Chap1

Data arises from observations

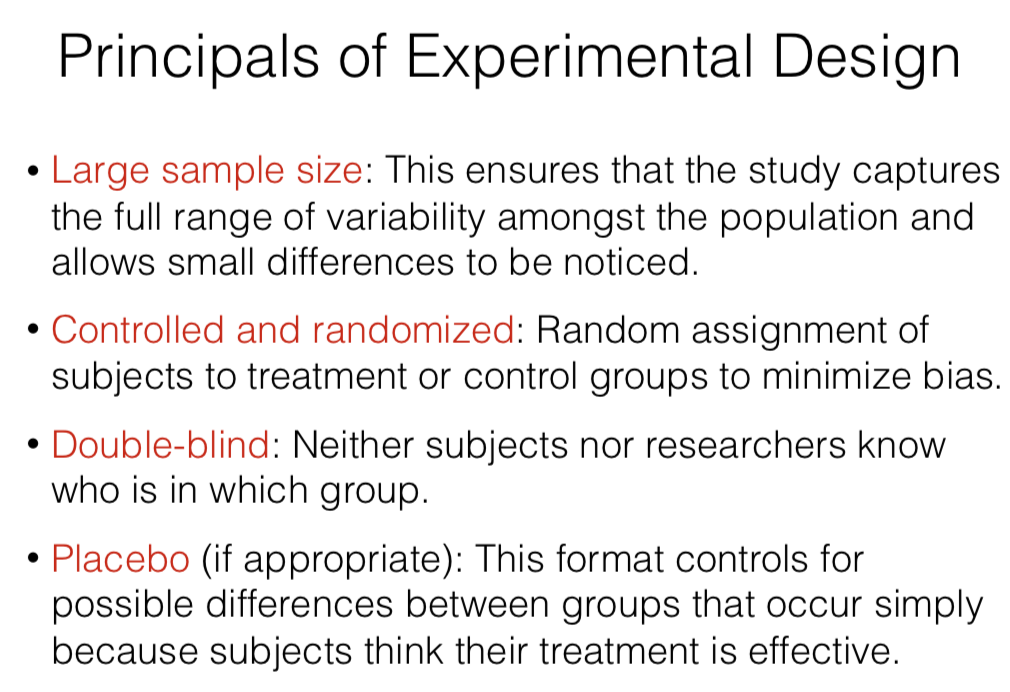
**Treatment group:** individuals who receive the treatment of interest in an experiment

**Control group**: individuals NOT receiving treatment

**Bias:** the tendency to over/underestimate a population parameter due to a measurement process.

Random assignment helps minimize bias

**Blinding:** helps prevent bias by ensuring that the participants (and sometimes the researchers) do not know who is assigned to which study group.

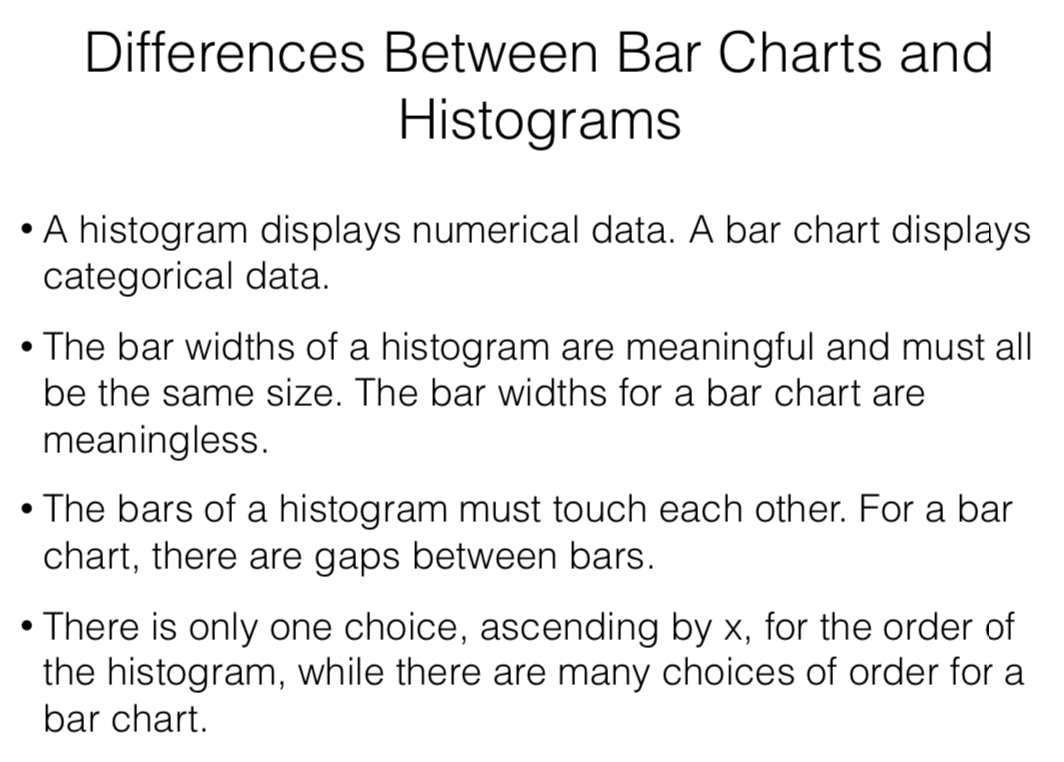


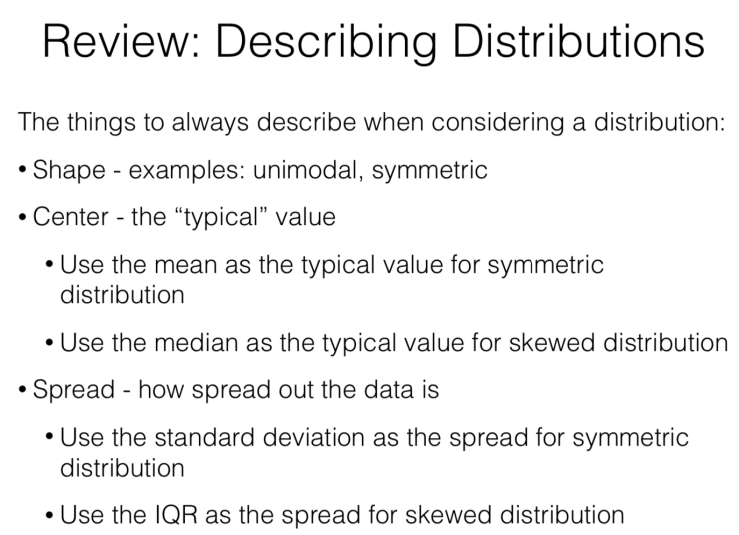
We can conclude causation if we use a controlled experiment, in which we deliberately assign the subjects to different treatments (usually at random)

Two way table displays counts of 2 categorical vars

Chap2

Bar chart displays the counts of each category, thus is for categorical data





Chap3

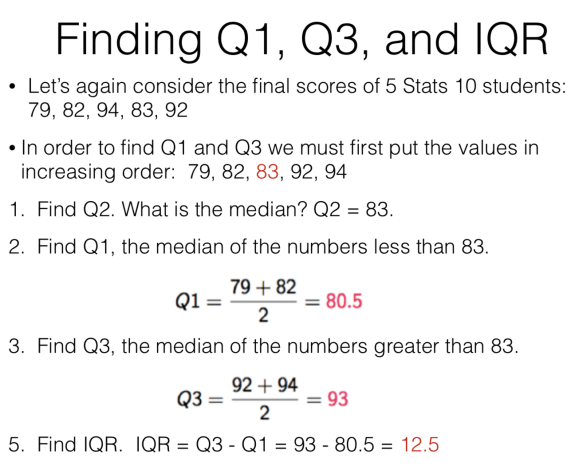
Remember to subtract 1 when calculating stdev!!

**Q1, Q3**

75% of data are below Q3

Q3-Q2<Q2-Q1 Easy Exam, clustered near high score, outliers are low

Q3-Q2=Q2-Q1 Mean is about the same as Median



Least square line

The sum of the square of residuals around the regression line is minimum

Calculating IQR

If the data set has an odd number of values, we omit the median(centermost) value of the set.

If even, just split the set

1,2,**2**,3,4|4,7,8,9,11 IQR=8-2=6

21,25,27,31,32,37,43,45,49

Q1=26, Q3=44, IQR=44-26=18

Five number Summary: Min. Q1, Median, Q3, Max

Outlier

Left limit = Q1 - 1.5 x IQR

Right limit = Q3 + 1.5 x IQR

Outliers have NO effects on IQR, but increase stdev

Unusual

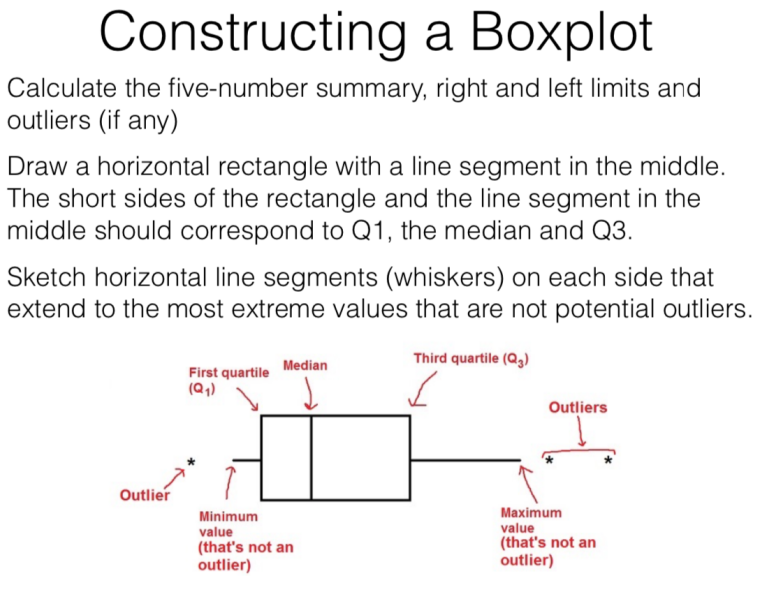
More than 2 stdev (Z-score exceeds +2 or below -2)

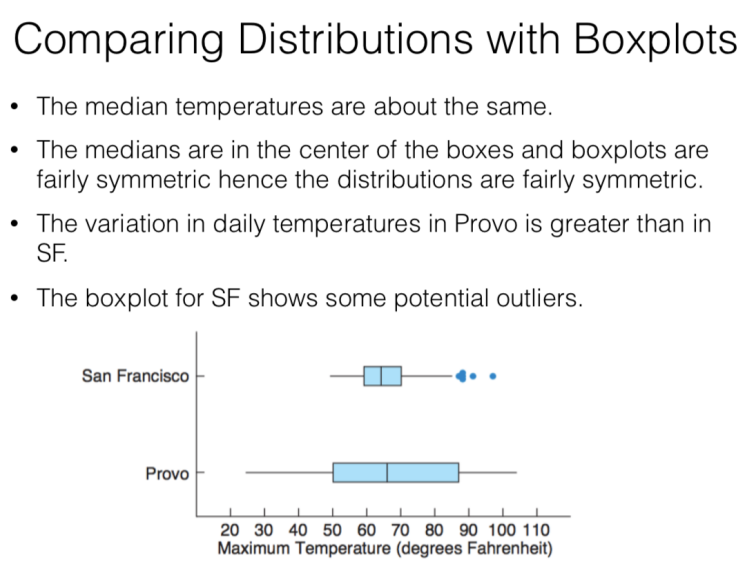
Use the mean and standard deviation when the distribution is symmetric and unimodal

Use the median and IQR when the distribution is left or right skewed

Boxplot

Should not use on very small data set(>=5)





Chap4

x-axis variable: predictor, independent, explanatory variable

y-axis variable: predicted, outcome, dependent, response variable

**Correlation coefficient(r):** measures the strength of linear relationship between 2 numerical variables.

Slope=r\*(sy/sx)

**Coefficient of determination (r2):** percentage of the variable of y is explained by x.

Interpretation of intercept: When x equals 0, we expect y to equal the intercept.

y-Intercept=y-bx

**Extrapolation**

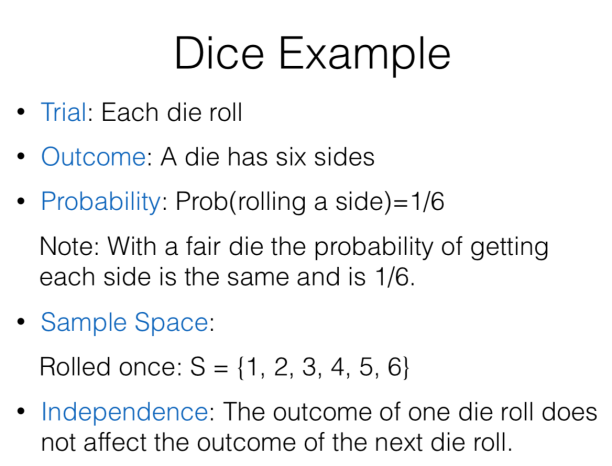
do not make a prediction for a x value outside the range of the data - the linear model may no longer hold outside that range.

**Chap5**

Simulate randomness: coin toss, random number table, roll the dice

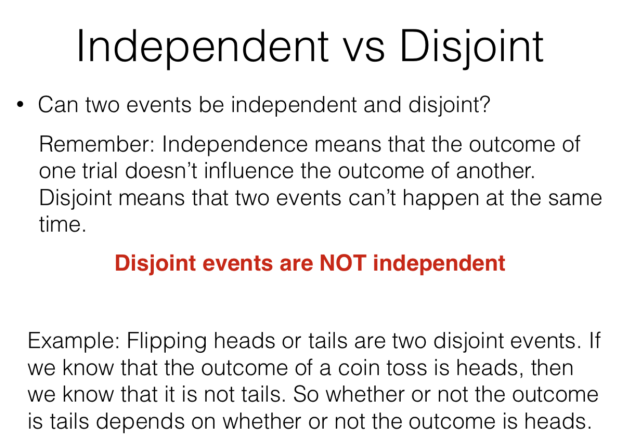
**Law of Large Numbers:** long-run relative frequency of repeated independent events gets closer to the true relative frequency as the number of trials increases.

E.g. If a coin is tossed many times, the overall percentage of heads should settle down to about 50% as the number of tosses increases



**Disjoint events:** Events that have no outcomes in common(thus cannot occur together P(A and B)=0

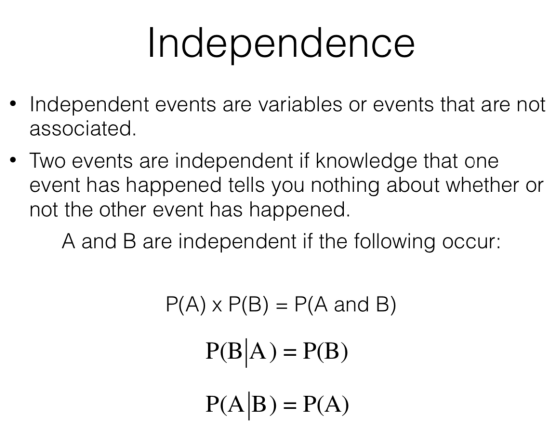
Disjoint events are NOT independent



Non-disjoint events may/may not be independent

E.g. pick two random people in the class who scored 100% on the midterm. They could be complete strangers whose performance had nothing to do with each other, or they could be close friends who studied together a lot.

Complement Rule: Prob(Ac)=1-Prob(A)



Sample space: collection of all possible outcomes of a trial. Two events have probabilities sum up to 1 **if they make up the sample space**.