

Biostat 200

Introduction to Biostatistics

Lecture 1

Course instructors

Course director

Judy Hahn, M.A., Ph.D.

Associate Professor in Residence

Phone: (415) 476-5815

Office: 550 16th St., 3rd Floor, 3550

Judy.hahn@ucsf.edu

TAs

- Mohannad Awad
- Nelson Kalema
- Ping Wang
- Hyeri You

- Lectures: Tuesdays 10:30-12:30
 - Videos of lectures are posted usually within a few hours
 - 11 lectures (Sept 13-Nov 29, no lecture Nov. 22)
 - Attendance encouraged but optional
- “Labs”: Thursday 10:30-12
- Hahn office hours:
 - Tuesday 12:30-1:30 Room MH 1400 and by appointment
- Course credits: 3
- Some STATA in class – Please bring your laptop

- Readings
 - Required readings will be from ***Principles of Biostatistics*** by M. Pagano and K. Gauvreau. Duxbury. 2nd edition.
 - Please read the assigned chapters *before* lecture, and review them after lecture
 - Lectures will closely follow book chapters

Assignments

- 9 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
- Assignment schedule in the syllabus file
- Assignments will consist of:
 - Data analysis and interpretation
 - Exercises in the book
 - 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 2_Hahn.doc” . Include your name within the Assignment as well

Assignments

- No late assignments accepted
- You will be able to drop your lowest grade. So if you are unable to complete an assignment on time you can drop that one.

“Labs”

- Labs will be every Thursday 10:30 -12
- No lab 11/24
- Lab content
 - Labs 1-2: Stata familiarity, exploratory data analysis
 - These lab exercises will not be turned in or graded
 - “Labs” 3-11
 - Lecture review
 - Examples from TAs own work or the literature
 - Review of more challenging homework problems
 - Time to get started and ask questions on new assignment
- Attendance encouraged but optional

Forum

- <https://courses.ucsf.edu/mod/forum/view.php?id=260051>
- Will be used for class communications, like assignment clarifications, etc.
- Please use for clarification questions that others in the class may also have

Grading

- Assignments (70%)
 - 9 assignments/problem sets
 - 8/9 assignments count towards your grade
- Late assignments will not be graded
- Answer keys will be posted soon after assignments are due.
- Graded assignments will be returned on the CLE 1-2 week after they are due.
- Grading is done by TAs who are students in the MCR program
- Final exam will be a take home exam (30% of grade)

TICR Professional Conduct Statement

Clarifications for this class

- 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?



- MA in Biostatistics, PhD in epidemiology
- At UCSF since 1992
- My research
 - Substance use and infectious disease
 - Injecting drug use and HCV in San Francisco
 - Unhealthy alcohol and HIV in sub-Saharan Africa
 - NIH funded grants since 2006



"And it was so typically brilliant of you to have invited an epidemiologist."

Who are you?

- Stage in training
 - Resident
 - Fellow
 - Faculty
 - PhD Student
- Outside of UCSF?
- 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 understand basic statistical analyses in published journals

Course goals – deeper understanding

- Have you read a journal article that reports p-values or 95% confidence intervals?
- Have you calculated a p-value or a 95% confidence interval?
 - Do you know where that p-value came from?

What will we cover

- Syllabus

Today's topics

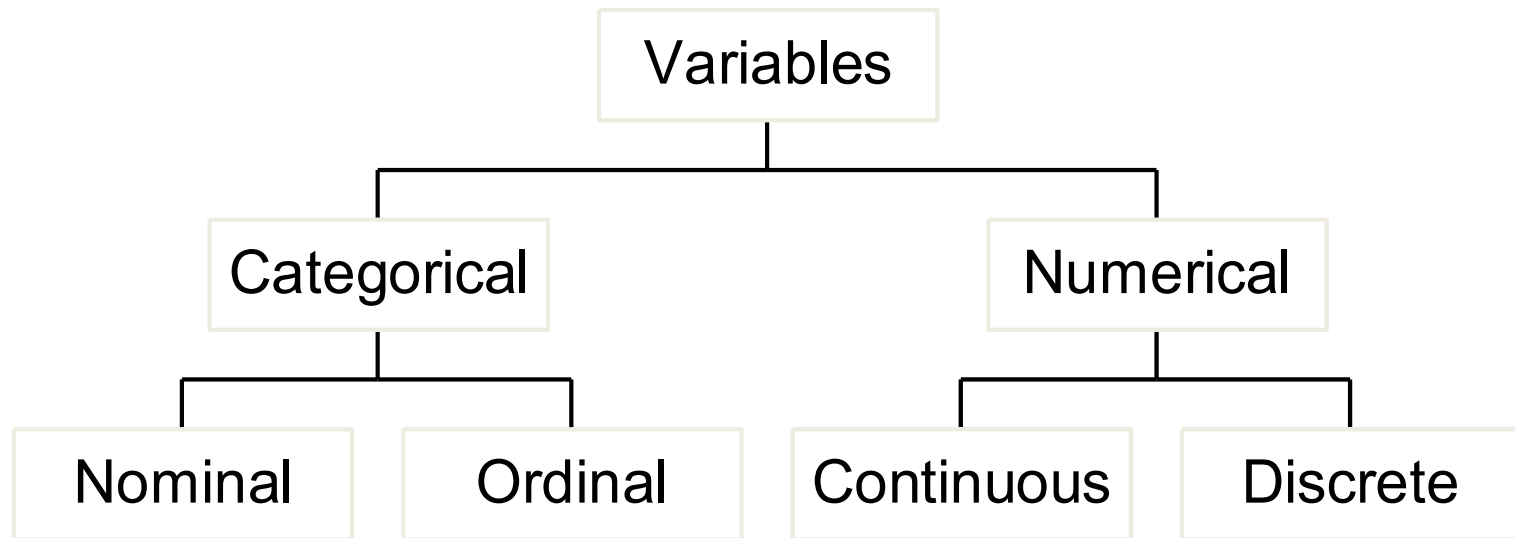
- Variables - numerical versus categorical
- Tables (frequencies)
- Graphs (histograms, box plots, scatter plots, line graphs)
- Summaries of numerical variables

But first, our data collection

https://ucsf.co1.qualtrics.com/jfe/form/SV_erm4RdujGxPWT9H

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 (from the French nom) 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

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

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...)

Variables in our data set

Categorical variables	Ordinal or nominal

Numerical variables	Cont or discrete?

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
 - 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

Tables to summarize data

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	
	n (%)
Male	
Female	
Trans or other	
Total	

Frequency tables

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

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 (%)

Frequency tables

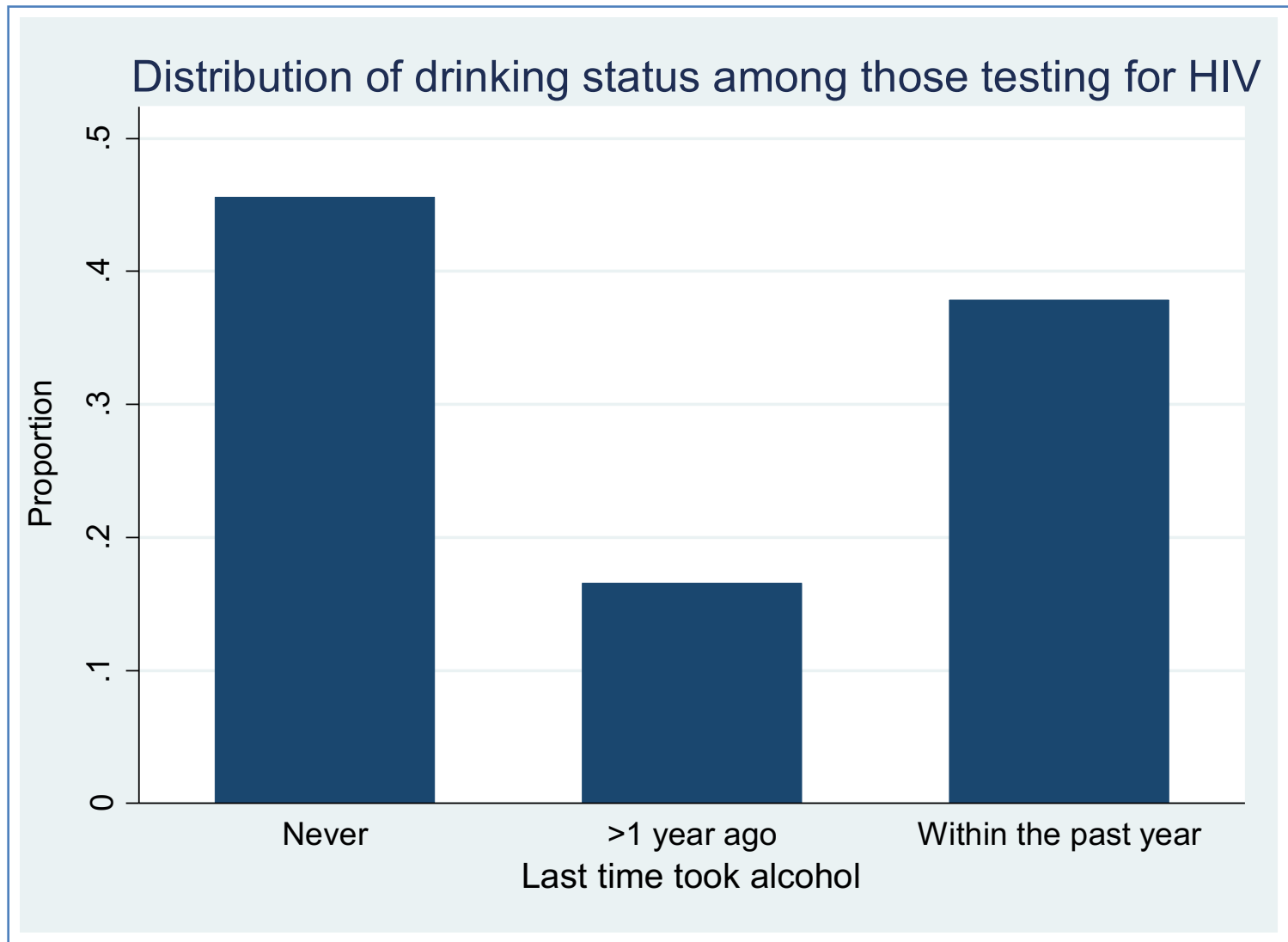
- The cumulative frequency is the percentage of observations up to and including the current category

Age of class participants		
	n (%)	Cumulative frequency (%)

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, example



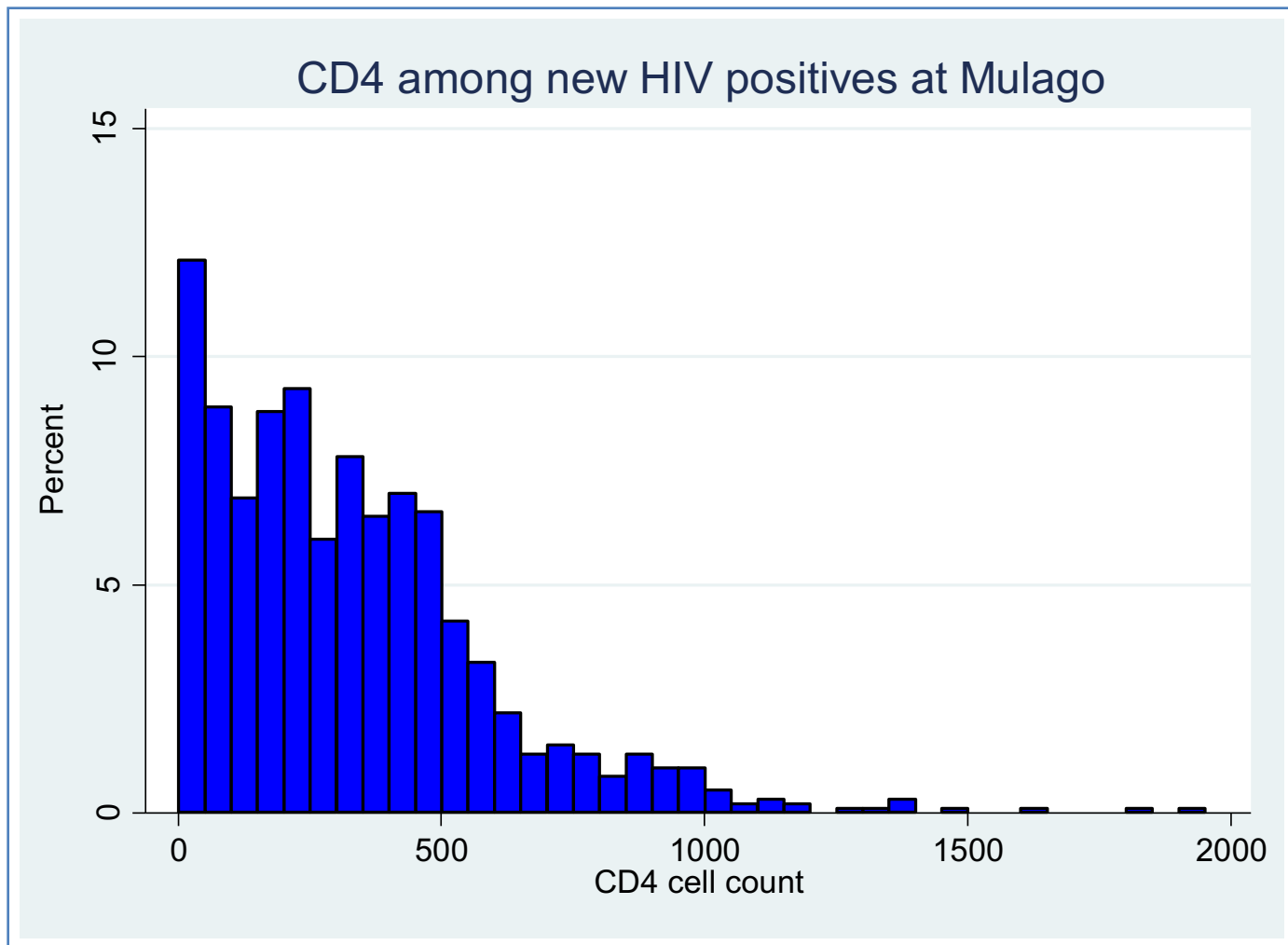
Bar charts, gender of class

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?

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

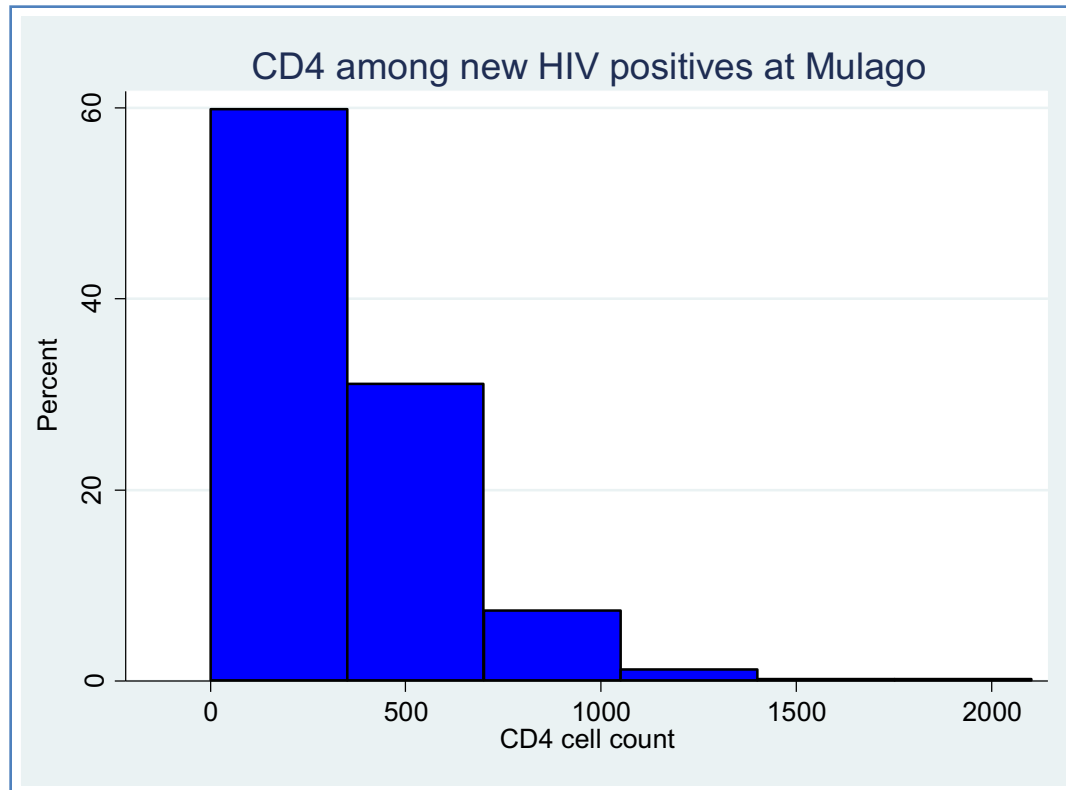


**** Stata code for this histogram ****

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

- Without specifying any options, the bin width and the color of your histogram be chosen automatically.
- To choose the number of bins – use `bin()` option
- To choose the width of the bins – use `width()` option
- To choose the color of the bars – use `fcolor()` option
- Add titles
- Use `freq` or `percent` so the y-axis makes sense

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

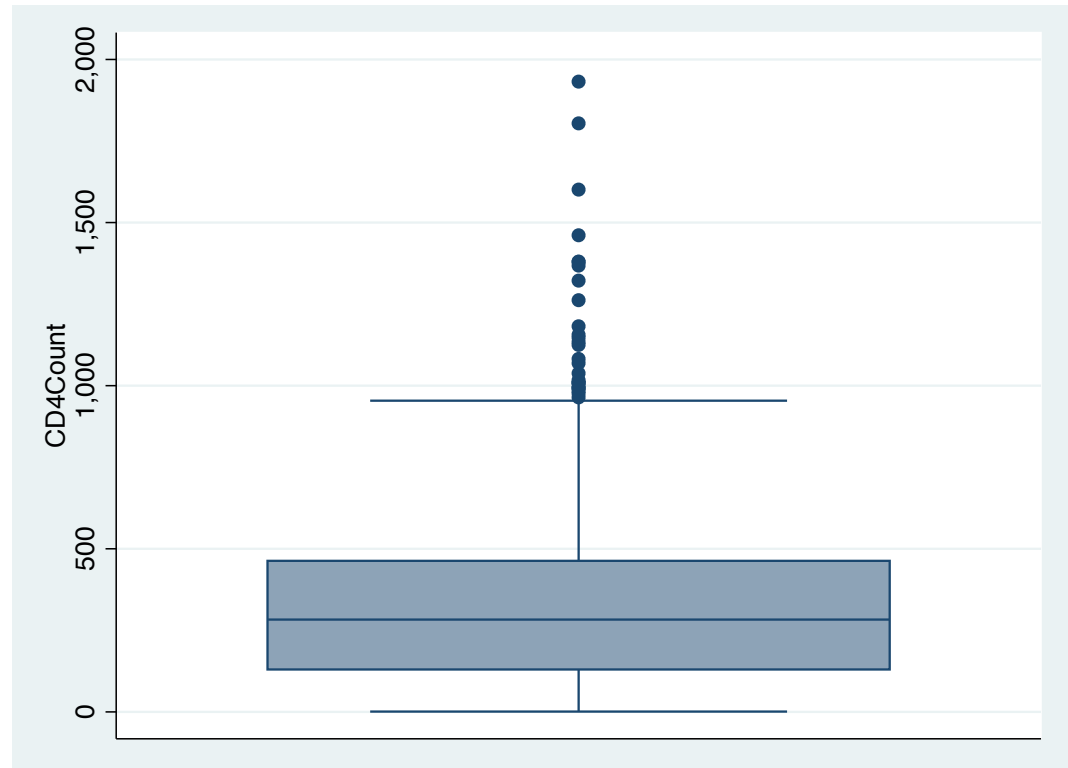


```
histogram cd4count, fcolor(blue) lcolor(black) width(350)
title(CD4 among new HIV positives at Mulago) xtitle(CD4 cell
count) percent
```

Histogram of class cash in pocket

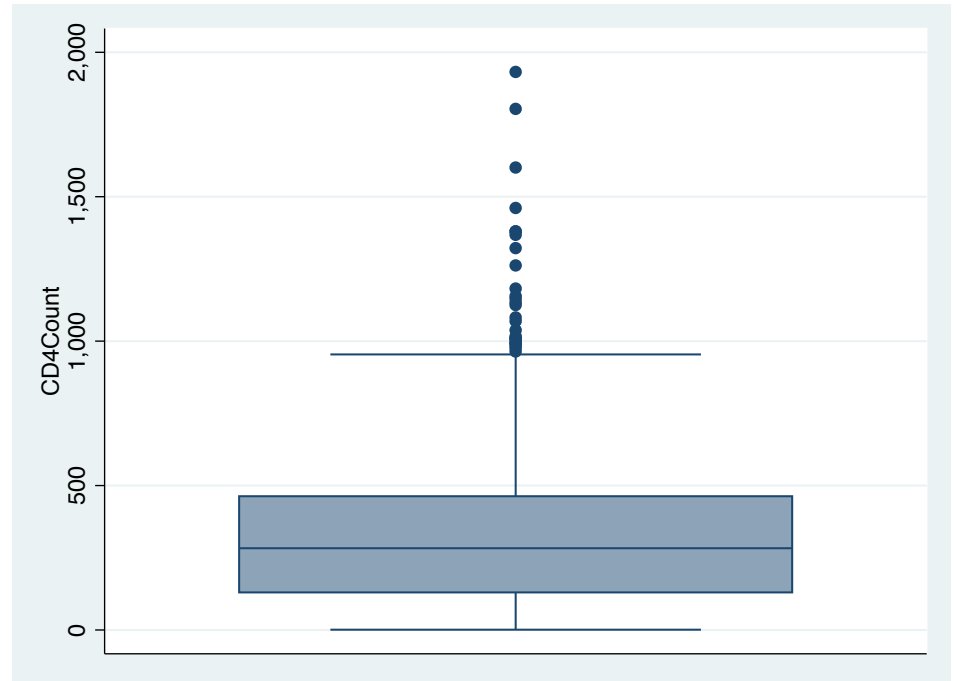
Box plots

- Middle line=median (50th percentile)
- Box covers the 25th to 75th percentiles (interquartile range)



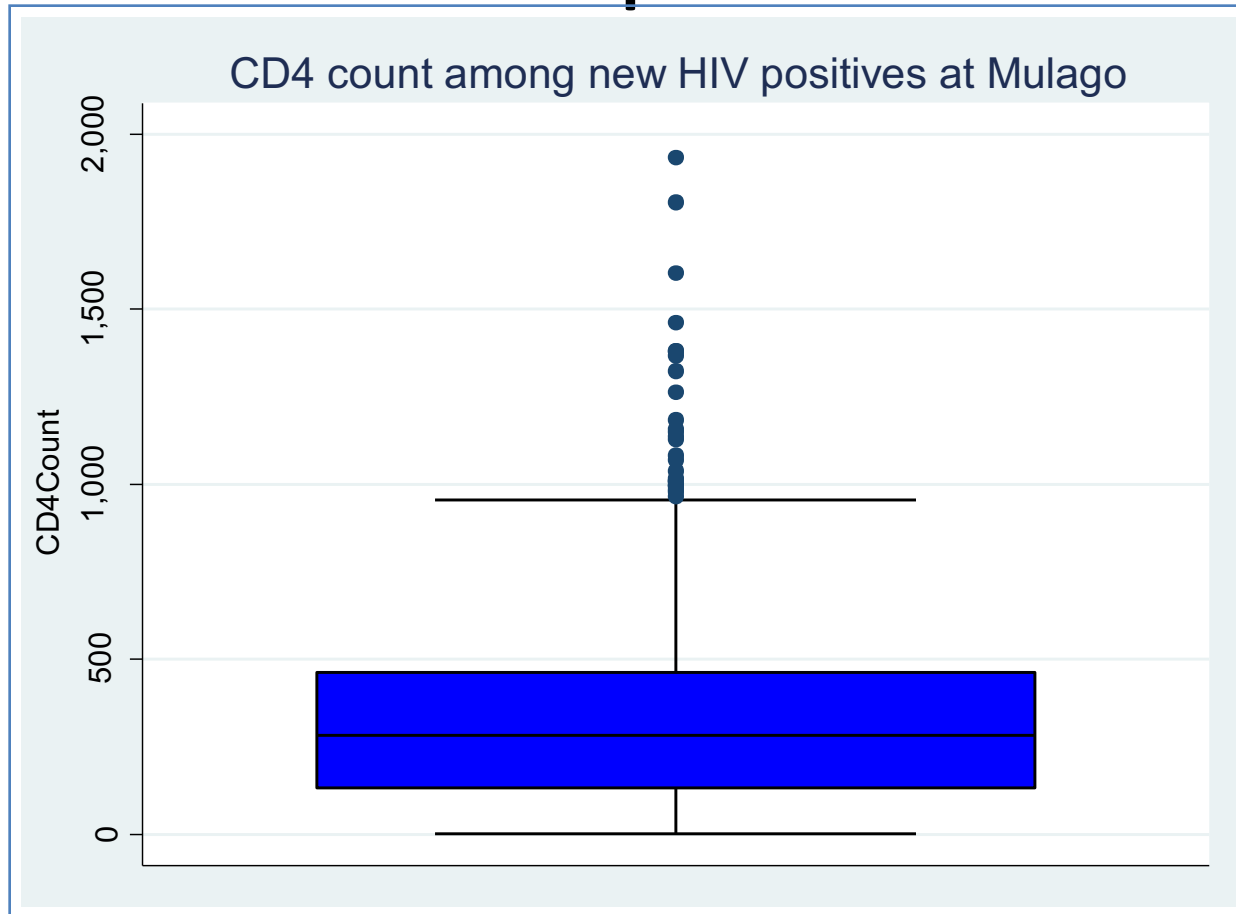
`graph box cd4count`

Box plots



- Bottom whisker: Data point at or above 25th percentile – 1.5*IQR or the minimum, whichever is greater
 - 25th %ile=130, IQR=333 → $130 - 1.5 * 333 = -369.5$
- Top whisker: Data point at or below 75th percentile + 1.5*IQR or the maximum, whichever is greater
 - 75th %ile=463, IQR=333 → $463 + 1.5 * 333 = 962.5$

Box plots



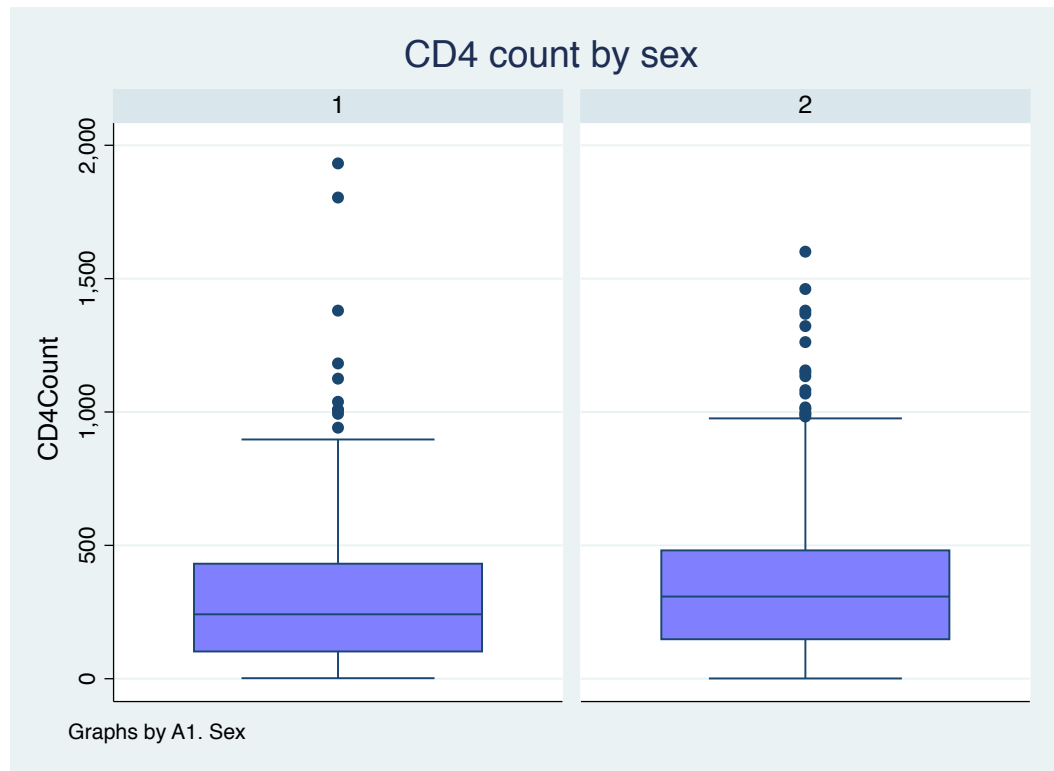
```
graph box cd4count, box(1, fcolor(blue) lcolor(black)  
fintensity(inten100)) title(CD4 count among new HIV  
positives at Mulago)
```

USE drop down menus in Stata to make your graphics look pretty!

Boxplot of cash on hand at Biostat 200

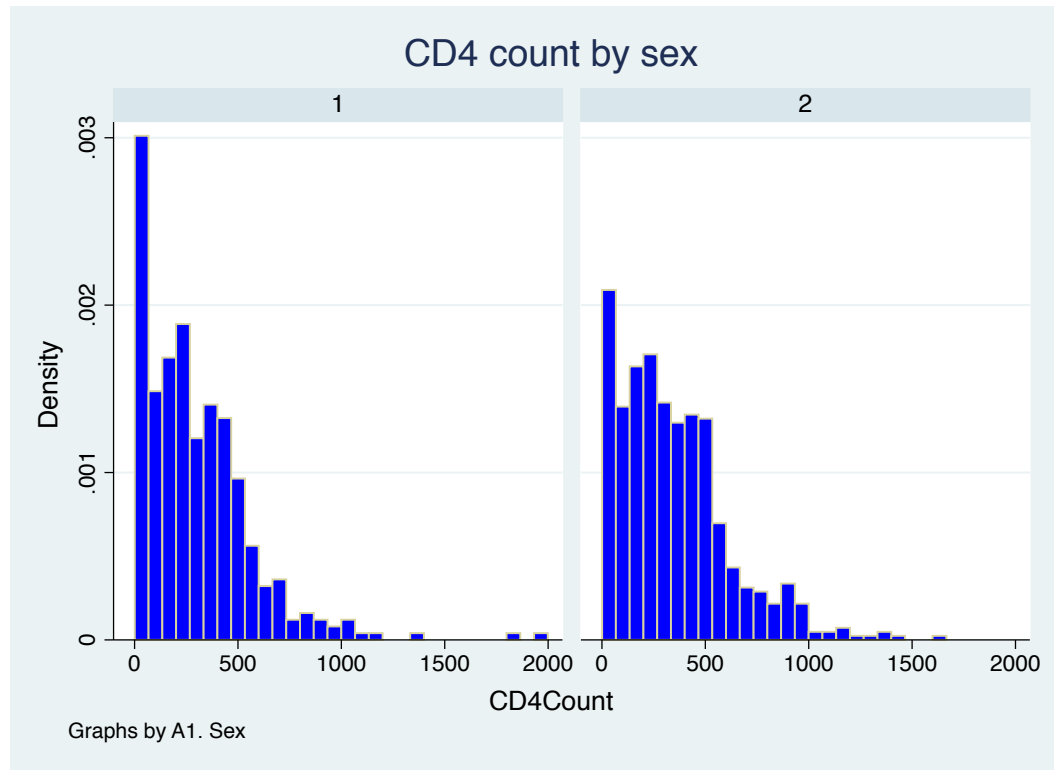
Box plots by another variable

- 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))
```

Histograms by another variable



```
hist cd4count, by(sex) fcolor(blue) by(,title(CD4 count by sex))
```

Boxplot of cash on hand at Biostat 200 by gender

Scatter plots – 2 numerical variables

Use scatter or lowess statements

Example: cash in pocket vs. number of apple products owned

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

CD4 cell count

```
. summarize cd4count, detail
```

CD4Count				

	Percentiles	Smallest		
1%	5	1		
5%	14	2		
10%	36	2	Obs	999
25%	130	2	Sum of Wgt.	999
50%	283		Mean	329.2332
		Largest	Std. Dev.	266.1177
75%	463	1461		
90%	659	1601	Variance	70818.64
95%	866	1804	Skewness	1.444705
99%	1182	1932	Kurtosis	6.518639

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

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

Interpreting the formula

- $\sum$ is the symbol for the sum of the elements immediately to the right of the symbol
- These elements are indexed (i.e. subscripted) with the letter i
 - The index letter could be any letter, though i is commonly used
- The elements are lined up in a list, and the first one in the list is denoted as x_1 , the second one is x_2 , the third one is x_3 and the last one is x_n .
- n is the number of elements in the list.

$$\sum_{i=1}^n x_i = x_1 + x_2 + \dots + x_n$$

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

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 do 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

Numerical variable summaries

- Coefficient of variation (CV)

$$CV = \frac{s}{\bar{x}} * 100\%$$

- For the same relative spread around a mean, the variance and standard deviation will be larger for a larger mean
- Uses: Can use CV to compare variability across measurements that are on a different scale (e.g. IQ and head circumference)

CD4 count

CD4 Count

Percentiles		Smallest		
1%	189	163		
5%	300	189		
10%	344	227	Obs	170
25%	415	232	Sum of Wgt.	170
50%	514.5		Mean	556.0765
		Largest	Std. Dev.	206.395
75%	672	1160		
90%	776.5	1239	Variance	42598.89
95%	907	1335	Skewness	1.307961
99%	1335	1449	Kurtosis	5.970502

$$CV = 206.4 / 556.08 = 0371 = 37.1\%$$

Soluble CD14

summ cd14, detail

CD14 ng/ml (high >3200 recoded to 3201)

	Percentiles	Smallest		
1%	321.075	244.073		
5%	663.132	321.075		
10%	864.413	395.812	Obs	170
25%	1111.232	480.851	Sum of Wgt.	170
50%	1517.295		Mean	1549.649
		Largest	Std. Dev.	589.4338
75%	1874.718	3001.133		
90%	2310.801	3120.826	Variance	347432.2
95%	2732.128	3201	Skewness	.487751
99%	3201	3201	Kurtosis	3.247377

$$CV = 589.43/1549.65 = 0.380 = 38.0\%$$

Grouped data

- Sometimes you are given data in aggregate form
- The data consist of frequencies of each individual value or range of values

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

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

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

(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

- Review today's material
 - Read Pagano and Gauvreau Chapters 1-3
- Next week's material (Probability)
 - Read Chapter 6
- Lab this Thursday – Stata review, exploratory data analysis. Assignment 1 will be posted this Thursday morning and due 9/22.