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 Randomizaton

- 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

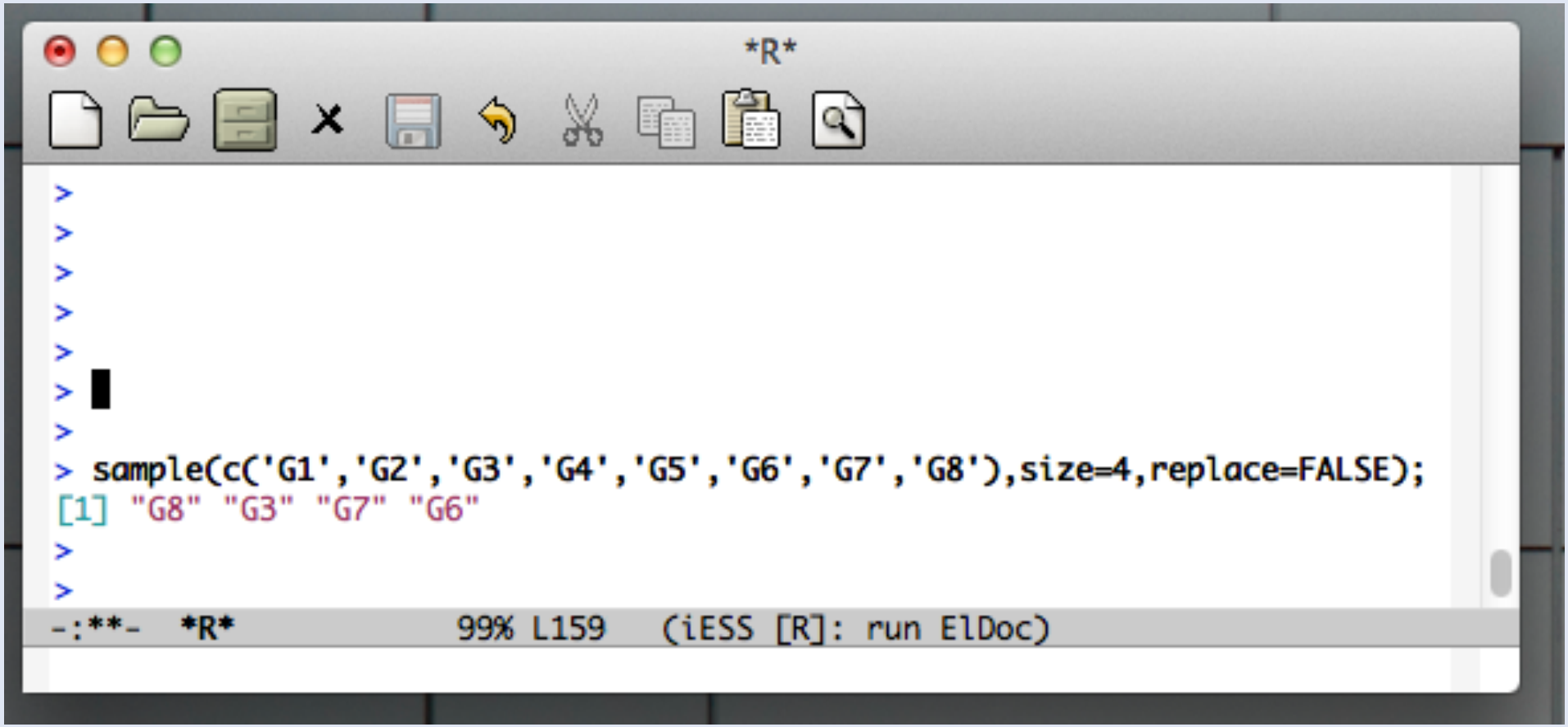
Block Randomization

- 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



www.randomization.com

randomization.com

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

www.randomization.com

GARZANTI | MEDLINE | Calender | 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

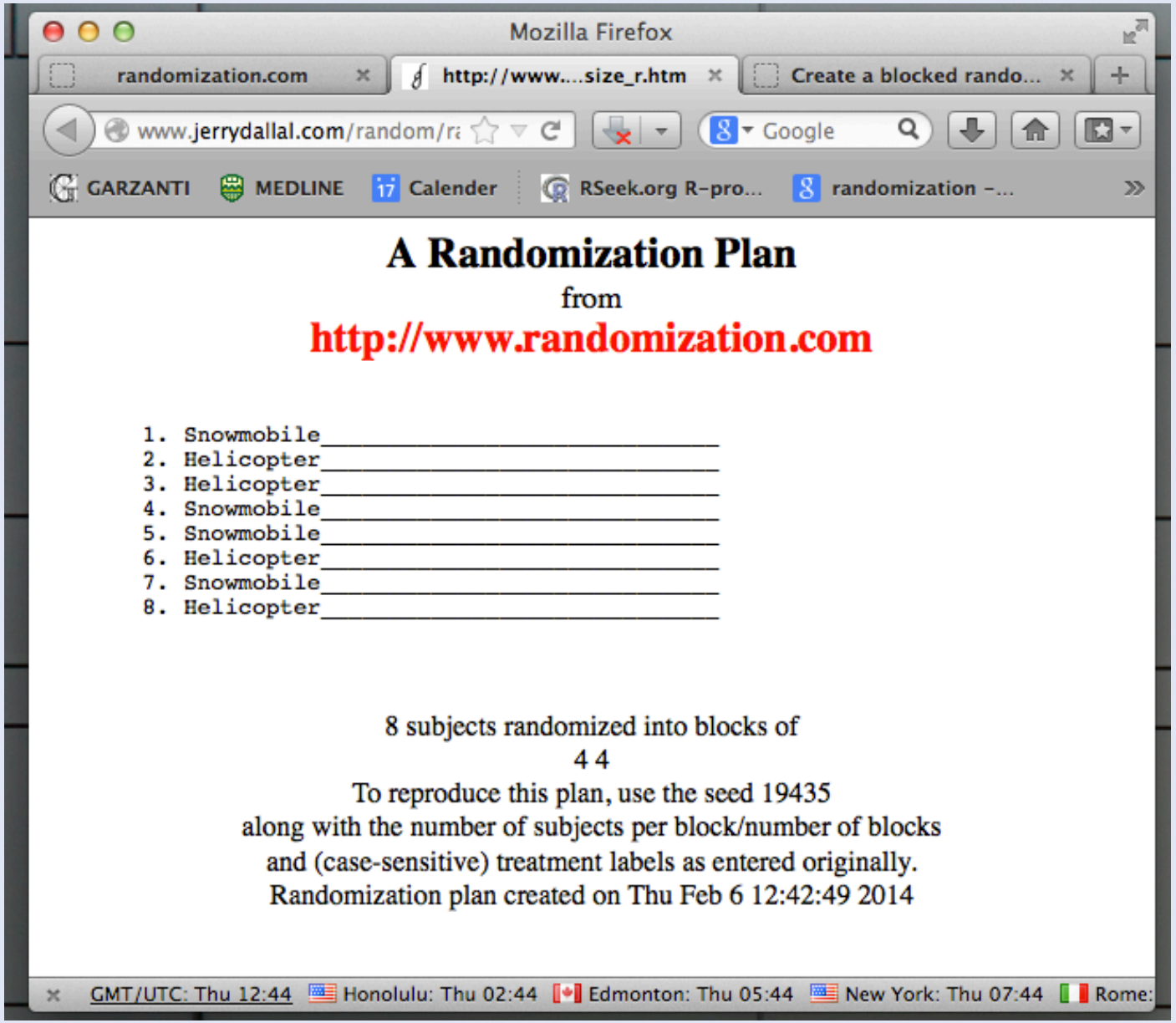
Initial subject ID number 1

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



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

Create a blocked randomisation list | Sealed Envelope | Randomisation (randomization) and database services for clinical trials

randomization.com | Create a blocked randomisation...

https://www.sealedenvelope.com/simple-randomiser/v1/lists

GARZANTI | MEDLINE | Calender | RSeek.org R-pro... | randomization -... | Edmonton - W... | Walmart PhotoC...

HOME | RANDOMISATION | RED PILL | TRIALS | PRICING | POWER CALCULATORS | HELP | CONTACT

SealedEnvelope

★ THE ORIGINAL INTERNET AND TELEPHONE RANDOMISATION SINCE 2001

CREATE A RANDOMISATION LIST

Use this tool to create a blocked randomisation list for your trial. The generated lists are suitable for use with our simple randomisation service

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?

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
```

GMT/UTC: Thu 12:26 | 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

	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			C
2	CASE	SCORE	TIME	TYPING TIME		SCORE	TIME	TYPING TIME		SCORE	TIME	TYPING TIME		SCORE	TIME	TYPING TIME		SCORE	TIME	TYPING TIME	S
3	1	4	144	not recorded		4	107	50		5	217	62		4	131	61		5	94	35	
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																					

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

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

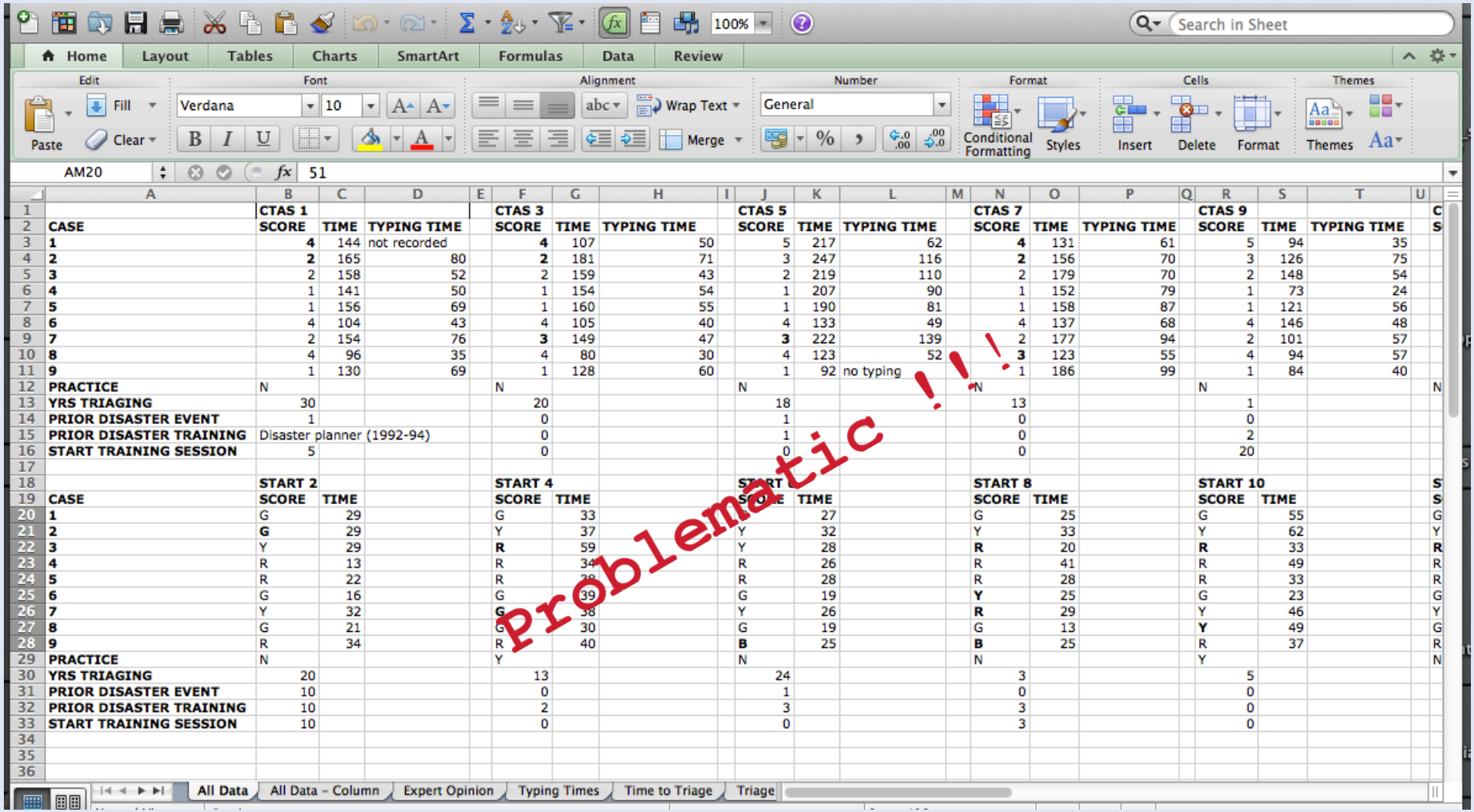
Which to Choose?

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



	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	
3	1	4	144	not recorded		4	107	50		5	217	62		4	131	61		5	94	35	
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			
19	CASE	SCORE	TIME			SCORE	TIME			SCORE	TIME			SCORE	TIME			SCORE	TIME		
20	1	G	29			G	33				27			G	25			G	55		
21	2	G	29			Y	37			Y	32			Y	33			Y	62		
22	3	Y	29			R	59			Y	28			R	20			R	33		
23	4	R	13			R	34			R	26			R	41			R	49		
24	5	R	22			R	[illegible]			R	28			R	28			R	33		
25	6	G	16			G	39			G	19			Y	25			G	23		
26	7	Y	32			G	38			Y	26			R	29			Y	46		
27	8	G	21			G	30			G	19			G	13			Y	49		
28	9	R	34			R	40			B	25			B	25			R	37		
29	PRACTICE	N				Y				N				N				Y			
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			

Problematic !!!

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

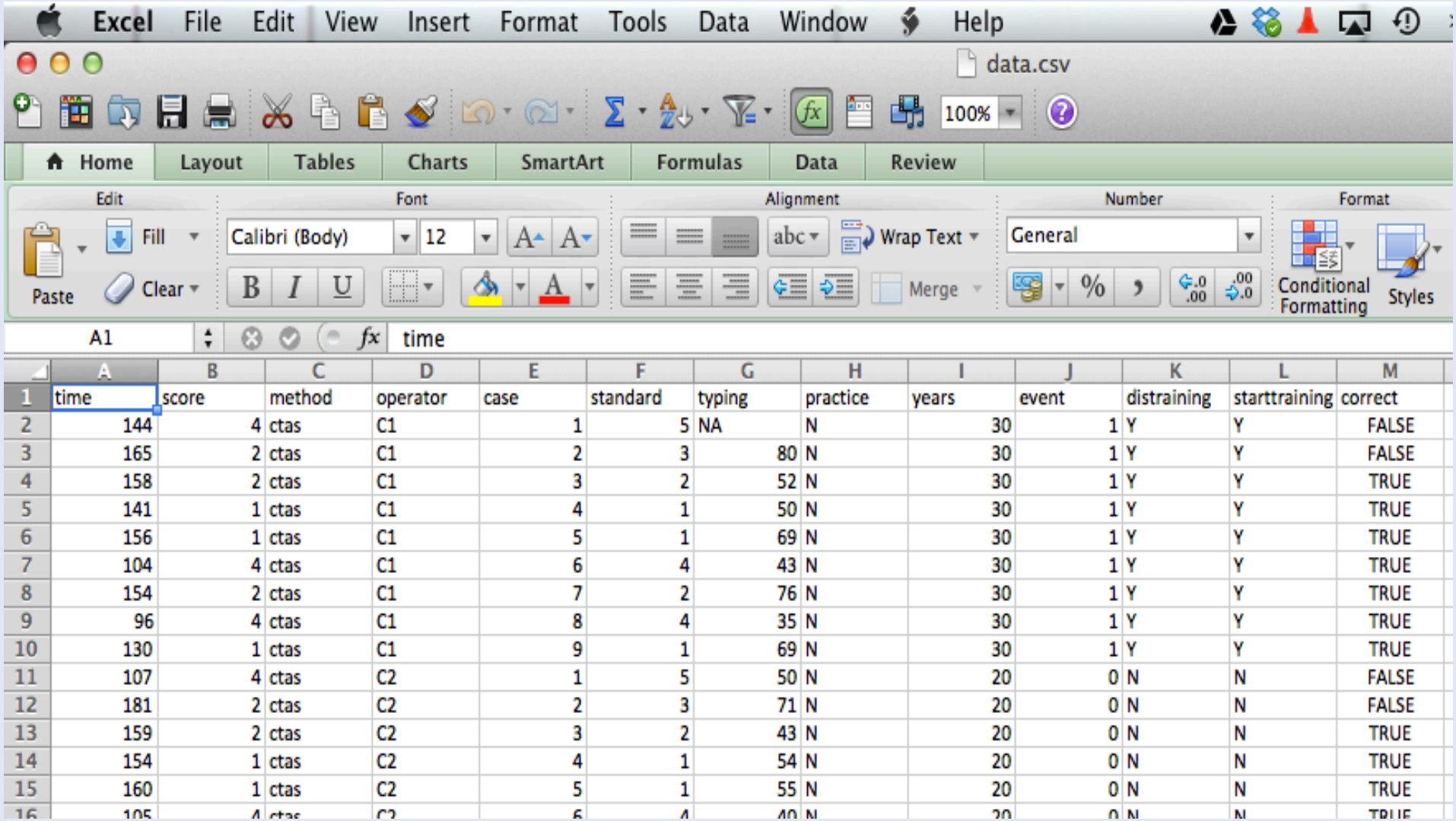
Constructing the Data Table With Excel

- 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

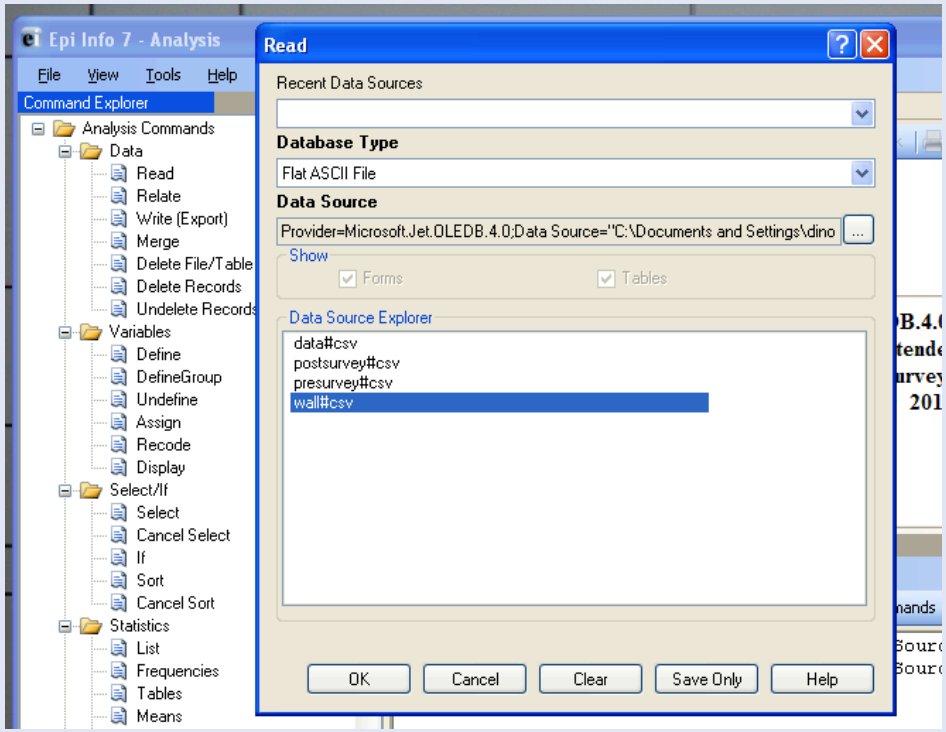


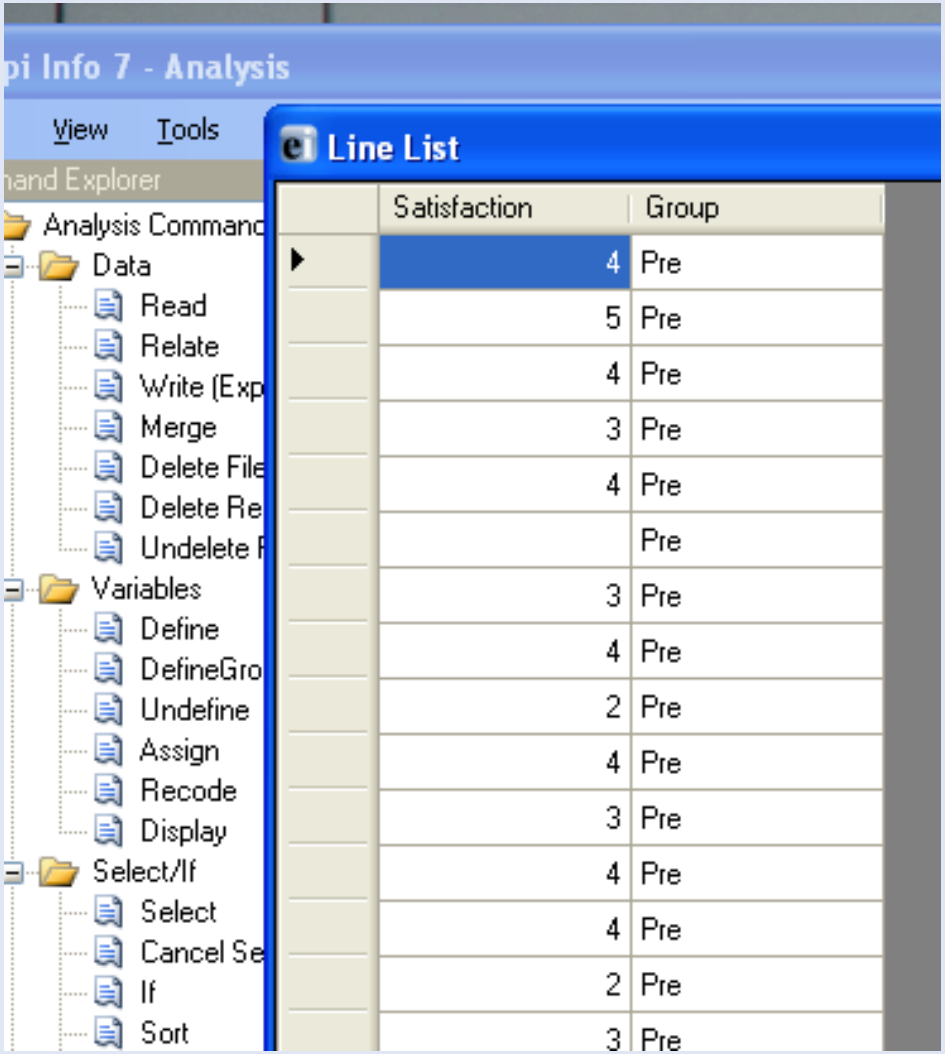
time	score	method	operator	case	standard	typing	practice	years	event	distraining	starttraining	correct
144	4	ctas	C1	1	5	NA	N	30	1	Y	Y	FALSE
165	2	ctas	C1	2	3	80	N	30	1	Y	Y	FALSE
158	2	ctas	C1	3	2	52	N	30	1	Y	Y	TRUE
141	1	ctas	C1	4	1	50	N	30	1	Y	Y	TRUE
156	1	ctas	C1	5	1	69	N	30	1	Y	Y	TRUE
104	4	ctas	C1	6	4	43	N	30	1	Y	Y	TRUE
154	2	ctas	C1	7	2	76	N	30	1	Y	Y	TRUE
96	4	ctas	C1	8	4	35	N	30	1	Y	Y	TRUE
130	1	ctas	C1	9	1	69	N	30	1	Y	Y	TRUE
107	4	ctas	C2	1	5	50	N	20	0	N	N	FALSE
181	2	ctas	C2	2	3	71	N	20	0	N	N	FALSE
159	2	ctas	C2	3	2	43	N	20	0	N	N	TRUE
154	1	ctas	C2	4	1	54	N	20	0	N	N	TRUE
160	1	ctas	C2	5	1	55	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 ElDoc)
```

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