AP EXAM REVIEW

CHAPTER 1: EXPLORING DATA

DEF'S:
Individuals: Objects described by a set of data
Variable: Any characteristic of an individual.
Categorical Variable: Records which of several groups of categories an individual belongs to.
Quantitative Variable: Has numerical values that measure some characteristic of each individual.
Outlier: An individual observation that falls outside the overall pattern of a graph or set of data.
Nonresistant: When data is sensitive to the influence of extreme observations. (mean, SD)

DISPLAYING DISTRIBUTIONS WITH GRAPHS:

*DOTPLOTS

*HISTOGRAMS

*STEMPLOTS

*TIME PLOTS

INTERPRETING GRAPHICAL DISPLAYS:

-Give center and spread
-Describe shape------symmetric...skewness
-Clusters and gaps
-Outliers and other unusual shapes

DESCRIBING DISTRIBUTIONS WITH NUMBERS:

-Measures of center: mean, median, mode(?)
-Measure of spread: range, interquartile range, standard deviation
-Measuring position: quartiles, percentiles

Use 5 number summary for skewed distributions; Use mean and SD for symmetric (normal) dist.

CHAPTER 2: NORMAL DISTRIBUTIONS

DEF'S:

Density Curve: (1) always on or above X-axis; (2) has area = 1 underneath it

Percentile: Percent of the distribution that is at or to the left of the observation.

*Locating mean and median on density curves

NORMAL DISTRIBUTIONS:

- -symmetric
- -single-peaked
- -bell-shaped
- -described by its mean μ, and its SD σ

68-95-99.7 Rule: In normal dist's with mean μ, and SD σ

68% of the observations fall within σ of the mean

95% of the observations fall within 2σ of the mean

99.7% within 3σ of the mean

STANDARDIZED OBSERVATIONS:

If x is an observation from a dist. with mean μ and SD σ, the **standardized value** of x is

$$z = \frac{x - \mu}{\sigma}$$

Use Table A to find Normal Proportions with your Z score.

ASSESSING NORMALITY

-Check histogram, stemplot, boxplot for shape....looking for bell shaped

-Construct and interpret normal probability plots.

CHAPTER 3: EXAMINING RELATIONSHIPS

DEF'S:

Response variable: Measure an outcome of a study
Explanatory variable: Attempts to explain the observed outcomes.
Scatterplot: Show the relationship between two quantitative variables measure on same individuals.
Correlation: Measures the strength and direction of the linear relationship between two quantitative variables.
Regression Line: A straight line that describes how a response variable y changes as an explanatory variable x changes.
Coefficient of determination r^2: Fraction of the variation in the values of y that is explained by least-squares regression of y on x.
Residual: Difference between observed and predicted values by the regression line.
Residual Plot: Plots residuals on the vertical axis vs. explanatory variable on horizontal axis.

SCATTERPLOTS:

Analyze patterns: Give direction, form, strength......clusters......outliers vs. influential points

Correlation: KNOW PROPERTIES ON PAGE 132 OF TEXT!!!!! Distinguish between r vs. r^2.

RESIDUAL PLOTS:

A Curved pattern----means not linear
Individual points with large residuals----Outliers

REGRESSION

BE ABLE TO EXPLAIN WHAT THE SLOPE AND INTERCEPT MEAN IN THE EQUATION: $y = a + bx$

The least square regression line is: $y = a + bx$

with **slope** $b = r\frac{s_y}{s_x}$ and **intercept** $a = y - bx$

CHAPTER 4: MORE ON TWO- VARIABLE DATA

DEF'S:

-A variable grows **linearly** over time if it adds a fixed increment in each equal time period.
-A variable grows **exponentially** if it is multiplied by a fixed number greater than 1 in each equal time period.
Extrapolation: The use of a regression line or curve for prediction outside the domain of values of the explanatory variable. Such predictions cannot be trusted.
Lurking variable: A variable that has an important effect on the relationship among the variables in a study but is not included among the variables studied.
Simpson's Paradox: The reversal of the direction of a comparison or an association when data from several groups are combined to form a single group.

***Understand TRANSFORMATIONS to achieve linearity: logarithmic and power transformations.

ASSOCIATION IS NOT CAUSATION!!!!

Understand the following relationships:

-Causation:

-Common Response:

-Confounding:

EXPLORING CATEGORICAL DATA:

-Marginal distributions and two way tables
-Conditional distributions

CHAPTER 5: PRODUCING DATA

DEF'S:

Census: A complete enumeration of an entire population.
Voluntary Response Sample: Consists of people who choose themselves by responding to a general appeal.
Two variables are **confounded** when their effects on a response variable cannot be distinguished from each other.
Population: An entire group of individuals we want information about.
Sample: A part of the population that we actually examine in order to gather information.
Design: The method used to choose the sample from the population.
Convenience Sampling: Method which chooses the individuals easiest to reach.
Bias: When a study systematically favors certain outcomes.
Simple Random Sample: Consists of *n* individuals from the population chosen in such a way that every set of *n* individuals has an equal chance of being selected.
To choose a **stratified random sample** divide the population into **strata**, groups of individuals that are similar in some way that is important to the response. Then choose a separate SRS from each stratum.
Undercoverage: Occurs when some groups in the population are left out of the process of choosing the sample.
Nonresponse: Occurs when an individual chosen for the sample can't be contacted or refuses to cooperate.

PLANNING AND CONDUCTING SURVEYS

-Know characteristics of a well designed and well conducted survey-----SRS's!!!!

-Be able to identify sources of BIAS in surveys
-undercoverage, nonresponse, response bias, wording of question, etc........

CHAPTER 5: PRODUCING DATA con't

DEF'S

Observational Study: Observes individuals and measure variables of interest but does not attempt to influence the responses.
Experiment: Deliberately imposes some treatment on individuals in order to observe their responses.
Experimental Units: Individuals on which an experiment is done.
Subjects: When the experimental units are human beings.
Treatment: A specific experimental condition applied to the units.
Factors: The explanatory variables used in an experiment.
Placebo Effect: A dummy treatment than can have no physical effect.
Control group: The group who receives the placebo...controls the effect of lurking variables.

PLANNING AND CONDUCTING EXPERIMENTS

-Know characteristics of a well designed and well conducted experiment

-use of randomization:

-Double Blind experiment:

-Block design:

-Matched Pairs:

3 PRINCIPLES OF STATISTICAL DESIGN OF EXPERIMENTS:

1) CONTROL........experiments need to compare 2 or more treatments to avoid confounding
2) RANDOMIZATION.......prevents bias.....creates treatments that are similar
3) REPLICATION.....reduces the role of chance variation

SIMULATIONS

-the imitation of chance behavior based on a model that accurately reflects the experiment under consideration
-know how to use TABLE OF RANDOM DIGITS AND CALCULATORS for simulated data.

CHAPTER 6: PROBABILITY: THE STUDY OF RANDOMNESS

DEF'S:

We call a phenomenon random if individual outcomes are uncertain but there is nonetheless a regular distribution of outcomes in a large number of repetitions.
Probability: of any outcome is the proportion of times the outcome would occur in a very long series of repetitions.
Independent Trials: When the outcome of one trial does not influence the outcome of any other.
Sample Space: (of a random phenomenon) The set of all possible outcomes.
Event: An outcome or a set of outcomes of a random phenomenon. Event is a subset of the sample space.
Disjoint: When two events have no outcomes in common (never occur simultaneously).
Complement: of an event consists of exactly the outcomes that are not in the event.
Union: of any collection of events is the event that at least one of the collection occurs.
Intersection: of any collection of events is the event that *all* of the event occur.
Joint event: The simultaneous occurrence of two events.
Conditional Prob.: Gives the probability of one event, under the condition of knowing another event

KNOW/UNDERSTAND THE FOLLOWING RULES:

Multiplication Principle: if you can do one task *a* number of ways and a second task *b* number of ways, then both tasks can be done in *a* x *b* number of ways.

Probability Rules:

-The probability P(A) of any event satisfies : $0 \le P(A) \le 1$
-If S is the sample space in a probability model, then P(S)=1.
-The **complement** of any event A is the event that A does not occur, written A^c
The **complement rule** states that P(A^c)= 1-P(A).
-If A and B are **disjoint** events, then P(A or B) = P(A) + P(B).
-Two events A and B are **independent** if knowing that one occurs does not change the probability that the other occurs. If A and B are independent, P(A and B) = P(A)P(B).
-For any two events A and B, P(A or B)= P(A) + P(B) - P(A and B).

Joint Prob: P(A and B) = $P(A)P(B|A)$ Conditional Prob: $P(B|A) = \dfrac{P(A \text{ and B})}{\text{P(A)}}$

ALSO: Should be able to use VENN DIAGRAMS and TREE DIAGRAMS to determine simple probabilities.
NOTE: Probabilities in a finite sample space must be numbers between 0 and 1 and sum to 1.

CHAPTER 7: RANDOM VARIABLES

DEF'S:
Random Variable: A variable whose value is a numerical outcome of a random phenomenon.
Probability Distribution: of a random variable tells us what the possible values are of the variable, and how probabilities are assigned to those values.
Discrete Random Variable: Has a countable number of possible values. The probability distribution lists the values and their probabilities.
Continuous Random Variable: Takes all values in an interval of numbers. The probability distribution is described by a density curve. (example: normal distributions)

PROBABILITY DISTRIBUTIONS:

-Every probability p_i is a number between o and 1
- $p_1 + p_2 + \ldots\ldots + p_k = 1$

Mean of a Discrete Random Variable: $\mu_x = x_1 p_1 + x_2 p_2 + \ldots\ldots + x_k p_k = \sum x_i p_i$ (expected value)

Variance of a Discrete Random Variable: $\sum (x_i - \mu_x)^2 p_i$ STANDARD DEVIATION IS $\sqrt{Var}$.

*****MUST understand the "Law of Large Numbers" concept!!!!

RULES FOR MEANS:

$\mu_{a+bX} = a + b\mu_X$
$\mu_{X+Y} = \mu_X + \mu_Y$

RULES FOR VARIANCES:

$\sigma^2_{a+bX} = b^2 \sigma^2_x$
$\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y$

$\sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y$

ALSO SHOULD BE ABLE TO:

-Construct a probability histogram for a random variable
-Determine probabilities of events as area under density curves
-Use the standard normal distribution to find probabilities of as events of areas under standard normal curves

CHAPTER 8: THE BINOMIAL AND GEOMETRIC DISTRIBUTIONS

CONDITIONS FOR A BINOMIAL SETTING:

1. Each observation falls into one of two categories...."success" or "failure"
2. There is a fixed number *n* of observations.
3. The *n* observations are all independent.
4. The probability of success, call it *p*, is the same for each observation.

The distribution of the count X of successes in the binomial setting is the binomial distribution with parameters *n* and *p*. *n* is the number of observations and *p* is the probability of a success on any one observation. The possible values of X are the whole numbers from 0 to *n*. We say that X is B(*n*,*p*).

If X has the binomial distribution with *n* observations and probability *p* of success on each observation, the possible values of X are 0,1,2,....,*n*. If *k* is any one of these values,

$$P(x = k) = nCk \cdot p^k \cdot (1 - p)^{n-k}$$

The mean and standard deviation of the binomial variable X are:

$$\mu = np$$
$$\sigma = \sqrt{np(1-p)}$$

CONDITIONS FOR A GEOMETRIC SETTING:

1. Each observation falls into one of two categories...."success" or "failure"
2. The probability of success, call it *p*, is the same for each observation.
3. The *n* observations are all independent.
4. The variable of interest is the number of trials required to obtain the first success.

If X has a geometric distribution with probability *p* of success and (1-*p*) of failure on each observation, the possible values of X are 1,2,3.....If *n* is any one of these values, then the probability that the first success occurs on the *n*th trial is: $P(X = n) = (1 - p)^{n-1} p$

The mean or expected value of a geometric random variable is $\mu = \frac{1}{p}$

The probability that it takes *more* than *n* trials to see the first success is: $P(x > n) = (1 - p)^n$

Given a random variable X, the cumulative distribution function of X calculates the sum of the probabilities for 0,1,2,....., up to the value X. That is, is calculates the probability of obtaining at most X successes in *n* trials.

NOTE: KNOW HOW TO DO BINOMIAL/GEOMETRIC PROB'S ON CALCULATOR!!!

CHAPTER 9: SAMPLING DISTRIBUTIONS

DEF'S:

Parameter: A number that describes the population. The value of a parameter is not known.
Statistic: A number that can be computed from the sample data without making use of any unknown parameters. We use *statistic* to estimate an unknown *parameter*.
Sampling Distribution: The distribution of values taken by a statistic in all possible samples of the same size from the same population.
A statistic used to estimate a parameter is unbiased if the mean of its sampling distribution is equal to the true value of the parameter being estimated.

The VARIABILITY OF A STATISTIC is described by the spread of its sampling distribution. This spread is determined by the sampling design and the size of the sample. Larger samples give smaller spread.

Sampling Distribution of a Sample Proportion

Choose an SRS of size *n* from a large population with population proportion *p* having some characteristic of interest. Let $\hat{p}$ be the proportion of the sample having that characteristic.

- The sampling distribution of $\hat{p}$ is approximately normal and is closer to a normal distribution when the sample size *n* is large.
- The mean of the sampling distribution is exactly *p*.
- The standard deviation of the sampling distribution is $\sqrt{\dfrac{p(1-p)}{n}}$

Rule of Thumb #1: Use the above for SD of $\hat{p}$ only when the pop. is at least 10 times as large as the sample.

Rule of Thumb #2: Use the normal approximation to the sampling distribution of $\hat{p}$ for values of *n* and *p* that satisfy $np \geq 10$ and $n(1-p) \geq 10$.

Note: Remember that μ and σ are **parameters** for mean and standard deviation; while $\bar{x}$ and s are **statistics** calculated from sample data.

Mean and SD of a Sample Mean:

Suppose that xbar is the mean of an SRS of size *n* drawn from a large population with mean μ and SD σ. Then the mean of the sampling distribution of xbar is μ and its SD is $\dfrac{\sigma}{\sqrt{n}}$.

CENTRAL LIMIT THM

Draw an SRS of size *n* from any population with mean μ and SD σ. When *n* is large, the sampling distribution of the sample mean xbar is close to the normal distribution $N(\mu, \dfrac{\sigma}{\sqrt{n}})$ with mean μ and SD $\dfrac{\sigma}{\sqrt{n}}$.

CHAPTER 10: INTRODUCTION TO INFERENCE

*In general, confidence intervals have the form: *estimate* $\pm$ *margin of error*, where our estimate is our guess for the value of the unknown parameter, and our margin of error is how accurate we believe our guess is, based on the variability of the estimate.
*Confidence intervals have two parts: an *interval* computed from the data, and a **confidence level** giving the probability that the method produces an interval that covers the parameter.
*A **level C confidence interval** for a parameter is an interval computed from sample data by a method that has probability C of producing an interval containing the true value of the parameter.

A level C CI (confidence interval) for μ and known standard deviation σ is $x_{bar} \pm z^* \frac{\sigma}{\sqrt{n}}$
(where z^* is the upper (1-C)/2 critical value for the standard normal distribution.)

Assumptions for above CI: (1) SRS; (2) Population is normal or $n \geq 30$; (3) Sigma known

**When using CI's we would like high confidence and small margin of errors (ME). The ME of a confidence interval gets smaller as:

(1) the confidence level C decreases
(2) the population standard deviation decreases
(3) the sample size *n* increases

*To determine the sample size *n* that will yield a confidence interval for a population mean with a specified margin of error *m*, set the expression for the ME to be less than or equal to *m* and solve for *n*: $z^* \frac{\sigma}{\sqrt{n}} \leq m$

CAUTIONS to keep in mind when using all confidence intervals:

(1) Data must be from a simple random sample.
(2) Outliers can have a large effect on confidence intervals
(3) If the sample size is small and the population is not normal, the true confidence level will be different from the value C used in computing the interval.

MAKE SURE YOU KNOW WHO TO INTERPRET A CONFIDENCE INTERVAL CORRECTLY:
"95% of all confidence intervals capture the value of the parameter";
"we are 95% confident the our sample produces one of the confidence intervals that contains the unknown parameter"

CHAPTER 10: INTRODUCTION TO INFERENCE *continued*

*A **test of significance** is intended to assess the evidence provided by data against a **null hypothesis** H_O in favor of an **alternative hypothesis.**
*The hypotheses are stated in terms of population parameters. Usually H_O is a statement that no effect is present, and H_a says that a parameter differs from its null value in a specific direction (one-sided) or in either direction (two-sided).
*The reasoning of a test of significance is as follows: Suppose for the sake of argument that the H_O is true. If we repeated our data production many times, would we often get data as inconsistent with H_O as the data we actually have? If the data are unlikely, when H_O is true, they provide evidence against H_O.

*A test of significance is based on a **test statistic**. The **P-value** is the probability, computed supposing H_O to be true, that the test statistic will take a value at least as extreme as that actually observed. Small P-values indicates strong evidence against H_O .
*If the P-value is as small or smaller than a specified value "alpha", the data are **statistically significant** at the "alpha" significance level.

Z test for a population mean :

1) State parameter of interest
2) State Choice of Test: **one sample z test for a population mean**
3) Check Assumptions: **a) SRS; b) normal population or n≥30; c) sigma known**
4) State Hypotheses
4) Calculate Test Statistic: $Z = \dfrac{x_{bar} - \mu}{\dfrac{\sigma}{\sqrt{n}}}$
5) Find the P-Value
6) Make Decisions at the "alpha" significance level: **Reject H_O or Fail To Reject H_O**
7) Interpretation in context of the problem.
(NOTE: You should also be aware of the rejection region approach!!)

In the case of testing H_O versus H_a , decision analysis chooses a decision rule on the basis of the probabilities of two types of error. A **type I error** occurs if we reject H_O when it is in fact true. A **Type II error** occurs if we accept H_O when in fact H_a is true. The **power** of a significance test measures its ability to detect an alternative hypothesis. The **power** against a specific alternative is the probability that the test will reject H_O when the alternative is true. In a fixed level "alpha" test, the level alpha is the probability of **type I error**, and the power against a specific alternative is 1 minus the probability of **type II error** for that alternative.

CHAPTER 11: INFERENCE FOR DISTRIBUTIONS

Standard Error: When the standard deviation of a statistic is estimated from the data, the result is called the **standard error** of the statistic. The standard error of a sample mean is the sample standard deviation divided by the square root of the sample size.

The T Distribution: A non-normal distribution used when the population Standard Deviation is not known. A *t* statistic has the same meaning as a *z* statistic. We specify a particular t distribution by giving its **degrees of freedom** (sample size minus 1).

-The density curves of the t distributions are similar in shape to the standard normal curve. They are symmetric about zero and are bell shaped.

-The spread of the t distributions is greater than that of the standard normal distribution.

-As the degrees of freedom increase, the density curve approaches the standard normal curve.

t CONFIDENCE INTERVALS:

$x_{bar} \pm t^* \frac{s}{\sqrt{n}}$, where t^* is the upper (1-C)/2 critical value for the t(n-1) distribution.

Assumptions for use: (1) SRS; (2) Population is normal or $n \geq 30$; (3) σ is unknown.

t test for a population mean :

1) State parameter of interest

2) State Choice of Test: **one sample t test for a population mean**

3) Check Assumptions: **a) SRS; b) normal population or $n \geq 30$; c) σ is unknown**

(note: it is a good idea to always plot your data if the sample size is <30....check for normality)

4) State Hypotheses

4) Calculate Test Statistic: $t = \frac{x_{bar} - \mu}{\frac{s}{\sqrt{n}}}$ **Matched Pairs:** $t = \frac{\bar{x}_d - \mu_d}{\frac{s_d}{\sqrt{n}}}$

5) Find the P-Value (**with n-1 degrees of freedom)**

6) Make Decisions at the "alpha" significance level: **Reject H_O or Fail To Reject H_O**

7) Interpretation in context of the problem.

(NOTE: You should also be aware of the rejection region approach!!)

Also be aware of using this procedure for a **matched pairs** situation. To compare the responses to the two treatments in a matched pairs design, apply the one-sample t procedures to the observed differences. The parameter μ is a matched pairs t procedure is the mean difference in the responses to the two treatments within matched pairs of subjects in the entire population.

CHAPTER 11: INFERENCE FOR DISTRIBUTIONS *continued*

* A confidence interval of significance test is called **robust** if the confidence level or P-Value does not change very much when the assumptions of the procedure are violated.

*When the goal of inference is to compare the responses of two treatments of to compare the characteristics of two populations (AND we have separate samples from each treatment or population) we will use two sample procedures.

When using two sample t procedures, we will use degrees of freedom calculated in one of the following ways:

(1) Use degrees of freedom calculated from your calculator (or technology)

(2) Use degrees of freedom equal to the smaller of n_1 -1 and n_2 - 1.

Draw an SRS of size n_1 from a normal population with unknown mean μ_1, and draw an independent SRS of size n_2 from another normal population with unknown mean μ_2. The confidence interval for $\mu_1 - \mu_2$ is: $(x_{bar1} - x_{bar2}) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

two sample t test for population means :

1) State parameter of interest: **"....the difference in population means.."**
2) State Choice of Test: **two sample t test for the difference in mean**
3) Check Assumptions: **a) two SRSs; b) two independent samples from normal populations OR $n_1+n_2>40$; c)both populations SD's unknown**

(note: it is a good idea to always plot your data if the sample size is <30....check for normality)

4) State Hypotheses: $H_o: \mu_1 = \mu_2 \quad \mathrm{H_a}: \mu_1 <, \neq, > \mu_2$

4) Calculate Test Statistic: $t = \dfrac{(x_{bar1} - x_{bar2})}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

5) Find the P-Value **(with degrees of freedom calculated in above ways)**

6) Make Decisions at the "alpha" significance level: **Reject H_O or Fail To Reject H_O**

7) Interpretation in context of the problem.

(NOTE: You should also be aware of the rejection region approach!!)

CHAPTER 12: INFERENCE FOR PROPORTIONS

Tests and confidence intervals for a population proportion *p* when the data are an SRS of size *n* are based on the **sample proportion** *phat*. When *n* is large, *phat* has approximately the normal distribution with mean *p* and standard deviation $\sqrt{\frac{p(1-p)}{n}}$.

Assumptions for inference about a proportion:

- The data are an SRS from the population.
- The population is at least 10 times larger than the sample.
- For a **confidence interval**, *n* is so large that both the count of successes *nphat* and the count of failures *n(1-phat)* are 10 or more.
- For an **inference test**, the sample is so large that both np_o and $n(1-p_o)$ are 10 or more.

Confidence interval for population proportion: $\hat{p} \pm z^* \sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}}$

One sample z proportion test :

1) State parameter of interest:
2) State Choice of Test: **one sample z test for a population proportion**
3) Check Assumptions: **see above**
4) State Hypotheses: $H_o: p = p_o$
4) Calculate Test Statistic: $z = \frac{\left(\hat{p} - p_0\right)}{\sqrt{\frac{\hat{p}_o\left(1-\hat{p}_o\right)}{n}}}$
5) Find the P-Value
6) Make Decisions at the "alpha" significance level: **Reject H_O or Fail To Reject H_O**
7) Interpretation in context of the problem.

(NOTE: You should also be aware of the rejection region approach!!)

To determine the sample size *n* that will yield a level C confidence interval for a population proportion *p* with a specified margin of error *m*, solve the following for *n*: $z^* \sqrt{\frac{p^*\left(1-p^*\right)}{n}} \le m$

where p^* is a guessed value based on either past experience with similar studies, OR a conservative guess of .5.

CHAPTER 12: INFERENCE FOR PROPORTIONS con't

When we want to compare the proportions of successes in two populations, the comparison is based on the difference between the sample proportions of successes. When the two sample sizes are large enough, we can use Z procedures because the sampling distribution of the difference in sample proportions is close to normal.

Confidence interval for comparing two proportions:

Draw an SRS of size n_1 from a population having proportion p_1 for successes and draw an independent SRS of size n_2 fro another population having proportion p_2 of successes. When both samples are large the Confidence Interval is:

$$\left(\hat{p}_1 - \hat{p}_2\right) \pm z^* \sqrt{\frac{\hat{p}_1\left(1-\hat{p}_1\right)}{n_1} + \frac{\hat{p}_2\left(1-\hat{p}_2\right)}{n_2}}$$

Use this confidence interval when (assumptions!!) both populations are are least 10 times as large as the samples; and when $n_1\hat{p}_1$, $n_1\left(1-\hat{p}_1\right)$, $n_2\hat{p}_2$, $n_2\left(1-\hat{p}_2\right)$ are all 5 or more.

When doing a test to compare two proportions, we use the **pooled sample proportion**.

$\hat{p}$ = (count of successes in both samples combined)/(count of observations in both samples combined)

Two sample z proportion test :

1) State parameter of interest: **...."interested in the difference between proportions..."**
2) State Choice of Test: **two sample z test for the difference in population proportions**
3) Check Assumptions: **same as confidence interval, but use the pooled proportion**
4) State Hypotheses:
4) Calculate Test Statistic: $z = \dfrac{\left(\hat{p}_1 - \hat{p}_2\right)}{\sqrt{\hat{p}\left(1-\hat{p}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$
5) Find the P-Value
6) Make Decisions at the "alpha" significance level: **Reject H_0 or Fail To Reject H_0**
7) Interpretation in context of the problem.

(NOTE: You should also be aware of the rejection region approach!!)

CHAPTER 13: INFERENCE FOR TABLES: CHI-SQUARE PROCEDURES

The **chi-square distributions** are a family of distributions that take only positive values and are skewed to the right. A specific chi-square distribution is specified by one parameter, called the **degrees of freedom**.

Chi square density curves have the following properties:

1) The total area under a chi-square curve is equal to 1.
2) Each chi-square curve begins at 0 on the horizontal axis, increases to a peak, and then approaches the horizontal axis asymptotically from above.
3) Each chi-square curve is skewed to the right. As the number of degrees of freedom increase, the curve becomes more and more symmetrical and looks more like a normal curve.

A **Goodness of Fit Test** is used to help determine whether a population has a certain hypothesized distribution, expressed as percents of population members falling into various outcome categories. Suppose that the hypothesized distribution has *n* outcome categories. To test the hypothesis H_0: the actual population percents are *equal* to the hypothesized percentages, first calculate the chi-squared statistic:

$$X^2 = \sum (Observed - Expected)^2 / Expected$$

Then X^2 has approximately a "chi-squared" distribution with (n-1) degrees of freedom.
For a test of H_0 against H_a: the actual population percentages are *different from* the hypothesized percentages the P value is P("chi-squared" $\geq X^2$). You may use this test when no expected counts are < 5.

Chi-Square Goodness of Fit Test:

1) State parameter of interest: **...."interested in the hypothesized distribution of...."**
2) State Choice of Test: **chi-square goodness of fit test**
3) Check Assumptions: **no expected counts < 5**
4) State Hypotheses:
4) Calculate Test Statistic: $X^2 = \sum \frac{(O-E)^2}{E}$
5) Find the P-Value **with n-1 degrees of freedom, where n is the number of categories**
6) Make Decisions at the "alpha" significance level: **Reject H_O or Fail To Reject H_O**
7) Interpretation in context of the problem.

(NOTE: You should also be aware of the rejection region approach!!)

CHAPTER 13: INFERENCE FOR TABLES: CHI-SQUARE PROCEDURES con't

*The **chi square test** for a two way table tests the null hypothesis that there is no relationship between the row variable and the column variable.
*One common use of the chi-square test is to compare several population proportions. The null hypothesis states that all of the population proportions are equal. The alternative hypothesis states that they are not all equal but allows any other relationship among the population proportions.

The **expected count** in any cell of a two way table when the Null is true is:
expected = (row total x column total) / (table total)

The chi-square test compares the value of the chi-squared statistic with critical values from the chi-square distribution with (r-1)(c-1) **degrees of freedom**. Large values of X^2 are evidence against the Null Hypothesis.

Chi-Square Test for Independence:

1) State parameter of interest: **...."interested in whether these two categorical variables are related..."**
2) State Choice of Test: **chi-square test for independence**
3) Check Assumptions: **no expected counts < 5**
4) State Hypotheses:
4) Calculate Test Statistic: $X^2 = \sum \frac{(O-E)^2}{E}$
5) Find the P-Value **with (r-1)(c-1) degrees of freedom, (r-rows; c-columns)**
6) Make Decisions at the "alpha" significance level: **Reject H_O or Fail To Reject H_O**
7) Interpretation in context of the problem.
(NOTE: You should also be aware of the rejection region approach!!)

SECTION 14.1: INFERENCE ABOUT A REGRESSION MODEL

Assumptions for regression inference:

1) For any fixed value of x, the response y varies according to a normal distribution
   Repeated responses y are independent of each other.
2) The mean response μ_y has a straight line relationship with x: $\mu_y = \alpha + \beta x$
   The slope β and the intercept α are unknown parameters.
3) The standard deviation of y (call it σ) is the same for all values of x. It is unknown.

The slope *b* of the LSRL is an unbiased estimator of the true slope β.
The intercept a of the LSRL is an unbiased estimator of the true intercept α.

A level C confidence interval for the slope β. of the true regression line is: $b \pm t^* SE_b$, where

$$SE_b = \frac{s}{\sqrt{\sum\left(x - \bar{x}\right)^2}}$$

Chi-Square Test for Independence:

1) State parameter of interest: **...."interested in the slope of the regression line..."**
2) State Choice of Test: **t test for regression slope**
3) Check Assumptions: ***see above***
4) State Hypotheses: $H_o: \beta = 0 \quad \mathrm{H_a}: \beta <, \neq, > 0$
4) Calculate Test Statistic: $t = \dfrac{b}{SE_b}$
5) Find the P-Value **with n-2 degrees of freedom**
6) Make Decisions at the "alpha" significance level: **Reject H_O or Fail To Reject H_O**
7) Interpretation in context of the problem.

(NOTE: You should also be aware of the rejection region approach!!)

YOU MUST BE ABLE TO READ/INTERPRET COMPUTER OUTPUT FOR THESE PROBLEMS!!!

MAKE SURE YOU CAN INTERPRET THIS COMPUTER OUTPUT CORRECTLY.

The regression equation is
C2 = 1.77 + 0.0803 C1

Predictor	Coef	Stdev	t-ratio	p
Constant	1.76608	0.03068	57.57	0.000
C1	0.080284	0.001617	49.66	0.000

s = 0.009068 R-sq = 99.8% R-sq(adj) = 99.8%