

MR. STEIN'S WORDS OF WISDOM

You are about to take a difficult exam. 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.

So, above all remember three things:

- 1) You are well prepared for this test.
- 2) No one expects you to get all the questions right.
- 3) Stay calm!!! My students who have not passed this test are, for the most part, those who gave up midway through the test and started to leave questions blank

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 – what I call the trees. I am not saying the details are unimportant. Obviously, without them you cannot solve some 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.

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 mean (for symmetrical data) or median (for non-symmetrical data) Spread can be standard deviation (for symmetrical data) or IQR (for non-symmetrical data). Shape can be symmetrical, approximately normal, are skewed either left or right. Make sure your descriptions are in the context of the problem.

Remember that means and standard deviations are less resistant- they move towards outliers and influential points.

If the problem tells you a population is normally distributed, get out the normal table and draw a sketch.

Know what 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. $Q1 - 1.5 IQR$ or $Q3 + 1.5 IQR$

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 y -hat
- Y-intercept- The y -hat, 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 - coefficient of determination. the percent of variation in y that can be explained by the regression of y on x
- Residual- $y - y$ -hat. A positive residual means the line is underestimating that point, a negative residual means it is overestimating that point

Remember that there are some formulas on your formulas sheet that give you way to calculate r and b . If they seem to be asking you to calculate one of these, check your formulas sheet

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 appropriate of the model should be mostly determined by looking at the data and looking at the residuals.

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.

Non-Linear Regression:

This procedure is basically taking some clearly non-linear data, doing some mathematical transformation on the x list and/or the y list to make it linear. We then can do linear regression and then undo the transformation.

Power regression using the transformation $(\log x, \log y)$
Exponential regression uses the transformation $(x, \log y)$

Experiment and Survey Design

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

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 that we have discussed. Study their names and, more importantly what they mean.

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.

Understand that if asked to carry out an experiment you must include (in detail) randomization, control, and replication. Be sure that you leave nothing to 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.

Probability

Independence means that the outcome of one event will not affect the outcome of the others.

Mutually Exclusive (or disjoint) means that the two events cannot happen simultaneously.

$P(A \text{ and } B) = P(A) * P(B)$ if and only if the two events are independent Use this formula to test if two events are independent

$P(A \text{ or } B) = P(A) + P(B)$ only if the two events are mutually exclusive. Otherwise,
 $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

Don't forget how to calculate conditional probabilities. First calculate the space, then ask yourself, of these, how many meet some condition.

Never forget that tree diagrams can really help with some complicated probability problems, particularly conditional probabilities

Probability Distributions

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

- The mean of a probability distribution (the expected value) is the sum of each outcome times its probability. $\mu_x = \sum x * P(x)$
- The variance of a probability distributions is $\sigma_x^2 = \sum (X - \mu_x)^2 * P(X)$
- The standard deviation is a probability distribution is the square root of the variance

Simulations

If you are asked to perform a simulation, you will have to use a random number table. 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.

Combining Means and Standard Deviations

$$\begin{aligned} \mu_{X+Y} &= \mu_X + \mu_Y \\ \mu_{X-Y} &= \mu_X - \mu_Y \\ \mu_{bX} &= b\mu_X \\ \mu_{X+a} &= \mu_X + a \\ \sigma_{X+Y} &= \sqrt{\sigma_X^2 + \sigma_Y^2} \\ \sigma_{X-Y} &= \sqrt{\sigma_X^2 + \sigma_Y^2} \\ \sigma_{bX} &= |b|\sigma_X \end{aligned}$$

Here are the formulas. There are a few basic rules. Combining means operates exactly the way common sense would dictate. Combining standard deviations does not. To add standard deviations, you must add variances and take the square root. Subtracting standard deviation you still add the variances and take the square root. Adding a constant to the data does not affect the standard deviation (adding two points to everyone's weight does not affect the spread of the group's weights.)

Binomial and Geometric Distributions

The formulas on your sheet are:

Binomial: $P(x = k) = \binom{n}{k} p^k (1-p)^{n-k}$, where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$

$\mu_x = np$ $\sigma_x = \sqrt{np(1-p)}$

Geometric: $P(x = k) = p(1-p)^{k-1}$ $\mu_x = \frac{1}{p}$

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. 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 calculators. Pdf calculates the probability of exactly some X, cdf calculates some X or less. To calculate some X or more, you need to use $1 - \text{cdf}$.

Sampling Distributions

Know the difference between a parameter and a statistic. A parameter is from the population; a statistic is from the sample.

Understand that a sampling distribution is the distribution of statistics from all possible samples. In inference, when we calculate the Standard Error, this is the standard deviation of the sampling distribution.

For proportions, as long as n is large enough, the mean of the sampling distribution is p and the standard deviation is $\sqrt{\frac{p(1-p)}{n}}$

For means, understand the Central Limit Theorem:

Central Limit Theorem

When n is sufficiently large, the sampling distribution from any population (regardless of shape) is approximately normal and the standard deviation of the sampling distribution (Standard Error) approaches $\frac{\sigma}{\sqrt{n}}$

Remember, if the population is known to be normal, the size of n is irrelevant. If it is not, n must be fairly large. The CLT lets us calculate using a normal table, the probability of obtaining a x-bar of a certain size. Remember that to do this calculation, you must use the standard error instead of sigma in the denominator of the z-score

Confidence Intervals

Be sure you completely understand the formula for a confidence interval

Confidence interval = estimate \pm critical value * standard error

And also that you 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.

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 , the narrower the interval

Hypothesis Tests

We made a separate note sheets with all the details of each hypothesis test (and Confidence interval, for that matter) so I will not go through that again here.

Keep in mind that all the hypothesis tests (except the chi-square tests) are basically the same. What we trying to estimate is different, the assumptions are different and the formulas for standard error are different. Otherwise the procedure is always the same.

When deciding on which test go through the following decision process

Means	or	Proportions
Do you know sigma? (Z-test) (you almost certainly won't)		Is there one sample? (1-Proportion Z-test)
Otherwise (some kind of t-test)		Are there two samples? (2-Prop Z-test)
Is there one-sample? (T-Test)		Are there more than two samples?(Chi-Sq ToI)
Otherwise (matched pair or 2 sample)		
Is there a reason to match data from one group to the other? (matched pair T-Test)		
For categorical data, you should be using either the chi-square Goodness of Fit or the Chi-Square Test of independence.		

Make sure each of your hypothesis tests in part II contains each of the following steps.

I Assumptions

II Null and alternative hypotheses

III Sketch and work (standard error and test statistic)

IV Decision

V Conclusion within the context of the problem

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 - the probability of a Type I error
- Beta - The probability of a type II error
- Power - The complement of beta ($1 - \beta$)
- 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. This requires more work on the experimenter's part, so that also is a trade-off.

Conclusions:

Here are some more general pointers for you:

- On part II, do question #1 first. This should be easy and will give you a confidence boost. Then turn to Question #6, the investigative task. Give yourself 25 minutes to do as much of #6 as you can. After 25 minutes, go back and do numbers 2, 3, 4 and 5. If you have time, go back and finish #6.
- 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.
- Skip a multiple choice question if you have no clue. If you can eliminate one or more answer choices as obviously incorrect, guess from the rest
- 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_0 and 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

And finally, I'll end where I started:

Keep telling your 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!