Mr. Stein’s Words of Wisdom

I am writing this review essay for two tests — the AP Stat exam and the Applied Stat BFT. The topics are more or less the same, so reviewing for the two tests should be a similar process. Either way, you are about to take a comprehensive exam. The AP exam will probably be easier than our final, but regardless, you have a great number of ideas that need to be up front in your brain.

That is the plain truth. While I hope this generates some degree of healthy nervousness in you, it should not scare you. In a way, I think you should be looking forward to this test. You are well prepared!!! You have studied every topic that will be tested. Your job will be to look at some questions that initially may seem confusing to you and to figure out which of the topics you know apply to that question.

1. The Forest and the Trees

I wanted to give you what I think are the important ideas of this course. These ideas make up the forest. In doing so, I will leave out a lot of the details or what I call the trees. I am not saying the details are unimportant. Obviously, without them you cannot solve some of the problems. But I think most of you know most of the details (although some hard-core studying over the next few days will surely help). I want to make sure that the big ideas are front and center in your mind as you study for the last few days and go in to take your exam.

2. Describing Distributions

Remember that if you are asked to describe a data set or, more likely, to compare two or more data sets, you must always comment on center, spread, and shape. Center can be expressed with the mean (for symmetrical, well-behaved data) or median (for non-symmetrical or otherwise naughty data). Spread can be expressed with the standard deviation (for symmetrical, well-behaved data) or IQR (for non-symmetrical or otherwise naughty data). Make sure you do not represent a data set as NORMAL. It can be approximately normal, but the normal distribution is theoretical only.

Don't confuse skewed left with skewed right. Remember "Skewness is Fewness". Remember that means and standard deviations are less resistant — they move towards outliers and influential points.

Make sure your descriptions are in the context of the problem.

If the problem tells you a population is normally distributed, get out the normal table and draw a sketch. *If the problem is about sampling more than one individual you must use the standard deviation of the sampling distribution, not the population (see the Central Limit Theorem on page 6).*

Know that a z-score is the number of standard deviations above or below the mean. Be able to calculate z-scores, find probabilities given a z-score and find a z-score given a probability.

Be able to make and describe a boxplot. Know what Q1 and Q3 are. Know that IQR $=$ Q3 $-$ Q1. Know the formula for determining outliers: The lower fence is Q1 $-$ 1.5IQR and the higher fence is Q3 $+$ 1.5IQR.

3. Regression

There are some basic terms that you should be prepared to identify and interpret:

- **Slope** — For every increase of one unit in x, there is a certain increase or decrease in $\hat{y}$ (in $\hat{y}$, not in y). It is how much our model predicts y will change with one unit increase in x.
- **y-intercept** — What our model predicts when x is zero
- **r** — Correlation coefficient. Indicates the strength of the linear relationship. Beware — r can mean nothing if the data is not linear to begin with.
- **R^2 (R-squared)** — the percent of variation in y that can be explained by the regression of y on x. The sum of the squares of the residuals from the regressions line are R^2% less than the sum of the square of the residuals from the mean line.
- **Residual** — $y - \hat{y}$. A positive residual means the line is underestimating that point, a negative residual means it is overestimating that point.

Know the formula $b = r\frac{S_y}{S_x}$. Understand the formula's relationship to the idea of **Regression to the Mean**. For an increase of one standard deviation in x, we predict an increase of less than one standard deviation in y (unless, of course, $r = \pm 1$, but that's not a real-life possibility).

Regression is useful in determining the degree to which a response variable can be predicted by an explanatory variable. It says nothing about whether an explanatory variable is causing a change in the response variable. To attempt to determine causation, you need to perform a controlled experiment. Understand the role lurking or confounding variables can play

The appropriateness of the linear model should be mostly determined by looking at the data and looking at the residuals. Remember that r and R-squared do not measure how linear the data is; in fact, they both assume linearity.

Don't forget that you when you are asked if there is a relationship between two quantitative variables you should do a linear regression t-test (if they are categorical, you should do a Chi-Square Test of Independence), Don't just stop with commenting on r or R-squared.

3.1. Transformations

This procedure is basically taking some clearly non-linear data, doing some mathematical transformation on the x-data and/or the y-data to make it linear. We can then do linear regression and finally undo the transformation. These transformations can be any mathematical operation; two of the more common ones are power regression using the transformation $(\log x, \log y)$ and exponential regression uses the transformation $(x, \log y)$

4. Experiment and Survey Design

Please know the difference between an experiment and an observational study. Experiments require a treatment.

Know what a **Simple Random Sample** (SRS) is: A sample in which each group of size n has an equal chance of being chosen. Know some other random sampling techniques: systematic, stratified, multistage, etc.

Understand the difference between variation and bias. Do you remember the analogy of the dart board? If you aim towards the center but don't hit it every time — that's variation! If your aim is off and all your shots are going right, that's bias. Think about the different types of bias (undercoverage, non-response, etc.) that we have discussed. Study their names and, more importantly, what they mean.

Understand that if asked to carry put an experiment you must include (in detail) randomization, control, and replication. Be sure that you leave nothing the grader's imagination. Make sure to discuss how you will analyze the results. If you are not asked to mention a specific hypothesis test, then at least mention which results will be compared. Be very clear on the concept of blocking. *We block to control for the variation caused by a variable other than the one we are studying.* We separate our sample into two or more blocks and then essentially carry out multiple, identical experiments. We then compare results within blocks.

5. Probability

Know the multiplication and addition rules, which are respectively

$$\begin{aligned} P(\text{A and B}) &= P(\text{A}) * P(\text{B} \mid \text{A}) \\ P(\text{A or B}) &= P(\text{A}) + P(\text{B}) - P(\text{A and B}) \end{aligned}$$

Independence means that the outcome of one event will not affect the outcome of the others or $P(\text{B} \mid \text{A}) = P(\text{B})$.

Disjoint (mutually exclusive) means that the two events cannot happen simultaneously. More formally, $P(\text{B} \mid \text{A}) = 0$.

Be able to substitute to get special cases of the two general rules. For example, if A and B are independent then $P(\text{A and B}) = P(\text{A})P(\text{B})$. On the other hand, if A and B are mutually exclusive (disjoint), then $P(\text{A and B}) = 0$.

Don‘t forget how to calculate conditional probabilities. First calculate the space, then ask yourself (“self”), of these, how many meet some condition? Never forget that **tree diagrams** can really help with some complicated probability problems, particularly conditional probabilities.

Make sure that you clearly state your model: which digits represent which outcome, how many digits you will take in each trial, how you will know when to stop each trial, and how many trials you will perform.

6. Probability Distributions of Random Variables

Know that a probability distribution consists of all possible outcomes and each outcome‘s probability.

Definition 6.1

The **mean (expected value)** of a probability distribution X is the sum of each outcome times its probability:

$$\mu_X = \sum xP(x).$$

Definition 6.2

The **variance** of a probability distribution X is denoted σ_X^2 and is defined by the formula

$$\sigma_X^2 = \sum (x - \mu_X)^2 P(x).$$

The standard deviation σ_X of a probability distribution is the square root of the variance.

Note that we can, with some arithmetic, calculate μ and σ for discrete probability distributions.

6.1. Rules for Combining Means and Standard Deviations

In general, means behave exactly as we would expect, while standard deviations can be a little trickier. Let X and Y be random variables and let a be a constant. The following relations hold:

Rule for μ	Rule for σ
$\mu_{X+a} = \mu_X + a$	$\sigma_{X+a} = \sigma_X$
$\mu_{aX} = a\mu_X$	$\sigma_{aX} = \lvert a \rvert \sigma_X$
$\mu_{X \pm Y} = \mu_X \pm \mu_Y$	$\sigma_{X \pm Y} = \sqrt{\sigma_X^2 + \sigma_Y^2}$ if X, Y are independent

Pay close attention to the last formula, known as the **Pythagorean Theorem of Statistics** (it is the basis for the two sample z and t tests). When adding X and Y, if X and Y are not independent, the standard deviation of X and Y will go in the direction of the correlation; it will be higher if X and Y are positively correlated and it will be lower if X and Y are negatively correlated. The opposite is true if X and Y are dependent. In fact, the actual theorem is $\sigma_{X \pm Y}^2 = \sigma_X^2 + \sigma_Y^2 \pm 2r\sigma_X\sigma_Y$.

6.2. Binomial and Geometric Random variables

Definition 6.3

The **binomial distribution** is defined by $P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$ with x an integer. The mean and standard deviation of the binomial distribution are given by $\mu_X = np$ and $\sigma_X = \sqrt{np(1-p)}$, respectively. Typically we also define $q = 1 - p$, so $P(X = x) = \binom{n}{x} p^x q^{n-x}$ and $\sigma_X = \sqrt{npq}$.

Definition 6.4

The **geometric distribution** is defined by $P(X = x) = pq^{x-1}$. The mean and standard deviation are given by $\mu_X = \frac{1}{p}$ and $\sigma_X = \frac{\sqrt{q}}{p}$, respectively.

Make sure you can identify the binomial and geometric settings. They both have either success or failure, independence, and the probability of success is always equal. *Note that success does not necessarily mean a positive outcome. It is up to you or the problem to define what a success is.* They differ in that binomial deals with how many successes in a fixed number of trials, whereas geometric deals with how many trials until you get a success. Be sure you can use the pdf and cdf features on your calculator. Pdf calculates the probability of X taking on a specific x, while cdf calculates some X or less. To calculate the probability of obtaining some x or more, you can use 1−cdf.

7. Sampling Distributions

Know the difference between a parameter and a statistic. A **parameter** is a numerical description of a population; a **statistic** is a numerical description of a sample Understand that a sampling distribution is the distribution of statistics from all possible samples from a given population. In inference, we are using the mean and standard deviation (or standard error if we are approximating). You need to know the following formulas for the means and standard deviations of the sampling distributions of sample means and sample proportions.

Rule for μ	Rule for σ
$\mu_{\overline{x}} = \mu$	$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$
$\mu_{\hat{p}} = p$	$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}$

Remember, if you don't know p or σ (a.k.a. real life), you will estimate them with $\hat{p}$ or s and calculate the standard error instead of the actual standard deviation.

7.1. CENTRAL LIMIT THEOREM

The **Central Limit Theorem (CLT)** says, in essence, that if the sample size is large enough, the sampling distribution will be approximately normal and will in fact be normal as the sample size approaches infinity.

This theorem is why we can calculate p-values and confidence intervals using a normal probability table.

8. Confidence Intervals

Be sure you completely understand the formula for a **confidence interval**.

In general, Confidence Interval = estimate ± critical value · standard error.

Make sure you can find the critical value (z^* or t^*) for any given confidence level with either a chart, your calculator, or your memory.

Completely understand the interpretation of a confidence interval. We are 95% (or whatever) confident that the true mean (or whatever) lies between # and #. If we take samples many, many times, 95% (or whatever) of our intervals will capture the true mean (or whatever).

Be clear in your mind that is different than saying that the true mean has a 95% chance of being in the interval or that this interval has a 95% chance of capturing the mean. The parameter doesn't change, the intervals do.

Above all, remember to phrase this statement in context. Don't just say true mean, actually explain what that means.

Make sure you remember that assumptions must be checked on confidence intervals. Understand that there are certain trade-offs with confidence intervals. The larger the confidence level, the wider the interval. The larger n is, the narrower the interval. Make sure you can calculate a minimum sample size given a confidence level and margin of error.

9. Hypothesis Tests

Keep in mind the logic of a **hypothesis test**. We are assuming the truth of the null hypothesis and, if so, calculating the likelihood of this sample occurring. If the likelihood is low, we will reject the null hypothesis; if it is not low, we will fail to reject the null hypothesis.

When deciding on which test go through the following decision process:

First, ask yourself ("self"), "Is the data categorical or quantitative?".

1. Categorical
    - One sample, two categories — One proportion z-test
    - One sample, more than two categories —- Chi square (χ^2) Goodness of Fit
    - Two samples (or treatments), two categories — 2 proportion z–test
    - Two (or more) samples (or treatments), more than two categories — χ^2 test of Homogeneity
    - One sample, two variables, more than two categories — χ^2 test of Independence

2. Quantitative — if you know the population standard deviation, you can use z-tests for means, but this extremely unlikely, so you almost certainly should be using a t-test
    - One sample t-test
    - Difference between two independent samples — 2 sample t-test
    - Difference between two linked samples — Matched Pair t-test
    - Relationship between two samples — Linear Regression t-test

Make sure each of your hypothesis tests contains all of the following steps

1. Check (not just state) assumptions
2. Null and alternative hypotheses (H_0, H_a, respectively)
3. Correct formulas with correct numbers filled in appropriately
4. p-value and decision
5. Interpretation in the context of the problem. Make sure this interpretation explains the conditional probability that you have calculated.

There are some definitions that are very important for you to understand.

- p-value — The probability of getting results equal or more extreme as the sample assuming the null hypothesis is true.
- p-value — The probability of falsely rejecting the null hypothesis.
- p-value — The probability of making a Type I error.
- Type I error — Rejecting the null hypothesis when it is true.
- Type II error — Failing to reject the null hypothesis when it is false.
- Alpha (α) — The probability of a Type I error.
- Beta (β) — The probability of a type II error.
- Power — The complement of β, or in other words, $1 - \beta$.
- Power — The probability of rejecting the null hypothesis when it is false.

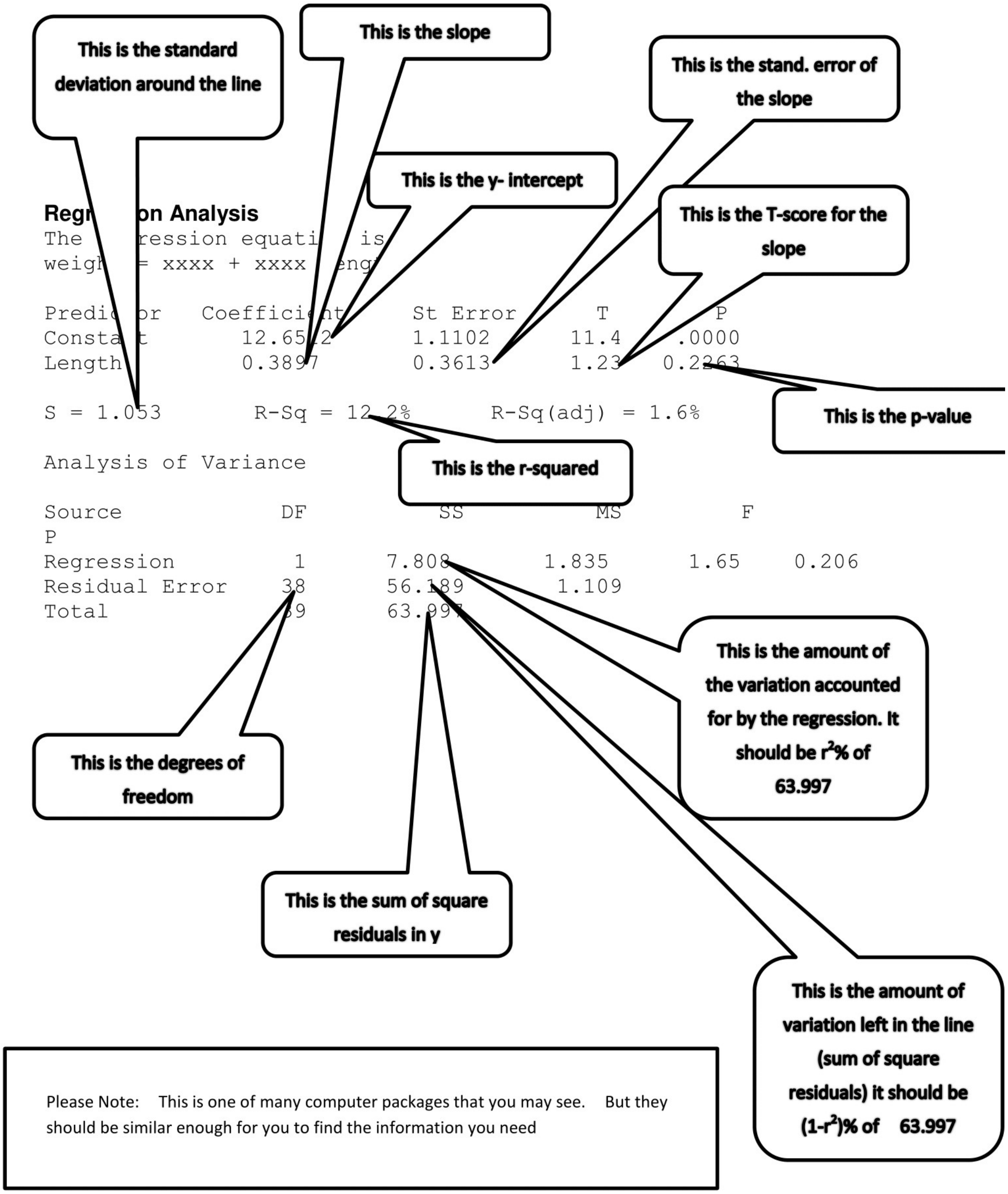
Keep in mind that there are some trade-offs here. The lower the alpha, or the significance level, the higher the beta. We can lower beta (and raise power) without adjusting alpha by increasing the sample size n. This requires more work on the experimenter‘s part, so that also is a trade-off.

Be able to calculate beta for a given alternative value of the parameter (not an AP topic) Here are all of the details of the various procedures and an example of a Minitab output. If you want a more detailed treatment of the assumptions and their importance, refer to the appendix.

	t-test	2-sample t-test	Matched Pair t-test
Assumptions	Check for SRS. Check for independence. Check if pop is normal, n is large enough, or check the data	Check for two indep. SRS's or random assignment of treatments. Pops are normal, or $n_1 + n_2$ is large enough, or check data.	Check for 2 linked SRS's or treatment on same individuals. Pop is normal, or n is large, or check data.
Hypotheses	$H_0 : \mu = \#$, H_a depends on question	$H_0 : \mu_1 - \mu_2 = 0$, H_a depends on question	$H_0 : \mu_{\text{diff}} = 0$, H_a depends on question
Picture	Put H_0 in the middle of the sampling distribution, shade from estimate: $\overline{X}$	Put 0 in the middle, shade from estimate: $\overline{X}_1 - \overline{X}_2$	Put 0 in the middle, shade from estimate: $\overline{X}_{\text{diff}}$
Formulae and Test Statistic	$\sigma_{\overline{x}} = \frac{s}{\sqrt{n}}$; $t_{\text{df}} = \frac{\overline{x}-\mu}{\sigma_{\overline{x}}}$	$\sigma_{\overline{x}_1-\overline{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$; $t_{\text{df}} = \frac{(\overline{x}_1-\overline{x}_2)-0}{\sigma_{\overline{x}_1-\overline{x}_2}}$	$\sigma_{\text{diff}} = \frac{s_{\text{diff}}}{\sqrt{n_{\text{diff}}}}$; $t_{\text{diff}} = \frac{\overline{x}_{\text{diff}}-0}{\sigma_{\overline{x}}}$
Conf. Interval	$\overline{x} \pm t^*\left(\frac{s}{\sqrt{n}}\right)$	$(\overline{x}_1 - \overline{x}_2) \pm t^*\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$	$\overline{x}_{\text{diff}} \pm t^*\left(\frac{s_{\text{diff}}}{\sqrt{n}}\right)$
Notes	To find t^*, use $n-1$ for the degrees of freedom (df). If you know σ, you can do a z-test.	For df, either use exact value from calculator or use the smaller of the df's.	ALL work should be done on differences. Do not use the original data.

	Linear Regression t-test	1-proportion z-test	2-proportion z-test
Assumptions	Check for SRS. Check scatterplot (linear model is appropriate). Independence of data (check residuals plot). Standard deviations of resids around LSRL is constant (look at residuals plot). Residuals are normal (check histogram, or n is large enough to apply CLT).	Check for SRS. Check independence and make sure pop size is greater than $10n$; Check that $np > 10$ and $nq > 10$ to approximate the sampling distribution as normal.	Check for two independent SRS's or random assignment of treatments. Check that the pop size is greater than $10n$ for respective samples unless experiment. Check that n_1p_1, n_1q_1, n_2p_2, and n_2q_2 are all greater than 10 to approximate each sampling distribution as normal.
Hypotheses	$H_0 : \beta = 0$, H_a depends on question.	$H_0 : p =$ decimal between 0 and 1, H_a depends on question	$H_0 : p_1 - p_2 = 0$, H_a depends on question
Picture	Put 0 in the middle of the sampling distribution, shade from estimate: b	Put H_0 in the middle, shade from estimate: $\hat{p}$	Put 0 in the middle, shade from estimate: $\hat{p}_1 - \hat{p}_2$
Formulae and Test Statistic	$s_e = \sqrt{\frac{\sum(y-\hat{y})^2}{n-2}}$; $SE_b = \dfrac{s_e}{s_x\sqrt{n-1}}$; $t = \frac{b-0}{SE_b}$	$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}$; $z = \frac{\hat{p}-p}{\sigma_{\hat{p}}}$	$\hat{p}_{\text{pool}} = \frac{n_1\hat{p}_1+n_2\hat{p}_2}{n_1+n_2}$; $\sigma_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{\hat{p}\hat{q}}{n_1} + \frac{\hat{p}\hat{q}}{n_2}}$ where $\hat{p} = \hat{p}_{\text{pool}}$
Conf. Interval	$b \pm t^*\left(\dfrac{s_e}{s_x\sqrt{n-1}}\right)$	$\hat{p} \pm z^*\sqrt{\frac{\hat{p}\hat{q}}{n}}$	$(\hat{p}_1 - \hat{p}_2) \pm$ $z^*\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$
Notes	$\text{df} = n-2$. You will almost certainly do this from computer output. Be familiar with output (see page 13).	For confidence intervals, assumptions and SE have $\hat{p}$ instead of p.	For assumptions and SE formula, use $\hat{p}_{\text{pool}}$ for hypothesis tests and $\hat{p}_1$ and $\hat{p}_2$ for confidence intervals.

	χ^2 GOF	χ^2 Test of Independence	χ^2 Test of Homogeneity
Assumptions	Check for SRS. All the expected cells should be greater than 5 (n is big enough).	Check for SRS. All the expected cells should be greater than 5 (n is big enough).	Check for SRS. All the expected cells should be greater than 5 (n is big enough).
Hypotheses	H_0 : The data fits the expected model. H_a : It doesn't.	H_0 : The two variables are independent. H_a : The two variables are related.	H_0 : Proportions are equal across different samples. H_a : Proportions aren't equal across different samples.
Picture	Put 0 to the left, shade to the right of the χ^2 statistic.	Put 0 to the left, shade to the right of the χ^2 statistic.	Put 0 to the left, shade to the right of the χ^2 statistic.
Formulae and Test Statistic	$\chi^2 = \sum \frac{(O-E)^2}{E}$	$E = \frac{\text{Row Tot} \times \text{Column Tot}}{\text{Group Tot}}$; $\chi^2 = \sum \frac{(O-E)^2}{E}$	$E = \frac{\text{Row Tot} \times \text{Column Tot}}{\text{Group Tot}}$; $\chi^2 = \sum \frac{(O-E)^2}{E}$
Conf. Interval	None.	None.	None.
Notes	df = # of categories − 1.	df = $(R-1)(C-1)$, where R and C are the number of rows and columns, respectively, of the contingency table.	df = $(R-1)(C-1)$. Note that the calculations for this test are exactly the same as the calculations for χ^2 Test of Independence. Both tests go by the same name "χ^2" on the TI-84 calculator



```
Regression Analysis
The regression equation is
weight = xxxx + xxxx length

Predictor   Coefficient     St Error      T        P
Constant       12.6512      1.1102       11.4    0.0000
Length         0.3897       0.3613       1.23    0.2263

S = 1.053       R-Sq = 12.2%      R-Sq(adj) = 1.6%

Analysis of Variance

Source            DF          SS          MS          F
P
Regression          1       7.808       1.835       1.65    0.206
Residual Error     38       56.189       1.109
Total              39       63.997
```

This is the standard deviation around the line

This is the slope

This is the stand. error of the slope

This is the y- intercept

This is the T-score for the slope

This is the p-value

This is the r-squared

This is the amount of the variation accounted for by the regression. It should be r^2% of 63.997

This is the degrees of freedom

This is the sum of square residuals in y

This is the amount of variation left in the line (sum of square residuals) it should be $(1-r^2)$% of 63.997

Please Note: This is one of many computer packages that you may see. But they should be similar enough for you to find the information you need

10. Final Remarks

Here are some more general pointers for you:

- On the AP, do problem #1 first, it should be easy and straight forward. Then move to question #6; it is longer and worth twice the points. It also may contain parts that are not exactly what you have done before. You should be able to do these, but they might require a little more thought. Then proceed to finish problems 2-5.
- Do not assume that you know what the test is asking you. Read the entire question (all parts) and the answer choices before you answer. Pay attention to key words (normally distributed, prediction, independence, null hypothesis, etc.).
- On multiple choice, use process of elimination. Focus on the difference between the answer choices.
- Don‘t be scared off by long and wordy multiple choice questions. Usually several of the answer choices are obviously incorrect. Focus on key words to decide between the remaining choices.
- Do not skip multiple choice questions. There is no additional penalty for getting them wrong.
- On part II, answer exactly what is being asked. Give solid statistical reasoning for your answers. Use hypothesis tests, confidence intervals, regression, etc.
- When using formulas, write down the formula, show how the numbers are plugged in, and then use the calculator to come up with the final answer.
- If you are running out of time, skip the calculations and include assumptions, H_o, H_a, and conclusions. This will get you most of the points.
- Try not to leave any part II questions blank. If you don‘t understand what to do, try to get at least one point. If you need an answer from a previous part which you couldn’t do, make up an answer and continue with the subsequent parts
- Keep telling yourself (“self”): “I am well prepared. I have taken a rigorous college-level statistics course. I know what I am doing. I just have to connect some piece of knowledge in my head with the question in front of me.”

GOOD LUCK!

A. Assumptions

Before we start talking about assumptions, you should know about **sampling distributions**. This is the distribution that contains all of the statistic of interest when collected from samples of a certain size. As a concrete example, suppose our population is all U.S. adults. If we took a SRS of 500 of them and asked them if they liked cake, then the sampling distribution would consist of the proportions from all possible samples of 500 U.S. adults.

The assumptions are arguably the most important part of the tests. Without the assumptions, the tests are invalid. The calculations, while important, are easy to get a handle of. Writing assumptions requires more care and practice.

Roughly speaking, there are three general classes of assumptions.

1. The **No Bias Assumption** — In an ideal world, we would know the parameters, always. But we never know the parameters, so what do we do? We take samples, measure the statistic, and then try to make an educated guess about the parameter. In fact, since means behave as we would expect (pun intended), if there is no bias, the mean of the sampling distribution is equal to the true mean. That is, $\mu_{\text{sampling distribution}} = \mu$. When we talk about proportions this would be $\mu_{\hat{p}} = p$; for means it's $\mu_{\overline{x}} = \mu$. But why is this true? Simple: The definition of bias means a deviation from the parameter.

   But how do we know if there's no bias? Simple: We don't. What we can do, is think to yourself (self) and see if it's reasonable to assume there's no bias. The typical way we approach this is to check if there's randomization; usually a SRS is what we're looking for. Randomization reduces variation due to hidden variables, which is why it's important to check for no bias. But just because a SRS does not mean that it's reasonable to assume no bias. Nothing else in the problem statement should jump out at you and scream "that's bias". Be on the lookout for the other types of bias we already looked at.

2. The **Independence Assumption** — If the responses of our sample are not independent, then we can't reasonably make inferences. For example, if I was sampling a friend group about whether they like a TV show, then independence could be violated since the friends would likely influence each other's opinions. If I wanted to characterize the sampling distribution, this would be problematic.

   Specifically, while the no bias assumption told us about the mean of the sampling distribution, the independence assumption tells us about the standard deviation of the sampling distribution. In particular, for proportions, we have that $\sigma_{\hat{p}} = \frac{pq}{n}$. For means, we have that $\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$. The former is something we'll be able to calculate by directly using the null hypothesis in a hypothesis test, or estimate using $\hat{p}$ and $\hat{q}$ in a confidence interval. The latter is more problematic; we usually don't know σ, so we have to approximate it with S. This has the unfortunate consequence of transforming an otherwise normal sampling distribution into a Student's t-distribution. There is a separate formula for Linear Regression t-tests: $\sigma_b = \frac{s_e}{s_x\sqrt{n-1}}$ where $s_e = \sqrt{\frac{\sum(y-\hat{y})^2}{n-2}}$.

So, how do we check independence? Again, we won't ever know for sure. But there are things we should at least think about. For example, what if the sample size is of comparable magnitude to the population size? Since most samples do not allow replacement, this would have affect independence. Consider this example.

In any case, we need to make sure the sample size is small enough compared to the population size for nonreplacement to have negligible effects on independence. How small is small enough? We make an arbitrary cutoff at 10% with the so-called 10% rule. If the sample size is less than 10% of the population size, then we can ignore the effects of nonreplacement on independence. Of course, there is more to independence than the 10% rule. You should make sure that the data are not related to each other in an obvious way. If you are working with paired data, you have to instead check that the differences are independent.

3. The **Normality Assumption** — Almost all of the tests rely on the shape of the sampling distribution being approximately normal. We care about this because it means that we can more easily do calculations. So, how do we check that the shape of the sampling distribution is approximately normal?

   One way is kind of silly. If you somehow already know that the population is normal, then certainly the sampling distribution is also normal. But that's not the majority of cases; in almost all cases you won't know that the population is normally distributed.

   If we're doing hypothesis testing or creating confidence intervals with proportions, we can use the "success/failure condition". If the sample size is n and the proportion is p, then the success/failure condition says that the sampling distribution of $\hat{p}$ is approximately normal if $np > 10$ and $nq > 10$. Note that this is an arbitrary cutoff, and in general it's better for both quantities to be significantly larger than 10.

   So, what about means? Here, we have to appeal to the Central Limit Theorem. The Central Limit Theorem states that as the sample size gets larger and larger, the sampling distribution approximates a normal distribution, regardless of the shape of the population. Again, larger is better, but we impose an arbitrary test for the sample size being large. If $n > 40$, then we say that the sample is large enough for the CLT to kick in. If $15 \leq n \leq 40$, we need to look at histograms and check that it's symmetric and unimodal. If $n < 15$, n is too small to infer anything meaningful, even with the histogram. Without more information about the population, it would be dangerous to make an assertion about normality.

   In any case, normality allows for ease of obtaining p-values and creating confidence intervals. There are of course some caveats. If we are using means and do not know the population standard deviation (which is almost always), the sampling distribution will end up being a t-distribution. However, we still need to check normality. If the sampling distribution wasn't already approximately normal if we knew the population standard deviation, then the new distribution will not be a t-distribution.

The above three assumptions account for most of the tests except for χ^2 and the linear regression t-test. We never really went into detail about why we care about the assumptions for χ^2. The SRS assumption seems obvious enough, but how about the expected frequency assumption? Roughly speaking, the reason we want E to be large enough is so that χ^2 isn't inflated.

For the linear regression t-test, there are even more assumptions.

1. The **Straight Enough Assumption** — Does the scatterplot of the data look straight enough to apply linear regression in the first place? Does the residuals plot not have any obvious pattern? If we can't justify using linear regression, making inferences with this test is nonsensical.
2. The **Independence Assumption** — Are the data independent? Again, you won't know, but you should at least think about possible relationships. Check the residuals plot; there shouldn't be any appreciable pattern.
3. The **Equal Variance Assumption/Does the Plot Thicken? (Homoscedasticity)** — We want to check that regardless of where we are on the line, the spread of the residuals should be equal. Hence, "equal variance" (homoscedastic is just a fancy word for this). If the residuals are roughly homoscedastic, then we can use one number to approximate the spread around the whole line. In order to check this assumption, look at the residuals plot. Are there any places where the variance is lower (or higher)?
4. The **Normality Assumption** We need to check that the residuals are normally distributed with respect to the line. Unless we know something extra, we usually do this by checking the histogram of the residuals. Here, we can appeal to the CLT, if n is large enough. Otherwise, the histogram of the residuals might be well-behaved enough for us to reasonable conclude that the residuals are approximately normally distributed around the line.

Remark

When performing the linear regression t-test, you need to have three pictures. The first is simply the scatterplot of the data. The second is the residuals plot. Finally, the third is the histogram of the residuals.