**AP Statistics**

**Semester One – Study Topics**

Data

* Who, What, Where, Why of the Data
* Categorical and Quantitative Data

Displaying Categorical Data

* Bar Charts and Segmented Bar Charts
* Pie Charts
* Make sure charts meet the area principle
* Contingency Tables – Two way data tables
* Conditional Distributions – Percentages of each category shown in the table

Displaying Quantitative Data

* Histograms – Equal width bins
* Stem and Leaf Displays – Include a key
* Dot Plots
* ***Shape***
	+ Modes – Unimodal, Bimodal or Multimodal
	+ Uniform
	+ Outliers
* ***Center***
	+ Mean or Median for skewed data

Describing Distributions Numerically

* ***Spread***
	+ Range
	+ IQR : Interquartile Range = Q3 – Q1
	+ Standard Deviation
* Five-Number Summary
	+ Minimum, Q1, Median, Q3, Maximum
	+ Box-Plots – Box is Q1, Median, Q3, Whiskers are 1.5 +/- IQR, or Min Max, whichever is closer to median

Standard Deviation – Normal Model

* Standardized results to compare equally with z-scores $z=\frac{(y-\overbar{y})}{s}$
* Shifting Data by adding a constant
	+ Center, percentiles, min and max increase by the same constant
	+ Spread is unchanged
* Rescaling Data by multiplying a constant
	+ Mean, Median and Percentiles and Spread are multiplied by the constant
* Normal Model – Z-score $z=\frac{(y-μ)}{σ}$
* Convert normal scores and percentages with your calculator, *normalcdf*, *invNorm*
* 68 – 95 – 99.7 Rule: Relates to one standard deviation, two and three respectively
* Normal Probability Plot – Straight data is normal

Scatterplots, Association and Correlation

* Scatterplots
	+ Direction – Positive or Negative
	+ Form – Linear, curved
	+ Strength – closely clustered in shape is strong
* Explanatory Variables and Response Variables
* Association – Direction, Form, Strength
* Correlation
	+ The strength of the linear association between two quantitative variables
	+ Correlation Coefficient $r=\frac{\sum\_{}^{}z\_{x}z\_{y}}{n-1}$
	+ Use LinReg(a+bx) to find on calculator

Linear Regression: $\hat{y}=b\_{0}+b\_{1}x$

* Residuals – Actual minus predicted, $y-\hat{y}$
	+ Negative residual = actual is less than predicted
* Line of Best Fit – Line that minimizes the square of all residuals
* Slope of line is the data’s association
* Distance in standard deviation of x moved is the distance moved in standard deviation of y
* Outliers may seriously affect the regression line
* Use calculator to find
* $R^{2}$ = Fraction of the data’s variation accounted for by the model
* Check residuals to make sure the regression is appropriate – no pattern
* Extrapolation should be questioned
* Leverage – A point with an x-value far from the mean x-value
* Influence – A point that changes the analysis if omitted
* Lurking Variables – An unobserved variable causing an apparent association

Re-expressing Data

* Goals
	+ To make the data more symmetrical
	+ Make the spread of side-by-side boxplots more alike
	+ Make the form of the scatterplot more linear
	+ Make the scatter spread out instead of fan shaped
* Ladder of powers
	+ Square, No Change, Square Root, Log, Neg/Square Root, Neg Reciprocal
	+ Test data moving up and down the ladder

Randomness

* Simulation – Random outcomes that model a situation – use a table on a test
	+ Complete several trials in a simulation

Sample Surveys

* Biased surveys under or over emphasize a characteristic in the population
* Randomize the selection
* Sample size is important, not the fraction of the population sampled
* Census – Surveying everyone in the population
* Statistics – A summary of the data
* Simple Random Sample
	+ Every combination of individuals has an equal chance of being selected
* Stratified Sample – Subjects taken from specific groups (strata)
* Cluster Sampling – Splitting the population into similar parts
* Multistage samples – A combination of the above
* Systematic Samples – Sampling every nth subject

Experiments and Observational Studies

* Observational Studies – Researchers observe, Retrospective or prospective
* Randomized Comparative Experiments – Random assignment of subjects to treatments
* Factor – An explanatory variable that is manipulated
* Levels – Specific values of factors chosen
* Experimental Design
	+ Control – Similar conditions for all trials
	+ Randomize – Experimental units assigned randomly
	+ Replicate – The outcome of the experiment can be repeated
	+ Block – Randomizing across treatment groups – not required in an experiment
* Statistically Significant – Measured differences that are greater than expected random variation
* Control group – Baseline treatment, or placebo
* Blinding – When subjects do not know which treatment they are receiving
* Double Blinding – Blinding, plus evaluators do not know who is receiving treatment
* Blocking – Groups of experimental units that are similar are grouped together
* Confounding – Levels of factors are associated with levels of another factor

Probability

* Probability – P(A) = Count of outcomes in A / Count of all possible outcomes
* Independent- Outcome of one trial does not influence another trial
* Law of Large Numbers – Long-run relative frequency settles down to the true probability
* Personal Probability – Subjective value
* Complement - 1 – probability event will not occur
* Disjoint – Two events with no outcomes in common
* Sample Space – Collection of all possible outcomes
* $P\left(A∪B\right)=P\left(A\right)+P\left(B\right)-P(A∩B)$
	+ Union =”or” ; Intersection = “and”
* “at least” use the complement
* $P\left(B|A\right)=\frac{P(A∩B)}{P(A)}$ $P\left(B|A\right)P(A)=P(A∩B)$
* Independent Events $P\left(B|A\right)=P(B)$
* Replacement – Selected item is returned and the next probability is unchanged