A. P. Statistics
Important Concepts

Here is a list of important concepts by chapter. You should skim the chapters that seem unfamiliar (based on this list), and look at the chapter reviews.

Chapter 1

- Distribution of a variable tells the values a variable attained and how often.
- Describe a distribution of a quantitative variable by describing shape, center, and spread
- Describe symmetry distributions using mean and standard deviation; use 5-number summary for skewed distributions
- Mean is not resistant and is always pulled toward the tail
- Standard deviation is always positive and equals zero only when all observations are identical
- Five number summary: Min, Q1, Median, Q3, Max. Q1 is the 25th percentile which means that 25% of observations are at or below that value. Q3 is the 75th percentile which means that 75% of observations are at or below that value. Median is 50th percentile.
- Frequency histogram has values of quantitative variable on one axis and frequency on other axis. *Relative* frequency histogram has values of quantitative variable on one axis and *proportion or percent* of observations on other axis.
- Cumulative frequency histogram or ogive gives the percent or frequency of observations *at or below* each value. Cumulative relative frequency histogram or ogive displays percentiles on one axis.
- Outliers may be identified using $1.5\times IQR$ rule, or by using a modified box plot on calculator.
- Mean and standard deviation are NOT resistant. Median and quartiles are resistant. Use median and IQR as measures of center and spread (respectively) if data is strongly skewed or has outliers.
- Graphs to display univariate, quantitative data: boxplot, stemplot, histogram, dotplot. (Note: box plot does not give information about individual observations.)

Chapter 2

- ***Density curve*** has area of 1 and is always on or above the x-axis
- Area under curve in a certain range is the same as the proportion of observations in that range. (Use area formulas for geometric density curves such as rectangles, triangles, and trapezoids.)
- ***Normal density curve*** is mound-shaped (or bell-shaped); mean=median; area can be found converting the observations to observation on the standard normal curve $\left(z=\frac{\text{statistic - parameter}}{\text{std. dev. of statistic}}\right)$ then use **Normalcdf(left bound, right bound)** or table.
- ***Empirical rule*** applies to all normal density curves.
- Use InvNorm to find the standardized observation associated with a certain percentile ranking: **InvNorm(%rank (as proportion)) = z-statistic**. Then use z-formula to convert from the z-value to the observation value.

- Use normal probability plot to determine if data can be modeled by a normal curve. The plot looks kind of like a scatterplot (last type of graph in the STAT PLOT menu) and if plot looks linear, then data can be modeled by a normal density curve. If plot shows definite curve, then data is skewed. If one observation is set apart from others on left or right, then the observation is possibly an outlier.

Chapter 3

- Bivariate data: Two measures recorded on each individual.
- Use a ***scatterplot*** to determine if there is a relationship between two quantitative variables.
- Positive association means positive slope; as values of explanatory variable increase, values of response variable increase. Or, above average values of one variable are associated with above average values of the other variable.
- To describe a plot give strength, form, and direction of relationship.
- ***Correlation*** is a measure of the strength of a ***linear*** relationship. Also gives direction (sign).
- Correlation coefficient has values: $-1 \leq r \leq 1$ where $r = 1$ is a perfect line with positive slope and $r = -1$ is a perfect line with negative slope. ***Properties of correlation are on p. 132.***
- Correlation of $r = 0$ or r close to zero could mean no association at all (randomly scattered points) or a *non-linear* association, such as a quadratic.
- ***Least squares regression line*** means that the line produces the smallest sum of squared residuals possible for the data.
- Least squares line always passes through $\left(\overline{x}, \overline{y}\right)$
- Least squares regression line can be obtained using LinReg on calculator or using means and standard deviations of the two data sets. (See formula on formula sheet.)
- Slope of the least squares lines tells the amount that the y-variable changes for each unit of change in the x-variable.
- ***Coefficient of determination***, r^2, is the percent of the <u>*variation*</u> in the response variable that is explained by the model on the explanatory variable.
- ***Residual*** is $observed - predicted$ or $y - \hat{y}$
- If residual plot has no pattern, then that is evidence that the model selected is a good fit for the data. (Residuals are plotted against x-values or against y-values.)
- ***Influential point*** sharply affects regression line if removed; points that are extreme in the x-direction can be influential; influential points may have small residuals. (problems on p. 168 good to look at for practice!)
- ***Outliers*** in a regression do not fit the pattern of the data; generally have large residuals
- ***Correlation does not mean causation***!! Even if you have a perfect correlation, that does not mean that the x-variables causes changes in the y-variable. The correlation could be due to lurking variables.

Chapter 4

- To transform exponential data so that it's linear, plot $\left(x, \log y\right)$ or $\left(x, \ln y\right)$. Linreg equation will be of the form: $\ln \hat{y} = a + bx$
- To transform data from power model, plot $\left(\log x, \log y\right)$ or $\left(\ln x, \ln y\right)$. Linreg equation will be of the form: $\ln \hat{y} = a + b \ln x$
- Predictions from ***extrapolation***, or predicting outside the range of the data, are not reliable as pattern may not continue.
- A ***lurking variable*** is a variable that has an important effect on the relationship between the variables in a study, but it is not included among the variables studied
- Relationship between two variables can be due to: ***causation*** – x causes y (very hard to demonstrate…only with controlled experiment); ***common response*** – both x and y variable are responding to some third variable; ***confounding*** – the effect of the explanatory variable on the y variable is hopelessly mixed up with the effects of other variables.
- ***Simpson's Paradox*** refers the reversal of the direction of a relationship when data from several groups are combined to form a single group.
- Categorial data can be analyzed using a two-way table.
- Graphs for categorical data: bar graph, pie graph, segmented bar graph
- To see if there is a relationship between two categorical variables, compare conditional probabilities.
-

Chapter 5

- A ***census*** is when every individual in a population is used in a study; a ***sample*** is when only a portion of a population is used in a study.
- A study is ***biased*** if one outcome is systematically favored over other outcomes.
- ***Sampling bias*** may be due to: ***voluntary response***, ***undercoverage, convenience sampling***, Random sampling reduces the chance of bias.
- Sample size is NOT bias!! Yes, larger sample sizes are more accurate (less spread), but sample size does not result in one outcome being systematically favored over another.
- ***Non-sampling bias*** cannot be corrected by random sampling. Sources of non-sampling bias are: ***poor wording of questions***, or ***nonresponse.***
- A sample is a ***simple random sample (SRS)*** if every group of *n* individuals has an equal chance of being selected.
- To create an SRS, number the individuals in the population from 1 to whatever and then using a random number table select individuals. Must use the same number of digits for each number, so made need to label individuals as 01, 02, etc, or 001, 002, etc.
- A ***stratified sample*** (which is NOT an SRS) is one in which individuals are first divided into separate groups, or strata, and then an SRS is selected from each stratum.
- An ***observational study*** occurs when the experimenter observes individuals and measures variables of interest but does not attempt to influence the responses.
- In an ***experiment***, the researcher deliberately imposes a treatment on individuals in order to observe how they respond to the treatment.
- The basic principles of ***experimental design*** are: control – ***control*** the effects of lurking variables (done by comparison of groups like control group and treatment group, blindness); ***randomization*** – randomly assigning subjects to groups; ***replication*** – perform the experiment on many subjects to reduce chance variation in results. **(Good idea to read p. 275!!)**

- ***Completely randomized experiment*** is one in which all experimental units are assigned completely at random to groups. (This is the opposite of a block design!!)
- ***Blind*** experiment means that subject does not know whether he is receiving the real treatment or the individual interacting with the experimental unit does not know. ***Double blind*** experiment means that neither the subject nor the people in contact with the subject know which treatment the subject is receiving.
- Blindness and double-blindness are used to control the placebo effect. The placebo effect is the phenomenon that humans will always respond to a treatment.
- ***Block design*** is one in which experimental units that are similar are grouped together and assigned to treatment groups within the block. Only subjects within a block are compared. When blocking, put LIKE THINGS together so that the variable being controlled is constant within the block.
- *Blocking is used to reduce variation and to control lurking variables* by grouping subjects according to those lurking variables.
- ***Matched pairs*** is a special case of blocking in which each block consists of only two experimental units, or one experimental unit receiving two treatments. Order of treatments must be randomized since there is not randomization within groups. Either one subject receives both treatments (in random order) or two subjects that are alike in every important way are compared with one subject randomly receiving one treatment and the other randomly receiving a treatment.
- ***Simulation*** can be used to represent results of a certain situation. To simulate a situation, assign numbers to outcomes so that the number occurs with the same frequency (percent) as the outcome. Use a random number table (or calculator) to find outcomes according to the numbers that come up.
- When designing a simulation: describe phenomenon to be simulated; described how numbers will be assigned to outcomes; state stopping condition; state assumptions (independence, special cases such as repeated numbers); carry out many repetitions.

Chapter 6

- ***Probability*** is the proportion of times that a certain event occurs in a long series of repetitions.
- The ***Law of Large Numbers*** describes the phenomenon of the more trials we do, the closer the ratio of occurrences to trials becomes closer to the true probability.
- The ***complement*** of an event is $1-P(event)$. Complement is denoted $P\left(event^c\right)$
- ***Two events are mutually*** exclusive if they cannot occur simultaneously; that is, the joint probability is zero, $P(A\cap B)=0$. (Two events that are mutually exclusive cannot be independent.)
- Two events, A and B, are independent if $P(A)=P(A\mid B)$. Events that can occur simultaneously may or may not be independent.
- The ***joint probability*** of two (or more) events is the probability that both occur simulataneously (AND). If two events are ***independent***, then joint probability is: $P(A\cap B)=P(A)P(B)$
- The ***union*** or ***probability that at least one event occurs*** is (on formula sheet) $P(A\cup B)=P(A)+P(B)-P(A\cap B)$

- ***Conditional probability*** of two events is the probability that one event occurs given that the other event has already occurred (on formula sheet). ($P(A|B)=\dfrac{P(A\cap B)}{P(B)}$
- Probability of *at least one* is $1-P(none)$

Chapter 7

- ***A random*** variable is a variable whose value is a numerical outcome of a random phenomenon.
- A ***discrete random variable*** is one in which there are a countable number of outcomes. The distribution of a discrete random variable is a table (or histogram) showing each possible outcome along with the probability of that outcome. To find the probability of a discrete random variable, add the probabilities of all of the outcomes in the range.
- A ***continuous random variable*** is one in which the variable takes on every possible value in an interval. The distribution of a continuous random variable is a density curve. To find the probability for a continuous random variable, find the area under the density curve.
- A normal distribution is a special distribution of a continuous random variable.
- The ***mean*** of a random variable or ***expected value*** is (on formula sheet) $E(X)=\sum x_i p_i$. (Multiply each outcome by its probability and add them all up.) This gives the average outcome per game if the phenomenon were repeated many times.
- The ***variance*** of a random variable is (on formula sheet) $VAR(X)=\sum(x_i-\overline{x})^2 p_i$; to get standard deviation, you square root the variance.
- You can find the mean and standard deviation of a random variable on your calculator by entering outcomes in L1 and probabilities in L2 and then doing 1-Var Stats L1,L2.
- Special rules:

a) If a constant is added to every number in a distribution, then the mean of the new distribution is the old mean plus the constant and the standard deviation stays the same. (Adding does not change the spread; it just shifts distribution.)

b) If every number in a distribution is multiplied by a constant, then the mean of the new distribution is the old mean times the constant and standard deviation of the new distribution is the old standard deviation times the constant. (The variance is the square of the standard deviation so the variance is multiplied by the square of the constant.)

c) If a new distribution is formed by adding or subtracting randomly selected individuals from two existing distributions then the mean of the sums (or differences) is the sum (or difference) of the means. That is,

$\mu_{x+y}=\mu_x+\mu_y$ or $\mu_{x-y}=\mu_x-\mu_y$

This calculation works regardless of whether the observations are independent.

d) If a new distribution is formed by adding or subtracting randomly selected individuals from two existing distributions then the variance of the sums (or differences) is the sum of the variances, <u>provided that the observations are independent</u>. That is,

$\sigma^2_{x+y}=\sigma^2_x+\sigma^2_y$. To the standard deviation, square root this result. Note that we always ADD variances.

Chapter 8

- ***Binomial distribution*** is a special probability distribution in which: there are two outcomes for the event; there is a fixed number of observations; the observations are independent; probability of success is the same for each observation
- To find the probability of exactly ***k*** success in ***n*** trials of a binomial phenomenon either use the formula (on formula sheet): $\boldsymbol{P(X=k)={}_nC_kp^n(1-p)^{n-k}}$ or use Binompdf on your calculator: Binompdf(num observations, probability of success, number of successes we want)
- Binomial distribution is symmetrical if p = 0.5; skewed right if p close to zero; skewed left if p close to 1.
- Mean of a binomial distribution is (on formula sheet) $\mu_X = \boldsymbol{np}$ and standard deviation of binomial distribution is $\boldsymbol{\sigma_X = \sqrt{np(1-p)}}$.
- ***Geometric distribution*** is a special probability distribution in which: there are two outcomes, success or fail, for the event; the observations are independent; the probability of success is the same for each observation; the variable of interest is the number of trials before we see the first success.
- To find the probability of the first success on the *kth* observation either use the formula: $\boldsymbol{P(X=k)=(1-p)^{n-1}p}$ Geometric distribution is always skewed right.
- The mean of a geometric distribution is (NOT on formula sheet) $\boldsymbol{\mu_X = \frac{1}{p}}$

Chapter 9

- Larger sample sizes are more accurate; increasing sample size decreases sampling variability.
- The ***sampling distribution of a statistic*** is the distribution of the statistic in all possible samples of a certain size.
- The ***distribution of the sample mean,*** $\bar{x}$**,** has mean $\mu_{\bar{x}} = \mu$ (where μ is the mean of the population from with the sample is drawn) and standard deviation $\boldsymbol{\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}}$.
- The ***Central Limit Theorem*** gives the mean and standard deviation of the sampling distribution of sample means as state above, and ***more importantly*** says that if the sample size is large ($n \geq 30$) the sampling distribution of sample means will be approximately normal regardless of the shape of the distribution of the population. (If $n < 30$, then the sampling distribution of sample means will mimic the shape of the population, and will be more like it the smaller the sample size is. The sampling distribution of sample means is normal if the distribution of the population is normal.)
- The **sampling distribution of sample proportions,** $\hat{p}$, has mean $\boldsymbol{\mu_{\hat{p}} = p}$ (where p is the population proportion) and standard deviation $\boldsymbol{\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}}$.
- The sampling distribution of sample proportions will be approximately normal if $\boldsymbol{np \geq 10}$ **and** $\boldsymbol{n(1-p) \geq 10}$

Chapter 10

- A ***confidence interval*** is used to estimate a population parameter.
- ***An N% confidence interval is interpreted as follows***: N% of all intervals that could be obtained contain the population parameter, so we're fairly confident that our interval contains the population parameter.
- Increasing the confidence level will increase the margin of error. Decreasing confidence and/or increasing sample size will decrease the margin of error.
- The ***P-value*** of a statistic is the probability of obtaining a statistic as extreme as the one you got, if the null hypothesis is true.
- A ***statistic is significant*** if it is unlikely to occur by random chance; the statistic (or sample) is unusual or rare. The smaller the P-value (closer to zero) the more significant the statistic and the stronger the evidence against the null hypothesis.
- A ***Type I error*** is the probability that we incorrectly reject the null hypothesis when it is really true. The probability of a Type I error is α (significance level). A Type I error will occur α of the time by random chance.
- A ***Type II error*** is the probability that we incorrectly accept the null hypothesis when the alternate is really true. The probability of a Type II error is called β and is the area under the "true" distribution that falls in the acceptance region of the hypothesized distribution.
- The ***power of a test*** is the probability that the test will reject the null if the alternate is really true. ($1-\beta$); power is the complement of Type II error.
- If sample size is increased, the probability of a Type II error decreases and power increases. (Probability of Type I error is still alpha.)
- If significance level (alpha) is made smaller (from 0.05 to 0.01) then probability of Type II error is increased and power decreased.

Chapter 11

- A ***t-distribution*** is used when we don't know the population standard deviation and we want to estimate a population mean. Mean of the distribution is always zero.
- A t-distribution is bell-shaped, but is more variable (wider and flatter) than a standard normal curve for small sample sizes. It is more variable because the standard deviation is calculated from the sample so varies with each sample. As sample size increases without bound, the t-distribution becomes closer to a normal distribution.
- For dependent samples (Matched Pairs) do a one-sample t-test on the list of differences. The null hypothesis would be $\mu_{diff} = 0$
- For two independent samples, use a two-sample t-test with null hypothesis $\mu_1 = \mu_2$
- For 2-sample t-test, the degrees of freedom is the smaller sample size minus 1.
- SE is standard error. For sample means, SE is an estimate of the standard deviation of the sampling distribution ($\sigma/\sqrt{n}$) and is $SE = \frac{s}{\sqrt{n}}$.
- SE for the difference of two means is $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

Chapter 12

- The standard error of p-hat is $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. Use this when creating a confidence interval to estimate the population proportion.
- Sampling distribution of sample proportions is approximate normal if $np \geq 10$ and $n(1-p) \geq 10$. For a 1-prop z-test, use the "p" from the null hypothesis. For a confidence interval, use p-hat.
- Use a one-proportion z-test to compare an unknown population proportion to a known population proportion. The standardized statistic is: $z = \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}}$ where p is the known proportion stated in the null hypothesis (since we assume the null is true). You may use
  1-Prop Z test on calculator to test if a population proportion is equal to a given value but you still must show all steps including the appropriate calculation for the z-statistic.
- The SE for a difference of two proportions is $\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$. Use this SE when estimating the difference between two population proportions. (Do NOT pool the p-hats when calculating a confidence interval because we have no null that assumes the proportions are equal.)
- The sampling distribution for the difference of two proportions is approximately normal if $n(\hat{p}) \geq 5$ and $n(1-\hat{p}) \geq 5$ for BOTH p-hats. If checking "nearly normal" for a two-proportion z-test, you may use the pooled $\hat{p}$ in the place of both p-hats when checking the nearly normal condition.
- Use 2-Prop Z-test on calculator to see if two population proportions are equally where you have two independent samples.

Chapter 13

- The *Chi-square distribution* is always skewed to the right. Also, the mean shifts to the right as the degrees of freedom increases.
- Use a Chi-square Goodness of Fit test if you want to see if the distribution of a single categorical variable "fits" some idealized distribution. Create the chi-sq statistic using lists. Remember to always use COUNTS not percents in your lists!
- Use a Chi-square Test of independence if you want to see if two variables measured on one set of individuals (sample or population) are independent---mainly a two-way table. Null hypothesis is that the variables are independent, or that there is no relationship between the variables.
- Use Chi-square test of homogeneity if you want to see if the values of a categorical variable are distributed the same way for two or more populations. The test works the same way as a test of independence except the null hypothesis is that the populations are homogeneous.
- Expected counts for two-way table: $\frac{\text{row total} \times \text{column total}}{\text{grand total}}$

Chapter 14

- To test to see if a linear model is appropriate, we make an inference about the parameter β where beta is the slope of the population line.
- Null: Assume there is no linear relationship, so beta is zero. Ho: $\beta = 0$
- Degrees of freedom is $n - 2$ where n is the number of observations.
- To make a decision, look at the t-statistic and p-value from the computer output.
- To create a confidence interval for the population slope, use the t-statistic with $n - 2$ degrees of freedom and use the SE from the computer output.

Hypothesis test decisions:

Quantitative variable(s)

Use 1-sample t-test if comparing *known* population mean to unknown population mean.

Use 2-sample t-test if comparing two unknown population means if the two samples are independent.

Use 1-sample t-test for matched pairs if you want to know if there's a difference in two means given that the two samples are NOT independent---matched on some variable.

Use 1-sample t-interval to estimate unknown population mean.

Use 2-sample t-interval to estimate the difference between two unknown population means.

Categorical variable(s)

Use 1-proportion z-test if comparing a known population proportion to an unknown population proportion.

Use 2-proportion z-test if comparing two unknown population proportions if the samples are independent.

Use 1-proportion z-interval to estimate an unknown population proportion.

Use 2-proportion z-interval to estimate the difference between two unknown population proportions.

Use Chi-square goodness of fit test if testing to see if the distribution of one categorical variable for one population matches some given distribution.

Use Chi-square test of homogeneity if you want to know if the distribution of one categorical variable is the same for two or more independent samples.

Use Chi-square test of independence if you want to determine if there is a relationship between two categorical variables.