Stat 104: Quantitative Methods

Class 5: Descriptive Statistics, Part II

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

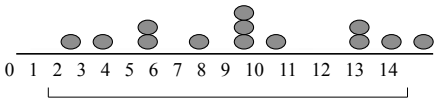
Measures of variation give information on the **spread** or **variability** of the data values.

Range

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

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

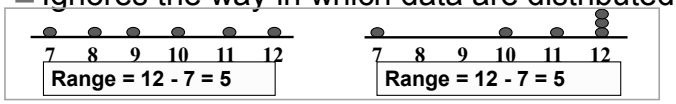
Example:



Range = 14 - 1 = 13

Disadvantages of the Range

- Ignores the way in which data are distributed



- 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

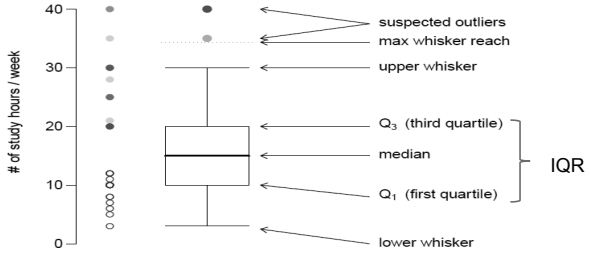
Range = 120 - 1 = 119

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 different, 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

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

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

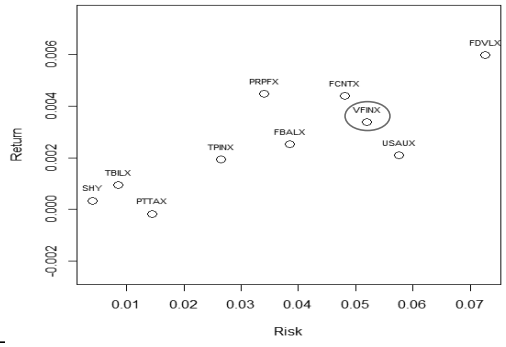
- Manager 1 makes a 2% return every month.
- Manager 2 makes a -2% return every month.
- If you compare them using standard deviation, who is better?

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



Some Tools so Far

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

Shifting and Rescaling Data

- Original data $x_1, x_2, \ldots x_n$
- Linear Transformation

$$Y_i = a + bX_i$$

Shifts data by a — **Changes scale**

- Linear Transformations do not change the shape of the data distribution, but do change the center and spread.

Examples

Examples: Changing

1. from feet (x) to inches (y): y=12x
2. from dollars (x) to cents (y): y=100x
3. from degrees celsius (x) to degrees fahrenheit (y): y = 32 + (9/5)x
4. from ACT (x) to SAT (y): y=150+40x
5. from inches (x) to centimeters (y): y = 2.54x

Linear Transformations (a+bX rule)

- The mean and variance of a data set have two interesting properties.
- These properties occur when one shifts a data set, or multiplies by a value (expands or contracts a data set).

$$Var(a+bX) = b^2 Var(X)$$

$$Average(a+bX) = a + b[Average(X)]$$

Effects of Linear Transformations

- Your Transformation: y = a + b*x
- $mean_{new}$ = a + b*mean
- $median_{new}$ = a + b*median
- $stdev_{new}$ = |b|*stdev
- IQR_{new} = |b|*IQR

Example

- Winter temperature recorded in Fahrenheit
  - mean = 20
  - stdev = 10
  - median = 22
  - IQR = 11
- Convert into Celsius C=(5/9)F-17.78
  - mean = -17.78 + 5/9 * 20 = -6.67 C
  - stdev = 5/9 * 10 = 5.56
  - median = -17.78 + 5/9 * 22 = -5.56 C
  - IQR = (5/9)(11)=6.11

Example

Worksheet 1 ***

	C1	C2	C3	C4	C5	C6	C7	C8
	X	2X	-4X	X+2	X-1	-2X+1	0.5X-2	
1	0.419335	0.83867	-1.67734	2.41934	-0.580665	0.161329	-1.79033	
2	0.650765	1.30153	-2.60306	2.65076	-0.349235	-0.301529	-1.67462	
3	0.569212	1.13842	-2.27685	2.56921	-0.430788	-0.138424	-1.71539	
4	0.370595	0.74119	-1.48238	2.37059	-0.629405	0.258811	-1.81470	
5	0.313381	0.62676	-1.25353	2.31338	-0.686619	0.373237	-1.84331	
6	0.584198	1.16840	-2.33679	2.58420	-0.415802	-0.168395	-1.70790	
7	0.652854	1.30571	-2.61142	2.65285	-0.347146	-0.305708	-1.67357	
8	0.130632	0.26126	-0.52253	2.13063	-0.869368	0.738736	-1.93468	
9	0.693629	1.38726	-2.77452	2.69363	-0.306371	-0.387259	-1.65319	
10	0.679339	1.35868	-2.71736	2.67934	-0.320661	-0.358679	-1.66033	
11	0.792006	1.58401	-3.16802	2.79201	-0.207994	-0.584012	-1.60400	
12	0.446753	0.89351	-1.78701	2.44675	-0.553247	0.106494	-1.77662	

Descriptive Statistics: X, 2X, -4X, X+2, X-1, -2X+1, 0.5X-2

Variable	N	Mean	Median	TrMean	StDev	SE Mean
X	50	0.4695	0.4428	0.4685	0.2880	0.0407
2X	50	0.9389	0.8856	0.9370	0.5760	0.0815
-4X	50	-1.878	-1.771	-1.874	1.152	0.163
X+2	50	2.4695	2.4428	2.4685	0.2880	0.0407
X-1	50	-0.5305	-0.5572	-0.5315	0.2880	0.0407
-2X+1	50	0.0611	0.1144	0.0630	0.5760	0.0815
0.5X-2	50	-1.7653	-1.7786	-1.7657	0.1440	0.0204

The Most Common Linear Transformation

- The Z-score is a common linear transformation

$$z = \frac{X - \overline{X}}{S}$$

- **By "z scoring" a data set, the new data set will have mean 0 and variance 1**.
- The number of standard deviations a raw score (individual score) deviates from the mean.

Using Z's to compare values

- Since z-scores reflect how far a score is from the mean they are a good way to standardize scores.
- We can take any distribution and express all the values as z-scores (distances from the mean). So, no matter the scale we originally used to measure the variable, it will be expressed in a standard form.
- This standard form can be used to convert different scales to the same scale so that direct comparison of values from the two different distributions can be directly compared.

Used for Comparison Purposes

- Mary's ACT score is 26. Jason's SAT score is 900. Who did better?
- The mean SAT score is 1000 with a standard deviation of 100 SAT points.
- The mean ACT score is 22 with a standard deviation of 2 ACT points.

Calculate the Z-scores

Jason: $\frac{900-1000}{100} = -1$

Mary: $\frac{26-22}{2} = +2$

- From these findings, we gather that Jason's score is 1 standard deviation below the mean SAT score and Mary's score is 2 standard deviations above the mean ACT score.
- Therefore, Mary's score is relatively better.

Interpreting standard deviation

We now have the two summaries

$\bar{x}$ s_x

where the data is

how spread out, or variable the data is

The mean is pretty easy to understand. What are the units?

We know that the bigger s_x is, the more variable the data is, but how do we interpret the number ?

What is a big s_x , what is a small one ?

What are the units of s_x ?

Rule of thumb

The most basic analysis is to simply compare the value of the mean to the value of the standard deviation.

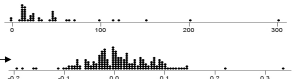
Intuitively, what do you think the following data sets look like ?

	$\bar{x}$	s	spread			
Data Set 1	50	0	none	small	medium	large
Data Set 2	50	3	none	small	medium	large
Data Set 3	50	14	none	small	medium	large
Data Set 4	50	42	none	small	medium	large
Data Set 4	50	1000	none	small	medium	large

The empirical rule will help us understand s_x and relate the summaries back to the dotplot (or histogram).

Empirical rule:

no



For "mound shaped data":

Approximately 68% of the data is in the interval

$$(\bar{x} - s_x, \bar{x} + s_x) = \bar{x} \pm s_x$$

Approximately 95% of the data is in the interval

$$(\bar{x} - 2s_x, \bar{x} + 2s_x) = \bar{x} \pm 2s_x$$

What good is the empirical rule again ?

Empirical Rule Example

- A survey of 1000 U.S. gas stations was conducted and you were told the average price charged for a gallon of regular gas was $3.90 with a std dev of $0.20.
- You were also told the data is mound shaped.
- What can you deduce?

You find +/- 2s in Many Places

Bollinger bands are $\bar{x} \pm 2s_x$ (based on a moving window of 20 time periods)

See http://en.wikipedia.org/wiki/Bollinger_Bands or take Stat 107

Don't fall in love with +/- 2s

- Standard deviation is a good measure of spread if your data is symmetric; if your data is not symmetric it really isn't interpretable.
- If your data is not symmetric, one needs to use Chebyshev's rule for interpreting the spread of your data.

Chebyshev's Rule

- For **any set of data** and for any number k, greater than one, the proportion of the data that lies within k standard deviations of the mean is at least:

$$1 - \frac{1}{k^2}$$

So for $(\bar{x} - 2s_x, \bar{x} + 2s_x) = \bar{x} \pm 2s_x$

- According to Chebyshev's Theorem, at least what fraction of the data falls within "k" (k = 2) standard deviations of the mean?
- At least $1 - \frac{1}{2^2} = \frac{3}{4} = 75\%$ **of the data falls within 2 standard deviations of the mean.**

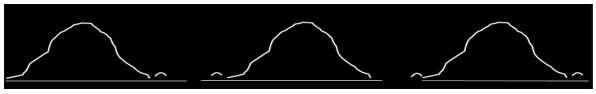
Hey, that's not 95% of the data. Exactly!

Detecting Outliers

- The detection of outliers is important for a variety of reasons.
- One rather mundane reason is that they can help identify erroneously recorded results.
- We have already seen that even a single outlier can grossly affect the sample mean and variance, and of course we do not want a typing error to substantially alter or color our perceptions of the data.
- So it can be prudent to check for outliers, and if any are found, make sure they are valid.

Outliers are Naughty

- Outliers can lead to too-high, too-low or nearly correct estimates of the population mean, depending upon the number and location of the outliers (asymmetrical vs. symmetrical patterns)
- Outliers always lead to overestimates of the standard deviation



Mean estimate is "too high" & std is overestimated

Mean estimate is "too low" & std is overestimated

Mean estimate is "right" & std is overestimated

Effect of Outliers on Summary Stats

Resistant to Outliers:	Median, IQR
***Not* Resistant to Outliers:**	Mean, Standard Deviation, Variance, Range

Classic Outlier Detection

- A classic outlier technique is to simply Zscore the data and declare any point an outlier if

$$|Z| = \left|\frac{X - \bar{X}}{s}\right| \geq 2$$

- The value 2 is motivated by the normal distribution that we will see in a few classes.

Example

- Consider the values

$$2,2,2,2,2,3,3,3,3,3,4,4,4,4,4,1000$$

- For this data mean=65.94 and s=249.1
- The Z-score for the point 1000 is

$$\left|\frac{1000 - 65.94}{249.1}\right| = 3.75$$

- So 1000 is declared an outlier.

Example

- Consider the data

$$2,2,3,3,3,4,4,4,100000,100000$$

- For this data mean=20002.5 and s=42162.38
- The Z-score for the point 100000 is

$$\left|\frac{1000000 - 20002.5}{42162.38}\right| = 1.897$$

- So 100000 is NOT declared an outlier.

Yuck

- The classic method would not declare the value 100,000 an outlier even though certainly it is highly unusual relative to the other eight values.
- The problem is that both the sample mean and the sample standard deviation are sensitive to outliers, which can effect our detection ability.
- An outlier detection technique is said to suffer from **masking** if the very presence of outliers causes them to be missed.

Example

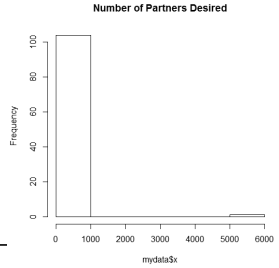
- Pedersen et al. (1998) conducted a study, a portion of which dealt with the sexual attitudes of undergraduate students.
- Among other things, the students were asked how many sexual partners they desired over the next 30 years.
- The responses of 105 males

Table 2.3 Responses by males in the sexual attitude study

6 1 1 3 1 1 1 1 1 1 6 1 1 1 4
5 3 9 1 1 1 5 12 10 4 2 1 1 4 45
8 5 0 1 150 13 19 2 1 18 3 1 3 1 11
1 2 1 1 1 12 1 1 2 6 1 1 1 1 4
1 150 6 40 4 30 10 1 1 0 3 4 1 4 7
1 10 0 19 1 9 1 1 1 5 0 1 1 15 4
1 4 1 1 11 1 1 30 12 6000 1 0 1 1 15

A Histogram


```
> mydata=read.csv("https://goo.gl/e8nYDF")
> hist(mydata$x,main="Number of Partners Desired")
```



Summary Statistics


```
> describe(mydata$x)
   vars   n  mean     sd median trimmed  mad min  max range skew kurtosis    se
X1    1 105 64.92 585.16      1    3.66 1.48   0 6000  6000 9.94    97.79 57.11
>
> summary(mydata$x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    1.00    1.00   64.92    6.00 6000.00
```

```
> sum(mydata$x<mean(mydata$x))
[1] 102
```

- The mean is not very typical since 102 of the 105 people surveyed gave a response less than the mean.

Outliers

- One participant surveyed responded he wanted 6000 sexual partners, over the next 30 years, which is clearly unusual compared to the other stat 100 students. Heck its unusual in general.
- Also, two gave the response 150, which again is unusual.
- For HW you will see that the 6000 is flagged as an outlier, but not the 150. Though it probably should be.

The Boxplot Rule

- One of the earliest improvements on the classic outlier detection rule is called the boxplot rule.
- It is based on the fundamental strategy of avoiding masking by replacing the mean and standard deviation with measures of location and dispersion that are relatively insensitive to outliers.

The BoxPlot Rule

- In particular, the boxplot rule declares the value X an outlier if

$$X < Q1 - 1.5(Q3 - Q1)$$

or

$$X > Q3 + 1.5(Q3 - Q1)$$

- So the rule is based on the lower and upper quartiles, as well as the interquartile range, which provide resistance to outliers.

Example

- Remember the sexual attitude data

```
> describe(mydata$x)
   vars   n  mean     sd median trimmed  mad min  max range skew kurtosis    se
X1    1 105 64.92 585.16      1    3.66 1.48   0 6000  6000 9.94    97.79 57.11
>
> summary(mydata$x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    1.00    1.00   64.92    6.00 6000.00
```

- Outlier if > 6+1.5(6-1)=13.5 so 12 points are flagged now instead of 1 as being outliers.

Outlier Detection in R

- Consider the following

```
> mydata=read.csv("http://people.fas.harvard.edu/~mparzen/stat104/cars10.csv")
> head(mydata)
           make price mpg headroom trunk weight length turn displacement
1   AMC Concord  4099  22      2.5    11   2930    186   40          121
2     AMC Pacer  4749  17      3.0    11   3350    173   40          258
3    AMC Spirit  3799  22      3.0    12   2640    168   35          121
4 Buick Century  4816  20      4.5    16   3250    196   40          196
5 Buick Electra  7827  15      4.0    20   4080    222   43          350
6 Buick LeSabre  5788  18      4.0    21   3670    218   43          231

> attach(mydata) ### this makes the variables directly available to us
```

Finding Outliers in R-Direct Method

- Consider the following

```
> summary(price)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   3291    4220    5006    6165    6332   15910

> price[price>6332+1.5*IQR(price)]
 [1] 10372 11385 14500 15906 11497 13594 13466 10371  9690  9735 12990 11995

> price[price<4220-1.5*IQR(price)]
integer(0)
```

Finding Outliers in R-Direct Method

- Consider the following

```
> summary(price)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   3291    4220    5006    6165    6332   15906

> price[(price-mean(price))/sd(price) < -1.96]
integer(0)

> price[(price-mean(price))/sd(price) > 1.96]
[1] 14500 15906 13594 13466 12990 11995
```

Finding Outliers in R-Easy Method

- Consider the following for finding outliers based on the boxplot rule.

```
> boxplot.stats(price)$out  #### easier way to get the outliers
 [1] 10372 11385 14500 15906 11497 13594 13466 10371  9690  9735 12990 11995
```

How to remove outliers from the data


```
> IQR(price)
[1] 2112
> outliers=boxplot.stats(price)$out
> cleanprice=setdiff(price,outliers)
> IQR(cleanprice)
[1] 1596.5

> sd(price)
[1] 2949.496
> sd(cleanprice)
[1] 1166.073

> mean(price)
[1] 6165.257
> mean(cleanprice)
[1] 5011.742
```

setdiff(a,b) removes b from a

Skewness

- A related idea to outliers is skewness(and one which we always wonder-do we really have outliers or is the data skewed, or both)?
- **Skewness** measures the degree of asymmetry exhibited by the data

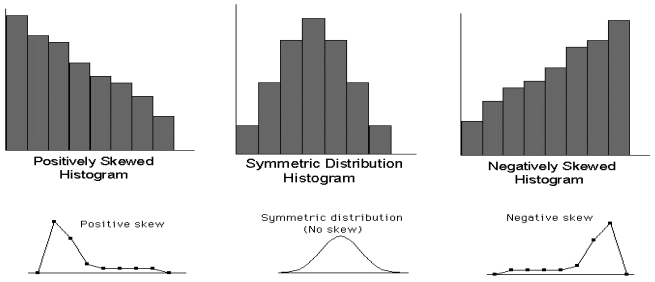
$$skewness = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{ns^3}$$

Never will calculate this by hand

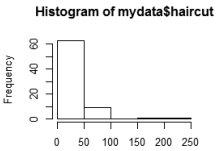
Values of Skewness

- A symmetric data set should have a skewness value near 0
- Negative values for the skewness indicate data that are skewed left and positive values for the skewness indicate data that are skewed right.
- By skewed left, we mean that the left tail is long relative to the right tail. Similarly, skewed right means that the right tail is long relative to the left tail.

Skewness

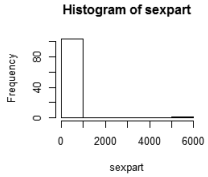


Example: Haircut Data




```
> describe(mydata$haircut)
   vars  n  mean    sd median trimmed   mad min max range skew kurtosis   se
X1    1 74 32.21 38.36   21.5   25.85 17.05   0 250   250 3.63     16.2 4.46
> describe(mydata$haircut[mydata$haircut<150])
   vars  n  mean    sd median trimmed   mad min max range skew kurtosis   se
X1    1 72 26.86 20.49     20   24.67 14.83   0  85    85    1     0.52 2.41
> describe(mydata$haircut[mydata$haircut<100])
   vars  n  mean    sd median trimmed   mad min max range skew kurtosis   se
X1    1 72 26.86 20.49     20   24.67 14.83   0  85    85    1     0.52 2.41
> describe(mydata$haircut[mydata$haircut<50])
   vars  n  mean    sd median trimmed   mad min max range skew kurtosis   se
X1    1 62 20.42 12.78     17   20.21 10.38   0  48    48 0.24    -0.77 1.62
```

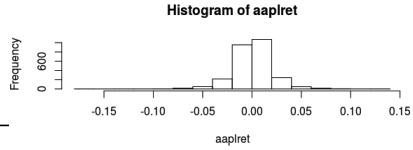
Example: Sexual Partners




```
> describe(sexpart)
   vars   n  mean     sd median trimmed  mad min  max range skew kurtosis    se
X1    1 105 64.92 585.16      1    3.66 1.48   0 6000  6000 9.94    97.79 57.11
> describe(sexpart[sexpart<150])
   vars   n mean   sd median trimmed  mad min max range skew kurtosis   se
X1    1 102 5.07 7.85      1    3.27 1.48   0  45    45 2.95     9.84 0.78
> describe(sexpart[sexpart<10])
   vars  n mean  sd median trimmed mad min max range skew kurtosis   se
X1    1 84  2.2 2.1      1    1.84   0   0   9     9 1.49     1.49 0.23
```

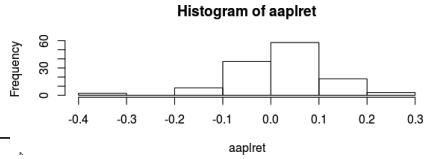
Remember data is time dependent


```
> library(quantmod)
> getSymbols("AAPL")
[1] "AAPL"

> aaplret=dailyReturn(Ad(AAPL))

> describe(aaplret)
             vars    n mean   sd median trimmed  mad   min  max range  skew kurtosis se
daily.returns    1 2630    0 0.02      0       0 0.01 -0.18 0.14  0.32 -0.19     6.29  0
```



Remember data is time dependent


```
> aaplret=monthlyReturn(Ad(AAPL))
> describe(aaplret)
                vars   n mean   sd median trimmed  mad   min  max range  skew kurtosis
monthly.returns    1 126 0.03 0.09   0.03    0.03 0.07 -0.33 0.24  0.57 -0.69     2.17
                  se
monthly.returns 0.01
```



Transforming Skewed Data

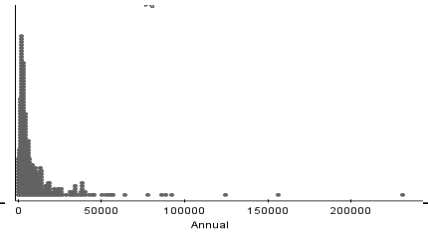
- When a distribution is skewed, it can be hard to summarize the data simply with a center and spread, and hard to decide whether the most extreme values are outliers or just part of the stretched-out tail.
- How can we say anything useful about such data? The secret is to apply a simple function to each data value.

Nonlinear Transformations

- Sometimes there is need to transform our data in a nonlinear way;
- Y=sqrt(X), Y=log(X), Y=1/x, etc….
- This is usually done to try to "symmetrize" the data distribution to improve their fit to assumptions of statistical analysis (will make more sense in a few weeks).
- Basically to reduce outliers in the data and/or reduce skewness.

Your dream job

- Consider the graph below which shows 2005 CEO data for the Fortune 500. The data is in thousands of dollars.

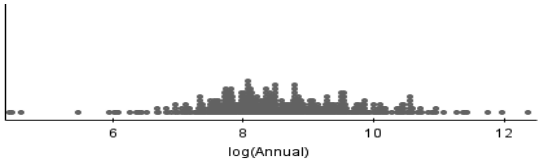


The data is heavily skewed

- Skewed distributions are difficult to summarize. It's hard to know what we mean by the "center" of a skewed distribution, so it's not obvious what value to use to summarize the distribution.
- What would you say was a typical CEO total compensation? The mean value is $10,307,000, while the median is "only" $4,700,000.

Log the data

- One way to make a skewed distribution more symmetric is to re-express, or transform, the data by applying a simple function to all the data values.



The Transform Cheat Sheet

- Calculate the skewness statistic for your data set
- If |skewness| < 0.8 data set is cool and unlikely to disrupt our analysis.
- Otherwise, try a transformation in the “ladder of powers”

λ	-2	-1	-1/2	0	1/2	1	2
y	$\frac{-1}{x^2}$	$\frac{-1}{x}$	$\frac{-1}{\sqrt{x}}$	$\log x$	$\sqrt{x}$	x	x^2

Today’s Tools

- New toolbox additions
  - Transformations, Skewness, Outliers
  - Empirical Rule

Things you should know

- Emprical Rule, Chebyshev’s Rule
- a+bX rule
- Z scoring
- Detecting Outliers
- Skewness and Transformations