Stat 104: Quantitative Methods
Class 4: Descriptive Statistics, Part I

Today's Topics

- Graphical Methods and Frequency Tables
- Center of a Distribution
- Quartiles
- Spread of a Distribution

Describing Data : Graphically and Numerically

- Assuming we have generated a *random sample of data from some population* (everyone in the population is equally likely to be selected), we need methods to analyze the data, both graphically and numerically.
- This section of the notes presents two graphical techniques (the **dotplot** and **histogram**) and several summary measures (**mean** and **median, standard deviation)** for dealing with **sample data**.

Types of (numeric) data

- Data are the statistician's raw material, the numbers we use to interpret reality. All statistical problems involve either the collection, description and analysis of data, or thinking about the collection, description and analysis of data.
- There are many aspects of data. Data may be univariate (one variable per case) or multivariate (more than one variable per case). There are also different types of data: discrete and continuous.

How many cats do you own ?

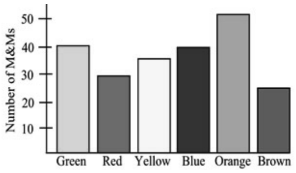
What is your weight ?

Pause: We care about numeric data

- Eventually we'll get to categorical data but for the first 80% of the class we deal with numerical valued data.
- A categorical variable is a variable that takes on a few discrete values that usually have no natural numerical coding

Example

Number of colors in bag of M+M Candies



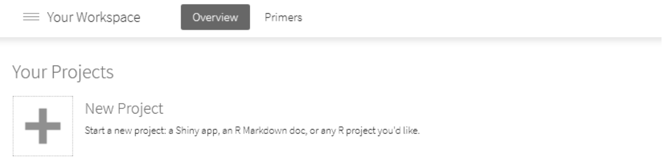
General Characteristics:

- Column label is categorical variable (colors).
- Column height is size of the group.
- Columns separated by space.
- Since this data is categorical, the only possible calculation is finding the mode (the category that occurs the most).
- Order of categories doesn't matter.

Example and Quick R Intro

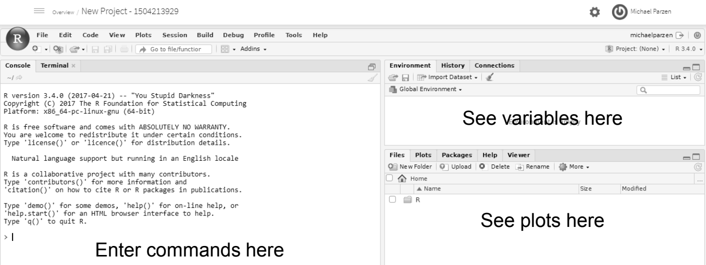
- Go to rstudio.cloud, create account and and log in
- Rstudio will be reviewed slowly in section this week and in Homework 0 and Homework 1. ☺

Log In and create a new project


Your Workspace
Overview
Primers
Your Projects
New Project
Start a new project: a Shiny app, an R Markdown doc, or any R project you'd like.

Click on the plus sign….then be a little patient as this is in alpha….

Eventually you will see this



```
R version 3.4.0 (2017-04-21) -- "You Stupid Darkness"
Copyright (C) 2017 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> |
```

See variables here

See plots here

Enter commands here

Let's do it

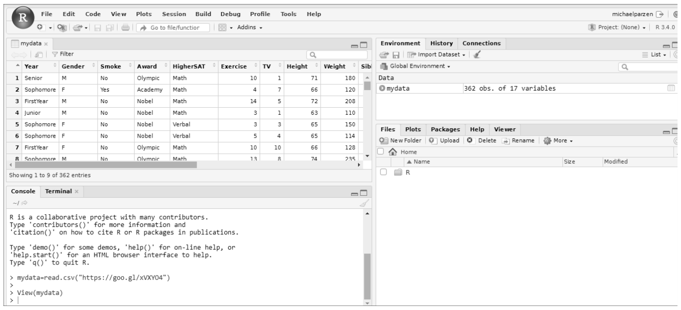
- All our data sets in this class are going to be read in via the internet (or typed in by hand if super small).
- Enter the command

```
mydata=read.csv("https://goo.gl/xVXYO4")
```

- Examine the data with the command

```
View(mydata)
```

Screen now looks like this



	Year	Gender	Smoke	Award	HigherSAT	Exercise	TV	Height	Weight	Sib
1	Senior	M	No	Olympic	Math	10	1	71	180	
2	Sophomore	F	Yes	Academy	Math	4	7	66	120	
3	FirstYear	M	No	Nobel	Math	14	5	72	208	
4	Junior	M	No	Nobel	Math	3	1	63	110	
5	Sophomore	F	No	Nobel	Verbal	3	3	65	150	
6	Sophomore	F	No	Nobel	Verbal	5	4	65	114	
7	FirstYear	F	No	Olympic	Math	10	10	66	128	
8	Sophomore	M	No	Olympic	Math	13	8	74	235	

Showing 1 to 9 of 362 entries

Data

mydata	362 obs. of 17 variables

```
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> mydata=read.csv("https://goo.gl/xVXYO4")
>
> View(mydata)
> |
```

Accessing Variables

- There are several ways to access variables as will be shown in section.
- For now we will use the convention

`mydata$varname`

Examine the variables

- It appears that Year (what year of college) is a categorical variable.
- It appears Award (would you rather win an Academy Award, Olympic Medal, or Nobel Peace Prize) is also a categorical variable.

Simple Command: `table`

- The `table` command tells us the number in each category.

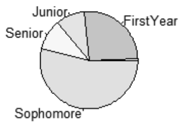
Note 2 are in a “blank” category-missing data

```
> table(mydata$Year)
         FirstYear     Junior     Senior Sophomore
       2        94         35         36       195

> table(mydata$Award)
Academy    Nobel Olympic
     31      149     182
```

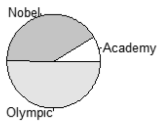
Can graph the data

- Pie chart

```
pie(table(mydata$Year))
```



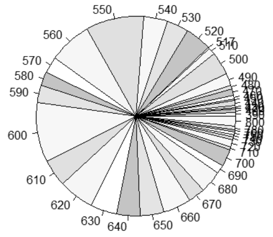
```
pie(table(mydata$Award))
```



See http://www.statmethods.net/graphs/pie.html for lots of options like adding the values to the chart.

Why is this silly?

```
pie(table(mydata$Verbal))
```



Some numeric data to play with

- A young college graduate in his first teaching assignment in a local school district decided to introduce his first grade class to statistics.
- As a class exercise, he brought in a scale and a tape measure and recorded the height (in inches) and weights (in pounds) of all 73 students in his two homeroom classes.

Here is the class data

62.00 49.00	37.50 42.00	56.00 46.00	38.00 40.25	43.50 45.50
46.00 51.00	38.50 39.00	40.00 42.00	59.00 42.00	48.00 43.75
48.50 42.00	47.00 44.50	48.00 46.00	52.50 43.00	34.00 40.00
32.00 40.00	39.75 42.50	46.50 44.50	42.75 44.00	50.00 47.25
63.50 45.50	60.00 46.00	72.00 45.50	31.50 43.50	35.00 38.50
41.25 46.25	41.00 42.00	31.00 42.00	43.50 44.00	49.00 46.00
40.00 42.25	41.00 45.25	48.00 48.00	40.00 44.00	46.25 46.25
34.25 40.75	30.00 40.25	36.50 43.25	40.50 43.50	43.50 44.50
34.75 41.25	45.00 43.00	43.75 45.75	60.00 44.00	37.25 40.25
43.50 42.00	51.00 45.00	34.25 41.50	57.50 39.50	39.00 31.00
46.00 44.50	35.25 43.25	41.25 45.25	48.75 45.25	34.50 39.00
42.50 43.25	40.50 46.25	41.75 41.00	44.50 45.50	47.50 44.50
53.00 49.00	39.50 44.25	45.25 44.50	49.50 46.00	42.00 49.00
43.50 46.00	36.00 40.00	43.50 43.50	33.75 42.50	45.50 44.50
36.50 43.50	53.00 43.00	38.50 41.00		

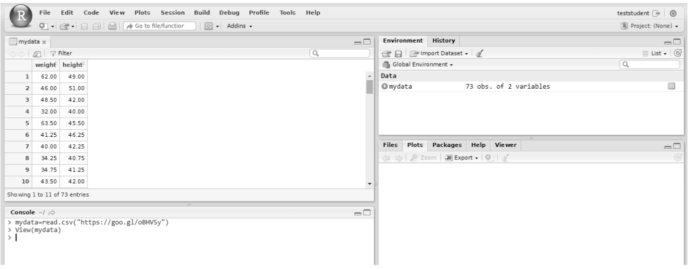
Without reading ahead, which column is weight and which is height ??

Back to Rstudio

- We’re just going to write over our previous dataset [or can say `rm(varname)`] to remove variables.
- We enter the commands

```
mydata=read.csv("https://goo.gl/oBHV5y")

View(mydata)
```

Our Rstudio screen looks as follows



	weight	height
1	62.00	49.00
2	46.00	51.00
3	48.50	42.00
4	32.00	40.00
5	63.50	45.50
6	41.25	46.25
7	40.00	42.25
8	34.25	40.75
9	34.75	41.25
10	43.50	42.00

Showing 1 to 11 of 73 entries

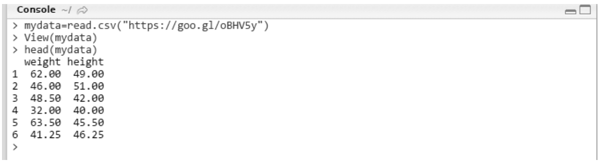
Data
mydata 73 obs. of 2 variables

Console ~/

```
> mydata=read.csv("https://goo.gl/oBHV5y")
> View(mydata)
> |
```

The head Command

- If you just want to see what your data looks like to verify it was loaded correctly:


Console ~/

```
> mydata=read.csv("https://goo.gl/oBHV5y")
> View(mydata)
> head(mydata)
  weight height
1  62.00  49.00
2  46.00  51.00
3  48.50  42.00
4  32.00  40.00
5  63.50  45.50
6  41.25  46.25
>
```

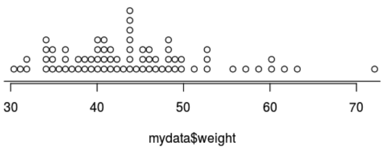
The Dotplot

- Clearly we need a graphical way to display the data so we can see what is going on.
- One simple graphical method is called the dotplot; one dot per student goes over each students reported weight and height.

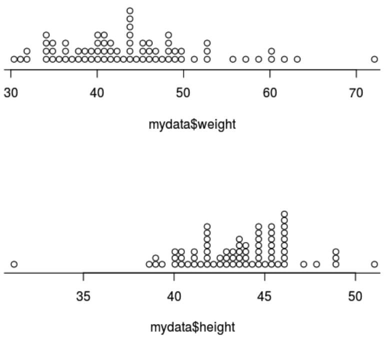
The Dotplot

- Note to do a dotplot in R we need to load a package in (this is typical)

```
> install.packages("BHH2")##only need to do this once
> library(BHH2) ## only need to do this once
> dotPlot(mydata$weight)
```



Dotplots of Weight and Height

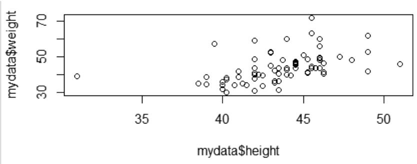


Any comments about skewness or symmetry?

Scatter Plot

- Eventually we will study how to find relationships between 2 (or more) variables. Visually we start with what are called scatter plots.

```
plot(mydata$height,mydata$weight)
```



Frequency Table

- We can also summarize the data with a frequency table.
- Divide the number line into intervals and count the number of student weights within each interval. The frequency is the count in any given interval.
- The relative frequency is the proportion of weights in each interval; i.e. it's the frequency divided by the total number of students.
- The cumulative relative frequency is the cumulative sum of the relative frequencies up to some class interval.

Here is the set-up for the weight data:

Class Interval		Freq.	Rel. Freq.		Cum. Rel. Freq.	
25-30		1	0.0137		0.0136	
30-35		10	0.137		0.1505	
35-40		15	0.2055		0.3562	
40-45		19	0.2603		0.6164	
45-50		16	0.2192		0.8356	
50-55		4	0.0548		0.8904	
55-60		5	0.0685		0.9589	
60-65		2	0.0274		0.9863	
65-70		0	0		0.9863	
70-75		1	0.0137		1	
Total		73	1		1	

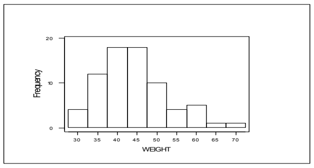
When possible, use equal width intervals.

Histogram

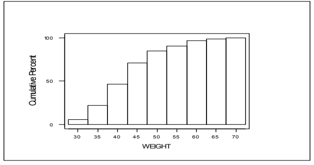
- In the frequency table, we are showing how many data points are in each interval. We can graph this information also.
- The resulting bar graph is called a histogram. Each bar covers an interval and is centered at the midpoint. The bar's height is the number of data points in each interval.

Histogram

Histogram

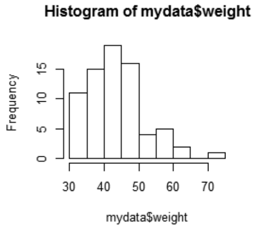


Cumulative frequency plot

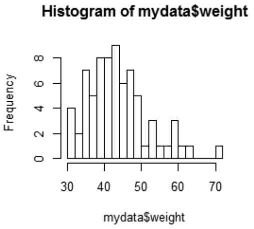


Doing this in R: `hist(varname)`

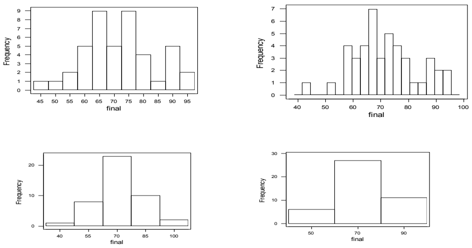
```
hist(mydata$weight)
```



```
hist(mydata$weight,breaks=20)
```



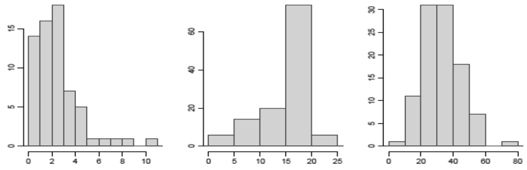
Histogram Issues



All 4 histograms are for the same data set!

Data Skewness

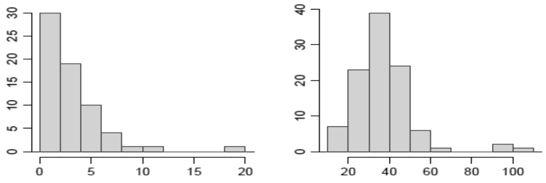
Is the histogram *right skewed*, *left skewed*, or *symmetric*?



We will define these ideas more formally in the next class

Unusual Observations

Are there any unusual observations or potential *outliers*?



We will define these ideas more formally in the next class

Summary Measures

Describing Data Numerically

Center and Location

Mean

Median

Other Measures of Location

Quartiles

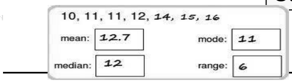
Variation

Range

Interquartile Range

Variance

Standard Deviation


10, 11, 11, 12, 14, 15, 16

mean: 12.7 mode: 11

median: 12 range: 6

Data set for location examples

- The data we shall use for examples comes from last year’s class survey, where we asked how much money student’s spent on their last haircut, including tip.

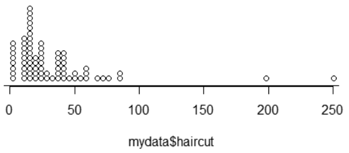
Make money on campus!

Get a flowbee!

```
mydata=read.csv("https://goo.gl/f4EYpG")
```

Dotplot of the data

```
dotPlot(mydata$haircut)
```



A little notation before we begin

- Suppose we collect n pieces of data. We need some way of describing the data. We write

$$x_1, x_2, \cdots, x_n$$

as the values we observe. Thus, n is the total number of data points and x_4, say, is the value of the fourth data point.

Example

We ask 5 people how many hours of tv they watch a week and we obtain the following data:

Person Number	1	2	3	4	5
Hours of TV	5	7	3	38	6

Then

$$x_1 = 5, x_2 = 7, x_3 = 3, x_4 = 38, x_5 = 6.$$

Mean (Arithmetic Average)

- The (sample) Mean is the arithmetic average of data values

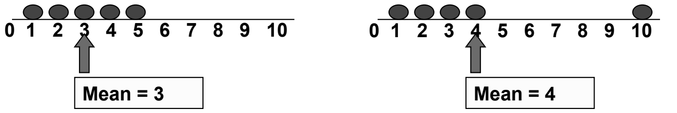
n = Sample Size

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

See website for explanation of summation notation

Mean (Arithmetic Average)

- The most common measure of central tendency
- Mean = sum of values divided by the number of values
- **Affected by extreme values (outliers)**



$$\frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$

$$\frac{1+2+3+4+10}{5} = \frac{20}{5} = 4$$

Calculating the mean in R

- Since R is a programming language there are several ways to calculate the mean.

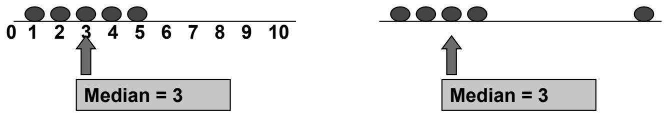
```
> mean(mydata$haircut)
[1] 32.21284

> summary(mydata$haircut)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   15.00   21.50   32.21   40.00  250.00

> library(psych)
> describe(mydata$haircut)
   vars  n  mean    sd median trimmed   mad min max range skew kurtosis   se
X1    1 74 32.21 38.36   21.5   25.85 17.05   0 250   250 3.63     16.2 4.46
```

Think About It

- Can the mean be larger than the maximum value or less than the minimum value?
- Can the mean be the minimum value? Can the mean be the maximum value.
- Can the mean be not equal to any value in the sample?

The Median

- Not affected by extreme values



Median = 3

Median = 3

- In an ordered list, the median is the “middle” number
  - ❑ If n is odd, the median is the middle number
  - ❑ If n is even, the median is the average of the two middle numbers

Calculating the median in R

- Since R is a programming language there are several ways to calculate the mean.

```
> median(mydata$haircut)
[1] 21.5

> summary(mydata$haircut)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   15.00   21.50   32.21   40.00  250.00

> library(psych)
> describe(mydata$haircut)
   vars  n  mean    sd median trimmed   mad min max range skew kurtosis   se
X1    1 74 32.21 38.36   21.5   25.85 17.05   0 250   250 3.63     16.2 4.46
```

For the haircut data set, the median is a lot less than the mean. Why is this ??

Mean versus Median

Although both the mean and median are good measures of the center of a distribution of measurements, the median is *less sensitive* to extreme values. The mean shifts towards extreme values in the data set.

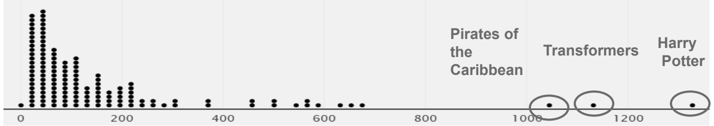
The median is not affected by extreme values since the numerical values of the measurements are not used in its computation.

Example:

1,2,3,4,5 Mean = 3 Median = 3

1,2,3,4,100 Mean = 22 Median = 3

Outliers



2011 Movies World Gross

(in millions)

Resistance or Robust

A statistic is ***resistant*** (robust) if it is relatively unaffected by extreme values.

- The median is resistant while the mean is not.

	Mean	Median
With Harry Potter	$150,742,300	$76,658,500
Without Harry Potter	$141,889,900	$75,009,000

Which measure of location is the “best”?

- **Mean** is generally used, unless extreme values (outliers) exist
- Then **median** is often used, since the median is not sensitive to extreme values.
  - ❑Example: Median home prices may be reported for a region – less sensitive to outliers

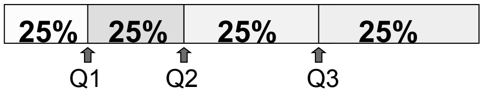
But is being average useful?

How useful are centers alone for conveying the true characteristics of a distribution?

http://www.youtube.com/watch?v=4B2xOvKFFz4

Other Location Measures: Quartiles

- Quartiles split the ranked data into 4 equal groups



Calculating Quartiles in R

- Aside-You shouldn’t care about this but there is no fully accepted way to calculate quartiles (think-is the median unique???).

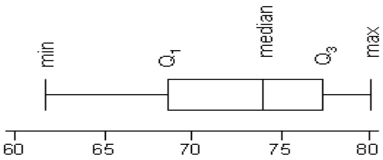
```
> summary(mydata$haircut)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   15.00   21.50   32.21   40.00  250.00
```

Interp: 25% of students pay less than $15 for a haircut, while 25% pay more than $40

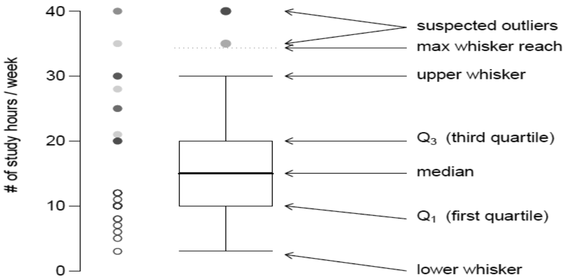
Box and Whisker Plot

- A Graphical display of data using a 5-number summary:

Minimum -- Q1 -- Median -- Q3 -- Maximum



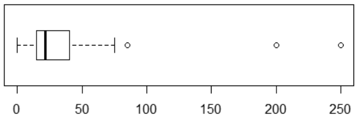
Anatomy of a Fancy Box Plot



We’ll discuss next time how the outliers are defined.

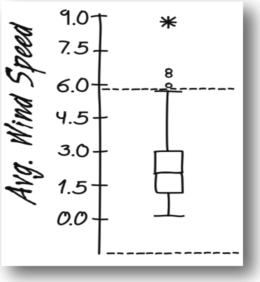
Boxplots in R

```
> boxplot(mydata$haircut,horizontal=TRUE)
```



What Do Boxplots Tell Us?

- The **center of the boxplot** shows us the middle half of the data between the quartiles.
- The **height of the box** is equal to the **IQR**.
- If the **median is roughly centered** between the quartiles, then the **middle half of the data is roughly symmetric**. Thus, if the **median is not centered**, the **distribution is skewed**.
- The **whiskers** also show the **skewness** if they are not the same length.
- **Outliers** are out of the way to keep you from judging skewness, but give them special attention.



Measure of Dispersion

The mean and median give us information about the central tendency of a set of observations, but these numbers shed no light on the dispersion, or spread of the data.

Example: Which data set is more variable ??

5,5,5,5,5 Mean = 5

1,3,5,8,8 Mean = 5

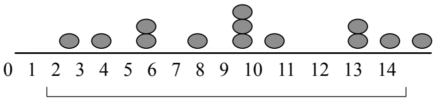
Measures of variation give information on the **spread** or **variability** of the data values.

Range

- Simplest measure of variation
- Difference between the largest and the smallest observations:

Range = $x_{maximum} - x_{minimum}$

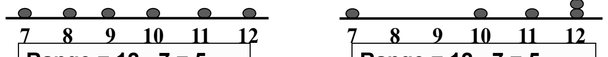
Example:



Range = 14 - 1 = 13

Disadvantages of the Range

- Ignores the way in which data are distributed



Range = 12 - 7 = 5 **Range = 12 - 7 = 5**

- Sensitive to outliers

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,5

Range = 5 - 1 = 4

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,120

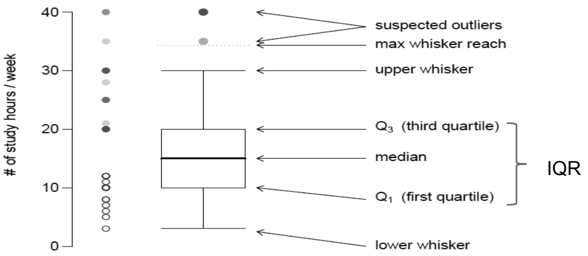
Range = 120 - 1 = 119

Interquartile Range (IQR)

- Can eliminate some outlier problems by using the **interquartile range**
- Eliminate some high-and low-valued observations and calculate the range from the remaining values.
- Interquartile range = 3rd quartile – 1st quartile

Interquartile Range

•Developed by John Tukey, the founder of EDA (exploratory data analysis)
•Doesn't take into account all your data-not used that much



Example: Haircut Data Again

```
> summary(mydata$haircut)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   15.00   21.50   32.21   40.00  250.00
```

IQR=40-15 = 25

```
> IQR(mydata$haircut)
[1] 25
```

How should we measure variability ?

The basic idea is to view variability in terms of distance between each measurement and the mean.

A natural measure of dispersion is to calculate the average distance all the observations are from the center of the data:

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

Is this a good measure of dispersion ? No, its horrible. Any idea why ?

Distance from a Fixed Point

- We can think of a measure of spread as average distance-like what is the average distance everyone lives from the Science Center.
- Say this average value is 1 mile. Then if you live less than 1 mile from the Science Center you realize you are closer than a lot of your fellow students, and if you live 20 miles away you know you are an outlier.

Distance Has to be Positive

- We know that distance can't be negative-that is, if you live north of the SC you are positive miles away and south of the SC you are negative miles away.
- But this spread formula doesn't know that-it just takes the difference between each value and the mean, which could result in negative or positive numbers.

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

In fact, this formula always returns a value of 0!

Does anyone have a calculator ?

- We need 3 numbers
- X1 = X2 = X2 =
- Calculate the mean =
- Now calculate

$$spread = \frac{1}{3}\sum_{i=1}^{3}(x_i - \bar{x})$$

One Solution: Mean Absolute Deviation (MAD)

- One way to get rid of negative distances is by using absolute values.
- The Mean Absolute Deviation (MAD) of a data set is defined to be

$$MAD = \frac{1}{n}\sum_{i=1}^{n}| x_i - \bar{x} |$$

- What are the units of MAD?
- Do people use it?

MAD for haircut data

- The function `mad` in R is not our mad; it’s a complicated robust measure of location.
- We can compute our MAD in R as follows (don’t worry about the code)

```
mean(abs(mydata$haircut-mean(mydata$haircut)))
[1] 22.38523
```

- The MAD for the haircut data is then $22.39
- This is very close to the IQR=$25. Hmmm

Another Solution: The Variance

The variance of a set of data is defined as

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

We use n-1 instead of n for technical reasons that will be discussed later-you could divide by “n”; “n-1” is just better.

What practical significance can be attached to the variance as a measure of variability ? Large variances imply a large amount of variation, but what constitutes large ?

The answer will appear in a few slides.

The variance of the haircut data is 1471.86. Yikes!! That seems like a pretty big number.

```
> var(mydata$haircut)
[1] 1471.866
```

What are the units of this number anyway ??

A measure of spread should have the same units as the original data. In the salary example, the variance is measured in dollars squared.

What can we do to get back to our original units??

Standard Deviation

- Most commonly used measure of variation
- Shows variation about the mean
- Has the same units as the original data

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

The standard deviation for the haircut data is $38.36 which still seems large, reflecting the wide spread in the data.

```
> var(mydata$haircut)
[1] 1471.866
> sd(mydata$haircut)
[1] 38.3649
> describe(mydata$haircut)
   vars  n  mean    sd median trimmed   mad min max range skew kurtosis   se
X1    1 74 32.21 38.36   21.5   25.85 17.05   0 250   250 3.63     16.2 4.46
```

Actually, how we determine if a std dev is “large” or “small” is something we will discuss in the next class.

Why is the std dev a lot larger than MAD or IQR?

Standard Deviation-a Measure of Risk?

- Standard deviation measures spread of a data set, so it seems natural for financial instruments to say the higher the standard deviation the riskier the asset.
- This can work, in that generally the higher the standard deviation the riskier the investment, but it does have some problems and you should keep these issues in mind.

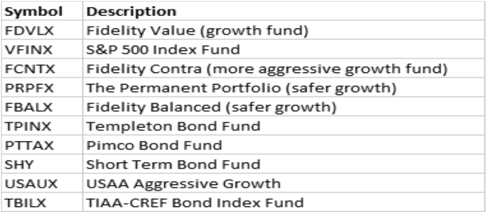
Comparing can be difficult

- Manager 1 makes a 2% return every month.
- Manager 2 makes a -2% return every month.
- If you compare them using standard deviation, who is better?

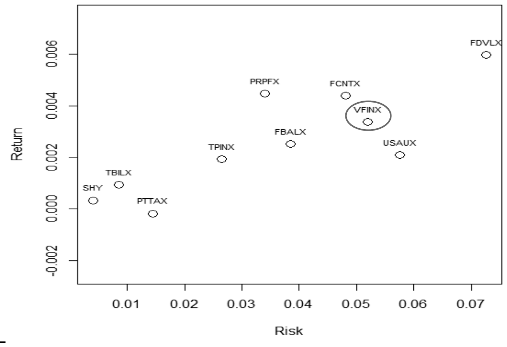
Finance Example : Comparing Mutual Funds

Let's use means and sd's to compare mutual funds. For 10 different assets we compute the mean and sd. Then plot mean vs sd.

The assets are:



Symbol	Description
FDVLX	Fidelity Value (growth fund)
VFINX	S&P 500 Index Fund
FCNTX	Fidelity Contra (more aggressive growth fund)
PRPFX	The Permanent Portfolio (safer growth)
FBALX	Fidelity Balanced (safer growth)
TPINX	Templeton Bond Fund
PTTAX	Pimco Bond Fund
SHY	Short Term Bond Fund
USAUX	USAA Aggressive Growth
TBILX	TIAA-CREF Bond Index Fund



Today's Tools

- New toolbox additions
  - ❑ Dotplot and Histograms
  - ❑ Summary Statistics (mean, median, std dev)

Things you should know

- Dotplot, Histogram
- Measures of center and location
  - ❑ Mean, median, mode
- Quartiles
- Variance and Standard Deviation
- Be wary of standard deviation