

Data Collection and Tabulation

Jeffrey Michael Franc
MD, MSc, FCFP.EM, Dip Sport Med, EMDM

Medical Director, E/D Management
Alberta Health Services

Associate Clinical Professor of Emergency Medicine
University of Alberta

Visiting Professor in Disaster Medicine
Universita' Degli Studi del Piemonte Orientale

MedStatStudio



Objectives

- Understand the principles of how to randomize and take a random sample
- Understand how to tabulate data for future statistical analysis

Randomization: Review

What are the major reasons for using randomization?

Randomization

1. Eliminates selection bias
2. Avoids confounding
3. Balances group with respect to covariates:
 - Known
 - Unknown
4. Allows statistical assessment of causality

How to Randomize?

How will you randomize your subjects?

Randomization of Subjects

What are possible methods to randomize subjects?

Randomization of Subjects

What are possible methods to randomize subjects?

- Flip a coin
- Random number lists
- Computer software
- Web based randomization
 - www.randomization.com
 - www.sealedenvelope.com

Randomization Methods

Method used will depend largely on the complexity of the randomization

Randomization

What are the basic types of randomization?

Randomization

What are the basic types of randomization?

- Simple
- Block
- Stratified
- Covariate Adaptive

Simple Randomization

- Equal probability of being randomized to any group
- Expect group sizes may be unequal
- Works well for large sample sizes and for ongoing studies where the sample size is unknown

Simple Randomization

A researcher is studying the effect of prednisone on the duration of infectious rhinosinusitis among patients with a history of seasonal allergies. Over the next one year, all patients meeting the study criteria will be randomized to receive prednisone or to the control group.

Coins

Obverse (National) =
Prednisone

Reverse (Euro) =
Control



Random Number Tables

Researcher will read
down column 'A'.
Odd numbers are
control, even
numbers are
prednisone

	A	B	C
01	46947	71735	94246
02	47417	72361	22495
03	63764	31439	69853
04	69586	04651	54047
05	64466	44369	54621
06	03030	85073	47591
07	97556	80617	38868
08	08858	18891	23055
09	58167	83419	52426
10	64238	97862	29802
11	68969	49254	93327
12	83410	76140	24855
13	65000	99048	91260
14	71572	87436	04552
15	91120	54017	26108
16	62834	51303	44974
17	19877	19006	52479

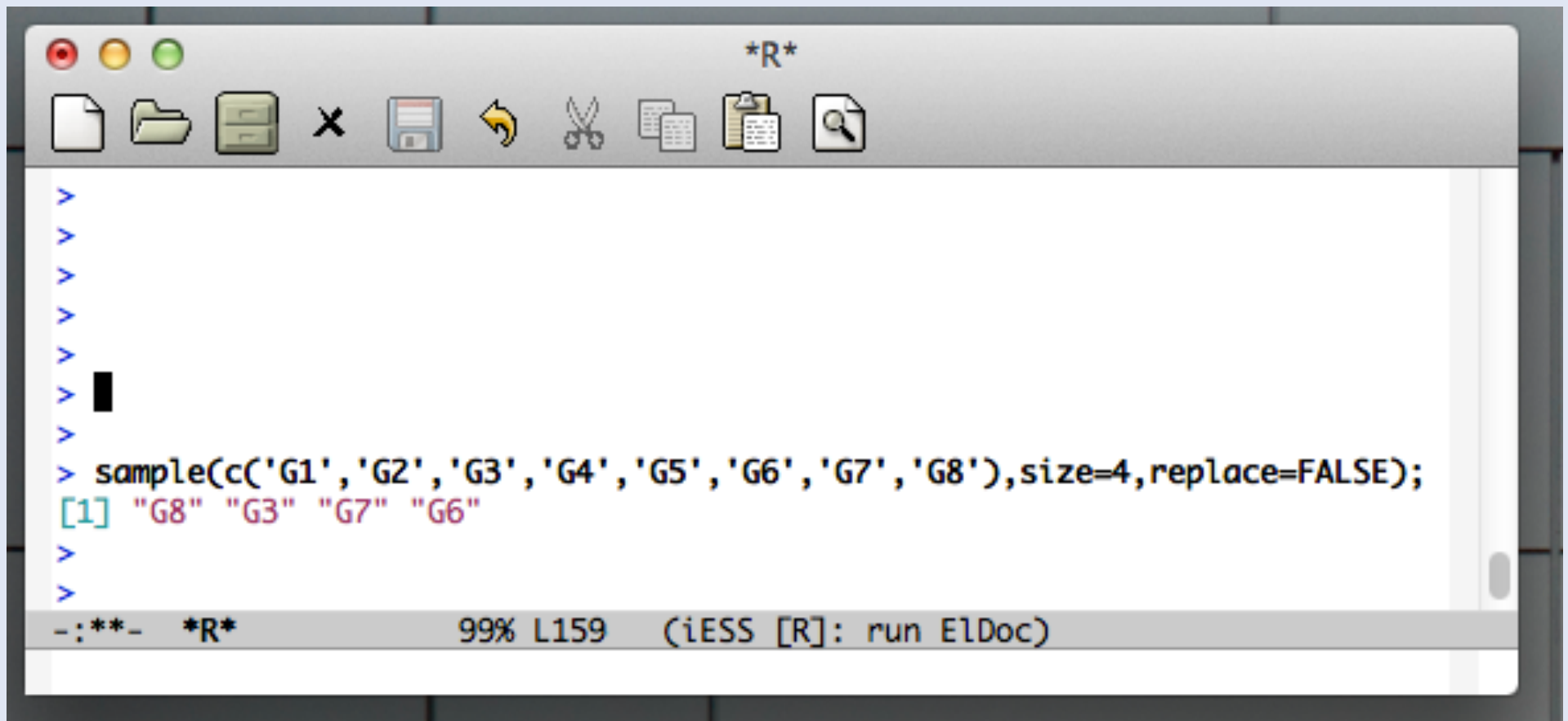
Block Randomization

- Used to randomize to equal size groups
- Must know size of sample at onset of experiment
- For small experiments maximizes the power by balancing group size

Block Randomization

A researcher is investigating the utility of helicopter versus snowmobile for backcountry rescue in winter. She has eight volunteer teams who will perform a simulated rescue. She wishes to assign 4 teams to the snowmobile group and 4 teams to the helicopter group.

R: Programming Language



The image shows a screenshot of an R console window. The window title is "*R*" and it has a standard macOS-style title bar with red, yellow, and green window control buttons. Below the title bar is a toolbar with icons for file operations: a document, a folder, a save icon, a close icon, a save icon, a redo icon, a scissors icon, a copy icon, a paste icon, and a search icon. The main area of the window is a text editor with a white background and a vertical scrollbar on the right. The text in the editor is as follows:

```
>  
>  
>  
>  
>  
>  
>  
>  
>  
> sample(c('G1', 'G2', 'G3', 'G4', 'G5', 'G6', 'G7', 'G8'), size=4, replace=FALSE);  
[1] "G8" "G3" "G7" "G6"  
>  
>
```

At the bottom of the window, there is a status bar with the following text: "-:***- *R* 99% L159 (iESS [R]: run ElDoc)".

www.randomization.com

The screenshot shows a web browser window with the URL `http://www...size_r.htm`. The page title is "Randomization Plans" and the sub-header is "Randomizing subjects to a single treatment". The form includes a section for "Treatment labels" with a table containing "Helicopter" and "Snowmobile". Below this is a section for "Number of subjects per block/number of blocks" with four rows, each containing the value "4" and "1". There is also a field for "Initial subject ID number" with the value "1". A "Generate Plan" button and a "Help" link are visible at the bottom of the form. The browser's address bar and search bar are also visible.

randomization.com

randomization.com x http://www...size_r.htm x Create a blocked rando... x +

www.randomization.com Google

GARZANTI MEDLINE Calendar RSeek.org R-pro... randomization -...

Randomization Plans

Randomizing subjects to a single treatment

Treatment labels: (enter as many as necessary)

Helicopter	Snowmobile		

Number of subjects per block/number of blocks 4 / 1

Number of subjects per block/number of blocks 4 / 1

Number of subjects per block/number of blocks / 1

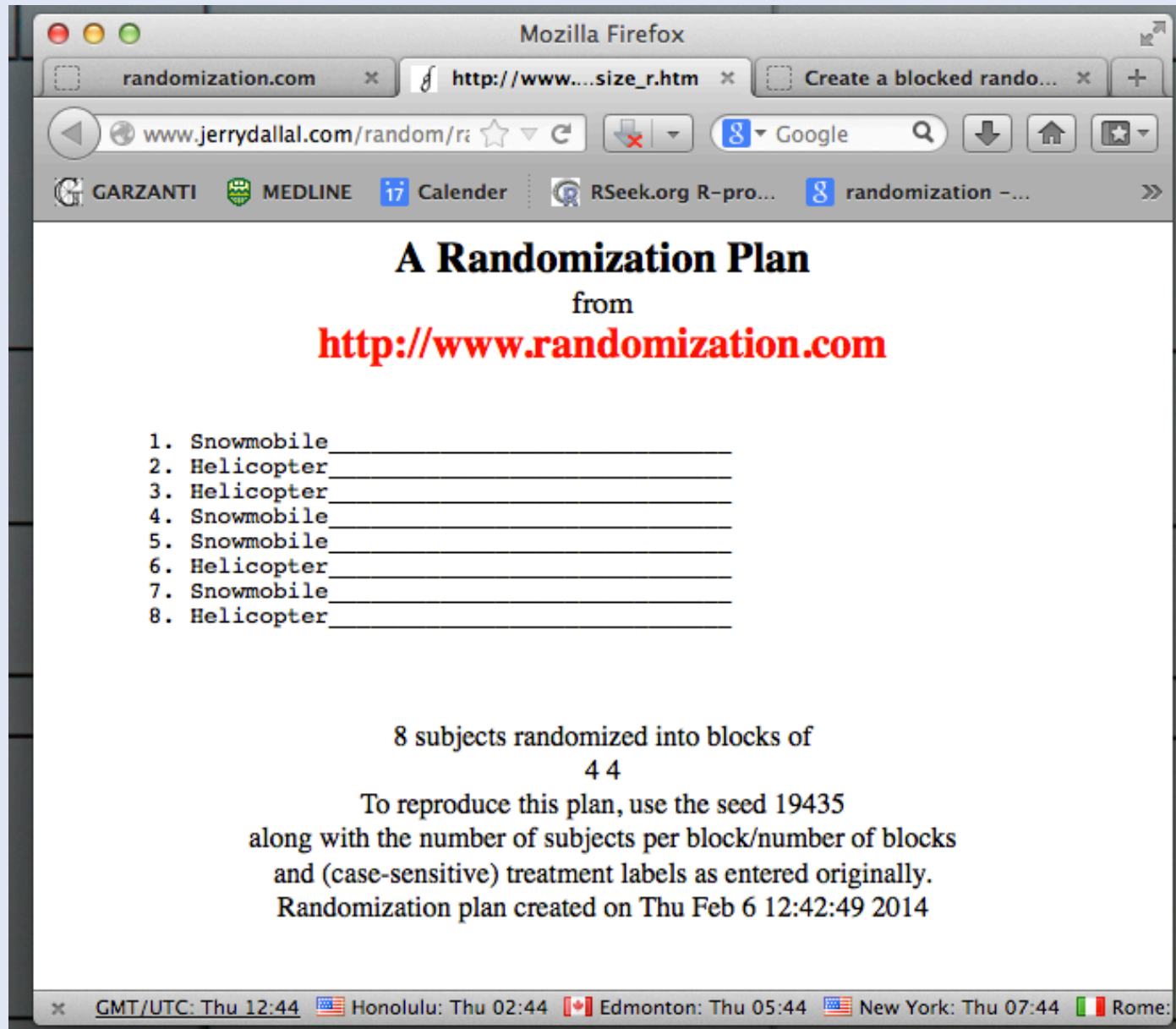
Number of subjects per block/number of blocks / 1

Initial subject ID number 1

Generate Plan Help

GMT/UTC: Thu 12:43 Honolulu: Thu 02:43 Edmonton: Thu 05:43 New York: Thu 07:43 Rome:

www.randomization.com



The screenshot shows a Mozilla Firefox browser window with the following details:

- Address bar: www.jerrydallal.com/random/r/
- Search engine: Google
- Bookmarks: GARZANTI, MEDLINE, Calender, RSeek.org R-pro..., randomization -...

The main content of the page is:

A Randomization Plan

from
<http://www.randomization.com>

1. Snowmobile _____
2. Helicopter _____
3. Helicopter _____
4. Snowmobile _____
5. Snowmobile _____
6. Helicopter _____
7. Snowmobile _____
8. Helicopter _____

8 subjects randomized into blocks of
4 4

To reproduce this plan, use the seed 19435
along with the number of subjects per block/number of blocks
and (case-sensitive) treatment labels as entered originally.
Randomization plan created on Thu Feb 6 12:42:49 2014

GMT/UTC: Thu 12:44 Honolulu: Thu 02:44 Edmonton: Thu 05:44 New York: Thu 07:44 Rome:

Stratified Randomization

- Balances a suspected covariate
- Must know the covariate at the outset of experiment
- Complicated if there are multiple covariates
- Should be considered when there are obvious significant covariates...especially if sample size is small

Stratified Randomization

Twenty students will be randomized to simulation training or no-training prior to a live exercise. The researchers will pick 4 students from each of the 5 years of the residency training, and wish to make sure that the groups are balanced but randomized.

www.sealedenvelope.com

The screenshot shows a web browser window with the URL <https://www.sealedenvelope.com/simple-randomiser/v1/lists>. The page title is "Create a blocked randomisation list | Sealed Envelope | Randomisation (randomization) and database services for clinical trials". The browser's address bar shows the URL and search engines like Google. The page header includes navigation links: HOME, RANDOMISATION, RED PILL, TRIALS, PRICING, POWER CALCULATORS, HELP, CONTACT. The main heading is "SealedEnvelope™" with the tagline "★ THE ORIGINAL INTERNET AND TELEPHONE RANDOMISATION SINCE 2001". Below this is a large blue and black heading: "CREATE A RANDOMISATION LIST". A sub-heading reads: "Use this tool to create a blocked randomisation list for your trial. The generated lists are suitable for use with our [simple randomisation service](#)".

The interface is divided into two main sections: "CREATE A LIST" and "YOUR LIST".

CREATE A LIST

- Seed:** 136454596625669
- Treatment groups:** Simulation, No Simulation
- Block sizes:** 4,4
- List length:** 18
- Strata (optional):** Year: Year1,Year2,Year3,Year4,Year5
- Generate unique randomisation code?
- Buttons: [Create list](#), [Download as CSV](#)

YOUR LIST

Seed: 136454596625669
Block sizes: 4,4
Strata: Year (Year1, Year2, Year3, Year4, Year5)

block identifier, block size, sequence within block, treatment, Year

```
1, 4, 1, No Simulation, Year1
1, 4, 2, No Simulation, Year1
1, 4, 3, Simulation, Year1
1, 4, 4, Simulation, Year1
2, 4, 1, No Simulation, Year2
2, 4, 2, Simulation, Year2
2, 4, 3, Simulation, Year2
2, 4, 4, No Simulation, Year2
3, 4, 1, Simulation, Year3
3, 4, 2, Simulation, Year3
3, 4, 3, No Simulation, Year3
3, 4, 4, No Simulation, Year3
4, 4, 1, No Simulation, Year4
4, 4, 2, Simulation, Year4
4, 4, 3, Simulation, Year4
4, 4, 4, No Simulation, Year4
5, 4, 1, Simulation, Year5
5, 4, 2, Simulation, Year5
5, 4, 3, No Simulation, Year5
5, 4, 4, No Simulation, Year5
```

The footer of the browser window shows the time in GMT/UTC: Thu 12:26 and other locations: Honolulu: Thu 02:26, Edmonton: Thu 05:26, New York: Thu 07:26, Rome: Thu 13:26.

Covariate Adaptive Randomization

- New participants are assigned to a particular group depending on the characteristics of those in the study.
- Useful when the subjects are not known before the trial starts
- Usually reserved to complex trials

Randomization

Questions?

Data Collection

20131207 - CTAS Data-3.xlsx

Search in Sheet

Home Layout Tables Charts SmartArt Formulas Data Review

Font: Verdana, 10. Alignment: abc, Wrap Text. Number: General. Format: Conditional Formatting, Styles. Cells: Insert, Delete, Format. Themes: Themes, Aa.

AM20 fx 51

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	
1		CTAS 1				CTAS 3				CTAS 5				CTAS 7			CTAS 9					
2	CASE	SCORE	TIME	TYPING TIME		SCORE	TIME	TYPING TIME		SCORE	TIME	TYPING TIME		SCORE	TIME	TYPING TIME	SCORE	TIME	TYPING TIME		C	
3	1	4	144	not recorded		4	107	50		5	217	62		4	131	61	5	94	35		S	
4	2	2	165	80		2	181	71		3	247	116		2	156	70	3	126	75			
5	3	2	158	52		2	159	43		2	219	110		2	179	70	2	148	54			
6	4	1	141	50		1	154	54		1	207	90		1	152	79	1	73	24			
7	5	1	156	69		1	160	55		1	190	81		1	158	87	1	121	56			
8	6	4	104	43		4	105	40		4	133	49		4	137	68	4	146	48			
9	7	2	154	76		3	149	47		3	222	139		2	177	94	2	101	57			
10	8	4	96	35		4	80	30		4	123	52		3	123	55	4	94	57			
11	9	1	130	69		1	128	60		1	92	no typing		1	186	99	1	84	40			
12	PRACTICE	N				N				N				N			N				N	
13	YRS TRIAGING	30				20				18				13			1					
14	PRIOR DISASTER EVENT	1				0				1				0			0					
15	PRIOR DISASTER TRAINING	Disaster planner (1992-94)				0					1			0			2					
16	START TRAINING SESSION	5				0				0				0			20					
17																						
18		START 2				START 4				START 6				START 8			START 10				S	
19	CASE	SCORE	TIME			SCORE	TIME			SCORE	TIME			SCORE	TIME		SCORE	TIME			S	
20	1	G	29			G	33			G	27			G	25		G	55			G	
21	2	G	29			Y	37			Y	32			Y	33		Y	62			Y	
22	3	Y	29			R	59			Y	28			R	20		R	33			R	
23	4	R	13			R	34			R	26			R	41		R	49			R	
24	5	R	22			R	38			R	28			R	28		R	33			R	
25	6	G	16			G	39			G	19			Y	25		G	23			G	
26	7	Y	32			G	38			Y	26			R	29		Y	46			Y	
27	8	G	21			G	30			G	19			G	13		Y	49			G	
28	9	R	34			R	40			B	25			B	25		R	37			R	
29	PRACTICE	N				Y				N				N			Y				N	
30	YRS TRIAGING	20				13				24				3			5					
31	PRIOR DISASTER EVENT	10				0				1				0			0					
32	PRIOR DISASTER TRAINING	10				2				3				3			0					
33	START TRAINING SESSION	10				0				0				3			0					
34																						
35																						
36																						

Normal View Ready Sum=406

Data Collection

How can we tabulate our data for future analysis?

Options

- How can we tabulate our data for future analysis?
 - Paper tables
 - Simple spreadsheets
 - Excel
 - Database
 - MySQL
 - Direct format for statistics software:
 - SAS
 - R
 - Others

Which to Choose?

Unless you are CERTAIN of what software will be used for the entire analysis, the best way to tabulate data is probably a spreadsheet.

Data will need to be read into a statistics software package

All common statistics software will be able to read the spreadsheet *if it is formatted properly*.

Spreadsheets

AM20 fx 51

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1																						
2		CTAS 1				CTAS 3				CTAS 5				CTAS 7				CTAS 9				
3	CASE	SCORE	TIME	TYPING TIME		SCORE	TIME	TYPING TIME		SCORE	TIME	TYPING TIME		SCORE	TIME	TYPING TIME		SCORE	TIME	TYPING TIME		
4	1	4	144	not recorded		4	107	50		5	217	62		4	131	61		5	94			35
5	2	2	165	80		2	181	71		3	247	116		2	156	70		3	126			75
6	3	2	158	52		2	159	43		2	219	110		2	179	70		2	148			54
7	4	1	141	50		1	154	54		1	207	90		1	152	79		1	73			24
8	5	1	156	69		1	160	55		1	190	81		1	158	87		1	121			56
9	6	4	104	43		4	105	40		4	133	49		4	137	68		4	146			48
10	7	2	154	76		3	149	47		3	222	139		2	177	94		2	101			57
11	8	4	96	35		4	80	30		4	123	52		3	123	55		4	94			57
12	9	1	130	69		1	128	60		1	92	no typing		1	186	99		1	84			40
13	PRACTICE	N				N				N				N				N				N
14	YRS TRIAGING		30				20				18				13				1			
15	PRIOR DISASTER EVENT		1				0				1				0				0			
16	PRIOR DISASTER TRAINING		Disaster planner (1992-94)				0				1				0				2			
17	START TRAINING SESSION		5				0				0				0				20			
18																						
19	CASE	START 2				START 4				START 6				START 8				START 10				S
20	1	SCORE	TIME			SCORE	TIME			SCORE	TIME			SCORE	TIME			SCORE	TIME			S
21	2	G	29			G	33			Y	27			G	25			G	55			G
22	3	G	29			Y	37			Y	32			Y	33			Y	62			Y
23	4	Y	29			R	59			Y	28			R	20			R	33			R
24	5	R	13			R	34			R	26			R	41			R	49			R
25	6	R	22			R	28			R	28			R	28			R	33			R
26	7	G	16			G	39			G	19			Y	25			G	23			G
27	8	Y	32			Y	38			Y	26			R	29			Y	46			Y
28	9	G	21			G	30			G	19			G	13			Y	49			G
29	PRACTICE		34			R	40			B	25			B	25			R	37			R
30	YRS TRIAGING	N				Y				N				N				Y				N
31	PRIOR DISASTER EVENT		20				13				24				3				5			
32	PRIOR DISASTER TRAINING		10				0				1				0				0			
33	START TRAINING SESSION		10				2				3				3				0			
34							0				0				3				0			
35																						
36																						

Problematic !!!

All Data All Data - Column Expert Opinion Typing Times Time to Triage Triage

Tips for Using Spreadsheet

- Best to use only one sheet per workbook
- Should be able to save in .csv
- Filename with no spaces, punctuation, or special characters
- Talk to the statistician in advance if you can

Constructing the Data Table With Excel

- Use a very simple table structure:
 - First column is the response variable
 - Subsequent columns are factors
 - First row is factor names
 - The names are short and will be used for the analysis
 - no spaces
 - No !@#\$%^&* except underscore (_)
 - Factor names start with a letter
 - Each row is an observation

Entering Data in Spreadsheets

- Leave BLANK if data not available
 - Do not use '0' to mean not available
 - Do not use any words to mean blank
- Be very consistent with formatting:
 - 'Y' vs 'Yes'
 - 'True' vs 'TRUE' vs 'true'
- Don't use Codes
 - Such as '999' for 'too high to measure'
- Use letters (not numbers) for nominal categories if possible

Example

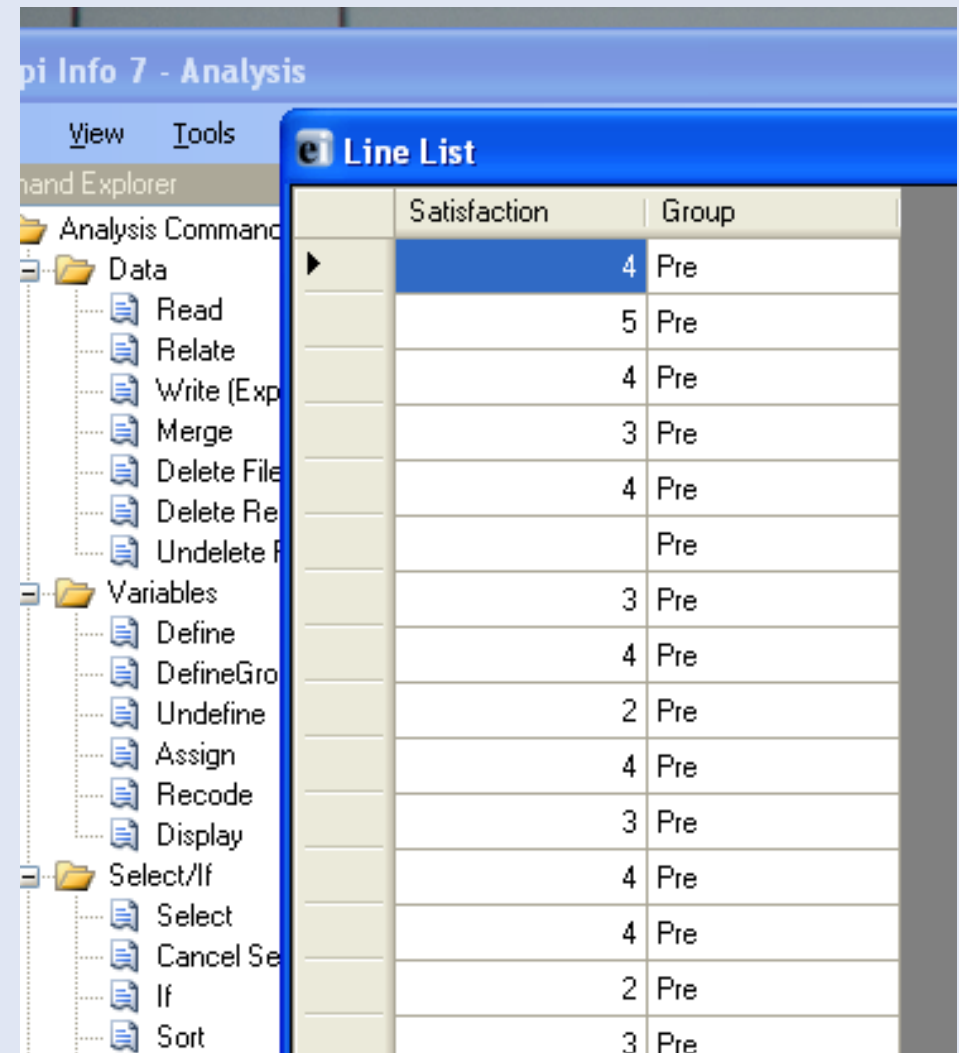
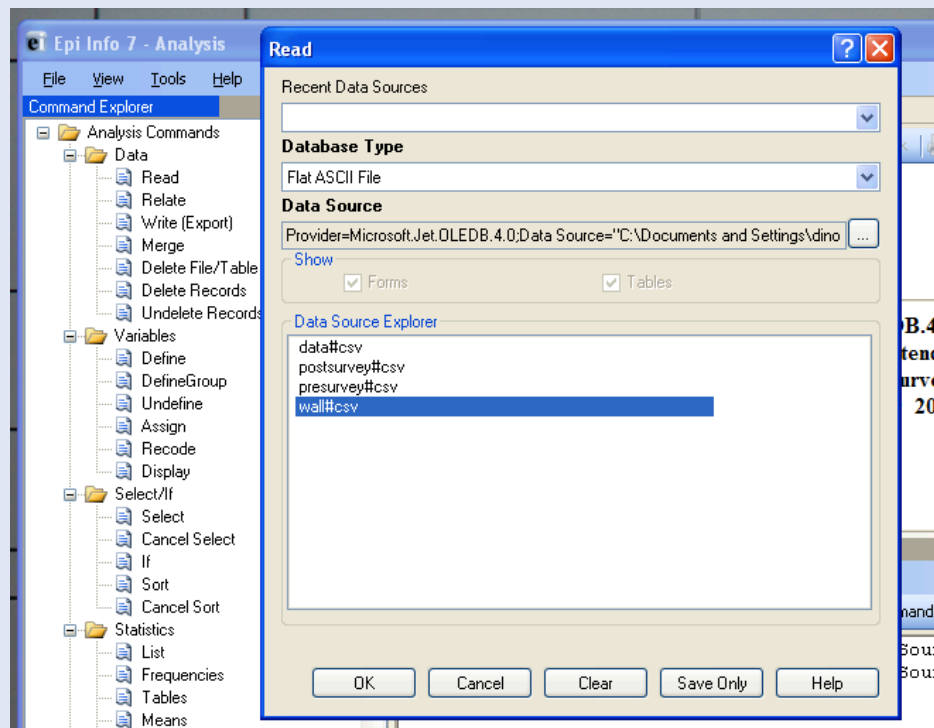
The image shows a screenshot of the Microsoft Excel application window. The title bar indicates the file is named 'data.csv'. The ribbon is set to the 'Home' tab, with the 'Edit' group selected. The active cell is A1, containing the text 'time'. The spreadsheet contains 16 rows of data with 13 columns. The columns are labeled: time, score, method, operator, case, standard, typing, practice, years, event, distraining, starttraining, and correct. The data rows contain numerical values, categorical text, and boolean values (TRUE/FALSE).

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	time	score	method	operator	case	standard	typing	practice	years	event	distraining	starttraining	correct
2	144	4	ctas	C1	1	5	NA	N	30	1	Y	Y	FALSE
3	165	2	ctas	C1	2	3	80	N	30	1	Y	Y	FALSE
4	158	2	ctas	C1	3	2	52	N	30	1	Y	Y	TRUE
5	141	1	ctas	C1	4	1	50	N	30	1	Y	Y	TRUE
6	156	1	ctas	C1	5	1	69	N	30	1	Y	Y	TRUE
7	104	4	ctas	C1	6	4	43	N	30	1	Y	Y	TRUE
8	154	2	ctas	C1	7	2	76	N	30	1	Y	Y	TRUE
9	96	4	ctas	C1	8	4	35	N	30	1	Y	Y	TRUE
10	130	1	ctas	C1	9	1	69	N	30	1	Y	Y	TRUE
11	107	4	ctas	C2	1	5	50	N	20	0	N	N	FALSE
12	181	2	ctas	C2	2	3	71	N	20	0	N	N	FALSE
13	159	2	ctas	C2	3	2	43	N	20	0	N	N	TRUE
14	154	1	ctas	C2	4	1	54	N	20	0	N	N	TRUE
15	160	1	ctas	C2	5	1	55	N	20	0	N	N	TRUE
16	105	4	ctas	C2	6	4	40	N	20	0	N	N	TRUE

Reading Data to Statistics Software

All popular software packages will be able to read your excel sheet if properly formatted!!

Read Spreadsheet: Epi Info



Read Spreadsheet: R

```
Emacs File Edit Options Tools iESS Complete In/Out Signals Bu
*R*
>
>
> ctasdata<-read.csv('ctasdata.csv');
> str(ctasdata);
'data.frame': 180 obs. of 13 variables:
 $ time      : int 144 165 158 141 156 104 154 96 130 107 ...
 $ score     : Factor w/ 9 levels "1","2","3","4",...: 4 2 2 1 1 4 2 4 1 4 ...
 $ method    : Factor w/ 2 levels "ctas","start": 1 1 1 1 1 1 1 1 1 1 ...
 $ operator  : Factor w/ 20 levels "C1","C10","C2",...: 1 1 1 1 1 1 1 1 1 3 ..
 $ case      : int 1 2 3 4 5 6 7 8 9 1 ...
 $ standard  : Factor w/ 8 levels "1","2","3","4",...: 5 3 2 1 1 4 2 4 1 5 ...
 $ typing    : int NA 80 52 50 69 43 76 35 69 50 ...
 $ practice  : Factor w/ 2 levels "N","Y": 1 1 1 1 1 1 1 1 1 1 ...
 $ years     : num 30 30 30 30 30 30 30 30 30 20 ...
 $ event     : int 1 1 1 1 1 1 1 1 1 0 ...
 $ distraining : Factor w/ 2 levels "N","Y": 2 2 2 2 2 2 2 2 2 1 ...
 $ starttraining: Factor w/ 2 levels "N","Y": 2 2 2 2 2 2 2 2 2 1 ...
 $ correct   : logi FALSE FALSE TRUE TRUE TRUE TRUE ...
> ctasdata
  time score method operator case standard typing practice years event
1  144     4   ctas      C1      1         5      NA          N 30.0     1
2  165     2   ctas      C1      2         3      80          N 30.0     1
3  158     2   ctas      C1      3         2      52          N 30.0     1
4  141     1   ctas      C1      4         1      50          N 30.0     1
5  156     1   ctas      C1      5         1      69          N 30.0     1
6  104     4   ctas      C1      6         4      43          N 30.0     1
7  154     2   ctas      C1      7         2      76          N 30.0     1
8   96     4   ctas      C1      8         4      35          N 30.0     1
9  130     1   ctas      C1      9         1      69          N 30.0     1
10 107     4   ctas      C2      1         5      50          N 20.0     0
11 181     2   ctas      C2      2         3      71          N 20.0     0
12 159     2   ctas      C2      3         2      43          N 20.0     0
-:***- *R* 70% L1012 (iESS [R]: run E1Doc)
```

Tips for Using Statistics Software

- Write reusable code (either scripts or functions).
- Your code should read the data file each time.
- Avoid proprietary data formats
- Most Important: Software makes it very easy to use multitudes of tests, make sure you know the right ones, and how to apply them

Data Tabulation

Questions?

Objectives

- Understand the principles of how to randomize and take a random sample
- Understand how to tabulate data for future statistical analysis