

Biostat 200

Introduction to Biostatistics

“No amount of experimentation can ever prove me right; a single experiment can prove me wrong.”

Albert Einstein (1879-1955)

Course instructors

Course director

Isabel Elaine Allen, PhD

Professor of Biostatistics & Epidemiology

Phone: 508-667-7737

Office: 550 16th St., 2nd Floor

Isabel.allen@ucsf.edu

TAs

- Shokoufeh Dianat
- Ilya Golovaty
- Scott Lu
- Patrick Yuan
- Rae Wannier

- Lectures: Tuesdays 10:30-12:00
 - Videos of lectures are posted usually within a few hours
 - 11 lectures (Tuesdays: Sept 18-Dec 4, no lecture Nov. 20)
 - Attendance encouraged but not required, video is available
- Labs: Thursdays 10:30-12 (no lab Nov 22)
- Allen office hours:
 - Tuesday after class, during lab, via e-mail, & by appointment
- Course credits: 3
- Some STATA in class – bring your laptop to play along

Assignments

- 9 homework assignments
- Each assignment will be posted at least one week before it is due
- Assignments will be due weekly on Thursdays at 10:30 a.m.
- Answers will be posted within one week
- Assignments will consist of:
 - Data analysis and interpretation
 - Reading and interpretation of scientific publications

Assignments

- All assignments will be posted on the Thursday lab section of the CLE. Turn in all assignments to the CLE where it is assigned.
- Turn in your assignments as a word document or a pdf. Please do not forget to put your name on the document.
- Name the file with assignment # and your last name: e.g. "Assignment #_YourName.docx". Include your name within the Assignment as well
- No late assignments accepted BUT I agree that there can be circumstances where you can't get an assignment in but you can drop your lowest grade (or skip an assignment)

Grading

- Assignments (70%)
 - 9 assignments/problem sets
 - 8/9 assignments count towards your grade
- Answer keys will be posted soon after assignments are due.
- Graded assignments will be returned on the CLE 1-2 weeks after they are due.
- Grading is done by TAs who are students in the program and you will each have a specific TA assigned
- Final exam will be a take home exam (30% of grade)

Labs

- Labs will be every Thursday 10:30 -12
- No lab 11/22
- Lab content
 - Labs 1-2: Stata familiarity, exploratory data analysis, nuts & bolts
 - Labs 3-11: Tied to the lecture material
- Attendance encouraged but optional (not recorded)
- These lab exercises will not be turned in or graded but come with questions

Forum

- <https://courses.ucsf.edu/mod/forum/view.php?id=371198>
- Will be used for class communications, like assignment clarifications, etc.
- Please use for clarification questions that others in the class may also have

TICR Professional Conduct Statement

- I will maintain the highest standards of academic honesty.
- I am allowed to collaborate with my classmates on assignments, however I will work through each problem myself and turn in my own work (no cutting and pasting from others).
- I will neither give nor receive help from other students on the final examination.
- I will not use questions or answer keys from prior years.

Who am I?

- PhD in statistics, MS in math
- UCSF since 2012
- Pre-UCSF in Boston @ Harvard & Babson & starting Biotech companies
- Part of DEB, GBHI, MAC, ICHS, QB3/MBC, CTSI & Catalyst @ UCSF



Who are you?

- Stage in training
 - Resident/Fellow
 - Faculty
 - PhD Student
 - Other
- Outside of UCSF?
- Your goals?
- Are you designing a study?
- Do you have a data set that you are analyzing?
- Are you in the process of collecting your own data?

Course goals

- Knowledge of basic biostatistics terms and notation
- Understanding of concepts underlying most statistical analyses, as a foundation for more advanced methods
- Ability to summarize data and conduct basic statistical analyses using STATA
- Ability to present and visualize data
- Ability to understand basic statistical analyses in published journals

Why Do Statistics?

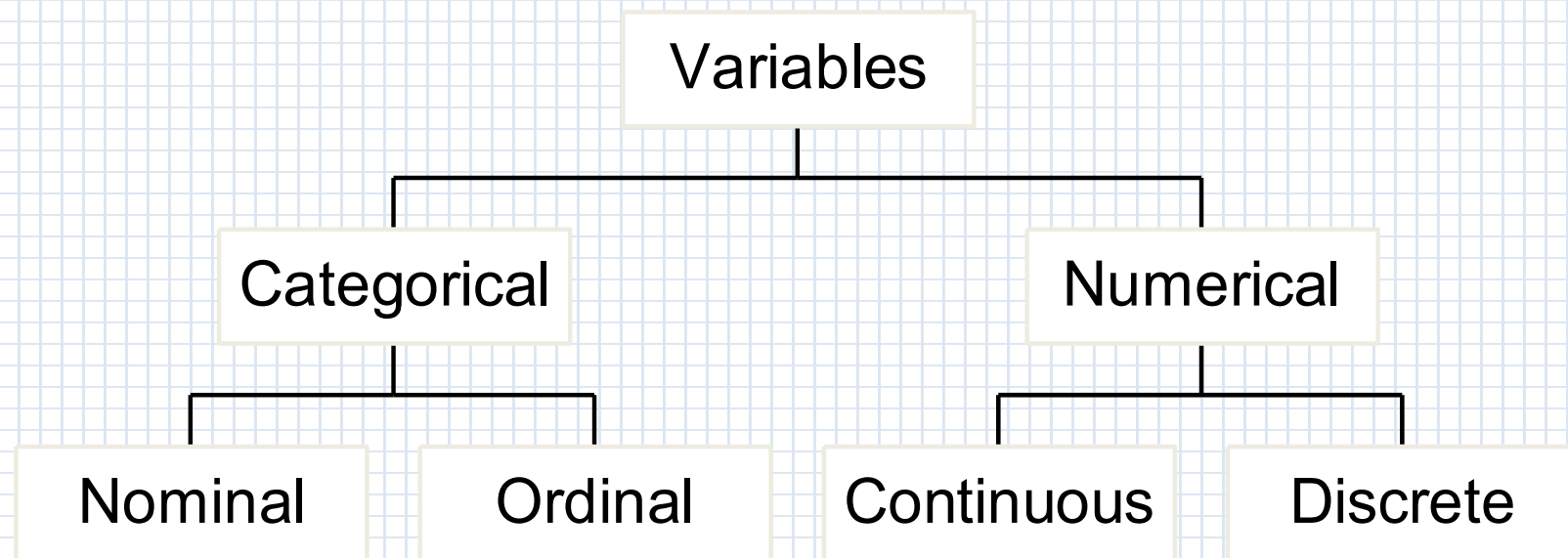
- Extrapolate from data collected to make general conclusions about a larger population
- Allows general conclusions to be made from limited amounts of data
- Often we must assume that data are randomly sampled from an infinitely large population, analyse this sample and use results to make inferences about the population
- Sometimes we don't have a random sample but use the data that are available

What we'll start to cover today

- Variables - numerical versus categorical
- Tables (frequencies and percents)
- Graphs (histograms, box plots, scatter plots, line graphs)
- Summaries of numerical variables
- What might be wrong with our data? How to clean our data.

Types of variables

- Variables are what you are measuring
- Data sets are made up of a set of measured variables



Types of variables

- Categorical variable: any variable that is not numerical (values have no numerical meaning)
- Examples: gender, race, drug, disease status

Types of variables

- Categorical variables
 - Nominal variables:
 - The data are unordered
 - For example: RACE: 1=Caucasian, 2=Asian American, 3=African American
 - A subset of these variables are binary or dichotomous variables
 - Binary variables have only two categories
 - For example:
 - » Biological sex: 1=male, 2=female
 - » 0=No 1=Yes

Types of variables

- Categorical variables
 - Nominal variables:
 - The data are unordered
 - Ordinal variables:
 - The data are ordered
 - For example:
 - AGE: 1=10-19 years, 2=20-29 years, 3=30-39 years
 - Likelihood of participating in a vaccine trial 1=Not at all likely 2=somewhat likely 3=very likely
 - Sometimes we will treat these as pseudo-continuous

Types of variables

- Numerical (quantitative) variables: naturally measured as numbers for which arithmetic operations are meaningful
- e.g. height, weight, age, salary, viral load, CD4 cell counts, hemoglobin A1c
 - Discrete variables: can be counted (e.g. number of goats owned by a household: 0, 1, 2, 3, etc.) but fractions do not make sense
 - Continuous variables: can take any value within a given range (e.g. weight: 2974.5 g, 3012.6 g)

Grey zone

- Dichotomous variables 0=No, 1=Yes
 - Doing arithmetic operations actually does make sense
 - If you take the mean of the 0's and 1's you get the proportion= yes
 - And even if you correlate dichotomous and continuous variables you get the point biserial coefficient so not impossible

Grey zone

- Ordinal or Discrete?
 - We may count the number of children you have (discrete)
 - We may rank families by number of children
 - As an ordinal variable you can get a correlation (Spearman) but you can't do regular arithmetic operations.

Grey zone

- Continuous variables are always truncated at the level of the precision of the measurement.
 - They may be truncated at integer values but if a fraction makes sense it is still a continuous variable
 - e.g. Age=33 years old (really 33 years, 17 days, 12 hours, 23 minutes, etc...)
 - Making continuous variables into categories

Why does it matter?

- Knowing what type of variable you are dealing with will help you choose your method of statistical analysis
- The most important/common distinction is between categorical and numerical

Manipulation of variables

- Continuous variables can be discretized (turned into categorical data)
 - e.g., age can be rounded to whole numbers
- Continuous or discrete variables can be categorized
 - e.g., age categories
- Categorical variables can be re-categorized
 - e.g., lumping from 5 categories down to 2

Manipulation of variables

- Why discretize/categorize a continuous variable or re-categorize a categorical variable?
 - Ease of interpretation
 - Ease of statistical methodology
 - Some groups are too small to make conclusions about
 - But discretizing/categorizing or lumping can have a statistical cost – loss of information
- We will do some of this in lab

Frequency tables

- Categorical variables are summarized by
 - Frequency counts – how many are in each category
 - Relative frequency or percent (a number from 0 to 100)
 - Proportion (a number from 0 to 1)

Gender of persons in Biostat 200 - 2016	
	n (%)
Male	
Female	
Trans or other	
Total	

```
. tab gender
```

gender – What is your gender?	Freq.	Percent	Cum.
Male	26	47.27	47.27
Female	29	52.73	100.00
Total	55	100.00	

Frequency tables

- Continuous variables can be summarized in frequency tables but must be categorized in meaningful ways beforehand

Frequency tables

- Choice of cutpoints for categories
 - Even intervals
 - e.g. 10-year age categories
 - Meaningful cutpoints related to a health outcome or decision
 - e.g. $CD4 < 50$ cells/mm³
 - Equal percentage of the data falling into each category
 - Tertiles – 33%
 - Quartiles – 25%
 - Quantiles – 20%

Frequency tables – categorizing a continuous variable

Age of class participants	
Category	n (%)

```
. tab age_cat
```

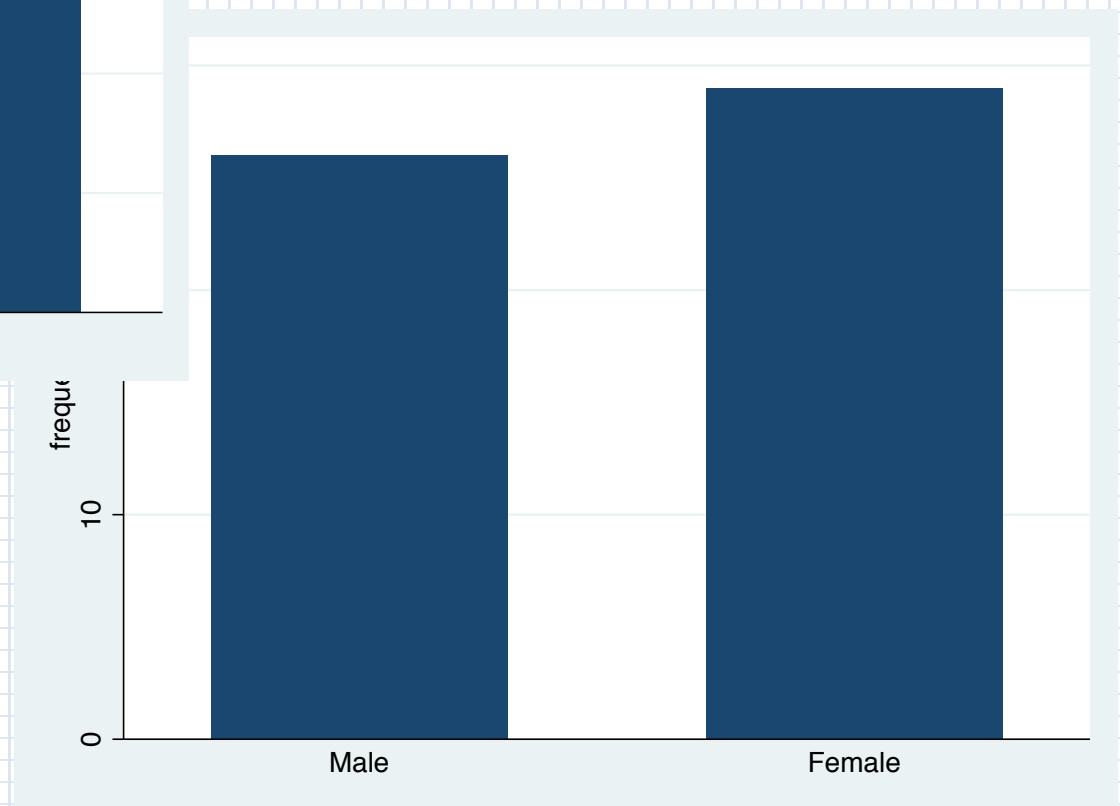
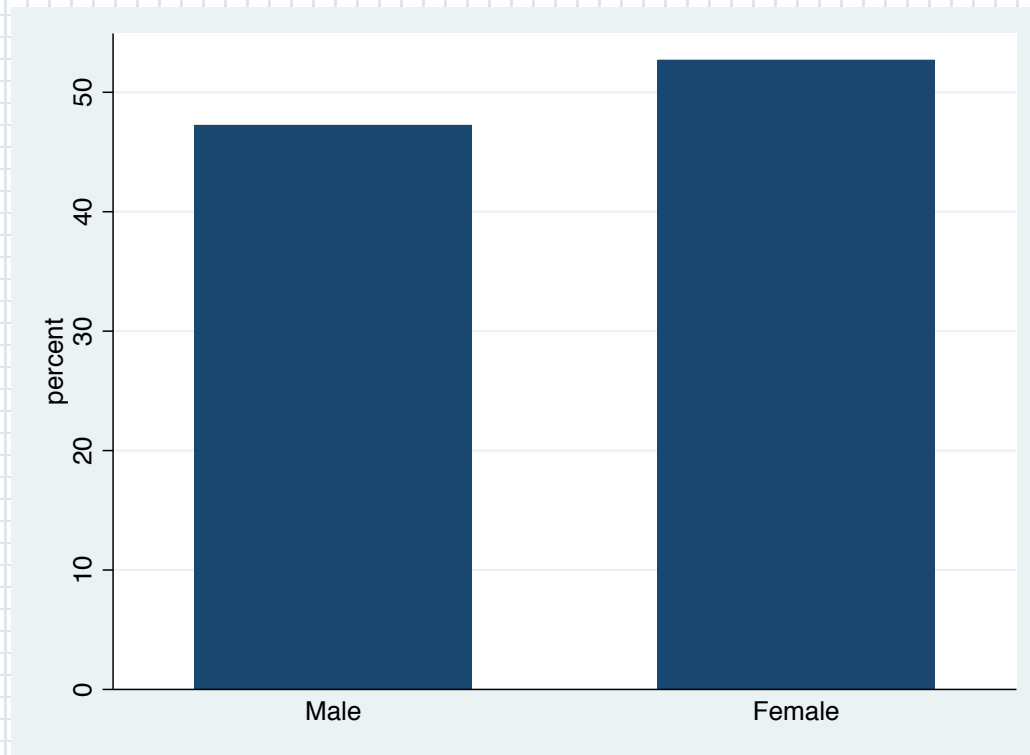
RECODE of age (age – What is your age?)	Freq.	Percent	Cum.
20–30	3	5.45	5.45
30–40	47	85.45	90.91
40–50	5	9.09	100.00
Total	55	100.00	

Do you think there is a better way to categorize this data?

Bar charts

- General graph for categorical variables
- Graphical equivalent of a frequency table
- The x-axis does not have to be numerical
- The height of the bars add up to 1 (if using percent options)

Bar charts, gender of class



graph bar (count), over(gender)

Summarizing Data: The mode is the most common value

- Used for nominal variables
- Can be found by making bar charts or frequency tables
- Mode of # of apple devices own by class members? Mode of age?

. tab apple_n

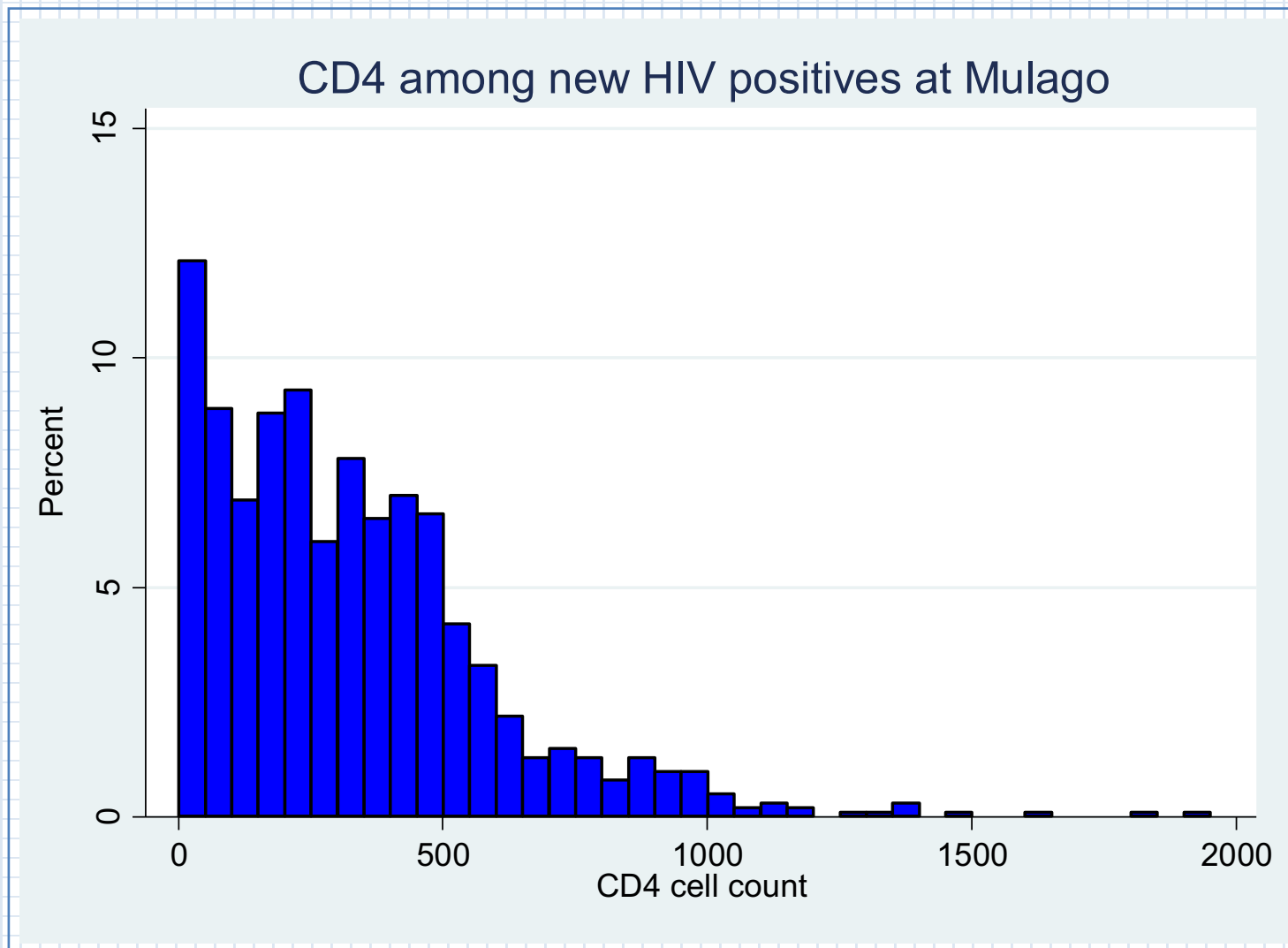
apple_n - How many currently functioning Apple devices do you own?	Freq.	Percent	Cum.
0	3	5.45	5.45
1	4	7.27	12.73
2	7	12.73	25.45
3	22	40.00	65.45
4	13	23.64	89.09
5	2	3.64	92.73
6	4	7.27	100.00
Total	55	100.00	

. tab age

age - What is your age?	Freq.	Percent	Cum.
23	1	1.82	1.82
24	1	1.82	3.64
27	1	1.82	5.45
30	6	10.91	16.36
31	7	12.73	29.09
32	5	9.09	38.18
33	6	10.91	49.09
34	9	16.36	65.45
35	8	14.55	80.00
36	3	5.45	85.45
37	3	5.45	90.91
40	2	3.64	94.55
43	1	1.82	96.36
44	2	3.64	100.00
Total	55	100.00	

Histograms

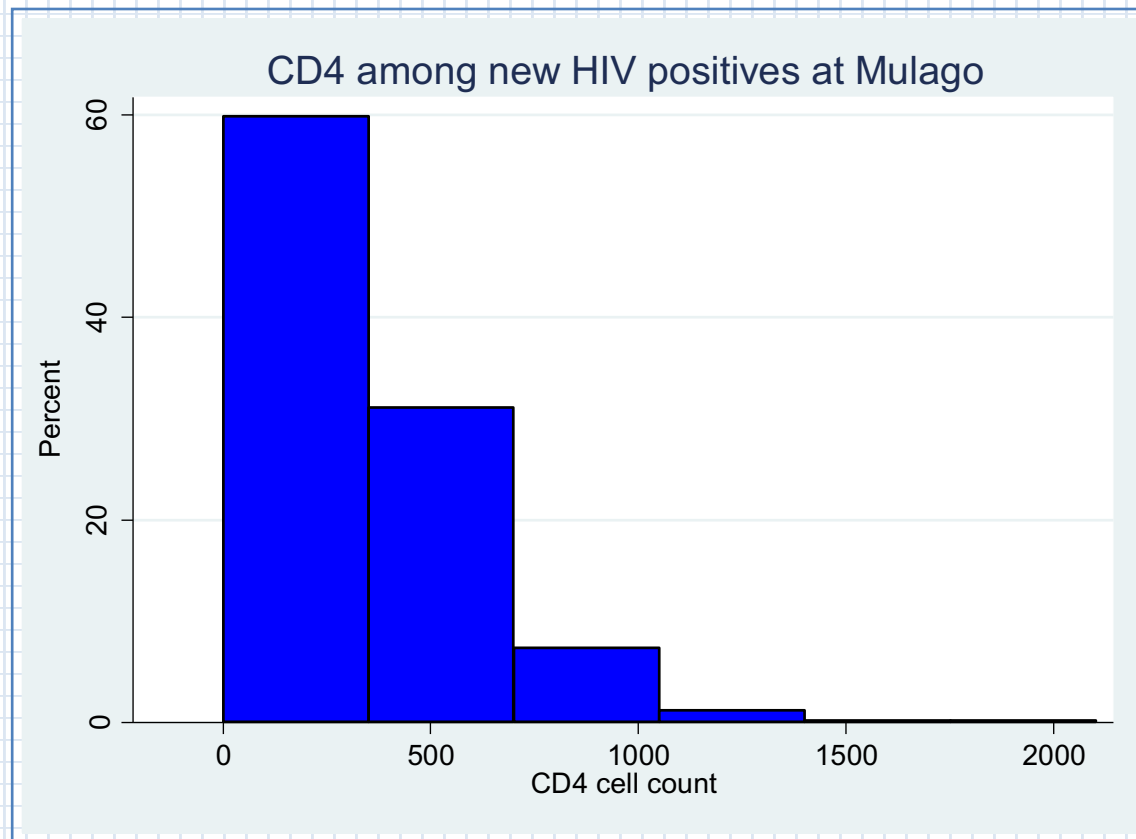
- Bar chart for numerical data
- The number of bins and the bin width will make a difference in the appearance of this plot
- Width and number of bins may affect interpretation
- Options such as percent, frequency will change the y-axis



**** Stata code for this histogram ****

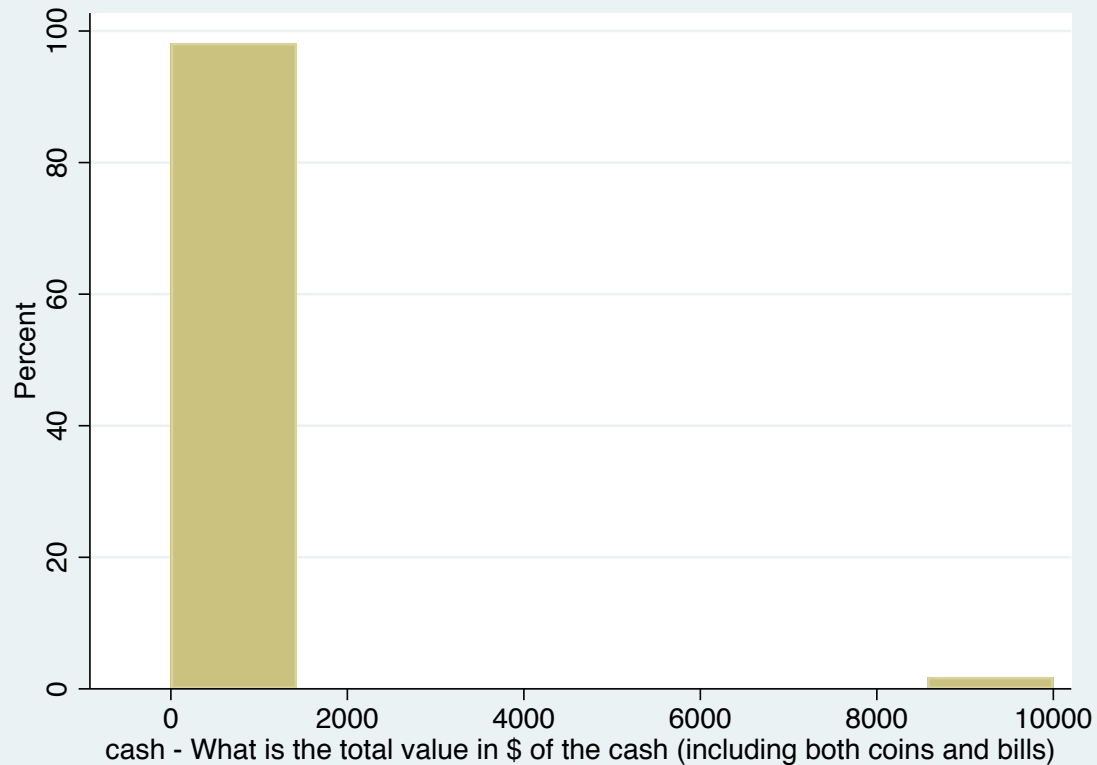
```
histogram cd4count, fcolor(blue) lcolor(black) width(50)  
title(CD4 among new HIV positives at Mulago) xtitle(CD4 cell  
count) percent
```

- This histogram has less detail but gives us the % of persons with CD4 <350 cells/mm³

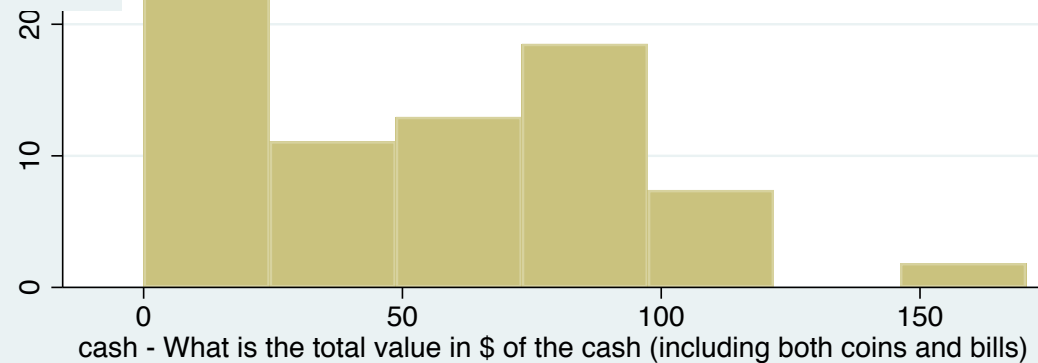


```
histogram cd4count, fcolor(blue) lcolor(black) width(350)  
title(CD4 among new HIV positives at Mulago) xtitle(CD4 cell  
count) percent
```

Histogram of class cash in pocket (Biostat 200-2016)

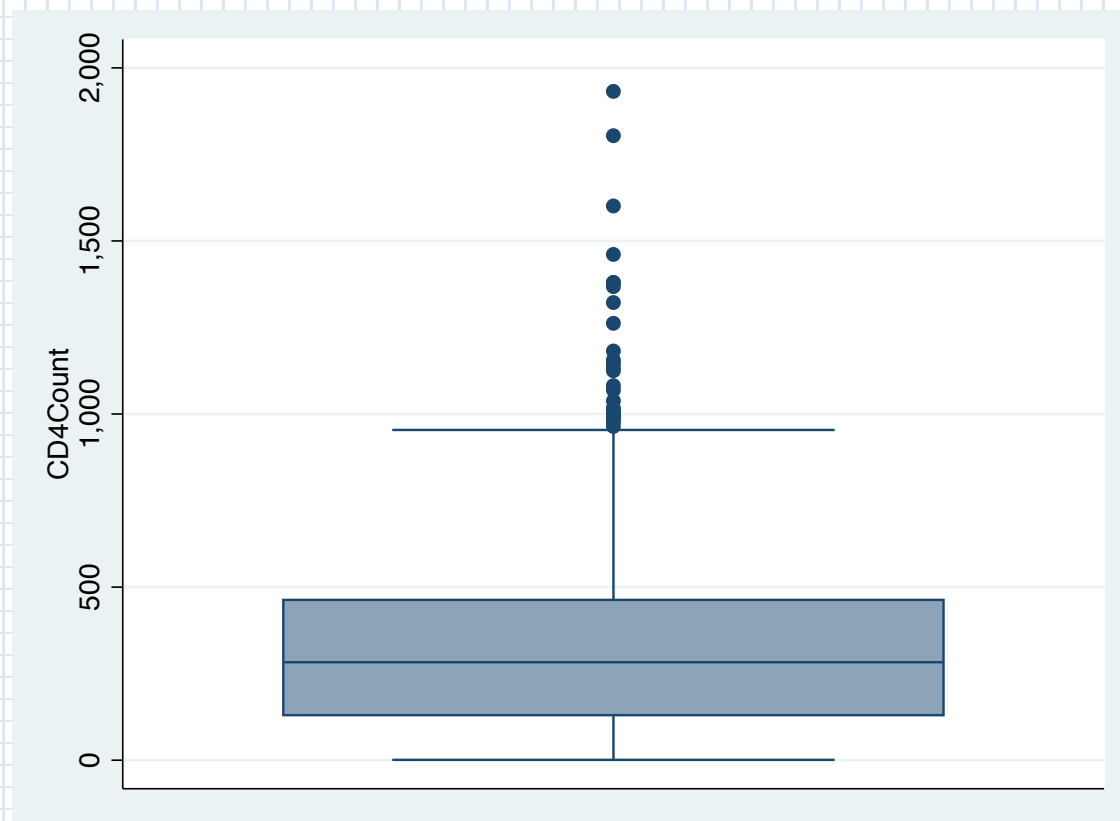


histogram cash, percent



Box plots – who invented them?

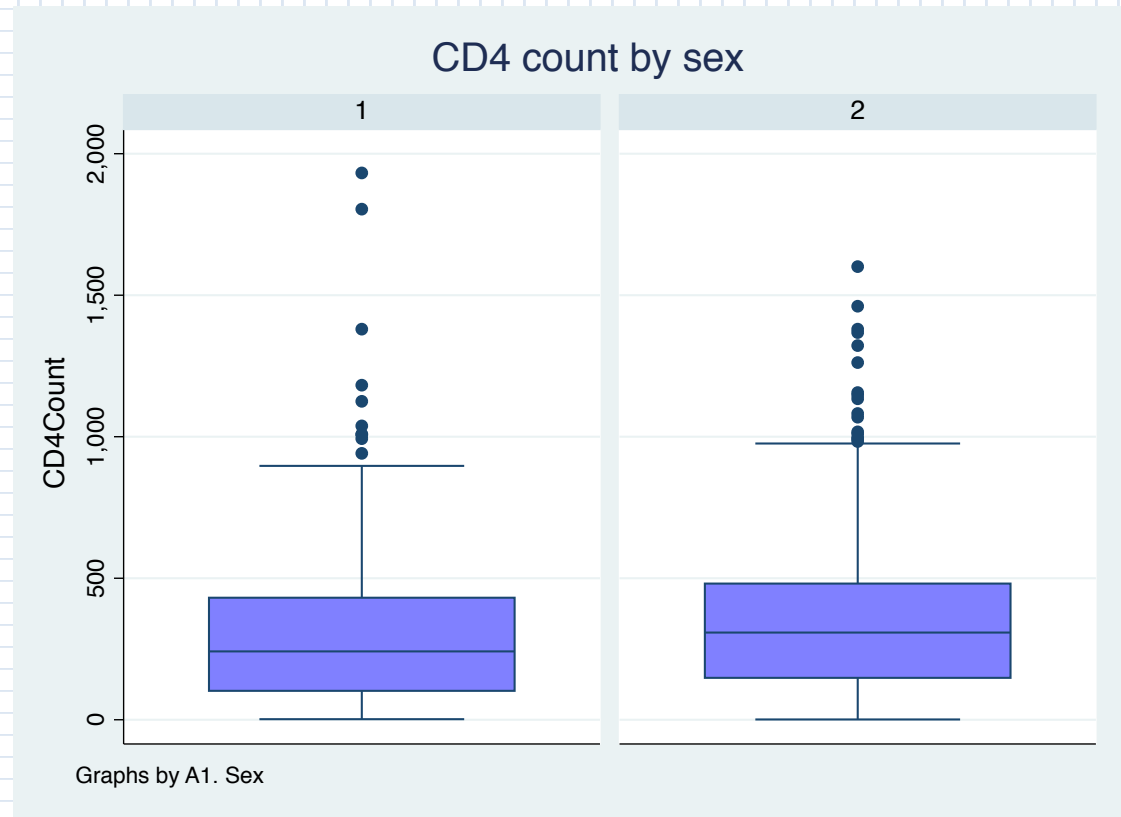
- Middle line=median (50th percentile)
- Box covers the 25th to 75th percentiles (interquartile range)
- Bottom whisker: Data point at or above 25th percentile – 1.5*IQR or the minimum, whichever is greater
 - 25th %ile=130, IQR=333 →
 $130 - 1.5 * 333 = -369.5$ (Hinge spread)
- Top whisker: Data point at or below 75th percentile + 1.5*IQR or the maximum, whichever is greater
 - 75th %ile=463, IQR=333 →
 $463 + 1.5 * 333 = 962.5$



`graph box cd4count`

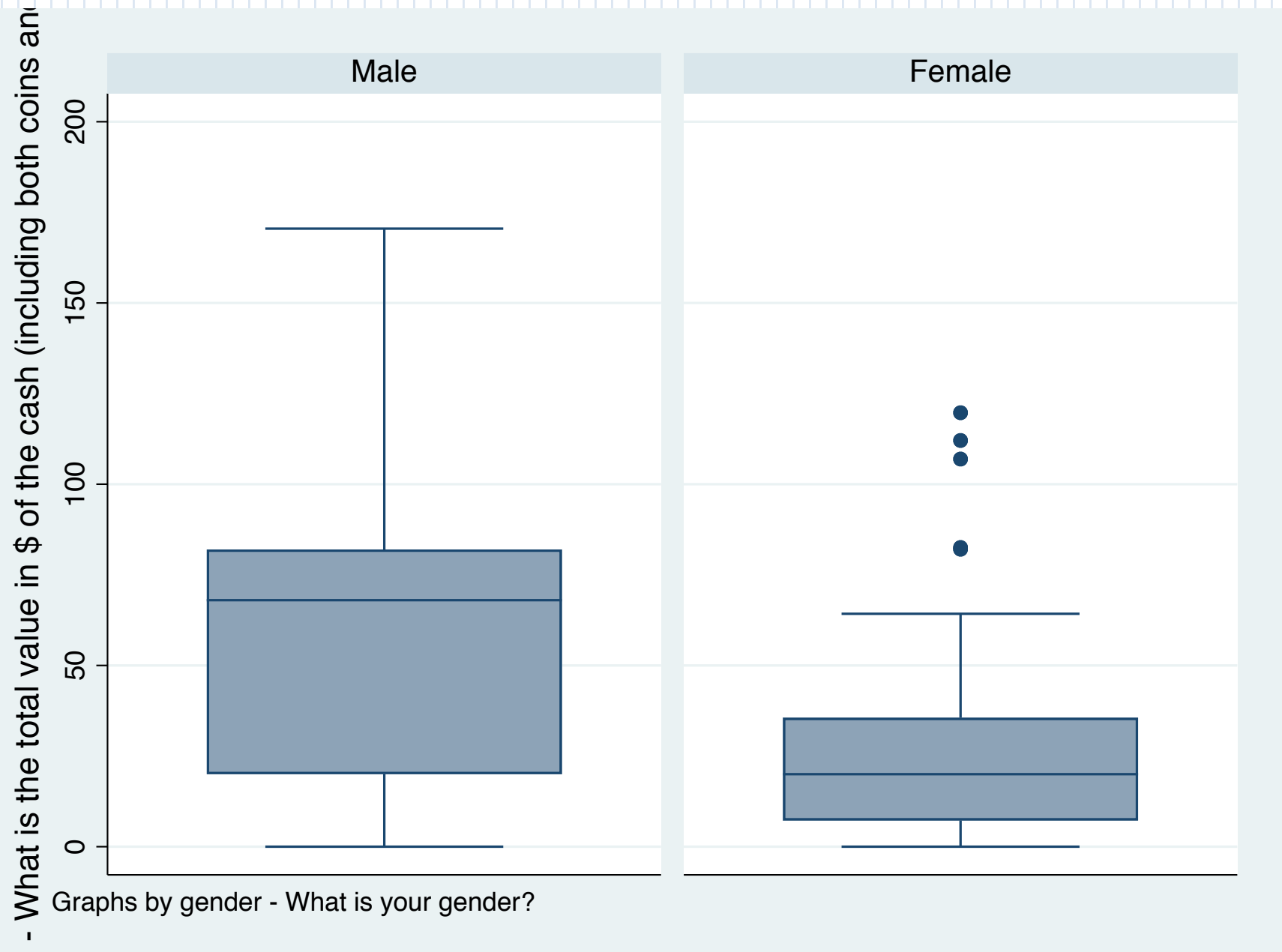
Box plots by another variable

- We can divide up our graphs by another variable
- A way to describe the relationship between a numerical and categorical variable



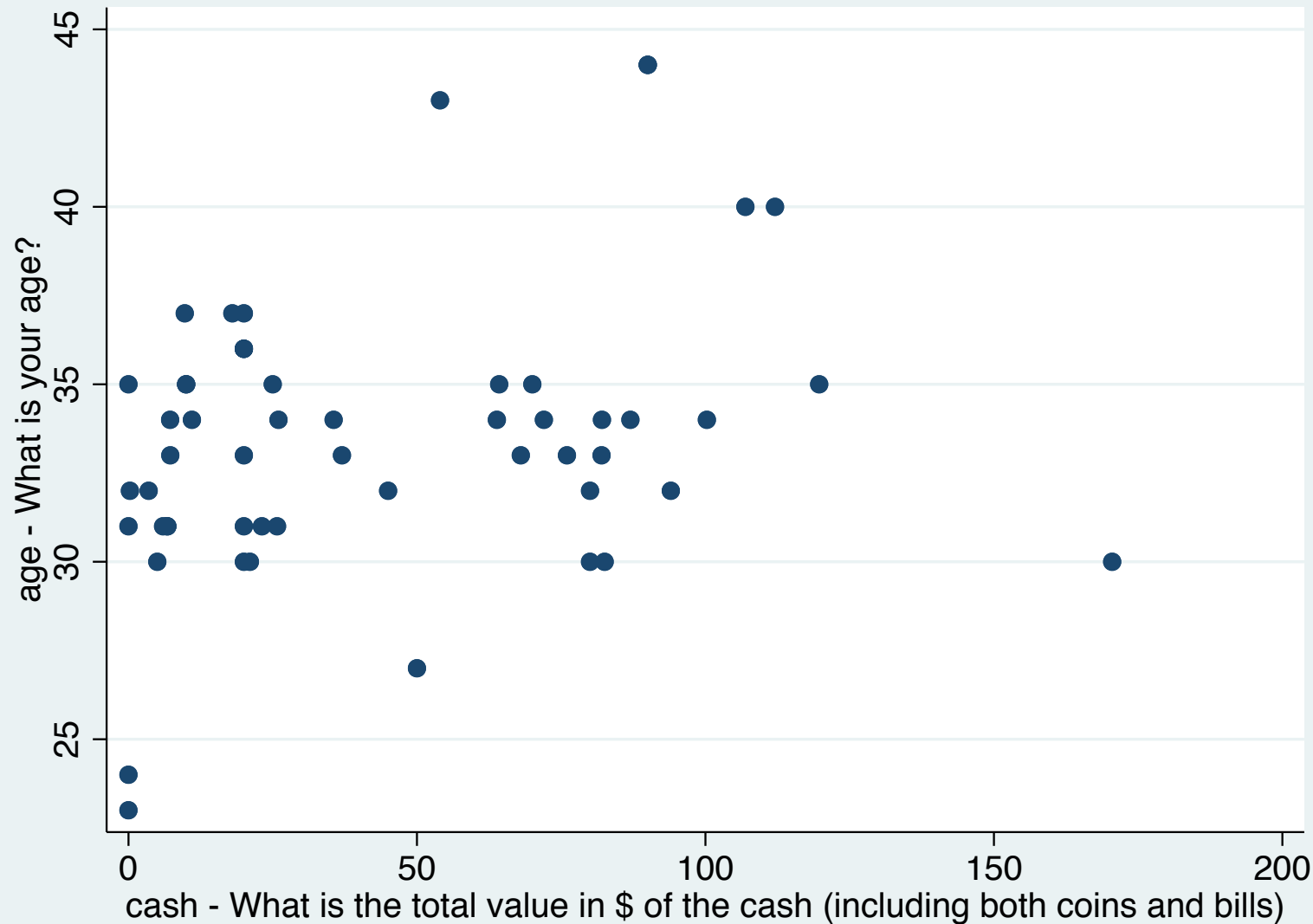
```
graph box cd4count, box(1, fcolor(blue)) by(, title(CD4  
count by sex)) by(sex) graphregion(fcolor(blue))
```

What is the total value in \$ of the cash (including both coins and



graph box cash, by(gender)

Scatter plots – 2 numerical variables



twoway (scatter age cash)

Numerical variable summaries

- Measures of central tendency – where is the center of the data?
 - Median – the 50th percentile == the middle value
 - If n is odd: the median is the $(n+1)/2$ observations (e.g. if $n=31$ then median is the 16th highest observation)
 - If n is even: the median is the average of the two middle observations (e.g. if $n=30$ then the median is the average of the 15th and 16th observation)

Median cash on hand

```
. summarize cash, detail
```

```
cash – What is the total value in $ of the cash
(including both coins and bills)
```

Percentiles		Smallest		
1%	0	0		
5%	0	0		
10%	3.5	0	Obs	54
25%	10	0	Sum of Wgt.	54
50%	25.375		Mean	43.98278
		Largest	Std. Dev.	39.85131
75%	80	106.93		
90%	94	112.06	Variance	1588.127
95%	112.06	119.72	Skewness	.8577528
99%	170.5	170.5	Kurtosis	3.121552

Numerical variable summaries

- Range
 - Minimum to maximum or difference
 - CD4 cell count range: (1-1932)
- Interquartile range (IQR)
 - 25th and 75th percentiles or difference
 - Less sensitive to extreme values
 - CD4 cell count IQR: (130-463)

Numerical variable summaries

- Measures of central tendency – where is the center of the data?
 - Mean – arithmetic average
 - Means are sensitive to very large or small values
 - Mean CD4 cell count: 329.2
 - Mean age: 31.7

$$\text{Mean: } \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

Numerical variable summaries

- Sample variance
 - Amount of spread around the mean

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

Why we divide by $n-1$ rather than n ?

“Because the sum of deviations is always zero, the last deviation can be found once we know the other $n-1$. So we are not averaging n unrelated numbers. Only $n-1$ of the squared deviations can vary freely, and we average by dividing by the total by $n-1$. So we have $n-1$ pieces of information. The $n-1$ is called the degrees of freedom of the variance or standard deviation.”

The Practice of Statistics Yates, Moore, McCabe

Numerical variable summaries

- Sample standard deviation (SD) is the square root of the variance
 - The standard deviation has the same units as the mean

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$$

- SD of CD4 cell count = 266.1 cells/mm³
- SD of Age = 9.9 years

Median cash on hand

```
. summarize cash, detail
```

cash – What is the total value in \$ of the cash
(including both coins and bills)

Percentiles		Smallest		
1%	0	0		
5%	0	0		
10%	3.5	0	Obs	54
25%	10	0	Sum of Wgt.	54
50%	25.375		Mean	43.98278
		Largest	Std. Dev.	39.85131
75%	80	106.93		
90%	94	112.06	Variance	1588.127
95%	112.06	119.72	Skewness	.8577528
99%	170.5	170.5	Kurtosis	3.121552

Grouped data

- Sometimes you are given data in aggregate form
- The data consist of frequencies of each individual value or range of values
- What do we do with the first group?

CD4 cell counts (per mm ³) of persons newly diagnosed with HIV at Mulago Hospital, Kampala (N=999)	
	n (%)
≤50	121 (12.1)
51-250	339 (33.9)
251-500	339 (33.9)
≥500	200 (20.0)

Grouped mean

- The mean uses the midpoint of each group
- For the highest group, the use the midpoint between the cutpoint and the maximum

- Grouped Mean $m_i = \text{the midpoint of the } i^{\text{th}} \text{ group}$
 $f_i = \text{the frequency in the } i^{\text{th}} \text{ group}$
$$\bar{x} = \frac{\sum_{i=1}^k m_i f_i}{\sum_{i=1}^k f_i}$$

$$= (25*121 + 150*339 + 375*339 + 1216*200) / 999$$

$$= 424.6 \text{ cells/mm}^3 \text{ (mean from original data was 329.2)}$$

Grouped standard deviation

- The standard deviation

$$s = \sqrt{\frac{\sum_{i=1}^k (m_i - \bar{x})^2 f_i}{(\sum_{i=1}^k f_i) - 1}}$$

$$= \text{sqrt} \left(\frac{(25-424.6)^2 * 121 + (150-424.6)^2 * 339 + (375-424.6)^2 * 339 + (1216-424.6)^2 * 200}{998} \right) = 413.9 \text{ cells/mm}^3$$

(SD from original data was 266.1)

Recap of today

- Class logistics
- Types of variables
- Tables to describe categorical data
- Graphs to describe categorical and numerical data
- Statistics to summarize numerical data

For next time

- Lab this Thursday – More Stata using descriptive statistics & exploratory data analysis. No Homework this week.
- Next Tuesday 9/25/18 we will cover the basics of probability and how we use this with distributions