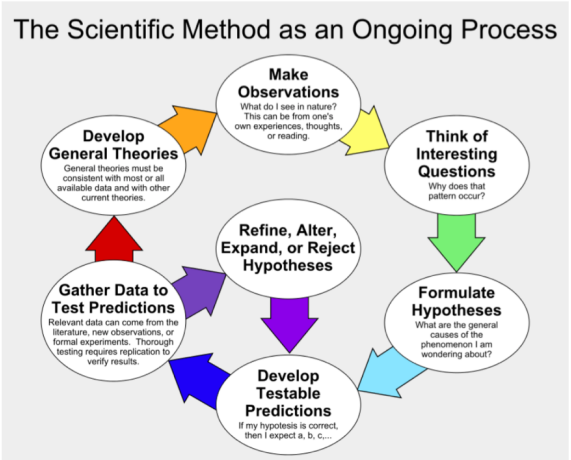
Stat 324 – Introduction to Statistics for Engineers

LECTURE 2: DESCRIPTIVE STATISTICS, AND MORE R/R STUDIO

What does it mean to “know” something?



Sample Data for Consideration

The following data represent the lifetimes (in hours) of 20 different incandescent lamps. The data was gathered as part of a routine quality control sample of lamps created at a large electronics manufacturer. They are ordered from smallest to largest for convenience:

612, 623, 666, 744, 883, 898, 964, 970, 983, 1003,
1016, 1022, 1029, 1058, 1085, 1088, 1122, 1135, 1197, 1201

Create a vector of data "lifetimes" in R so we can do some coding with it.

```{r}
lifetimes<-c(612, 623, 666, 744, 883, 898, 964, 970, 983, 1003, 1016, 1022, 1029, 1058, 1085, 1088, 1122, 1135, 1197, 1201)
```

Descriptive Statistics (graphically)

Graphical Summaries: Most appropriate options depends on the type and amount of data you have.

Numeric/Quantitative Data:

Univariate (each subject has 1 variable of information)

larger data sets: Histograms and Boxplots

smaller data sets: Stem & Leaf, dot plot

Bivariate (each subject has 2 variables with information)

scatterplot

Categorical/Qualitative Data: Bar Charts, Pie Charts (only Sometimes useful display)

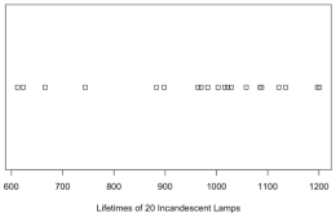
Descriptive Statistics (graphically) cont…

Since, our data set is borderline large, any of the graphs will do (so lets do them all!)

Lets first suppose that we are interested in the raw lifetime values, then this data is numeric.

1. To make a ***dotplot***, draw a number line and a point for each datum above the line.
   Something similar in R:

```
stripchart(lifetimes, xlab="Lifetimes of 20 Incandescent Lamps")
```



2. To make a ***stem and leaf***, organize data with same magnitude on the same "stem"
Each data point is represented by a different leaf. In this case, the stem is the hundreds place and the leaf is the (rounded) tens place of the observation

```
stem(lifetimes, scale=1)

stem(lifetimes, scale=2)  #increases the number of stems
```

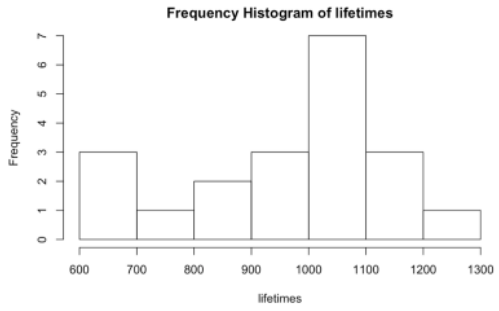
```
The decimal point is 2 digit(s) to the right of the |

 6 | 127
 7 | 4
 8 | 8
 9 | 0678
10 | 0223699
11 | 24
12 | 00
```

Descriptive Statistics (graphically) cont…

3. To make a ***histogram*** (by hand is optional),
   1. Make a *frequency table*
      *choose ~ 5-15 equal-length intervals covering [min, max] (number of bins matters)
      *count # points in each interval (include left end point, exclude right)

   2. Above each interval, draw bar whose height indicates its: (A) *frequency* (count)

```
hist(x=lifetimes, main="Frequency Histogram of lifetimes")
```



From this graph, we can see there were
3 lamps that had lifetimes between 600 and 700 (612, 623, 666)
and only 1 between 700 and 800 (744)

Descriptive Statistics (graphically) cont…

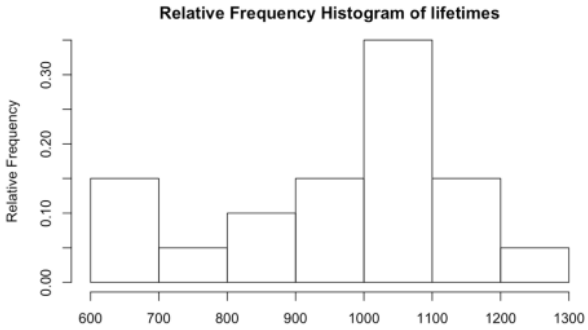
3. To make a ***histogram*** by hand (optional),

2. Above each interval, draw bar whose _____ indicates its

(B) *relative frequency*=$\frac{frequency}{total\ \#\ observations}$ (_____)

From this graph, we can see that 15% (3/20) of the lamps had lifetimes between 600 and 1200 (612, 623, 666)

and only 5% (1/20) between 700 and 800 (744)



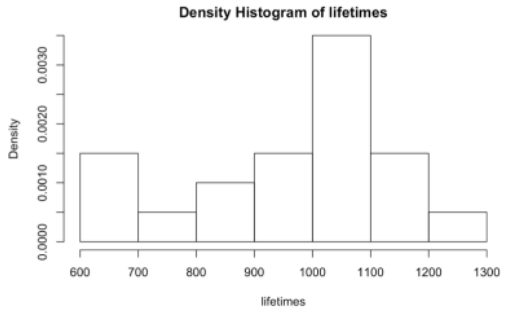

```
h<-hist(lifetimes, plot=F)
h$counts<-h$counts/sum(h$counts)  #we need to actually calculate the proportion in each bin
plot(h, freq=TRUE, ylab="Relative Frequency", main="Relative Frequency Histogram of lifetimes")
```

Descriptive Statistics (graphically) cont…

3. To make a ***histogram*** by hand (optional),

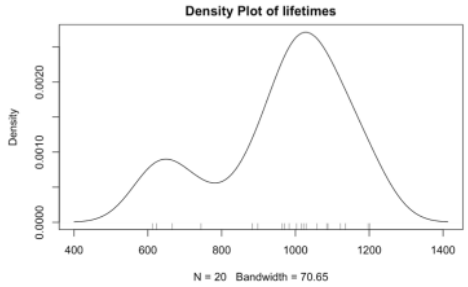
2. Above each interval, draw bar whose height indicates its

(C) *density*=$\frac{relative\ frequency}{width\ of\ bins}$ (so that total area=1) Density*Width= Rel. Freq




```
hist(x=lifetimes, freq=FALSE, main="Density Histogram of lifetimes")
```

*or a density plot "smooths" out the bars

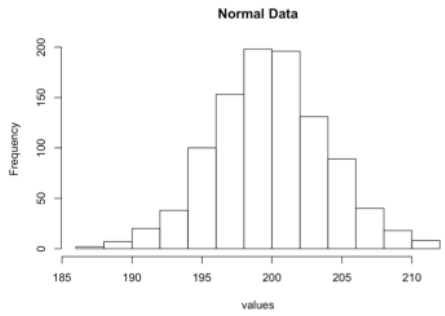


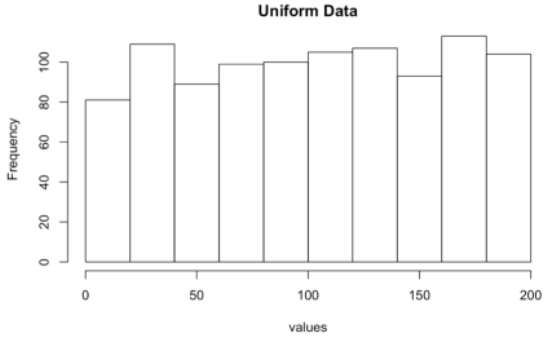

```
plot(density(lifetimes), main="Density Plot of lifetimes"); rug(lifetimes)
```

Descriptive Statistics (graphically) cont…

Graphing numeric data allows us to see the shape of the data

*most easily seen in **stem and leaf** or histogram





e.g.

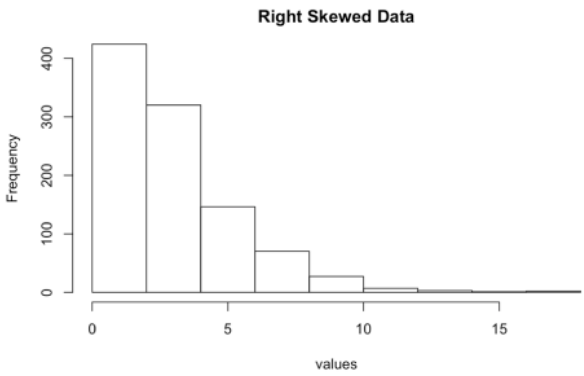
* repeated measures
* length of 2X4's in stockyard, etc

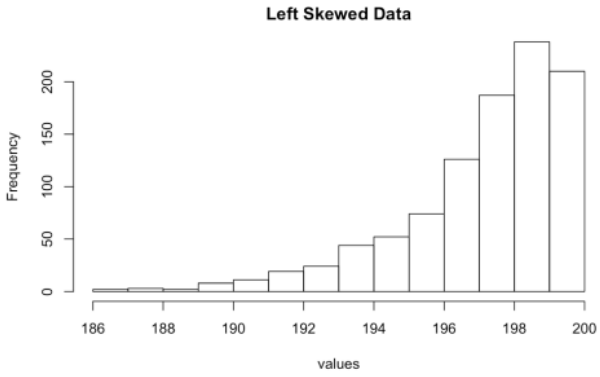
e.g.

*face that lands up when tossing dice
* serial # on a randomly selected product, etc

Descriptive Statistics (graphically) cont…

Graphing numeric data allows us to see the ***shape*** of the data





e.g.

* # of items purchased by each customer

*Score on hard test test

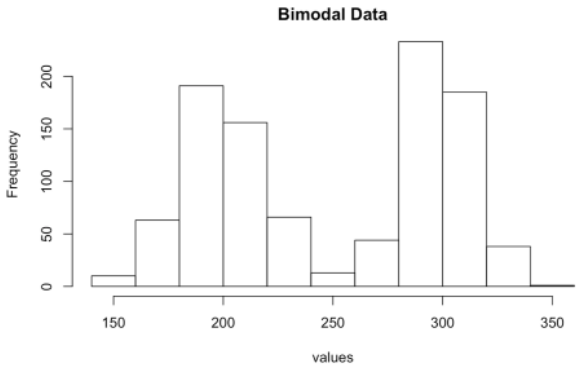
e.g.

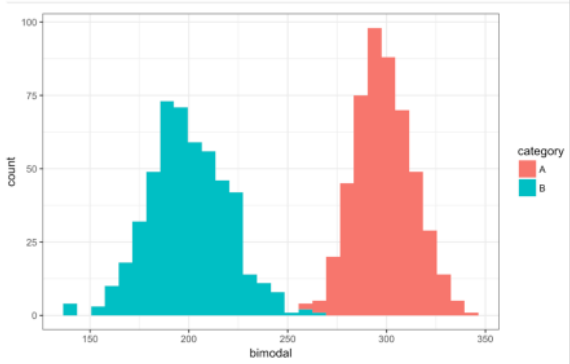
*age at death or retirement in developed countries

*score on easy test

Descriptive Statistics (graphically) cont…

Graphing numeric data allows us to see the ***shape*** of the data





e.g.

*Starting salary of lawyers

*Book prices (paper vs hard cover) or textbooks vs non?

*number of hours after opening that a customer comes into a restaurant

Descriptive Statistics (graphically) cont…

Coming Back To Boxplots…..

Descriptive Statistics (Numerically)

Numeric/Quantitative Data:

if we are interested in the numeric values of the hours, then we concentrate on the measures of center, **location** and spread of the data.

Measure of center 1:

Sample Mean $(\bar{X})$ of data $X_1, \ldots, X_n$ is their sum divided by the sample size:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

For the sample {6, 4, 7, 5, 3}, $\bar{X} =$

For the lifetime sample {612, 623, 666, 744, 883, 898, 964, 970, 983, 1003, 1016, 1022, 1029, 1058, 1085, 1088, 1122, 1135, 1197, 1201}

$\bar{X}$ =964.95 in R via mean(lifetimes)

Descriptive Statistics (Numerically) cont.

Measure of center 2:

Sample Median (M) of data $X_1, \ldots, X_n$ is the midpoint of a sample of size n

- If n is odd , then M is the center data point (at position $\frac{n+1}{2}$)
- If n is even, then M is the average of the two center points (at positions $\frac{n}{2}$ and $\frac{n}{2}+1$)

For the sample {6, 4, 7, 5, 3}, $Median = 5$ ~~3~~ ~~4~~ 5 ~~6~~ ~~7~~

For the sample { 6, 4, 7, 5, 3, 8}, Median= 5.5 ~~3~~ ~~4~~ 5 6 ~~7~~ ~~8~~

For the lifetime sample, we need an average of 1003 1016, which is 1009.5

$Median =$ 1009.5 in R via median(lifetimes)

The median is a more "accurate" measure of center when the data has outliers - extreme values or is skewed ...Why?

In bimodal data, measures of center are not real helpful for the full data set...why?

Other Measurements of Position:

Quartiles divide the data into 4 groups of ≈ equal sizes

**First Quartile (Q1)* is the median of the first (lower) half of the data.

**Third Quartile (Q3)* is the median of the second (upper) half of the data.

If the data set contains an odd number of observations, include the median in both the first half of the sorted list and the second half of the sorted list (when moving to calculate Q1 and Q3)

*There are other acceptable ways of computing quartiles but this is the one we'll employ in class (R has 9 (!) different ways of calculating "quartiles").

Descriptive Statistics (Numerically) cont.

Quartiles Considering the lamp lifetime data:

{612, 623, 666, 744, 883, 898, 964, 970, 983, 1003, 1016, 1022, 1029, 1058, 1085, 1088, 1122, 1135, 1197, 1201}

890.5 median = 1009.5 1086.5

Summary in R gives us something else!?
*Just be clear in your hwk/exam what Process you are using – R or by hand

```
summary(lifetimes)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   612.0   894.2  1009.5   965.0  1085.8  1201.0
```

Considering the lamp lifetime data (with highest value removed):

{612, 623, 666, 744, 883, 898, 964, 970, 983, 1003, 1016, 1022, 1029, 1058, 1085, 1088, 1122, 1135, 1197}

890.5 median 1071.5

```
lifetimes_new<-lifetimes[-20]  #20th datum from "lifetimes""
summary(lifetimes_new)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   612.0   890.5  1003.0   952.5  1071.5  1197.0
```

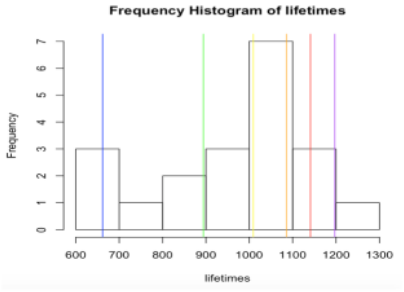
Descriptive Statistics (Numerically) cont.

Quantiles [Percentiles] (more general than quartiles):

Roughly, the pth quantile is the point p in [0,1] such that the proportion p of the data are smaller.

The .25 quantile is Q1,
the 0.5 quantile is Q2 (Median),
and the 0.75 quantile is Q3

Use R to find quantiles: 0.10, Q1, Med, Q3, 0.90, 0.95




```
#Quantiles Function
quantile(lifetimes, probs=c(.10, .25, .50, .75, .90, .95))
```

```
##     10%     25%     50%     75%     90%     95%
##  661.70  894.25 1009.50 1085.75 1141.20 1197.20
```

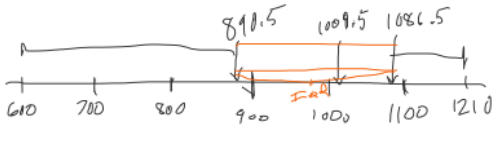
*The different methods of calculating quantiles/percentiles give less distinct values for large data set.

Descriptive Statistics (graphically) cont...

4. To make a ***boxplot*** by hand
    1. Draw and label a vertical or horizontal axis that spans the range of the data
    2. Draw longer lines at Q1, median, and Q3 perpendicular to axis
    3. Connect ends of Q1 and Q3 to create box.
        The length of this box is the IQR
    4. Identify any point outside [Q1-1.5*IQR, Q3+1.5*IQR] an outlier and plot each on the axis with a dot.
    5. Draw lines from the box to the largest non-outlier and from box to smallest

_____

```
#boxplot
boxplot(x=lifetimes, horizontal=TRUE, xlab="Lamp bulb lifetimes (hr)")
```



$Q_1 - 1.5 \times IQR,\ Q_3 + 1.5 \times IQR$

$894.5 - 294 \qquad 1086.5 + 294$

$(\overline{1}96.5, 1380.5)$

$IQR = 196$

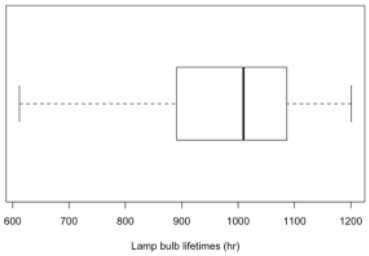
$1086.5 - 894.5 = 196$

Descriptive Statistics (graphically) cont...

Boxplot Pros & Cons:

+ Gives **Numeric and Graphically** summary in one plot

+ Efficient for large sets of data



- Hides number of values that are plotted (similar to histogram)

- Can hide gaps in the data or clusters of data

- Shape is slightly harder to distinguish

Descriptive Statistics (Numerically) cont.

Numeric/Quantitative Data:

if we are interested in the numeric values of the hours, then we concentrate on the measures of **center, location** and **spread** of the data.

Measures of SPREAD/VARIABILITY:

1. Range: Max Value- Min Value. - crude Measure

2. Interquartile Range: IQR:Q3-Q1. Difference between the third and first quartiles (range of middle 50% of data)- this is the length of the box

Q1= 890.5 M= 1009.5 Q3= 1086.5

612, 623, 666, 744, 883, 898, 964, 970, 983, 1003, 1016, 1022, 1029, 1058, 1085, 1088, 1122, 1135, 1197, 1201

IQR= 1086.5 − 890.5 = 196

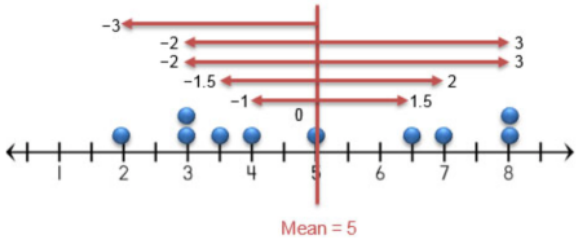
```
#IQR
IQR=quantile(lifetimes, prob=0.75)-quantile(lifetimes, prob=0.25)
IQR
```

```
##   75%
## 191.5
```

Descriptive Statistics (Numerically) cont.

Numeric/Quantitative Data:

2. Sample **Standard Deviation (s) :** "average deviation" how far a typical datum is from the mean



The ***deviation*** of the ith observation from the mean is :(observation)–(mean)=$X_i - \bar{X}$

*Sum of all deviations= 0

Standard Deviation (s) =$\sqrt{Variance}$

Sample $Variance\ s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

*Divide by n-1 so that our sample variance is a better estimate of variability for whole population

Sample $Standard\ Deviation\colon\ s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}.$

*SD units are the as the data

Descriptive Statistics (Numerically) cont.

Numeric/Quantitative Data:

2. Sample **Standard Deviation (s) :** "average deviation" how far a typical datum is from the mean

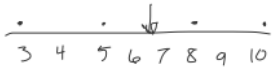
Ex: Calculate the sample standard deviation (s) of the numbers 3,5,8,10 by hand:

Data	Deviation from Mean	(Dev)^2
3	3-6.5 −3.5	(-3.5)^2=12.25
5	5-6.5 −1.5	$(-1.5)^2 = 2.25$
8	8-6.5 1.5	(1.5)^2=2.25
10	10-6.5 3.5	$(3.5)^2 = 12.25$
	0	12.25+2.25+2.25+12.25=29

Sum of Squared Deviations: $\sum_{i=1}^{n}(X_i - \bar{X})^2 =$

Sample $Variance\ s^2 = \frac{29}{3}$ (4−1) # observations, always 1

Sample $Standard\ Deviation\ s = \sqrt{29/3} = 3.11$



Check with R:

```
sd(sd_ex)
```

```
## [1] 3.109126
```

$mean = \bar{X} = \frac{3+5+8+10}{4} = \frac{26}{4} = 6.5$

Descriptive Statistics (Numerically) cont.

Numeric/Quantitative Data:

2. Sample **Standard Deviation (s) :** "average deviation" how far a typical datum is from the mean

Ex: Calculate the [sample] standard deviation for the lamp lifetime data

```
sd(lifetimes)
```

```
## [1] 178.2982
```

Mean and Standard Deviation are the most appropriate/useful measures of center and spread when:

*Data is fairly symmetric and highly concentrated around the mean value

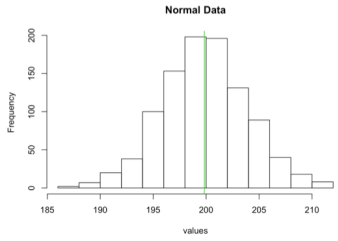
Median and IQR are the most appropriate/useful when:

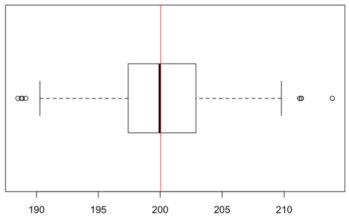
*Data is highly skewed or has extreme outliers

EXAM TIP:
Learn how to calculate SAMPLE MEAN and SAMPLE STANDARD DEVIATON on your calculator

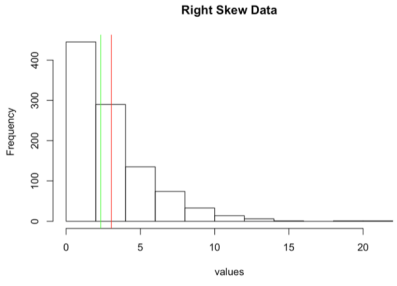
Descriptive Statistics (graphically) cont…

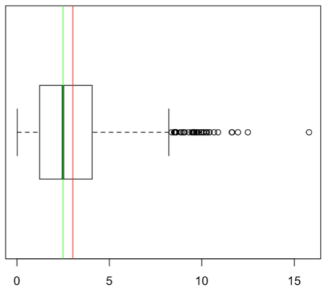
Relating Shape, Centers and Spreads in Histograms and Boxplots:





```
                sum
[1,] "range"      "25.3804733436331"
[2,] "sample sd" "3.91396121323502"
[3,] "IQR"        "5.47799699273062"
```





```
                sum
[1,] "range"      "21.9139832116258"
[2,] "sample sd" "2.5799746377832"
[3,] "IQR"        "2.96447363342925"
```

Descriptive and Numeric Statistics for Categorical Data

Numeric Summaries: Most appropriate options depends on the type of data you have and your questions of interest.

Categorical/Qualitative Data:

Percentage within each Category of interest

E.g: If our category of interest is "has lifetime over 1000 hours" compared to not, an interesting summary table might display the percent of success in each category:

	Lifetime under 1000 hours	**Lifetime over 1000 hours**
Count	9	
Relative Frequency	9/20=0.45	

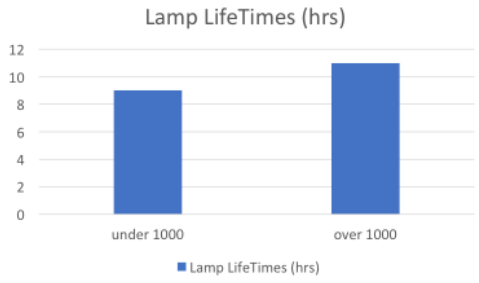
Graphical Summaries: Most appropriate options depends on the type of data you have and your questions of interest.

Categorical/Qualitative Data:

Bar Chart (Similar to Histogram, but bars do not typically touch)

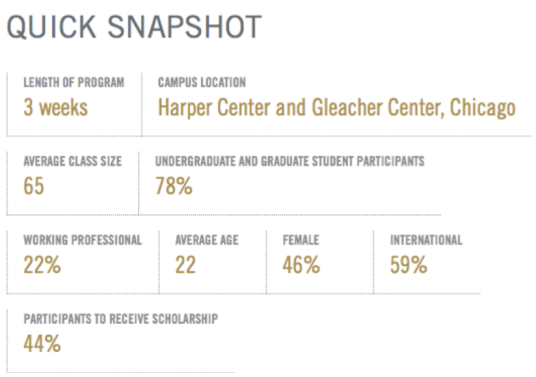
Stacked Bar Chart is one of my favorites

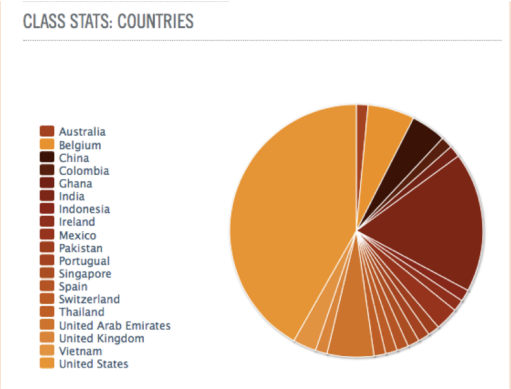
Pie Chart (Can get confusing if there are too many or too similar categories)



Descriptive and Numeric Statistics for Categorical Data (cont..)

Source: https://www.chicagobooth.edu/programs/summer-scholars





For Next Time

1. Run through R code from today's class and play around with it a bit – what happens when you change the lifetime data a bit? Add an outlier or two and see how the graphs and summary measures change.

2. Save off course files for tomorrow so you can work through R with me in class.