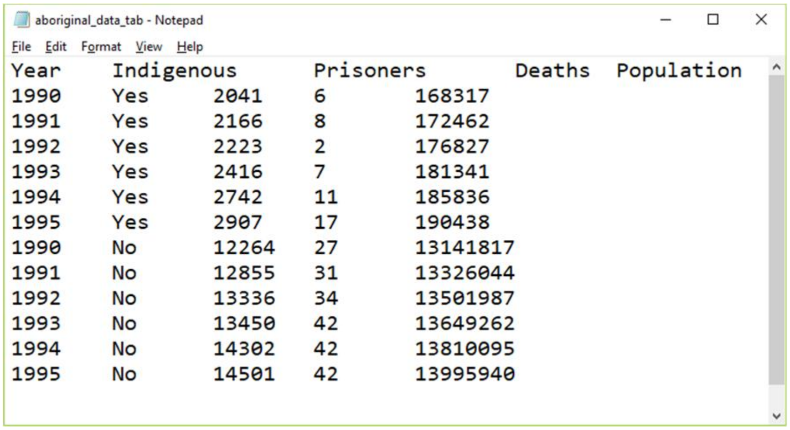
Spaced format



```
Year Indigenous Prisoners Deaths Population
1990        Yes      2041      6     168317
1991        Yes      2166      8     172462
1992        Yes      2223      2     176827
1993        Yes      2416      7     181341
1994        Yes      2742     11     185836
1995        Yes      2907     17     190438
1990         No     12264     27   13141817
1991         No     12855     31   13326044
1992         No     13336     34   13501987
1993         No     13450     42   13649262
1994         No     14302     42   13810095
1995         No     14501     42   13995940
```

Data set using a spaced format

Tab separated values



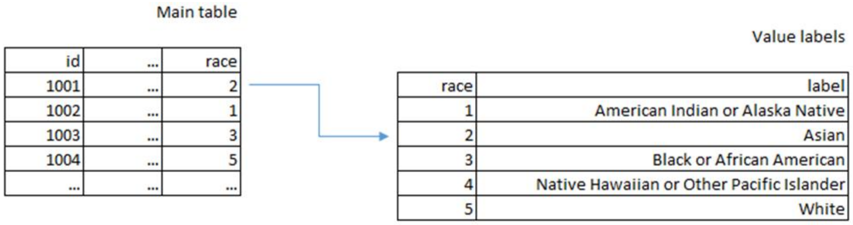
```
Year    Indigenous      Prisoners       Deaths  Population
1990    Yes     2041    6       168317
1991    Yes     2166    8       172462
1992    Yes     2223    2       176827
1993    Yes     2416    7       181341
1994    Yes     2742    11      185836
1995    Yes     2907    17      190438
1990    No      12264   27      13141817
1991    No      12855   31      13326044
1992    No      13336   34      13501987
1993    No      13450   42      13649262
1994    No      14302   42      13810095
1995    No      14501   42      13995940
```

Data set using a tab separated value format

Database systems

- Terminology
  - Tables
  - Fields
  - Records
  - Primary key
  - Foreign key

Database table for value labels



Main table

id	...	race
1001	...	2
1002	...	1
1003	...	3
1004	...	5
...	...	...

Value labels

race	label
1	American Indian or Alaska Native
2	Asian
3	Black or African American
4	Native Hawaiian or Other Pacific Islander
5	White

Database linkage between race code and race labels

REDCap

– Research Electronic Data Capture

– Not open source, but freely distributed by Vanderbilt

– Software components
  - PHP
  - JavaScript
  - MySQL

– Case report forms

– Strongly recommended

Conclusion

– Data dictionary
  - Variable names, Variable labels, Value labels, Missing value codes

– Managing complex files
  - Multiple response, Longitudinal/repeated measures data

– Storage options
  - Spreadsheet, Text file, Database, REDCap