AP STATISTICS REVIEW

Inference about:	Given info:	Test:
Population Means (μ)	One sample	One sample t test
	Matched pairs	One sample t test for differences
	Two samples	Two sample t test
Population Proportions (p)	One sample	One sample z test
	Matched pairs	One sample z test
	Two samples	Two sample z test
	Several samples	Chi Squared test
Relationships between two Categorical variables		Chi Squared test
Relationships between two Quantitative variables		Regression inference

Assumptions:

- ALL TOS assume data was collected by using an SRS.
- Assume or prove Normally distributed data if using a z or t test
- If you are using proportions then the ROT's are in effect. np and n(1-p) must EACH be greater than 10 (for each data set) and the population must be ten times the sample size.
- X^2 (Chi squared) tests have the assumptions that all expected values are greater than or equal to 1 and that no more than 20% of the expected values are less than 5. This test also uses degrees of freedom.

Degrees of Freedom are n-1 unless:

- Using X^2 matrix test then it is (r – 1)(c – 1)
- n – 2 if using a LinReg t Test or 2 sample test
- n – 1 if using t test
- n-2 if using two sample t test

Calculator notes:

Normpdf (x, μ,σ)	tpdf (x, df)	Binompdf (n, p, k)	Geompdf (p, k)
Normcdf (l, h, μ,σ)	tcdf (l, h, df)	Binomcdf (n, p, k)	Geomcdf (p, k)

General Notes

A confidence interval is a TOS if you include hypotheses and make a conclusion.

A **z test** can only be done if you have population information mean μ and standard deviation σ. When doing a z test with samples, be sure to use $\frac{\sigma}{\sqrt{n}}$ for the std dev.

A **t test** uses samples to approximate μ with $\bar{x}$ and also approximate σ with s_x. Know the conditions you need for using less than 30 samples.

A **LinReg t Test** is the TOS used to determined if two quantitative variables have a linear relationship

CLT deals with the distribution of samples (the more samples the closer to normally distributed).

LL#'s shows that as you collect more samples, the $\bar{x}$ approaches μ

Use a visual to help explain answer

Check for normality using NPP or 68-95-99.7 rule

Type I error is means you reject H_o but it is true. It is set by the user and it is the significance level α.

Type II error means you accept H_o when H_a is true. It is shown with two density curves one for H_o and the other for H_a. Use the cutoff value obtained from the H_o curve to calculate the Type II error on the H_a curve. Power is "1 – Type II error" and is the probability that you will make a correct decision.

Regression lines are for variables with an explanatory/response relationship only. Look at the residual plot to confirm model fits the data. Slim chance it is on there but look out for curved data. Log y data see if that helps. If it is still curved, log x data too.

Randomization, Repetition and Control are the key factors to an experiment. Blocking keys in on differences and you don't compare. Matched pairs key in on similarities and you do compare.

Probability is $\frac{favorable}{possible}$. Use a tree diagram to help get a visual. Think of all the paths there are to get to the same outcome.

H_o and H_a for X^2 and for regression inference are very specific. Make sure you know them.

Formulas Etc.

1. The z score puts two things into a similar context so that they can be compared fairly. The z score formula is: z = $\frac{x(bar)-\mu}{\sigma}$
2. To do a z test you do not need to know the population μ, but you must know the population σ. The z test statistic is z = $\frac{x(bar)-\mu}{\frac{\sigma}{\sqrt{n}}}$ Where n is the number of samples.

   The samples must have been obtained form an SRS of the population. A z test is basically a one sample t test.
3. A two sample z test is very similar is it z = $\frac{\left(x(bar)_1 - x(bar)_2\right)-\left(\mu_1-\mu_2\right)}{\sqrt{\frac{\sigma_1^2}{n_1}}+\sqrt{\frac{\sigma_2^2}{n_2}}}$
4. A t test uses the standard error "s" from the data set because the population σ is unknown. In fact the population μ is unknown also. All samples must be taken by an SRS from a population that is normally (or approximately normally) distributed. It can not have any outliers and must have minimal skewness. The t test statistic is found by: t = $\frac{x(bar)-\mu}{\frac{s}{\sqrt{n}}}$ Where n is the number of samples and n-1 is the number of degrees of freedom. Important assumptions to meet are if $n \le 15$ there must be no outliers. If $n \ge 15$ then it must have no outliers but a little skewness is okay. If $n \ge 40$ then it is good even if there is fairly strong skewness. This is a robust test and it gives a good, safe approximation.
5. A two sample t test is used when you have 2 SRS's from 2 different populations, they are independent samples, both populations are approximately normally distributed, and neither the population μ nor the population σ are known. The test statistic for this is: t = $\frac{\left(x(bar)_1 - x(bar)_2\right)-\left(\mu_1-\mu_2\right)}{\sqrt{\frac{s_1^2}{n_1}}+\sqrt{\frac{s_2^2}{n_2}}}$
6. An 80% confidence interval means we want 80% of the normal sample distribution of x(bar). This will leave 10% in each tail. The z^* is the point with 90 % of the data to the left and only 10% to the right in table "C" this is the value 1.282. If you look in table "A", you will find that 90% of the data falls to the right of a point at the z score value of 1.28. The two tables are saying the same thing, just in a different way.
7. The "P value" is the probability that the standard normal variable "z" takes on a value in the range simply by chance. Thus, when the "p value" is the evidence against the

H_0. So a low "p value" says that the "observed value" you were testing is unlikely to have happened just by chance and thus your H_0 is likely to be incorrect.

8. Type I error: reject H_0 when H_0 is true. AKA the significance level or α
   Type II error: accept H_0 when H_a is true

9. While the t test is used for a single item comparison, the X^2 test is used for "group" comparison to compare observed samples to the hypothesized population distribution. You can not do a X^2 test on percentages. The chi squared test statistic is: $X^2 = \frac{(O-E)^2}{E}$ and uses degrees of freedom. The more degrees of freedom, the more normal the data would look. There are two ways to describe this test. It is called the "goodness of fit test" when you have outcome categories. In this case you most often express the null hypothesis as H_0 : the actual population distribution is equal to the hypothesized distributions. You will always use the X^2 test statistic and compare observed to expected values with n-1 degrees of freedom. You may use this test when all individual expected counts are at least 1 and no more than 20% of the expected counts are less than 5. It is most often called the "chi squared test" when you are comparing categorical variables in a two-way table. Typically the null hypothesis will be: H_0: there is NO relationship between two categorical variables in the two-way table. You can use this test when: 1) you have an independent SRS from each of several populations with each individual classified according to one variable. 2) A single SRS with each individual classified according to both of two categorical variables. And 3) an entire population, with each individual classified according to both of two categorical variables.

10. Other:

a) Remember to use some sort of visual when appropriate. Visuals will help you to analyze and check for things like skewness, center, spread and shape.
b) Define an outlier as 1.5IQR
c) Normality follows the 68-95-99.7 rule, looks linear on a Normal Probability plot.
d) A regression line can only be used with linear data. This can be seen on as a scattered pattern on a residual plot. It will predict y values for given x values. It's slope tells of the change of one standard deviation in x corresponding to a change of "r" standard deviations of y.
e) If data appears to be non-linear try the only other two checks you know. Log the y data and plot it against the x data. If it then looks linear, you have exponential growth in your data. If it doesn't look linear, try to plot "log y" data against "log x" data. If that looks linear, then your data is part of a "power function". Once you get the data to look linear, you can find an LSRL and analyze.
f) Designing an experiment requires Control, Randomness and Replication. Show all three in your design. An experiment must impose a treatment. Use blocking and blind/double blind tests when appropriate. Have a reason (and state it) when you do this.

g) Probability: Use a tree diagram. Relate everything to: $\frac{favorable\_outcomes}{possible\_outcomes}$. Look for more than one path to get to a specific place and include all of them in your calculations.

Common Mistakes on AP Test

Students don't read and think about the **entire** question. Look at all answers for M/C. Read all of the parts of a F/R before answering. Students don't tie in answers within a free response question.

Students have poor communication skills and/or don't show all work (no "magic" answers). They also forget to put the answer in the **context** of the question.

Students are afraid to leave white space on answer sheet. Just because there is space, doesn't mean you have to fill it up with words or pictures.

In probability, the terms "at least" and "at most" give students some problems.

Don't use "calculator-speak" in your answers!

Don't forget your units!

Make sure your answer makes sense in the context of the problem.

Show your work – answers alone without appropriate justification will receive no credit.

Avoid Misgridding - It's a good idea to grid five or so multiple choice answers at a time to save time and avoid misgridding.

Guessing on Multiple-Choice Questions

In the multiple choice part of the exam, you can benefit by using test-smart strategies and techniques. Remember that there is a penalty for incorrect answers versus simply leaving an item blank. You receive 1 point for a correct answer, 0 points for no answer, and -1/4 for an incorrect answer. In general, you can eliminate one or two of the options on a multiple choice item, the odds shift in your favor to go ahead and guess. If you have absolutely no ideas, then it may not be wise to guess.

Good luck! ☺

Summary – Basic inference flow chart