# Introduction to Statistical Computing in Clinical Research

BIOSTAT 212, Summer 2016

Kristen Aiemjoy
Department of Epidemiology and Biostatistics
University of California San Francisco

# Outline today

- Course details
- Lecture 1
  - Why STATA?
  - Orientation to STATA
  - Immediate commands
- Mini-lecture: Sample size calculations in STATA
- Lecture 2
  - Opening data sets
  - Workflow: .do files and log files

# Introductions

- Who are you ?
- Teaching Assistants
  - Biftu Mengesha
  - Swaroop Mistry
  - Anjana Sharma
  - Yiwey Shieh
  - Qurratul Warsi

# Acknowledgements


- Mark Pletcher

# Course Objectives

- Introduce you to using STATA for:
  - Data management
  - Basic statistical and epidemiologic analysis
  - Turning raw data into presentable tables, figures and other research products
- Prepare you for fall courses
- Start analyzing your own data

# Course Objectives

- NOT statistical theory
  - Biostat200 in the fall is the first biostats class
  - I'm an epidemiologist, not a statistician

# Course details

- 1 Unit Course
- 7 Lectures
  - Tuesdays 1:15 – 2:45
- 6 Labs
  - Tuesdays 3:00 – 4:00 (except for today)
- 5 Assignments
- Additional online Lab Session/Office Hours
  - Wednesdays 10:00-11:00

# Course Details

- Course Website
  - https://courses.ucsf.edu/course/view.php?id=3039

- Recorded Lectures
  - Audio + video of lecturer + video of screen
  - Available same day for viewing
  - Links posted on website syllabus

# Course details

- Online forum
  - Post all questions here! (no email please)
    - TA's will staff Forum, reply by end of day on weekdays
  - Before you post, see if it's already there and answered
  - Look at previous year's forums for useful information
  - Settings for email notifications
    - https://learningtech.library.ucsf.edu/-/?s=forum+email+notifications

# Course details

- Course Requirements
  - Hand in all 5 Assignments (even if late)
  - Satisfactory Final Project

# Course details

- Grading (not relevant for all students)
  - Letter grades: Standard cutoffs
    - 90-100% A
    - 80-89% B
    - 70-79% C
    - 60-69% D
    - <60% or Course Requirements not met: F
  - Satisfactory/Unsatisfactory
    - >80% Satisfactory
  - TICR Grading policy
    - http://www.epibiostat.ucsf.edu/courses/masters/grading_policy.html

# Lecture Topics

1. Introduction, orientation and immediate commands
2. .do files, .log files, and workflow
3. Importing data, exploring data and generating variables
4. Cleaning data
5. Statistical analysis
6. Epidemiologic analysis
7. Tables and figures

# Mini-lectures (given by TAs)

1. Sample size calculations
2. Reshaping data
3. Dealing with dates
4. Formulas in excel
5. Loops
6. Add-on packages

# Lab Sessions

- Most important element of the course!
- Time to do your assignment and ask for help

# Assignments

- **5** Assignments
- Assignments are due 1 week after they are asigned prior to lecture.
  - Please upload to the course website. **By 1:00 PM**
- Assignments turned in late (less than 1 week) will receive only half credit; after that, no points will be awarded.
- ALL labs must be turned in to pass the class (even if no points are awarded).

# Final Project

- Create a table and a figure using your own data, document analysis using Stata.
- Due September 13 at 9 PM
- 30 points docked for each 1 day late.
- See 1-page description and example in syllabus
- Start looking for data!

# Lecture 1: Introduction to STATA

Introduction to Statistical Computing in Clinical Research
BIOSTAT 212, Summer 2016

Kristen Aiemjoy
Department of Epidemiology and Biostatistics
University of California San Francisco

# Many Software Packages

- SAS
- R
- SPSS
- Epi-Info
- STATA
- ….

# Why use STATA?

- Quick start, user friendly, but full-featured and powerful
- Immediate results
- Command line and Menu-driven option
- Good graphics
- Log and .do files
- Good manuals, help menu, customer service
- UCSF

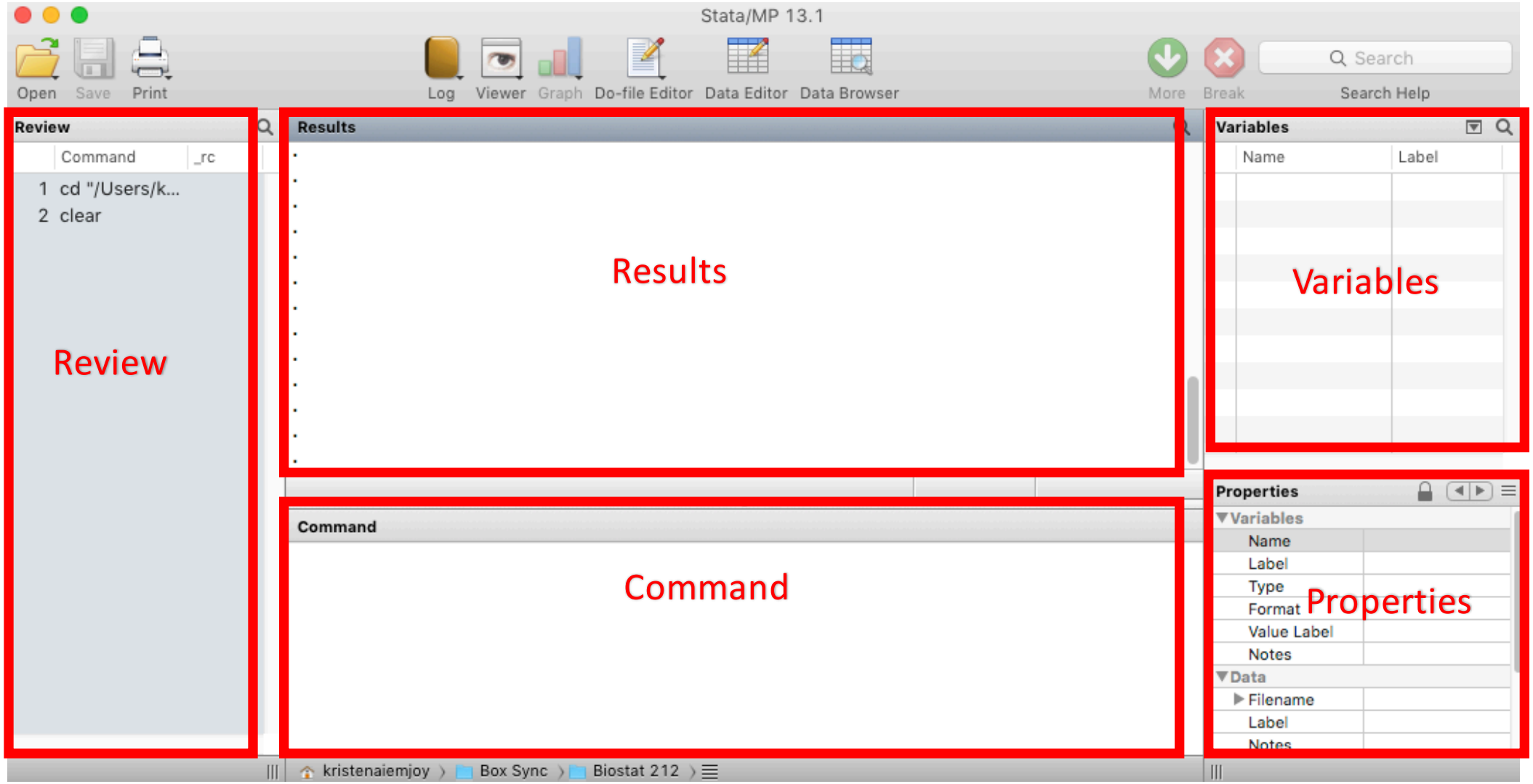
# Why not STATA?

- $$$
- R and Epi-info are free
- Programming easier in R
- Graphics can be better in R
- In STATA, you can only have one data frame open at a time

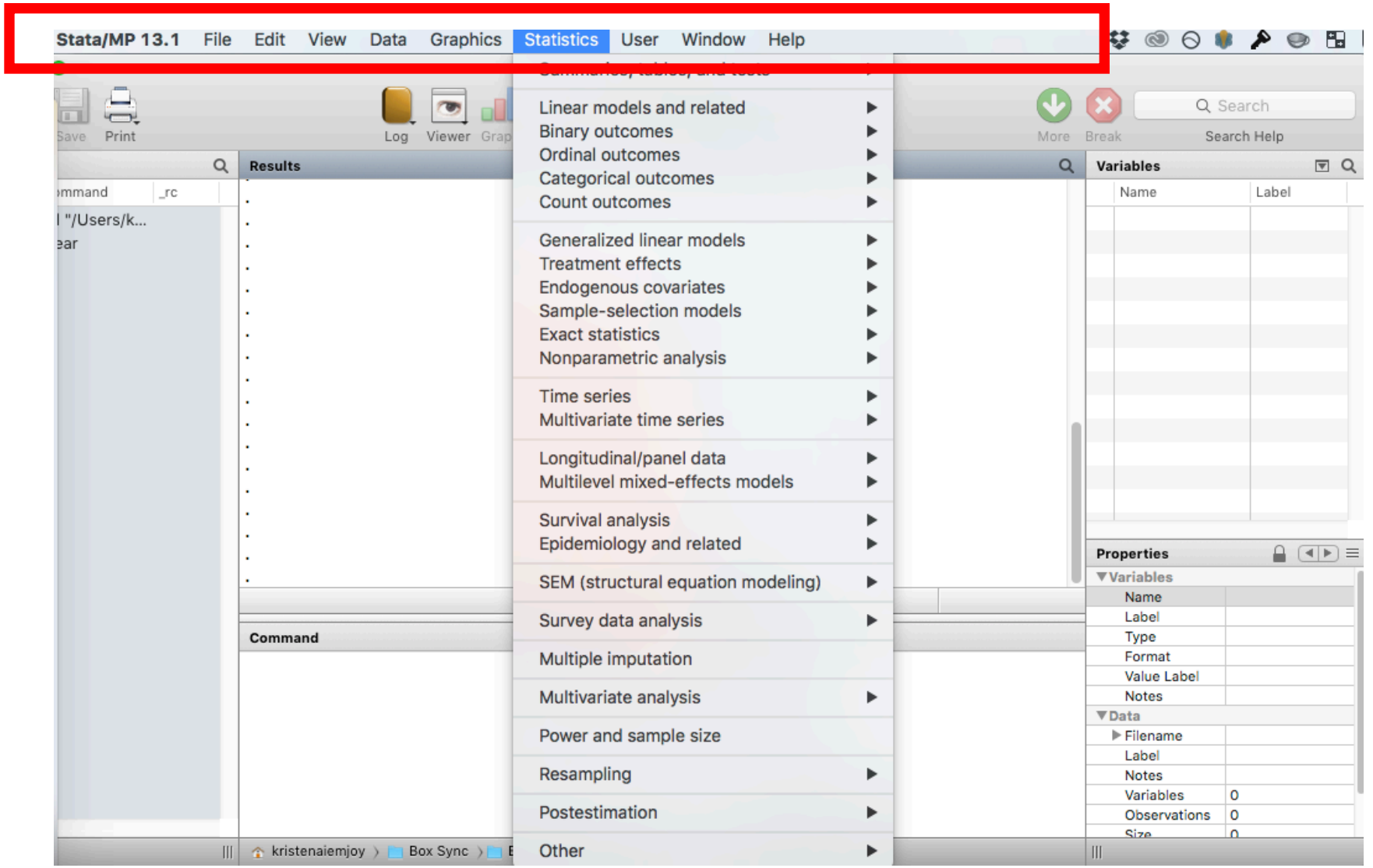
# STATA vs. other software packages

| Features | Stata | SPSS | SAS | R |
|---|---|---|---|---|
| **Learning curve** | Moderate | Gradual/flat | Pretty steep | Pretty steep |
| **User interface** | Programming/point-and-click | Mostly point-and-click | Programming | Programming |
| **Data manipulation** | Very strong | Moderate | Very strong | Very strong |
| **Data analysis** | Powerful | Powerful | Powerful/versatile | Powerful/versatile |
| **Graphics** | Very good | Very good | Good | Excellent |
| **Cost** | Affordable (perpetual licenses, renew only when upgrade) | Expensive (but not need to renew until upgrade, long term licenses) | Expensive (yearly renewal) | Open source |

# Windows


Stata/MP 13.1
Open
Save
Print
Log
Viewer
Graph
Do-file Editor
Data Editor
Data Browser
More
Break
Search
Search Help
Review
Command
_rc
1 cd "/Users/k...
2 clear
Review
Results
Results
Variables
Name
Label
Variables
Properties
Variables
Name
Label
Type
Format
Value Label
Notes
Data
Filename
Label
Notes
Properties
Command
Command
kristenaiemjoy
Box Sync
Biostat 212

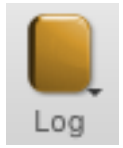
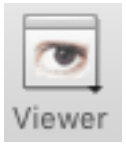
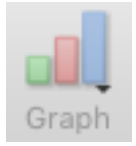
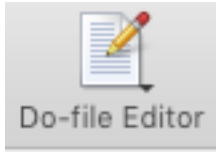
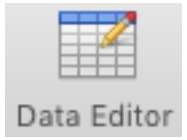
# Drop-down Menu / Taskbar



# Menu versus Command Line

- Menu advantages
  - Browse for commands you don't know already
  - See the options for each command in dialog boxes
  - Good way to learn syntax for complex commands
- Command line advantages
  - MUCH faster
  - ONLY way to write "do" files
    - Document and repeat analyses

# Buttons

| | | |
|---|---|---|
|  | **Log Begin/Close/Suspend/Resume** | Starts or closes a file to record results |
|  | **Viewer** | Opens a help window |
|  | **Graph** | Brings Graph Window to Front |
|  | **Do-file Editor** | Opens a new .do (command) file |
|  | **Data Editor** | Opens data window and allows edits |
|  | **Data Browser** | Opens data window to browse but not edit |

# Break and More

When there are more results than will fit in your result window use:

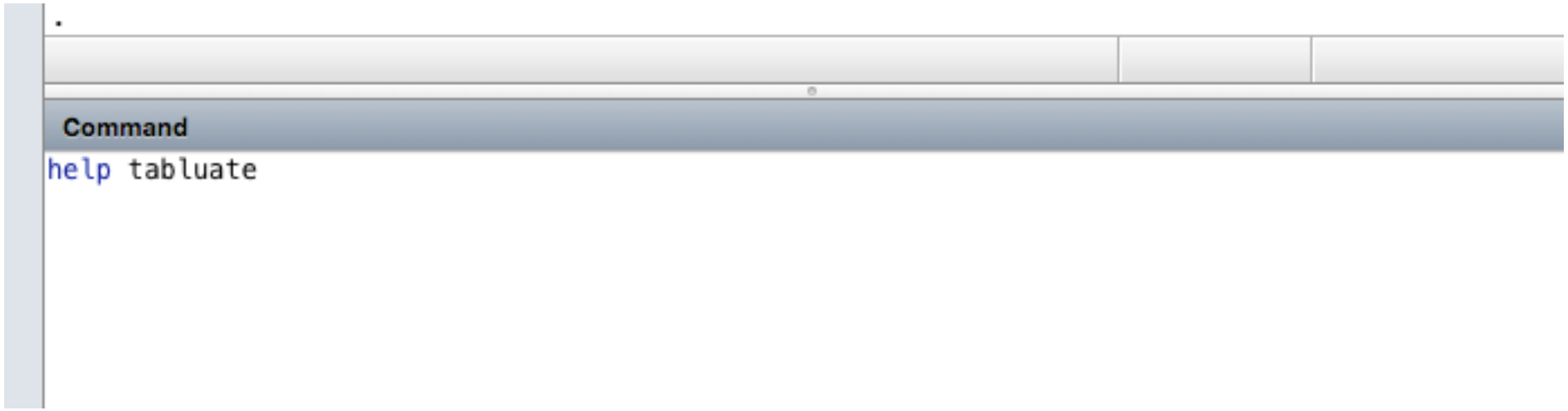

Or the


To see more

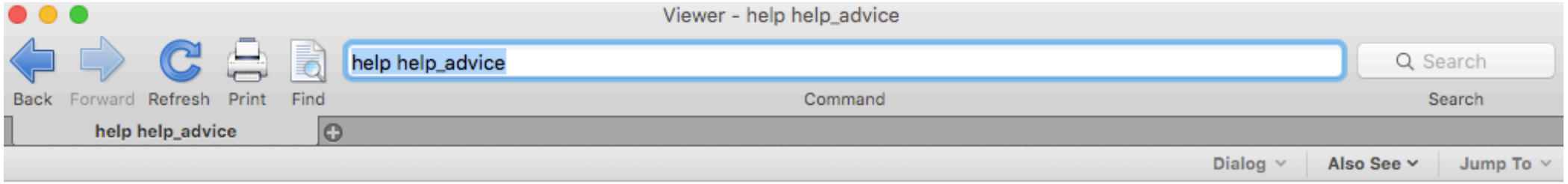

To stop

*note: you can also use the command `set more on`
`set more off`

# Help!



# Help!




**Advice on finding help**

Let's say you are trying to find out how to do something. With over 2,000 help files and 11,000 pages of PDF documentation, we have probably explained how to do whatever you want. The documentation is filled with worked examples that you can run on supplied datasets. Whatever your question, try the following first:

1. Select **Help** from the Stata menu and click on **Search....**

2. Type some keywords about the topic that interests you, say, "logistic regression".

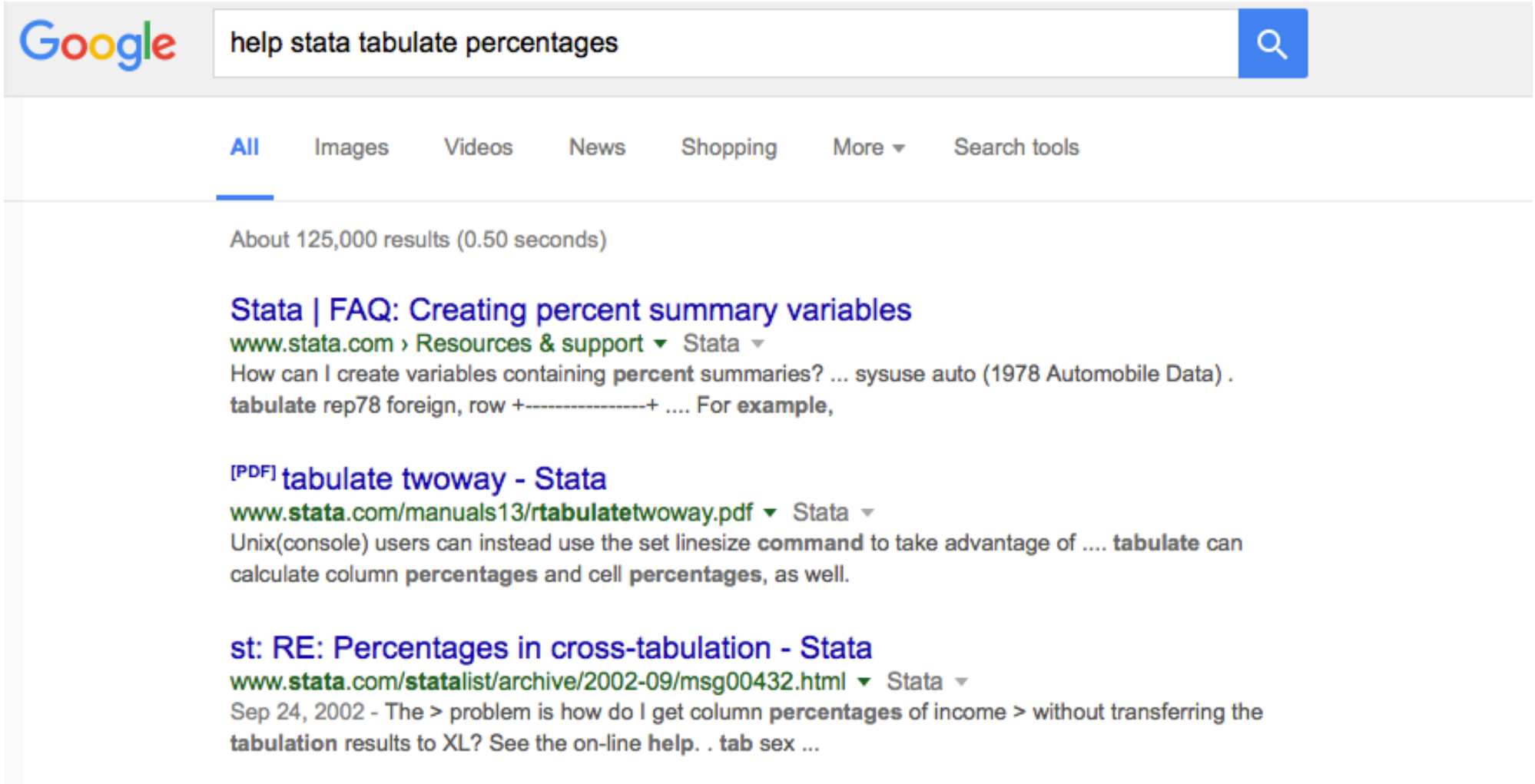
3. Look through the resulting list of available resources, including help files, FAQs, Stata Journal articles, and other resources.

4. Select the resource whose description looks most helpful. Usually, this description will be a help file and will include "(help <xyz>)", or, as in our example, perhaps "(help logistic)". Click on the blue link "logistic".

5. Let's assume you have selected the "(help logistic)" entry. You are probably not interested in the syntax at the top of the file, but you would like to see some examples. Select **Jump To** from the Viewer menu (in the top right corner of the help file now on your screen), and click on **Examples**. You will be taken to example commands that you can run on example datasets. Simply cut and paste those commands into Stata to see the results.

6. If you are new to the **logistic** command and want both an overview and worked examples with discussion, from the **Also See** menu, click on [R] logistic with the PDF icon. Or at the top of the help file, click on the blue title of the

# Help!


Google
help stata tabulate percentages
All
Images
Videos
News
Shopping
More
Search tools
About 125,000 results (0.50 seconds)
Stata | FAQ: Creating percent summary variables
www.stata.com › Resources & support ▾ Stata ▾
How can I create variables containing percent summaries? ... sysuse auto (1978 Automobile Data) . tabulate rep78 foreign, row +----------+ .... For example,
[PDF] tabulate twoway - Stata
www.stata.com/manuals13/rtabulatetwoway.pdf ▾ Stata ▾
Unix(console) users can instead use the set linesize command to take advantage of .... tabulate can calculate column percentages and cell percentages, as well.
st: RE: Percentages in cross-tabulation - Stata
www.stata.com/statalist/archive/2002-09/msg00432.html ▾ Stata ▾
Sep 24, 2002 - The > problem is how do I get column percentages of income > without transferring the tabulation results to XL? See the on-line help. . tab sex ...

# STATA syntax

1. I will use `Courier font` for commands
2. Words that are NOT meant to be typed in but instead are descriptive placeholders will be `blue`.
3. Parts of the command that are optional will be placed in `[brackets]`.
4. Shortcuts for commands are indicated by underlining
   - `summarize`

# STATA syntax

| | |
|---|---|
| `varlist` | 1 or more variables |
| `varname` | One variable |
| `var1, var2` | Exactly 2 variables |
| `#` | A number (in immediate commands)<br>*aka* `numlist` *in stata help files* |

# STATA Syntax

```
tabulate var1 var2, [row]
```

**tab male smoke, row**

| male | smoke 0 | 1 | Total |
|---|---|---|---|
| 0 | 2,723 | 1,466 | 4,189 |
| | 65.00 | 35.00 | 100.00 |
| 1 | 3,539 | 2,644 | 6,183 |
| | 57.24 | 42.76 | 100.00 |
| Total | 6,262 | 4,110 | 10,372 |
| | 60.37 | 39.63 | 100.00 |

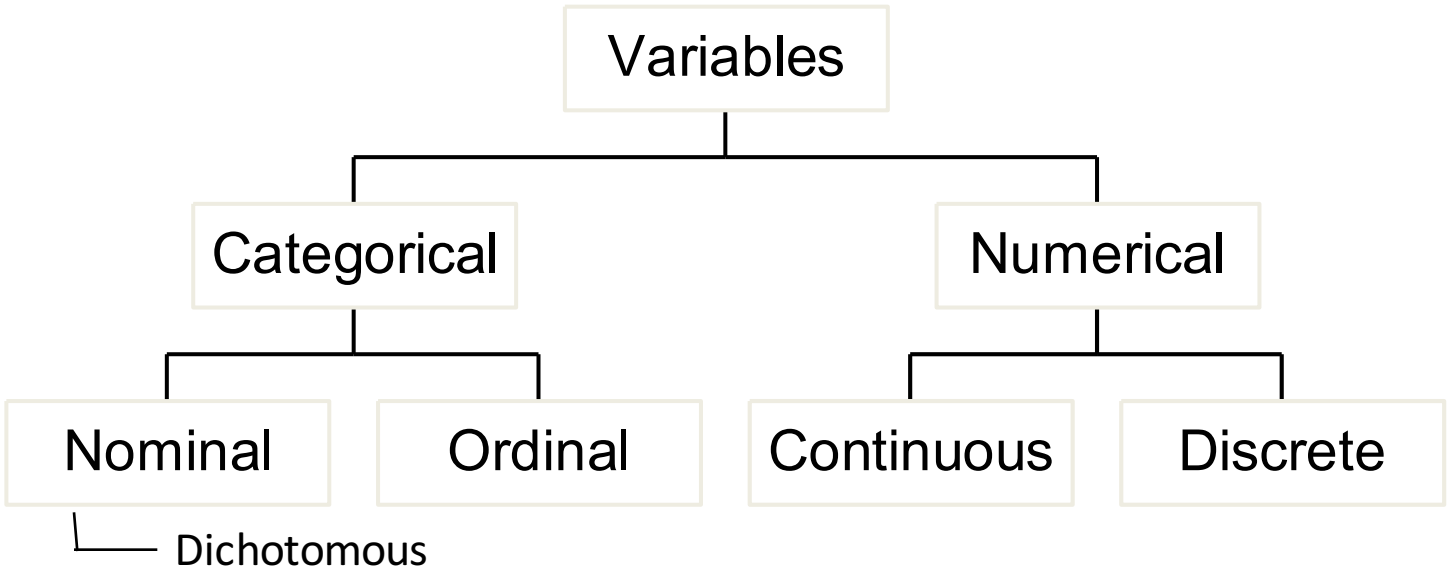
# STATA Syntax

- STATA is case sensitive
  - Commands are in lower case
  - If your variables have upper case letters, you need to remember this when typing them into the command line
    - Often easier to use all lower case

# Types of Files

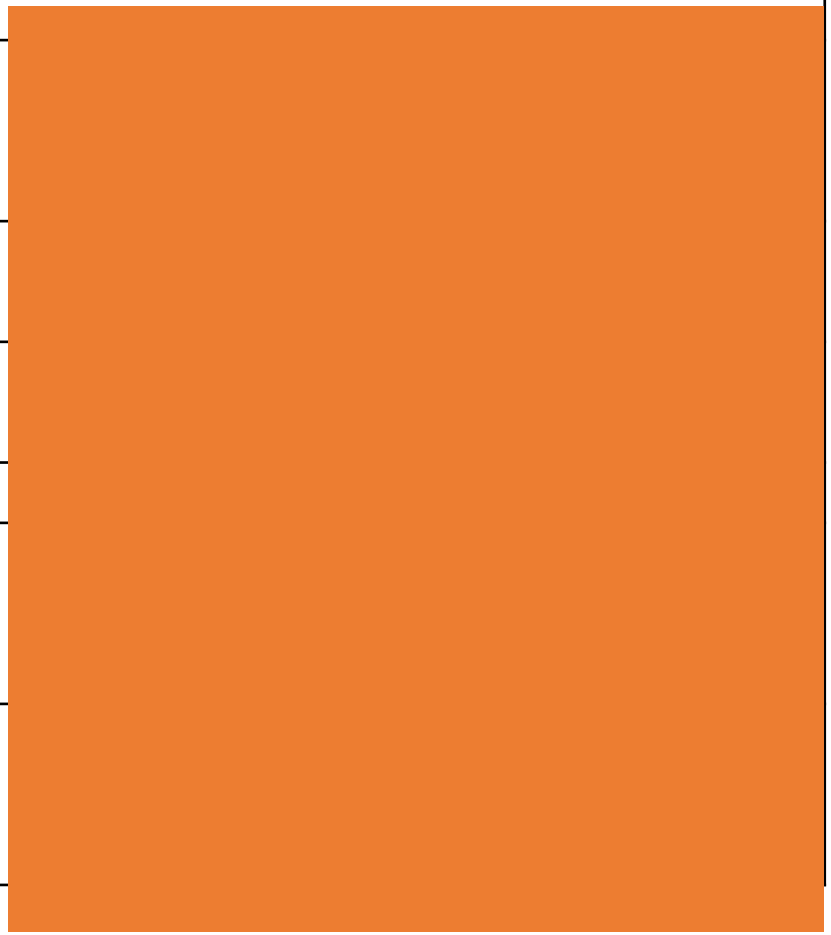
| | |
|---|---|
| .dta | data files |
| .do | command files |
| .ado | programs |
| .hlp | help files |
| .gph | graphs |
| .dct | dictionary files |

# **Preview: types of Variables



# Types of Variables

| | |
|---|---|
| | Categorical |
| 1. Nominal | No order |
| → 1a. Dichotomous | No order, only 2 categories |
| 2. Ordinal | Ordered |
| | Numerical |
| 1. Discrete | Can be counted |
| 2. Continuous | Can take any value within a given range |



# A quick note on binary

- 0 No
- 1 Yes

# STATA Variable Storage Types

| | | |
|---|---|---|
| str1 | string 1 character long | words |
| strx | string x characters long | |
| strL | string up to 2 billion characters long | |
| byte | dichotomous | true/false |
| int | integer values between -32,767 and 32,740 | numbers |
| long | integer values between -2,147,483,647 and 2,147,483,620 | |
| float | real numbers (i.e. numbers with decimal places) with 8 digits of accuracy | |
| double | real numbers (i.e. numbers with decimal places) with 16 digits of accuracy | |
| Missing | Missing data | no data |

# Variable storage types

| obstime | anycac | caccat | agecat | male | age_ge60 | sex | haircolor_~e |
|---|---|---|---|---|---|---|---|
| 1125 | 0 | 0 | <30 | 1 | 0 | male | blond |
| 1125 | 0 | 0 | <30 | 0 | 0 | female | black |
| 1064 | 0 | 0 | <30 | 0 | 0 | female | black |
| 1064 | 0 | 0 | <30 | 0 | 0 | female | black |
| 913 | 0 | 0 | <30 | 1 | 0 | male | black |
| 852 | 0 | 0 | <30 | 0 | 0 | female | blond |
| 699 | 1 | 1-99 | <30 | 1 | 0 | male | black |
| 1674 | 0 | 0 | <30 | 1 | 0 | male | blond |
| 1460 | 0 | 0 | <30 | 0 | 0 | female | black |
| 1460 | 0 | 0 | <30 | 0 | 0 | female | red |
| 1368 | 0 | 0 | <30 | 1 | 0 | male | brown |
| 1156 | 0 | 0 | <30 | 1 | 0 | male | blond |

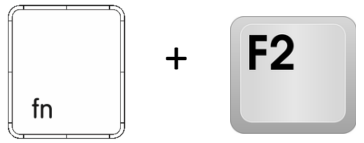
**Blue** Categorical with label
**Black** Numeric or categorical with no label
**Red** String

# Demo 1

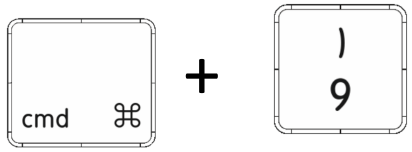
# Useful keyboard shortcuts

Mac

F2 Describe data

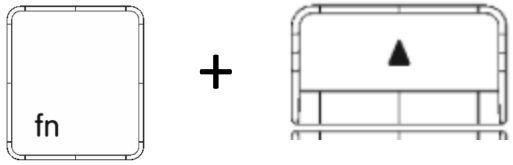



Ctrl + ( 9 Open a .do file




Ctrl + D Execute command from .do file (highlight text)

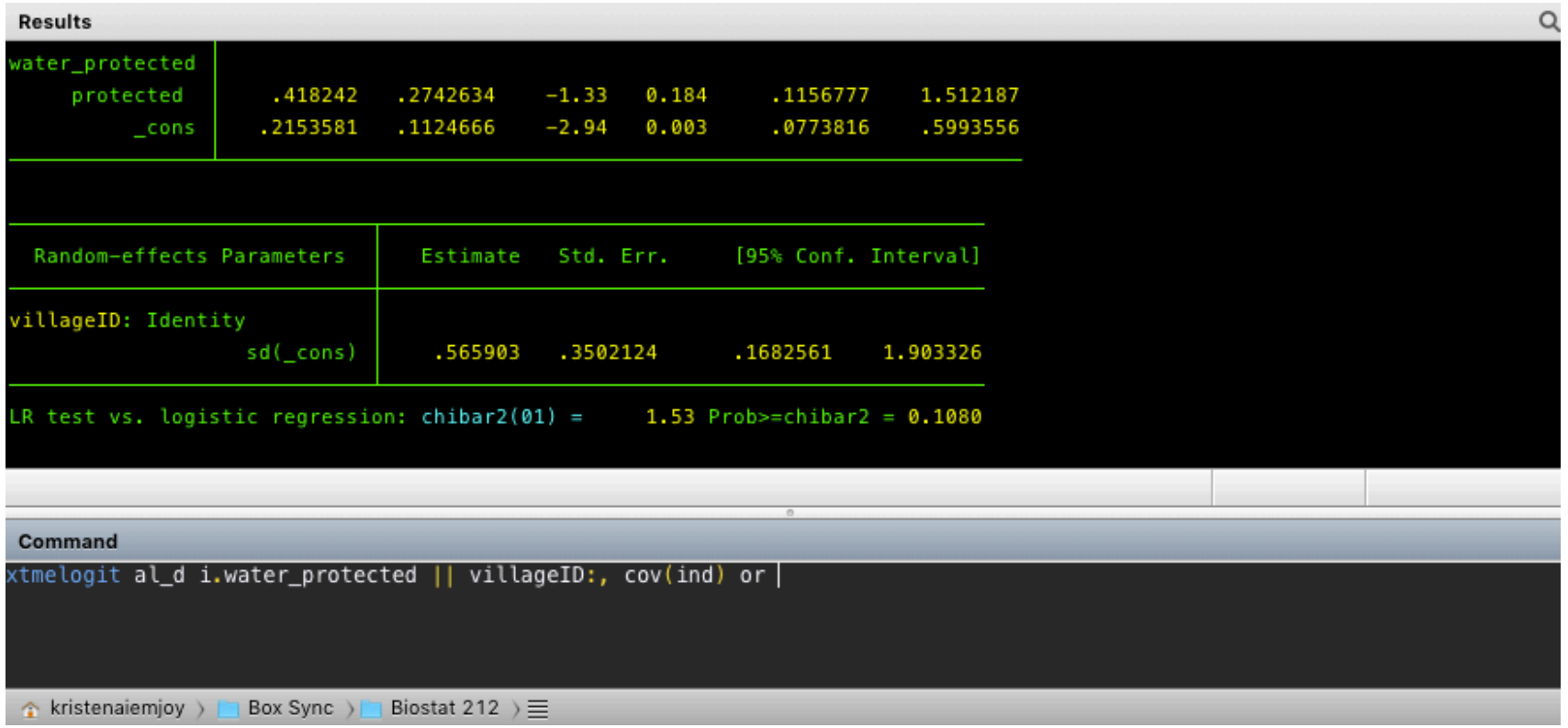

Page Up scroll through previous commands




Tab Autocomplete variable name after typing part of it

# Display preferences

- Drop down menu
  - STATA < General preferences < Syntax Highlighting < Color scheme
  - STATA < General preferences < Windows <Results <Color scheme



# Immediate commands

- An immediate command does not require stored data to operate

1. Calculator (`display` )
   - Random number generator
2. Sample size calculator (`power`) **[mini-lecture today]**
3. Statistical/epidemiological immediate commands
   - (end in the letter i)

# Immediate commands: Calculator

- Stata can work as a calculator using the `display` command

```
display 3+5
8

display 23*8
184

display 3+5
8
```

```
display (54.3-21)*83
2763.9

display 17^3
4913

display sqrt(156)
12.489996
```

# Mathematical operators in STATA

| English | Stata |
|---|---|
| Add | `+` |
| Subtract | `-` |
| Multiply | `*` |
| Divide | `/` |
| To the power of | `^` |
| Natural log of | `ln(`*`expression`*`)` |
| Base 10 log of | `log10(`*`expression`*`)` |
| Square Root | `sqrt(#)` |

# Immediate commands: random number generator

- To generate a random number between [a, b]
  - `display #a + (#b-#a)*runiform()`

- To generate a random integer between [a, b]
  - `display #a + int((#b-#a+1)*runiform())`

# QUIZ 1

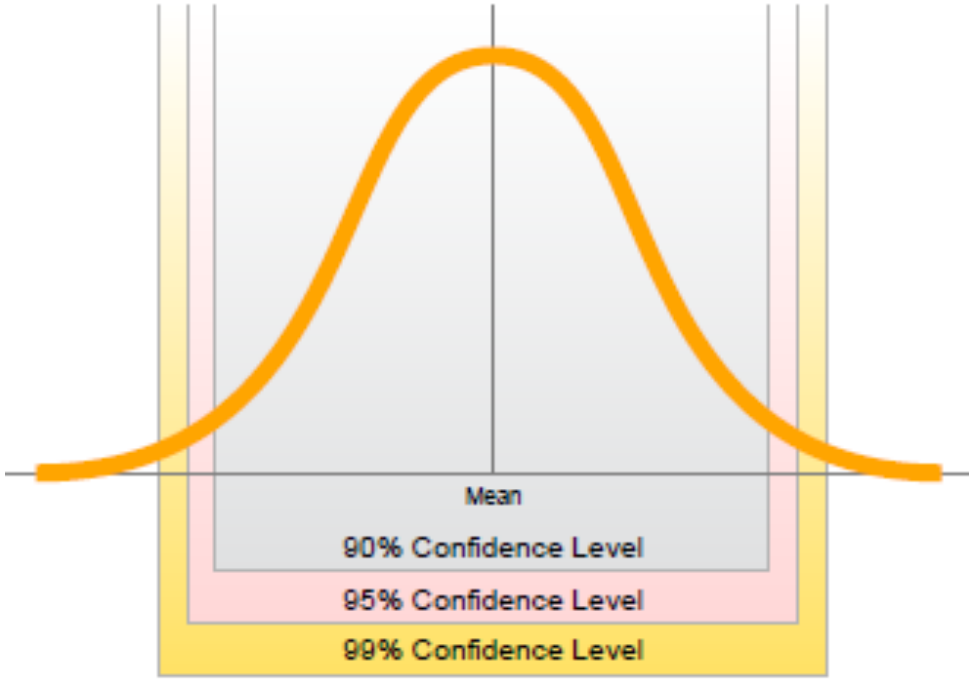
- Square Root of 457 _____
- (10,875*456)/(7^2) _____
- Natural log of 342 _____
- Random Number between 1 and 20 _____
  - `display a + int((b-a+1)*runiform())`

# Immediate commands: Statistics

- Examples:
  - `cii` confidence interval calculator
    - `cii #observations #mean #standard deviation`
  - `ttesti` T test calculator
    - 1 sample:
    - `ttesti #obs # reported mean #standard deviation #hypothesized mean`
    - 2 sample:
    - `ttesti #obs1 #mean1 #sd1 #obs2 #mean2 #sd2`

# **Preview: confidence intervals

- Over many repetitions of the study (with different samples) it is expected that 95% of the 95% confidence intervals will include the true (population) value.



# QUIZ 2


**Abstract**

**Background:** Despite high morbidity and mortality, the laboratory diagnosis of gastrointestinal infections is largely neglected in tropical African settings. This study aims to apply the Luminex multiplex PCR assay for the diagnosis of gastrointestinal pathogens in rural Ghana to evaluate its usefulness as a routine method.

**Methods:** A case–control study was conducted at the Agogo Presbyterian Hospital in Ghana. Stool samples were collected from children below 6 years of age with (cases) and without (controls) diarrhoea. Samples were screened for 15 different diarrhoeal pathogens by the Luminex xTAG GPP assay and associations between diarrhoea and gastrointestinal infections and fractions attributable to diarrhea (AF) were determined.

**Results:** The Luminex PCR assay identified organisms in 96.6 % ($n = 428$) of 443 cases and in 92.5 % ($n = 221$) of 239 selected controls. A mean of 2.5 (standard deviation [SD]: ±1.3) and 2.3 (SD: ±1.3) organisms per sample were detected in cases and controls respectively. An association with diarrhoea was found for rotavirus (adjusted odds ratio [aOR] = 7.2; 95 % confidence interval [CI]: 2.9–18.1), norovirus (aOR = 2.7; 95 % CI: 1.4–5.3) and *Shigella* spp. (aOR = 1.7; 95 % CI: 1.2–2.4) with respective AFs of 12.5 % (95 % CI: 9.6–15.3), 7.9 % (95 % CI: 3.8–11.7) and 16.9 % (95 % CI: 6.9–25.9).

**Conclusion:** The high proportion of pathogen-positive stool samples with a high number of co-infections in cases and controls suggests a substantial amount of transient or colonizing microorganisms for which treatment is not necessarily implicated. The use of sequential diagnostic algorithms with pathogen specific or quantitative PCRs might be most appropriate for diagnosing gastrointestinal infections.

**Keywords:** Ghana, West Africa, Multiplex Polymerase Chain Reaction, Gastrointestinal infections


1. Calculate the 95% confidence interval around the mean number of diarrhea pathogens per stool sample in <u>cases</u>

   _____

2. Calculate the 95% confidence interval around the mean number of diarrhea pathogens per stool sample in <u>controls</u>

   _____

```
cii #observations #mean #standard deviation
```


Eibach D, Krumkamp R, Hahn A, et al. Application of a multiplex PCR assay for the detection of gastrointestinal pathogens in a rural African setting. *BMC Infectious Diseases*. 2016;16:150. doi:10.1186/s12879-016-1481-7.

# Is the mean number of diarrhea pathogens different between cases and controls?

```
. ttesti 443 2.5 1.3 239 2.3 1.3

Two-sample t test with equal variances
```

| | Obs | Mean | Std. Err. | Std. Dev. | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| x | 443 | 2.5 | .0617649 | 1.3 | 2.378611 | 2.621389 |
| y | 239 | 2.3 | .08409 | 1.3 | 2.134344 | 2.465656 |
| combined | 682 | 2.429912 | .0498772 | 1.30255 | 2.33198 | 2.527844 |
| diff | | .2 | .1043361 | | -.0048597 | .4048597 |

```
    diff = mean(x) - mean(y)                                      t =   1.9169
Ho: diff = 0                     degrees of freedom =      680

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9722         Pr(|T| > |t|) = 0.0557          Pr(T > t) = 0.0278
```

# Immediate commands: Epidemiology

- Examples:
  - `iri` incidence rate calculator
    - `iri #exposed_cases #unexposed_cases #exposed_time #unexposed_time`
  - `cci` odds ratio calculator (**c**ase-**c**ontrol study)
    - `cci #exposed_cases #unexposed_cases #exposed_controls #unexposed_controls`
  - `csi` risk/ risk ratio calculator (**c**ohort **s**tudy)
    - `cci #exposed_cases #unexposed_cases #exposed_not_cases #unexposed_not_cases`

# **Preview: odds ratio

| Exposure \ Disease | Yes | No |
|---|---|---|
| Yes | a | b |
| No | c | d |

$$\mathbf{OR} = \frac{\frac{a}{b}}{\frac{c}{d}}$$

For case-control studies

# QUIZ 3

| | | Diarrhea: Cases | Diarrhea: Controls |
|---|---|---|---|
| Rotavirus | + | 64 | 5 |
| | - | 379 | 234 |

1. Calculate the odds ratio

_____

```
cci #cases_(+)rotavirus #cases_(-)rotavirus #controls_(+)rotavirus #controls_(-)rotavirus
```

Eibach D, Krumkamp R, Hahn A, et al. Application of a multiplex PCR assay for the detection of gastrointestinal pathogens in a rural African setting. *BMC Infectious Diseases*. 2016;16:150. doi:10.1186/s12879-016-1481-7.

```
. cci 64 379 5 234
```

| | Exposed | Unexposed | Total | Proportion Exposed | |
|---|---|---|---|---|---|
| Cases | 64 | 379 | 443 | 0.1445 | |
| Controls | 5 | 234 | 239 | 0.0209 | |
| Total | 69 | 613 | 682 | 0.1012 | |
| | Point estimate | | [95% Conf. | Interval] | |
| Odds ratio | 7.902902 | | 3.144012 | 25.4793 | (exact) |
| Attr. frac. ex. | .8734642 | | .681935 | .9607525 | (exact) |
| Attr. frac. pop | .126189 | | | | |

chi2(1) = 26.06 Pr>chi2 = 0.0000

**Table 2** Crude odds ratio (OR), age-adjusted OR (aOR) on associations between detected pathogens and diarrhoea, sorted by number of cases

| Pathogen[a] | Cases ($n = 443$) | Controls ($n = 239$) | Negative PCR-result | Crude OR | aOR |
|---|---|---|---|---|---|
| *Escherichia coli* LT/ST (ETEC) | 247 (55.8) | 144 (60.3) | 291 | 0.8 (0.6–1.2) | 0.9 (0.6–1.2) |
| *Giardia lamblia* | 228 (51.5) | 125 (52.3) | 329 | 0.9 (0.7–1.3) | 1.0 (0.7–1.3) |
| *Shigella* spp. | 185 (41.8) | 76 (32.6) | 421 | 1.5 (1.1–2.2) | 1.7 (1.2–2.4) |
| *Campylobacter* spp. | 153 (34.5) | 78 (32.6) | 451 | 1.1 (0.8–1.5) | 1.1 (0.8–1.6) |
| Rotavirus | 64 (14.5) | 5 (2.1) | 613 | 7.9 (3.1–25.5) | 7.2 (2.9–18.1) |
| *Escherichia coli* O157 | 59 (13.3) | 37 (15.5) | 586 | 0.7 (0.5–1.2) | 0.8 (0.5–1.2) |
| Norovirus | 55 (12.4) | 11 (4.5) | 616 | 2.9 (1.5–6.3) | 2.7 (1.4–5.3) |
| *Salmonella* spp. | 53 (12.0) | 37 (15.5) | 592 | 0.8 (0.5–1.3) | 0.8 (0.5–1.3) |
| *Cryptosporidium* spp. | 33 (7.5) | 12 (5.0) | 637 | 1.5 (0.7–3.3) | 1.4 (0.7–2.9) |
| Adenovirus (41/41) | 17 (3.8) | 14 (5.9) | 651 | 0.6 (0.3–1.4) | 0.6 (0.3–1.3) |
| *Escherichia coli* STEC | 9 (3.7) | 12 (2.7) | 661 | 0.7 (0.3–1.9) | 0.7 (0.3–1.8) |
| *Clostridium diffcile* toxin A,B | 2 (0.5) | 2 (0.8) | 678 | - | - |
| *Entamoaba histolytica* | 1 (0.2) | 2 (0.8) | 679 | - | - |

[a]*Vibrio cholerae* and *Yersinia enterocolitica* were not detected in any sample
*OR* odds ratio, *aOR* age adjusted odds ratio

Eibach D, Krumkamp R, Hahn A, et al. Application of a multiplex PCR assay for the detection of gastrointestinal pathogens in a rural African setting. *BMC Infectious Diseases*. 2016;16:150. doi:10.1186/s12879-016-1481-7.