Biostat 200: Introduction to Biostatistics

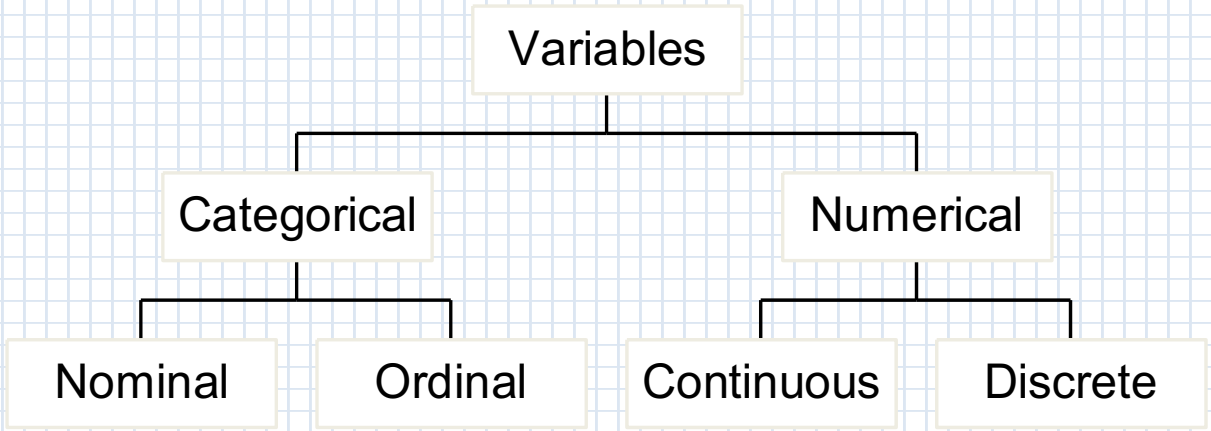
Module 1: Variables, frequencies, & summaries

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

"No amount of experimentation can ever prove me right; a single experiment can prove me wrong." Albert Einstein (1879-1955)

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 numeric (values have no numerical meaning)
  - Nominal variables: The data are unordered
    - For example:
      - gender, race, drug, disease status
  - 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

- Numeric (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)

The Grey zone

- Making continuous variables into categories
  - Age=33 years old (really 33 years, 17 days, 12 hours, etc…)
- 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
- Ordinal and/or Discrete?
  - We may count the number of children you have (discrete)
  - We may rank families by number of children (ordinal)

Why does it matter?

Manipulation of variables

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

Why would you do this?

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 - 2019	
	n (%)
Male	
Female	
Trans or other	
Total	

Class2019.dta

```
. 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

- 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%
    - Quintiles – 20%

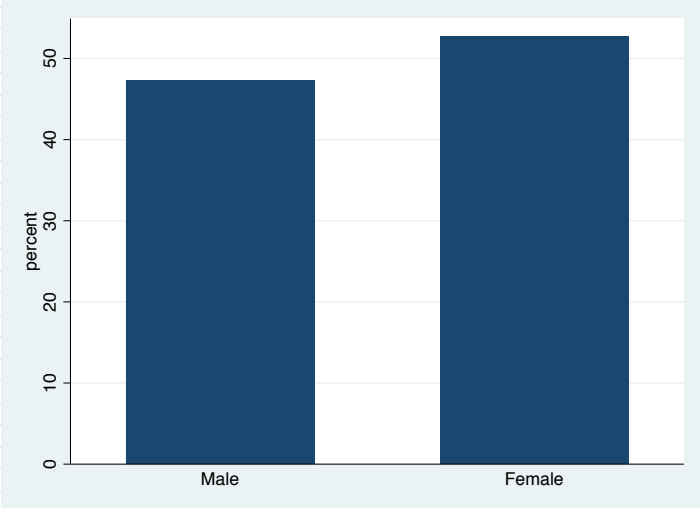
```
. tab age_cat
```

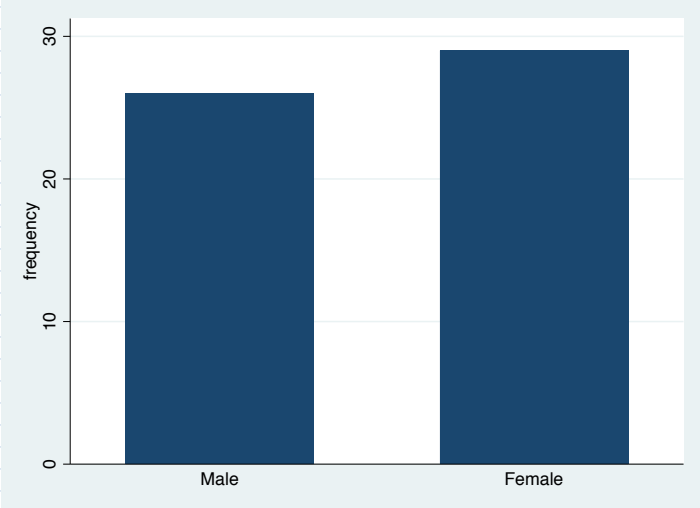
RECODE of age (age – What is your age?)	Freq.	Percent	Cum.
20–30	3	5.45	5.45
30–40	47	85.45	90.91
40–50	5	9.09	100.00
Total	55	100.00	

Are there better ways 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)





graph bar, 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 owned by class members? Mode of age?

- **Is it useful?**

Example of Modal Values

```
. tab apple_n
```

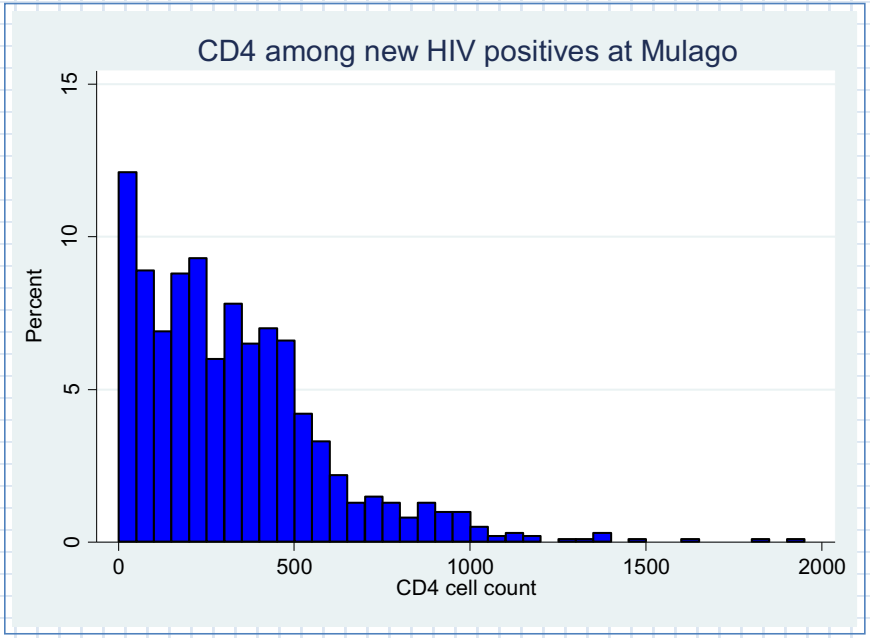
apple_n – How many currently functioning Apple devices do you own?	Freq.	Percent	Cum.
0	3	5.45	5.45
1	4	7.27	12.73
2	7	12.73	25.45
3	22	40.00	65.45
4	13	23.64	89.09
5	2	3.64	92.73
6	4	7.27	100.00
Total	55	100.00	

```
. 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 numeric data

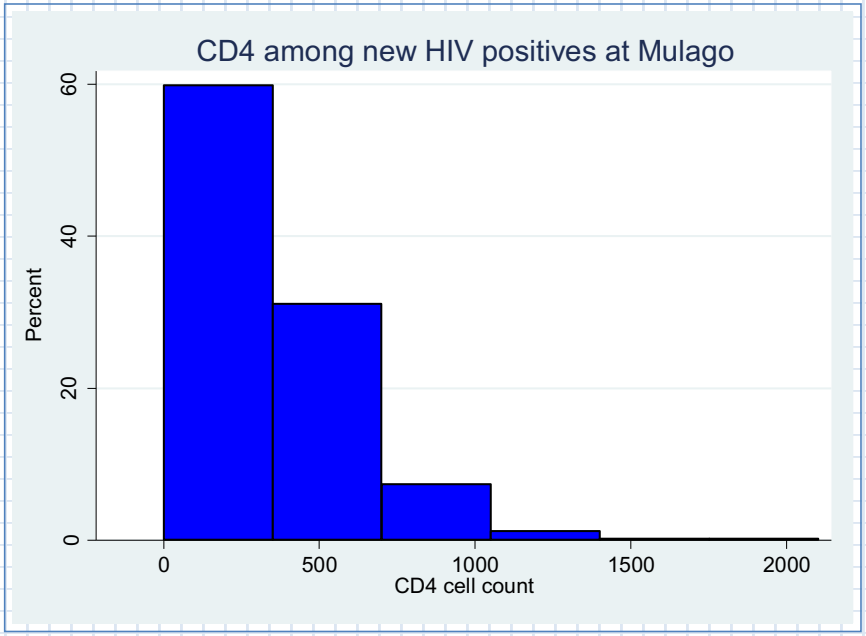
- The number of bins and the bin width will make a difference in the appearance of this plot
- Options such as percent, frequency will change the y-axis




```
USING vct.dta
** 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
```

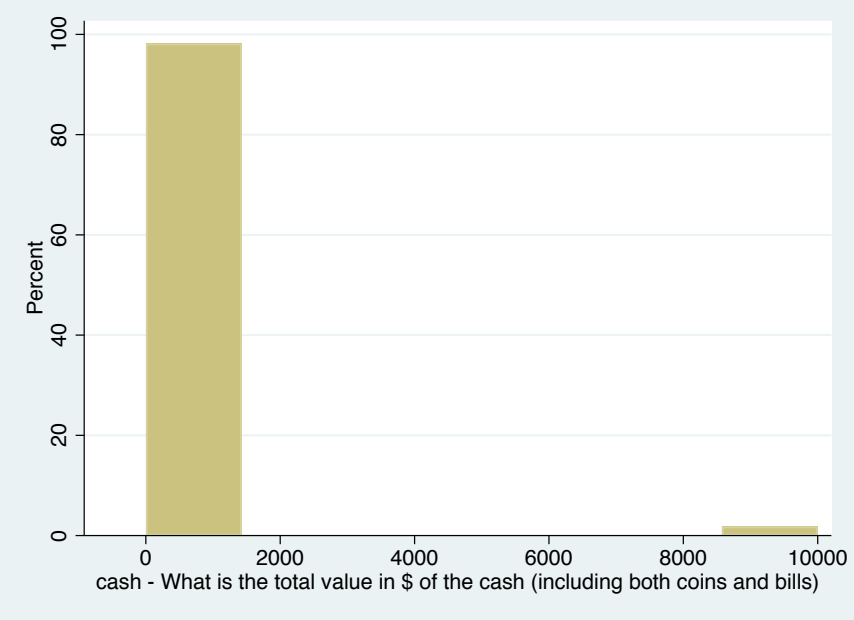
Histograms: changing the width

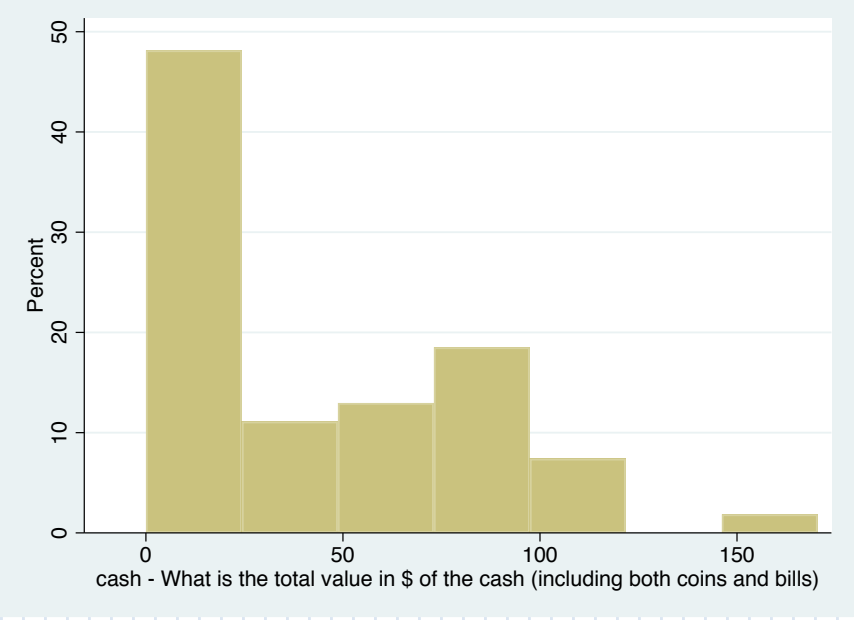
- This histogram has less detail but gives us the % of persons with CD4 <350 cells/mm^3




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

Histograms may be misleading
Cash in pocket (Biostat 200-2019)





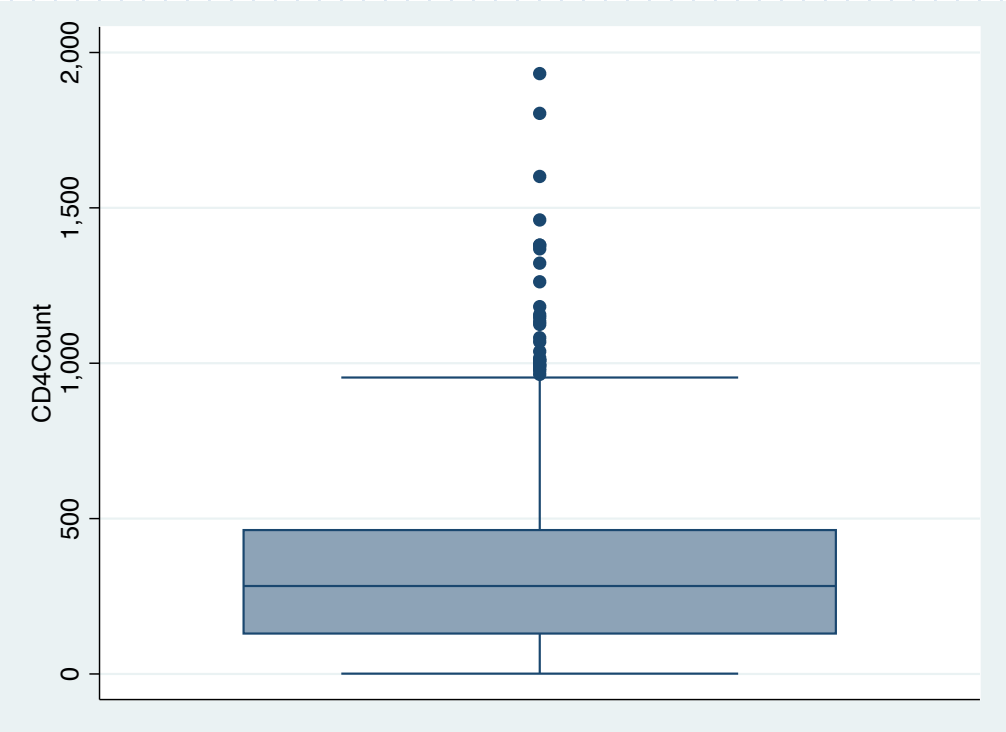
histogram cash, percent

Something went wrong here

A better way to categorize this data?

Box plots – who invented them?

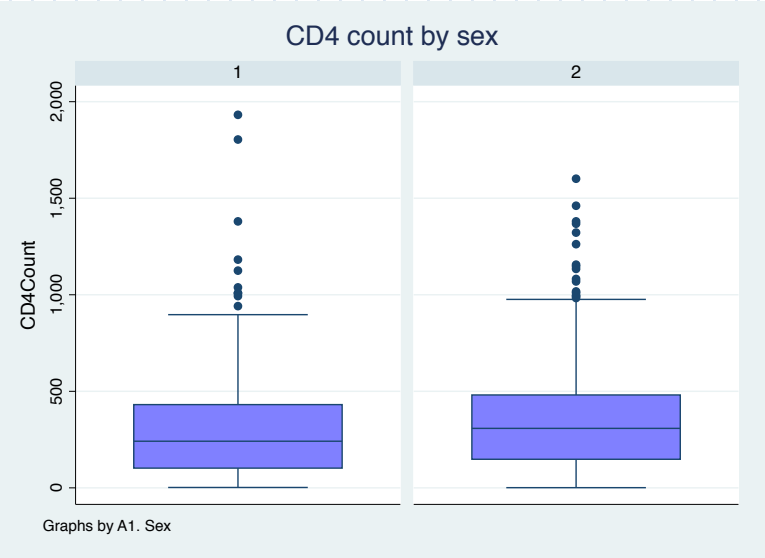
- Middle line=median (50th percentile)
- Box covers the 25th to 75th percentiles (interquartile range)
- Bottom whisker: Data point at or above 25th percentile – 1.5*IQR or the minimum, whichever is greater
  - 25th %ile=130, IQR=333 → 130-1.5*333=-369.5 (Hinge spread)
- Top whisker: Data point at or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



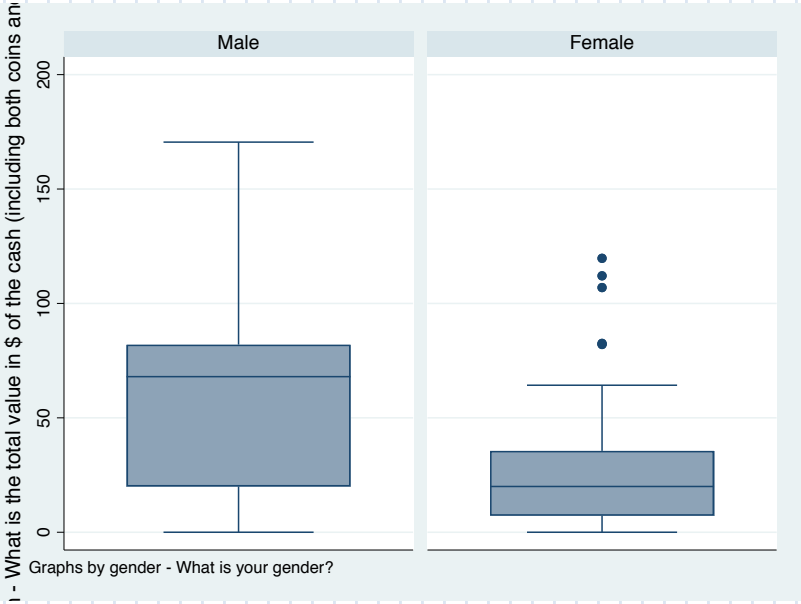

```
graph box cd4count
```

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

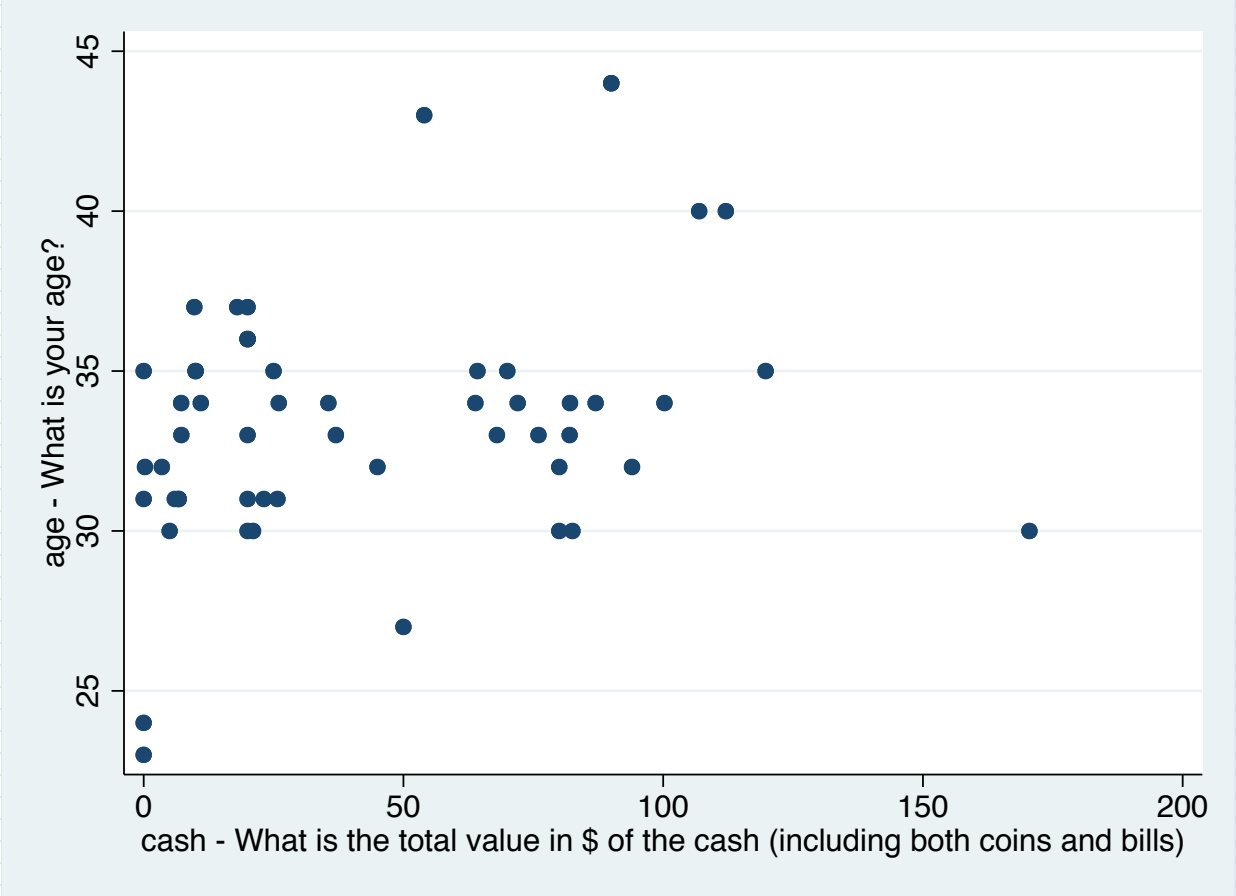


graph box cd4count, box(1, fcolor(blue)) by(, title(CD4 count by sex)) by(sex) graphregion(fcolor(blue))



graph box cash, by(gender)

Scatter plots – graphing 2 numeric variables



Are these two variables related to each other?

Can I predict age by your cash on hand?

twoway (scatter age cash)

Numeric variable summaries

- Measures of central tendency – where is the center of the data? Median/Mode/Mean

Median

50th percentile = 23.375

Range

Minimum to maximum or difference

0 to 170.50

Interquartile range (IQR)

25th and 75th percentiles or difference

Less sensitive to extreme values

10 to 80

```
. 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
```

Numeric 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

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

  - Sample variance
    - Amount of spread around the mean
    - How variable is the data?

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

Note that this sample variance divides by n-1, not n, the sample size

Numeric 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/mm^3
- SD of Age = 9.9 years

Back to the 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
- You can summarize but how do we find the values?
- How does it compare to the actual 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

- Use 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)

Recap & Next Up

- Types of variables and how to describe them
- Tables & graphs
- Statistics to summarize numerical data

NEXT!!

- Probability & Estimation
- Read the articles on probability for next tune
  - *The Collins Case*
  - *What are the chances?*