Biostat 200

Introduction to Biostatistics

Lecture 1

Course instructors

Course director

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

Associate Professor in Residence
Phone: (415) 597-4961
Office: 50 Beale St., Suite 1300
Judy.hahn@ucsf.edu

TAs

- Gabriel Brooks, MD
- Elaine Ku, MD
- Annie Narla, MD
- Aaron Spicer, MD

Course web site

- https://courses.ucsf.edu/course/view.php?id=1631

- Lectures: Tuesdays 10:30-12:30
  – Videos of lectures are posted usually within a few hours
- Labs: Thursday 10:30-12
- Hahn office hours: Tuesday 12:30-1:30 Room CB 6702 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

- 8 assignments
- Assignments will be due weekly on Thursdays at 10:30 a.m. starting 10/2
- Each assignment will be posted at least one week before it is due
- 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

Send all assignments to:
biostats200.2014@gmail.com

Send your assignments as a word document. Please do not forget to put your name on the document.

Name the file with assignment # and your last name:

e.g. "Lab 2_Hahn.doc"

Labs

- Labs will be every Thursday 10:30 -12
- No lab 9/25, 11/13, and 11/28
- Lab content
  - Lab 1: Stata familiarity, exploratory data analysis
  - Labs 2-6
    - 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
  - Labs 7-10 similar, but more emphasis on real-world problems

Forum

- https://courses.ucsf.edu/mod/forum/view.php?id=108486

- 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%)
  - 8 problem sets
- Late assignments will not be graded
  - You may earn 60% credit if complete
- Final exam (30%)
- Grading is done by TAs

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.

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

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

Alcohol and the HIV cascade

How to measure alcohol use?

Sexual risk behavior / biologic factors

HIV infection

HIV diagnosis

HIV care

ART receipt

ART adherence

Viral suppression

HIV transmission to others

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
- Have a strong foundation for understanding more complex material

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?

Who are you?

- Stage in training
  - Resident
  - Fellow
  - Faculty
  - PhD Student
- Do you have a data set that you are analyzing?
- Are you in the process of collecting your own data?

What will we cover

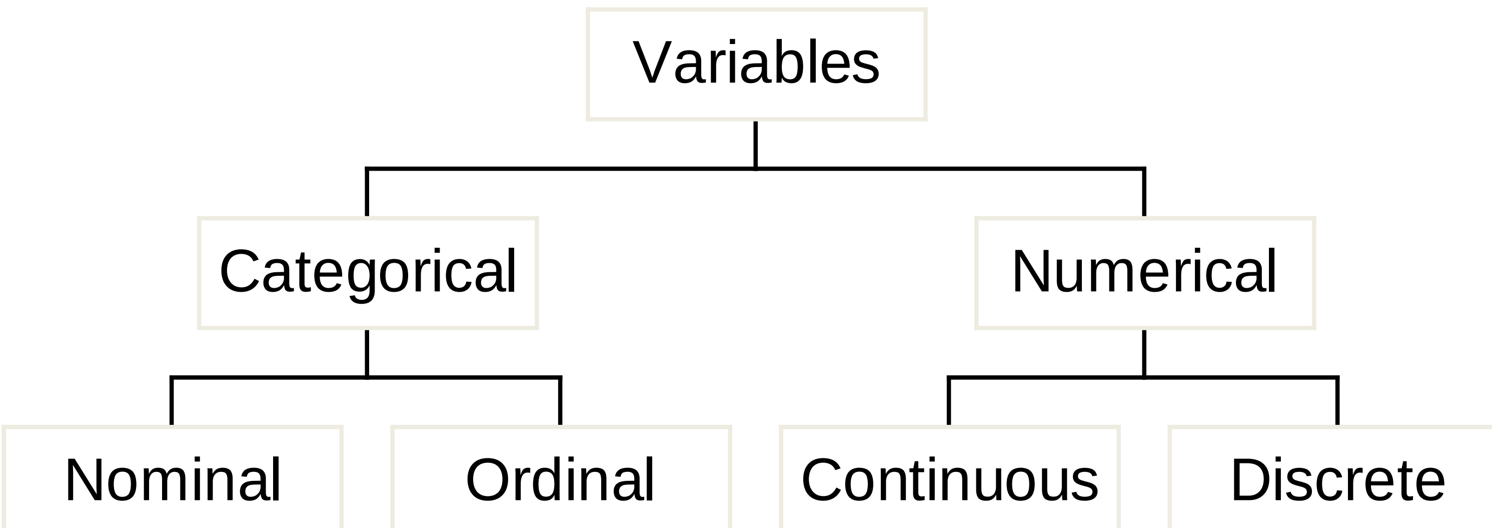
- Syllabus

Today's topics

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

Types of variables

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



Types of variables

- <u>Categorical variable</u>: 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:
        - » GENDER: 1=male, 2=female
        - » 0=No 1=Yes

Types of variables

- Categorical variables
  - <u>Nominal variables</u>:
    - The data are unordered
  - <u>Ordinal variables</u>:
    - 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
  - Discrete variables: can be counted (e.g. number of children in 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…)

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 receiving new HIV test, Mulago Hospital, Kampala, Uganda, 2008-2011.	
	n (%)
Male	1553 (46)
Female	1836 (54)
Total	3389 (100)

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/mm^3
  - Equal percentage of the data falling into each category
    - Tertiles – 33%
    - Quartiles – 25%
    - Quantiles – 20% each

Frequency tables

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)

Frequency tables

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

CD4 cell counts (per mm^3) of persons newly diagnosed with HIV at Mulago Hospital, Kampala (N=999)		
	n (%)	Cumulative frequency (%)
≤50	121 (12.1)	12.1
51-250	339 (33.9)	46.1
251-500	339 (33.9)	80.0
≥500	200 (20.0)	100.0

In Stata

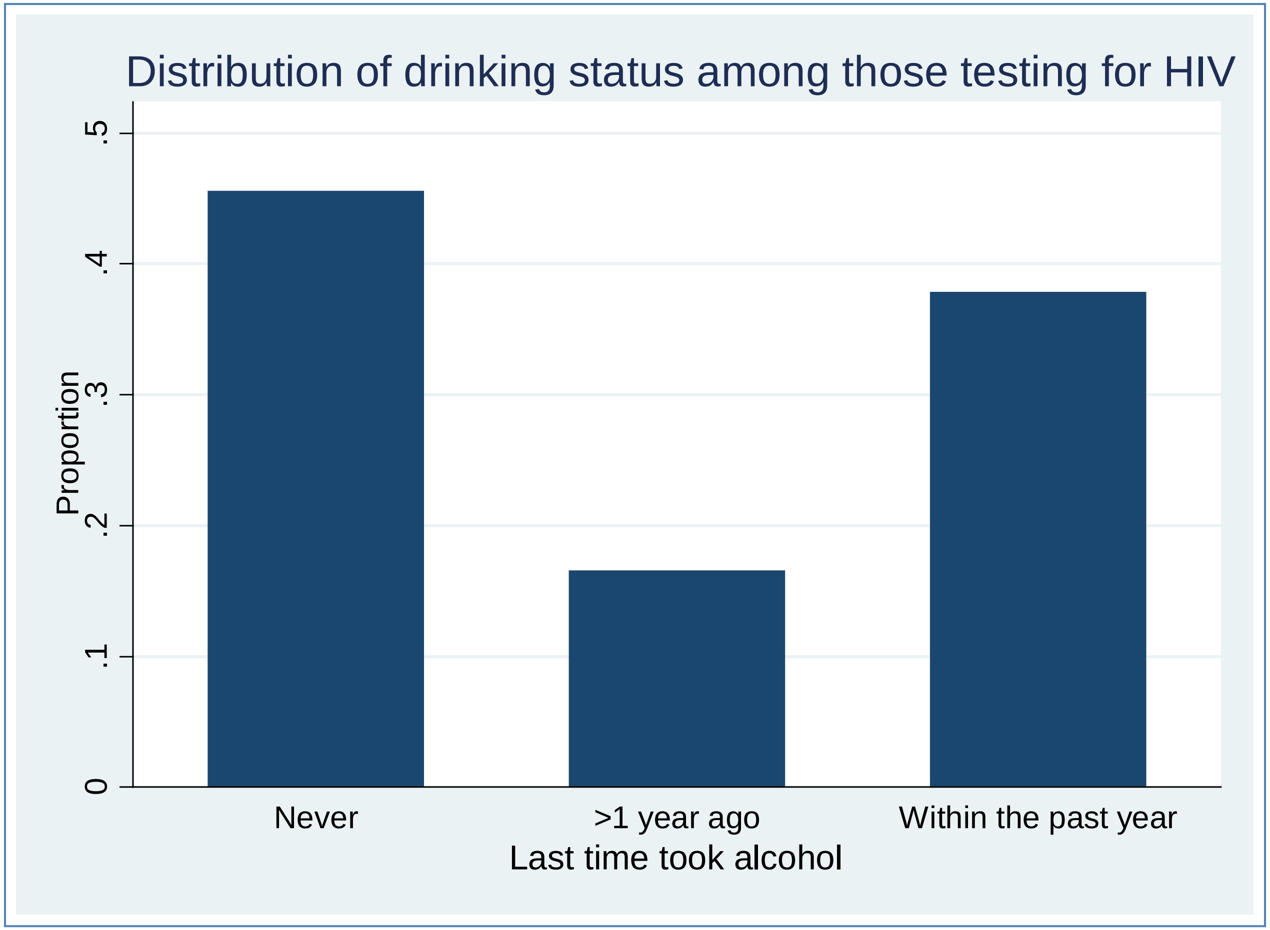
```
. tab cd4_cat

   RECODE of |
    cd4count |
  (CD4Count) |      Freq.     Percent        Cum.
-------------+-----------------------------------
     CD4<50 |        121       12.11       12.11
 CD4=51-250 |        339       33.93       46.05
CD4=251-500 |        339       33.93       79.98
    CD4>500 |        200       20.02      100.00
-------------+-----------------------------------
       Total |        999      100.00
```

These data are posted on the class website and called:
vct_baseline_biostat200_v1

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 should add up to 1

Bar charts

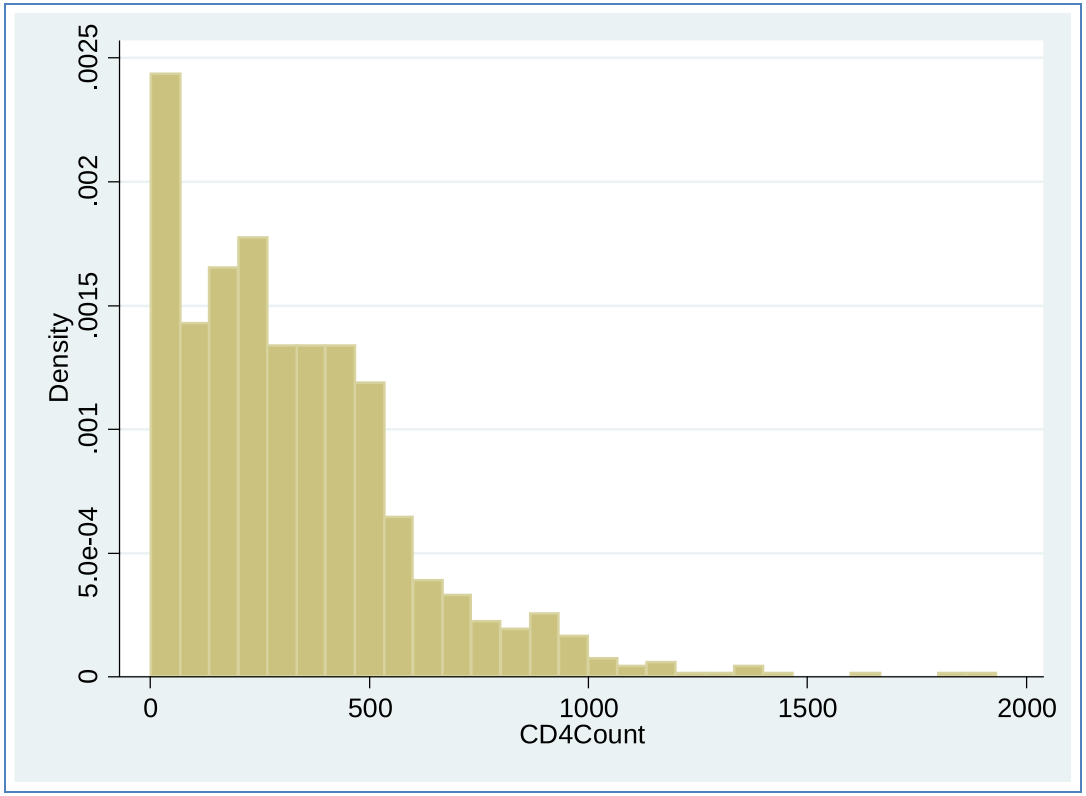


Distribution of drinking status among those testing for HIV
Proportion
0
.1
.2
.3
.4
.5
Never
>1 year ago
Within the past year
Last time took alcohol

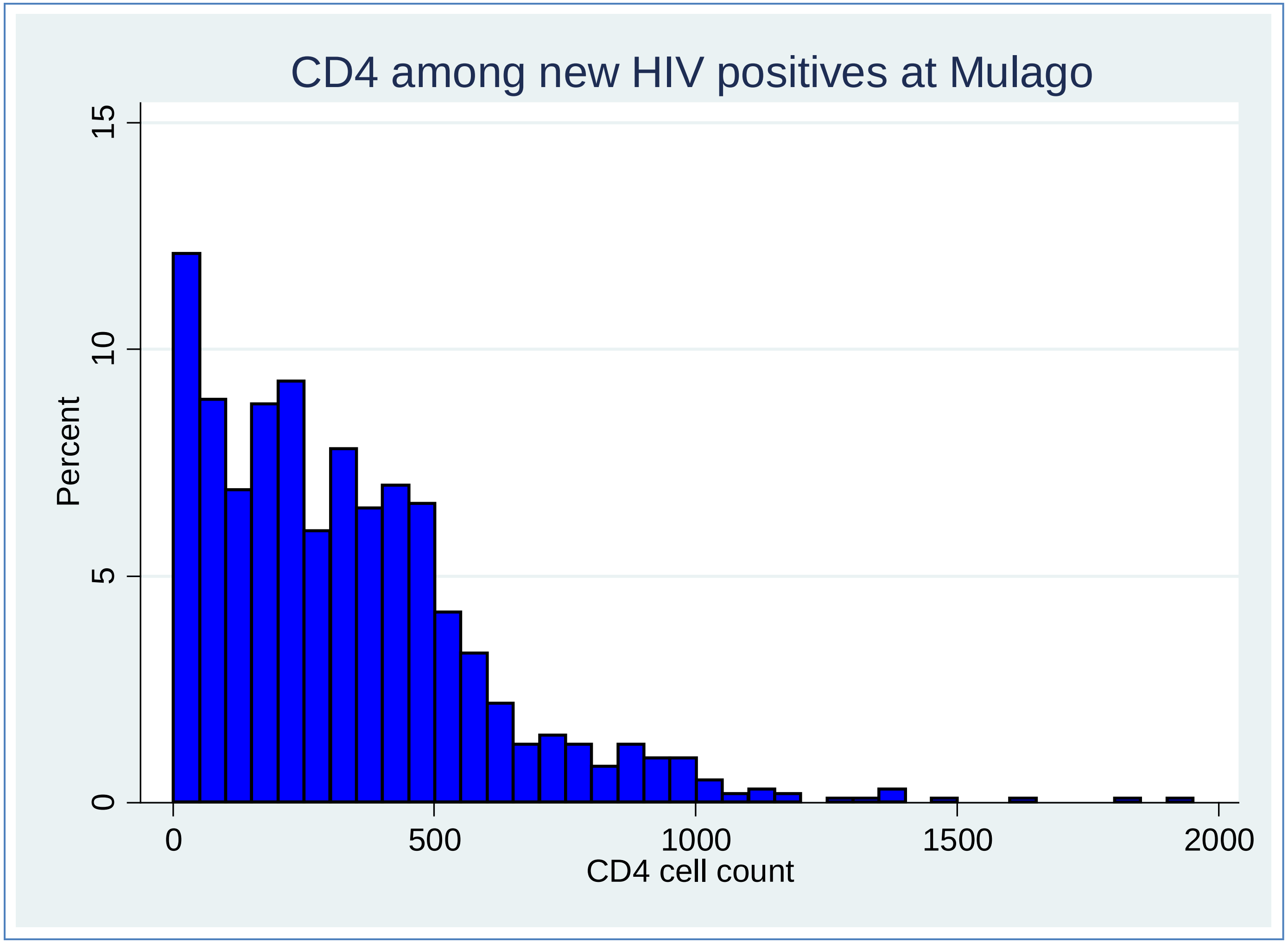
Histograms

- 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

- Without specifying any options, your histogram will look like this. The bin width will be chosen automatically.

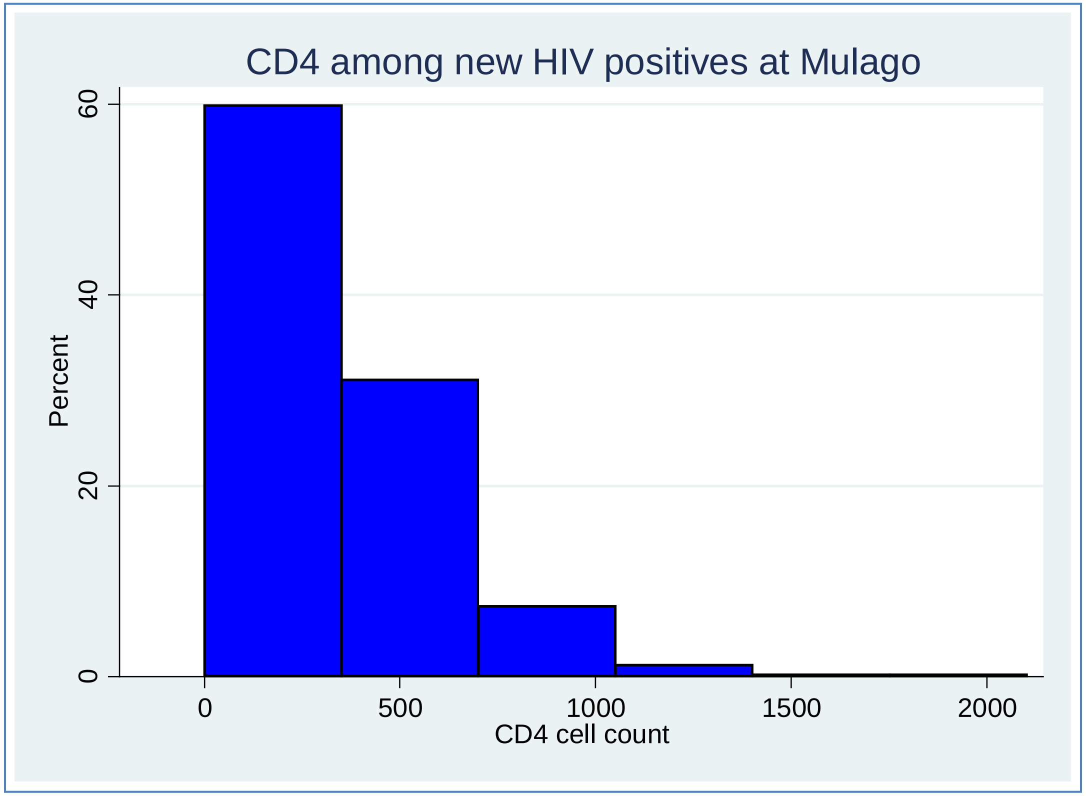



```
** Stata code for this histogram **
histogram cd4count
(bin=29, start=1, width=66.586207)
```




```
** 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
```

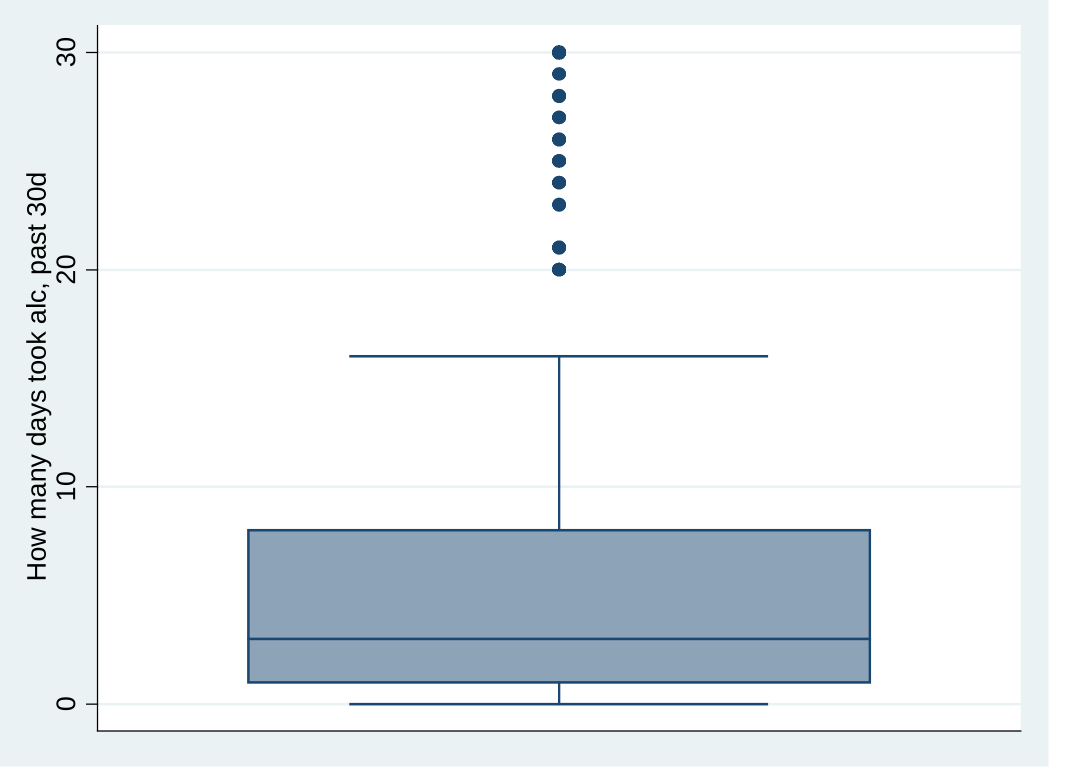
- This histogram has less detail but gives us the % of persons with CD4 <350 cells/mm^3




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

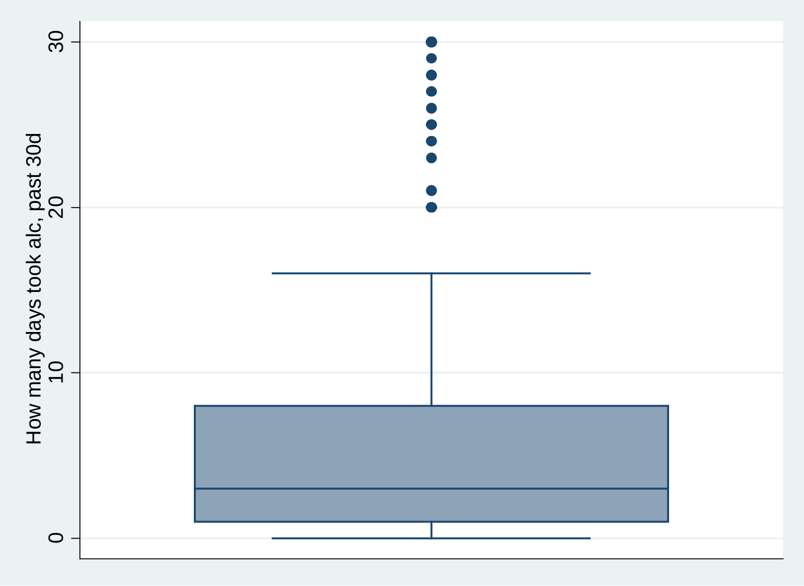
Box plots

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



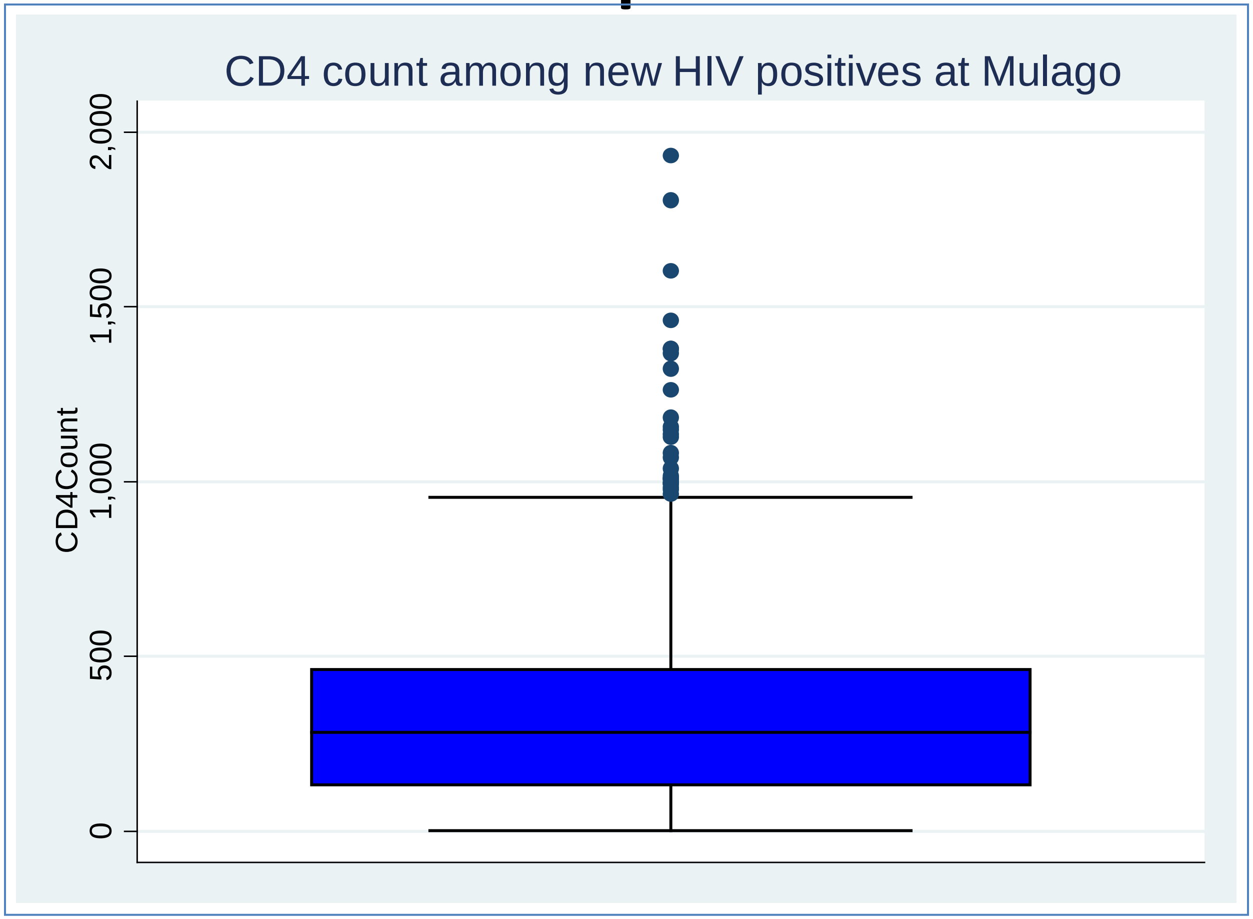
`graph box e5`

Box plots



- Bottom whisker: Data point at or above 25th percentile – 1.5*IQR or the minimum, whichever is greater
  - 25th %ile=1, IQR=7, 1-1.5*7=-9.5
- Top whisker: Data point at or below 75th percentile + 1.5*IQR or the maximum, whichever is greater
  - -75th %ile=8, IQR=7, 8+1.5*7=18.5

Box plots

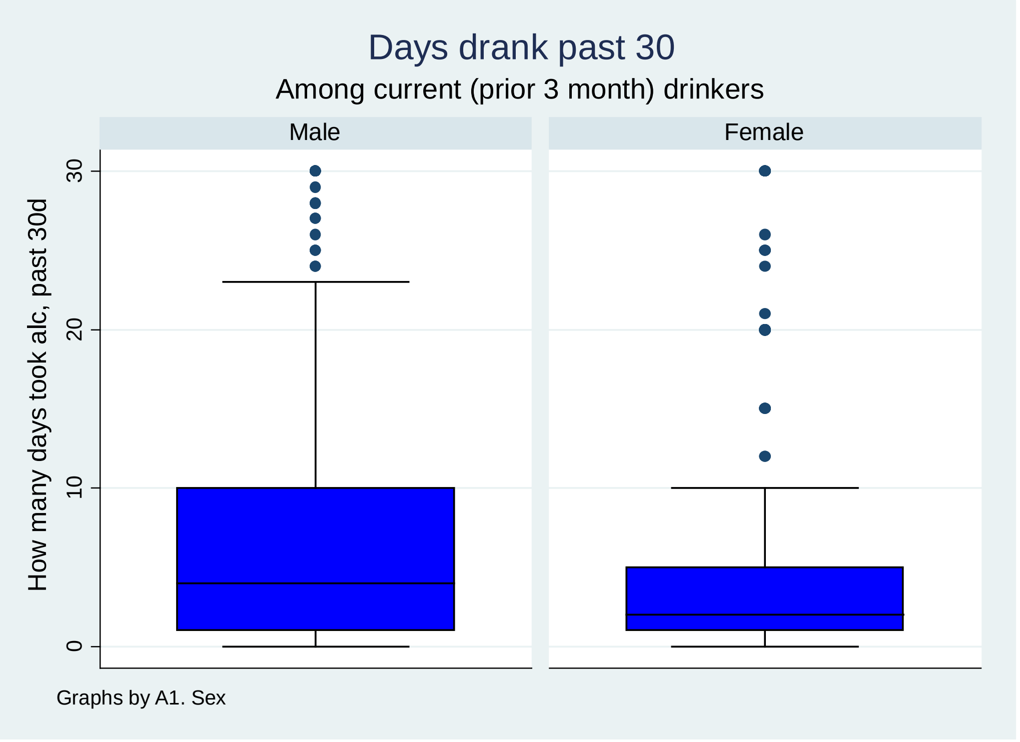



```
graph box cd4count, box(1, fcolor(blue) lcolor(black)
fintensity(inten100)) title(CD4 count among new HIV
positives at Mulago)
```

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

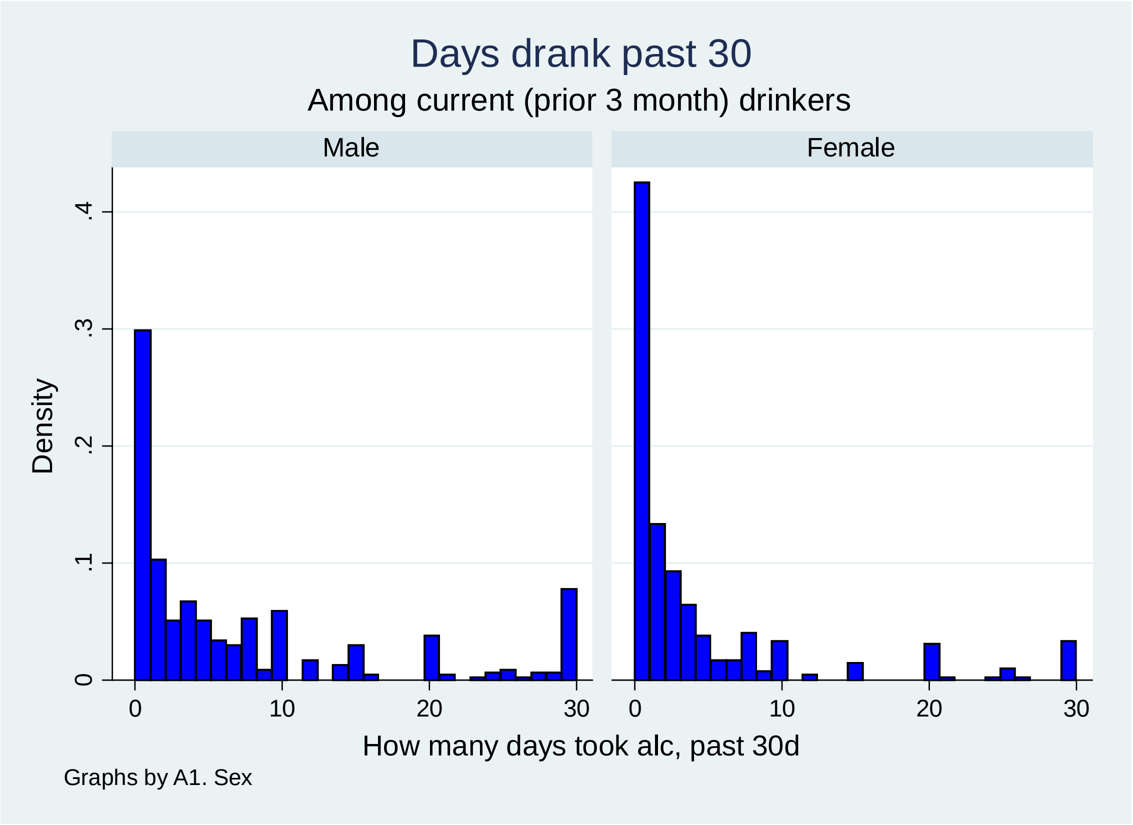
Box plots by another variable

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




```
graph box e5, by(, title(Days drank past 30)
subtitle(Among current (prior 3 month) drinkers)) by(sex)
box(1, fcolor(blue) lcolor(black) fintensity(inten100))
```

Histograms by another variable

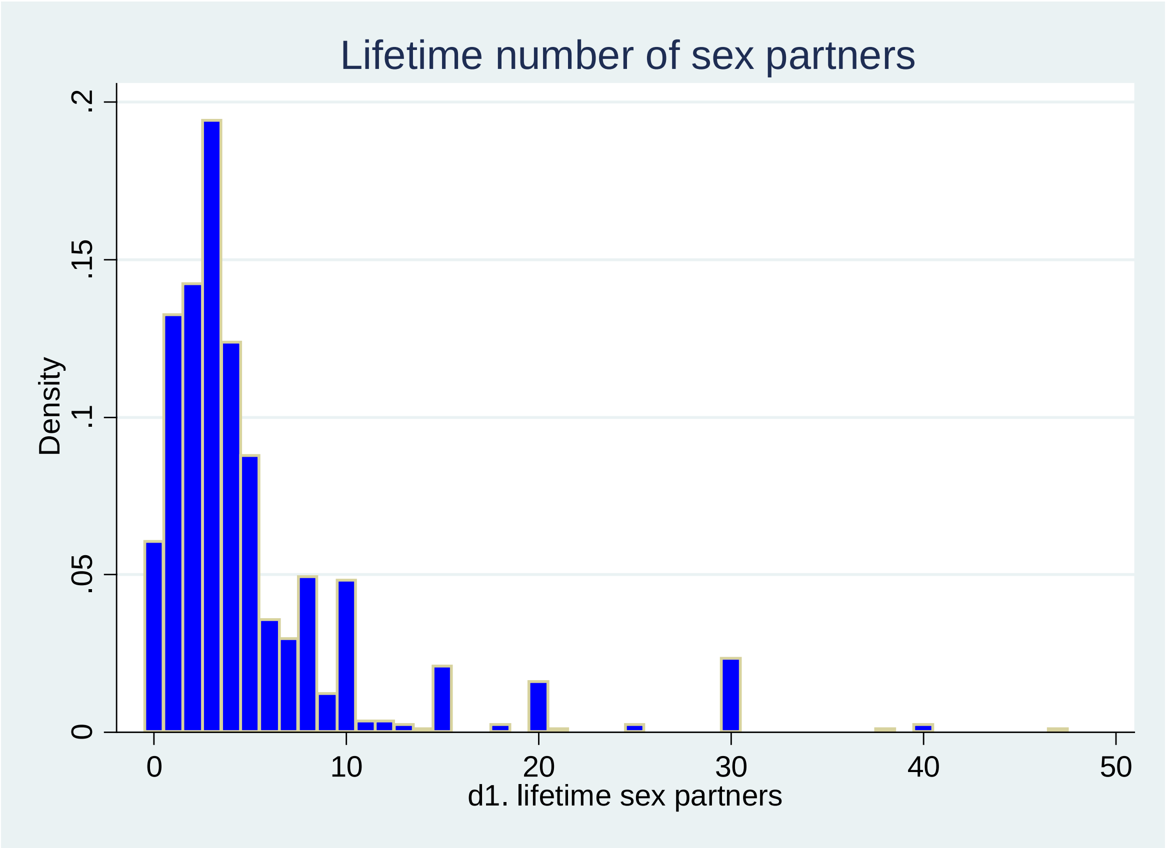



```
histogram e5, by(, title(Days drank past 30)
subtitle(Among current (prior 3 month) drinkers)) by(sex)
fcolor(blue) lcolor(black)
```

Histograms help find the mode

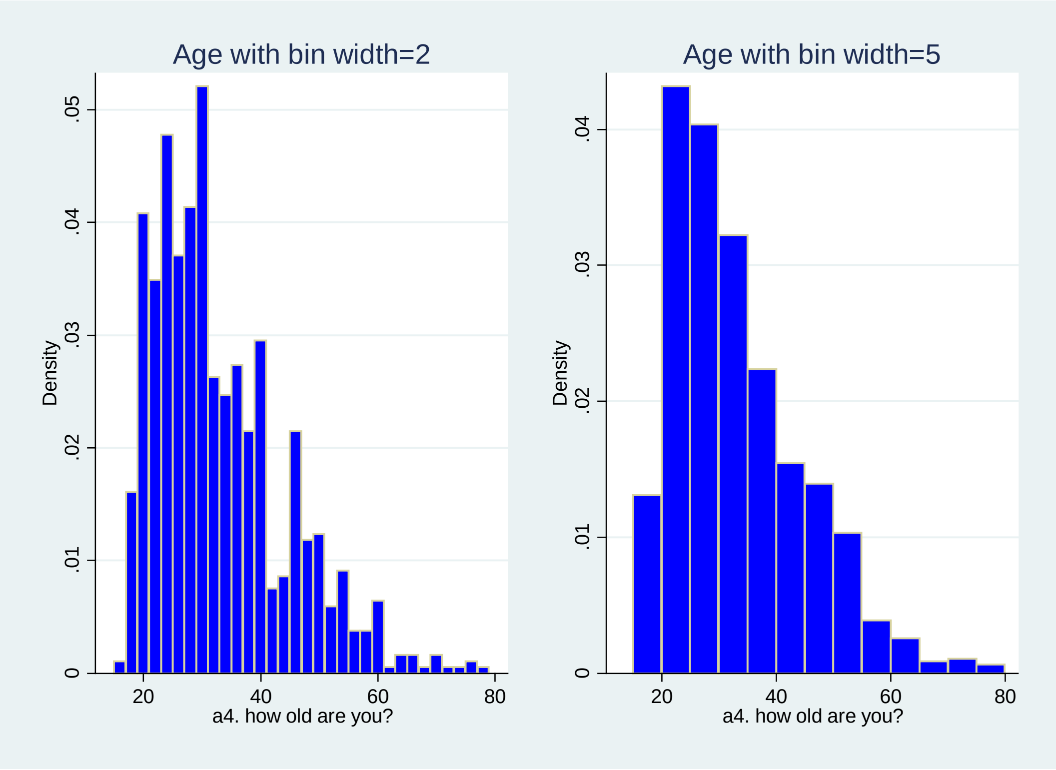
- Mode – the value (or range of values) that occurs most frequently
- Sometimes there is more than one mode, e.g. a bi-modal distribution (both modes do not have to be the same height)
- The mode makes most sense for categorical data
- For continuous data you can find the mode if you categorize the data

- What type of variable is this?
- What is the mode?
- Is the distribution of this variable bi-modal?



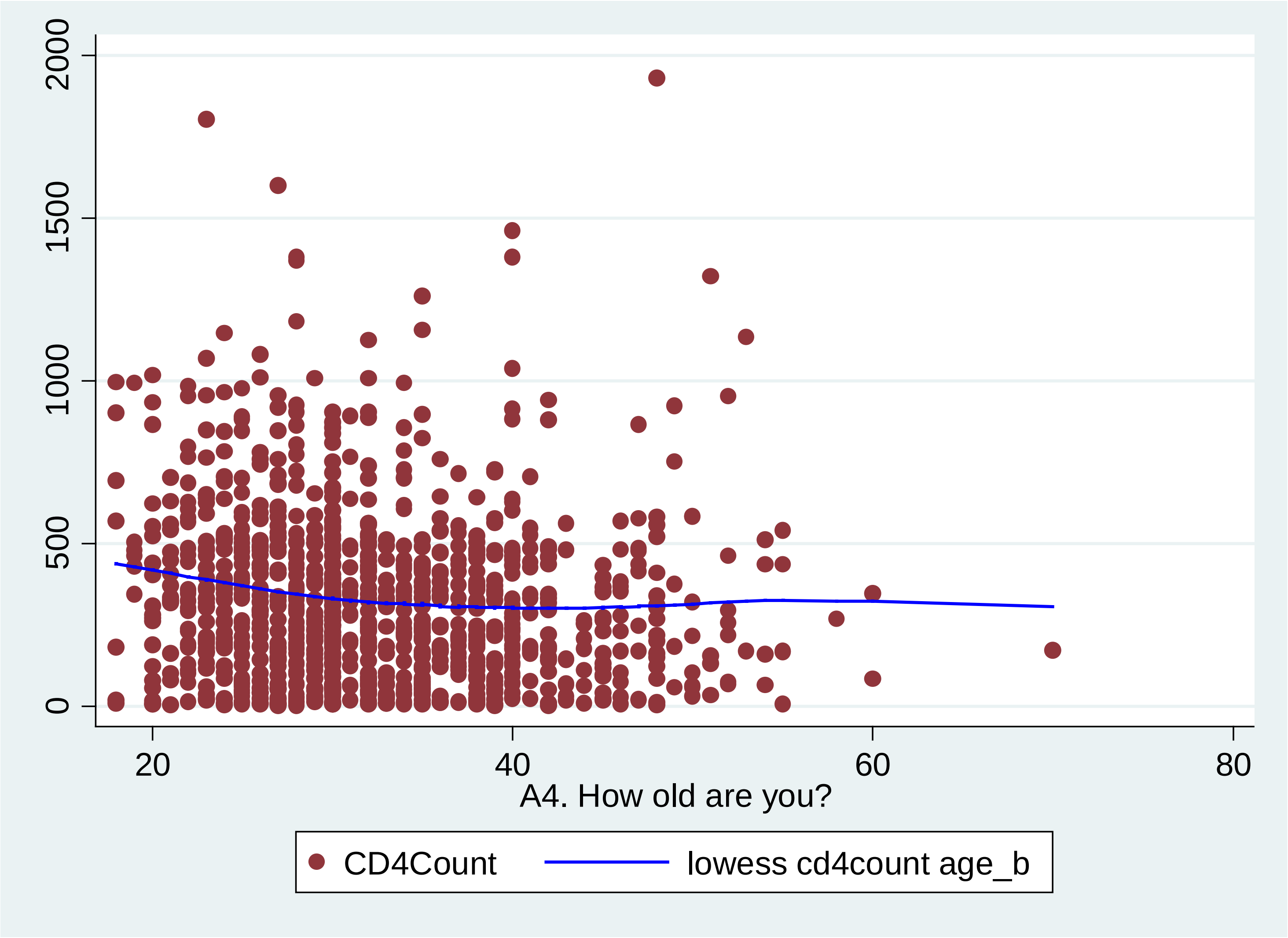

```
hist d1 if d1>=0 & d1<50, discrete fcolor(blue)
title(Lifetime number of sex partners)
```

- For numerical variables, the mode is dependent on the bin width




```
.hist a4, width(2) fcolor(blue) title(Age with bin width=2)
name(age_2, replace)
.hist a4, width(5) fcolor(blue) title(Age with bin width=5)
name(age_5, replace)
.graph combine age_2 age_5
```

Scatter plots – 2 numerical variables




```
twoway (scatter cd4count age, color(maroon)) (lowess
cd4count age, lcolor(blue))
```

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 and16th observation
  - Median CD4 cell count in previous data set = 283

In Stata

```
. 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 (e.g. age range 18-80 or range=62)
    - CD4 cell count range: (0-1932)
- Interquartile range (IQR)
  - 25th and 75th percentiles (e.g. IQR for age: 24-38) or difference (e.g. 14)
  - 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

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

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.

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

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

Numerical variable summaries

- Sample variance
  - Amount of spread around the mean

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

Note: $\sum_{i=1}^{n}(x_i - \bar{x})^2 = (x_1 + x_2 + x_3 + x_4 + \ldots - n\bar{X})^2$

Numerical variable summaries

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

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

- SD of CD4 cell count = 266.1 cells/mm^3
- 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)

CV for CD4 count

```
. summ 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
```

CV for age

```
. summm age, detail

                    A4. How old are you?
-------------------------------------------------------------
      Percentiles      Smallest
 1%           18             18
 5%           20             18
10%           21             18       Obs                3387
25%           24             18       Sum of Wgt.        3387

50%           30                      Mean           31.72808
                        Largest       Std. Dev.      9.850006
75%           38             75
90%           46             75       Variance       97.02261
95%           50             78       Skewness       1.030799
99%           60             80       Kurtosis       3.975972
```

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 = *the midpoint of the* i^{th} *group*

  f_i = *the frequency in the* i^{th} *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 cells/mm^3 (mean from original data was 329.2)

Grouped standard deviation

- The standard deviation

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

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

(SD from original data was 266.1)

Class survey data analysis

- https://ucsf.co1.qualtrics.com/SE/?SID=SV_3pG8FBLuMhMESfH
-
- Histogram , boxplot of coinvalue, cashvalue
- Mode, Median, 25^{th} percentile, 75^{th} percentile
- Mean, SD
- Differ by gender? Smartphone ownership? Height?

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