Preparation Pointers

[2] Review the previous test reviews.
[3] Review all your graded quizzes, tests & homework, with special attention to:
   - written feedback on steps & details of your work
   - understanding where you lost points and why
[4] Review all quizzes & tests from current & previous year posted on class website.
[6] Work through extra problems, beyond those assigned. Remember, the "Review" sections in the textbook have several additional problems, as does your ActivStats CD.

Included topics & syllabus

The final exam will be based on material from the entire syllabus. This includes the following portions of the textbook:

* Chapters 1-5  (part I)
* Chapters 6-7  (part II)
* Chapters 9-11  (part III)
* Chapters 12-13 (part IV)
* Chapters 15-18  (part V)
* Chapters 20-21  (part VI)

A more concise way to list the same thing would be:
   All chapters from 1-21, except 8, 14 and 19.

On Chapter 21, the final exam will only include what we cover in class through May 2.
**Distribution of questions on the test**

Approx. 40-50% from parts V and VI.

The rest will be split approx. equally between the other parts.

**Structure/format of the test**

* 4 questions that require complete and detailed solution. This will consist of 1-2 questions from parts V and VI above. The remaining questions will cover the other parts, or multiple parts.

* 6-8 questions that require short answers: a couple of sentences and, possibly, a sketch &/or supporting calculation.

**Reference materials**

You will be given a copy of the following materials from the textbook:


* The standard normal table.

* The student t-table.

No other reference materials will be allowed.

You are expected to bring your own calculator. No borrowing from others during the test.
Questions that require short answers

* These generally require understanding of key concepts.
* They are not difficult, but they're easy to mess up (often without even knowing it!) if you are not careful.
* Here are some examples from the textbook that could be "short answer" candidates:

Pg. 689-692: 8(a,c), 9(a,b), 23.
Pg. 571-584: 5, 9(a,b), 12, 24, 25(a,c), 27a, 38(a-c), 45, 52, 57(a,c), 58, 68.
Pg. 534-540: 5(a,b), 8, 9(a-c), 11, 13(a,b), 15, 20(a-c), 21(a,b), 24, 25(a,b), 41(a,b,d), 43, 44(b,c).
Pg. 393-398: 1(a,b), 6, 15(a,b), 20(a,b), 25(a-c), 31, 36, 42a, 44(a,b).
Pg. 315-320: 18, 22, 28, 31(a,b), 32(a,b), 41.
Pg. 139-148: 3, 6(a,b), 8(e,f), 12, 15b, 25(c,d), 31(a-c), 32(a-c), 34, 38(a,d).

* Here are a few extra probability "short" questions:

(1) Given the probabilities \( P(A) = 0.4 \), \( P(B) = 0.6 \), \( P(A \text{ or } B) = 0.65 \), find \( P(A \text{ and } B) \).

(2) Given \( P(A) = 0.7 \), \( P(B) = 0.3 \), and \( P(A \text{ and } B) = 0.28 \), determine, if possible, whether \( A \) and \( B \) are independent. Give reasons.

(3) If 1 card is drawn randomly from a standard deck of 52 playing cards, find the probability of getting a queen or clubs.

The Central Limit Theorem is the foundation of inferential statistics, and is critically important. The final exam will have at least one question that requires understanding the theorem's conceptual details and nuances, and/or requires you to state it precisely and fully. A complete statement of the theorem must include all the assumptions/conditions.
Problem solving hints for parts V-VI

It is helpful to take a "Preliminary inventory" of the following items for inference type problems:

* Is it about proportions or means?
* Does it involve 1 sample or 2 samples?
* If 2, are they independent or paired samples?
* What is the right sampling distribution model? [what is the SD or SE? what type of model - normal, student-t, df, etc.?]
* Does the problem involve conf. interval or hypothesis test?

[P.694: Ex. 32]

Preliminary inventory: This Q. is about proportions. Involves 2 independent samples. Should follow normal model for difference bet. 2 proportions.

Strategy1: Create confidence interval for the difference, and check whether it contains 0.

Strategy2: Hypothesis test with "null" saying there is no difference. Find conclusion based on P-value.

Confidence interval outline:

* Difference in sampled proportions: \( \hat{p}_1 - \hat{p}_2 = \frac{33}{53} - \frac{38}{76} = 0.1226 \) [\( \hat{p}_1 \) denotes left, \( \hat{p}_2 \) denotes right]
* Pick 95% confidence level: \( z^* = 1.96 \).
* Check assumptions: . . .

* \( SE = \sqrt{\frac{0.62 (1 - 0.62)}{53} + \frac{0.5 (1-0.5)}{76}} = 0.0879 \)

* Confidence interval = statistic ± \( z^* \) x SE = \([-0.05, 0.29]\)

* Conclusion: Based on these samples we're 95% confident that the true difference in musical ability between left- and right-handed people is between -0.05 and 0.29. Thus, there is no evidence of a statistically significant difference between these 2 groups.
Hypothesis test outline:
* H0: p1 - p2 = 0; HA: p1 - p2 ≠ 0 [note 2-tailed]
* Check assumptions: . . .
* Find SE - Must pool p1 & p2 data first, to get an average p:
  
  \[
  p_{\text{pool}} = \frac{33 + 38}{53 + 76} = 0.5504
  \]
  \[
  \text{SE} = \sqrt{\frac{0.5504(1-0.5504)}{53} + \frac{0.5504(1-0.5504)}{76}} = 0.0890
  \]

* Sampling distribution based on null hypothesis follows normal model N(0, 0.089)

* Z-score & P-value: z = 0.1226/0.089 = 1.38.
  Area above z=1.38 is 0.0838. P-value=2x0.0838=0.1676

* Conclusion: P-value of 16.76% is larger than even a 10% significance level, indicating there is more than a 10% probability that sampling variability caused the observed deviation from the null. So, we retain the null and conclude there is no evidence of a statistically significant difference between left- and right-handed people.
(a) The key here is to learn to distinguish between "distribution of data within a sample" Vs. "distribution of a particular statistic between different samples." Thus, bimodal data distribution within a sample will remain bimodal as the sample gets larger.

(b) We expect the sampling distribution of means to be equal to the true mean of the population. This is true for all sample sizes (as long as the assumption of minimum sample size is satisfied).

(c) The SD (i.e., variability) of the sampling distribution of means is $\sigma / \sqrt{n}$, where $\sigma$ is the true standard deviation of the population. The size of the sample matters.

(d) According to the Central Limit Theorem, the shape of the sampling distribution is approximately normal. From (c) above, larger sample sizes yield smaller variability. Thus, we expect the shape to become narrower and closer to normal as sample size increases.

Understanding the Central Limit Theorem & its implications is an important topic.
Preliminary inventory: question is about means; 2 samples; they're independent; should follow student-t model for diff. bet. 2 means.

(a) The mean pulse rates are almost equal, but there is greater variability in the female rates.

(b) Must check assumptions: (i) Independent data values within samples and independent samples? True, if we assume the samples are randomly drawn. Each is < 10% of the population. (ii) Approximately normal distribution within each sample? Boxplots appear symmetric & unimodal, so it satisfies the assumption. Thus, the inference methods of this Chapter are appropriate.

(c) Confidence interval has the form: \( \left( \bar{y}_1 - \bar{y}_2 \right) \pm t^*_{df} \times SE. \)

Here: \( \left( \bar{y}_1 - \bar{y}_2 \right) = 72.75 - 72.625 = 0.125. \)

\[
SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{5.372^2}{28} + \frac{7.7^2}{24}} = 1.871;
\]

\( t^*_{df} = 1.69 \) (for 90% confidence)

Therefore, CI is: \( 0.125 \pm 1.69 \times 1.871 = (-3.04, 3.28) \)

(d) Yes. Since 0 is contained within the CI, it says that the true value of our statistic (difference in mean pulse rate) could be 0.
Example: The Census Bureau reports that 26% of all U.S. businesses are owned by women. A Colorado consulting firm surveys a random sample of 410 businesses in the state and finds 115 of them have women owners. Should the firm infer the area is unusual?

Solution outline:

Preliminary inventory: question is about proportions; 1 sample; it should follow normal model for 1 proportion.

Strategy1: Hypothesis test with "null" saying prop. of women-owned businesses in CO is 0.26. Find conclusion from P-value.

Strategy2: Create confidence interval for sampled proportion, & check whether it contains 0.26.

* Null hypothesis: true proportion of women-owned businesses is 26%
  \[ H_0: \hat{p} = 0.26 \]

* Alt. hypothesis: prop. of women-owned businesses in CO is different
  \[ H_A: \hat{p} \neq 0.26 \]

* Check assumptions: (i) Independent sample; (ii) Large enough sample.

* Sampling distribution, observed proportion & its z-score:

  \[
  SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.26 \times 0.74}{410}} = 0.0217
  \]
  \[
  \hat{p} = \frac{115}{410} = 0.2805
  \]
  \[
  z = \frac{0.2805 - 0.26}{0.0217} = 0.9447
  \]

  * Find P-value for z=0.9447: Lookup standard normal table.
  
  This gives: (1-0.8276) x 2 = 0.3448 [Q: Why do we have x 2 here?]

  * Conclusion: There is a 34% chance that the observed proportion may have occurred due to sampling variability, if the null hypothesis is true. This is a fairly large probability, so we retain the null. Thus, the CO area is not different from the rest of the country in the proportion of women-owned businesses.
Confidence interval & hypothesis test solution requirements

* Must state & verify assumptions that apply.
* Must fully state the sampling distribution model you're using (its name, mean, SD, degrees-of-freedom, etc. - simply saying "normal model," or "student-t model" is not enough).
* For confidence intervals, clearly indicate confidence level & critical z* or t* value. State your final result in words (in conf. int. jargon), and infer an appropriate conclusion.
* Hypothesis tests must clearly: state the null & alt. hypotheses; show sketch & details of sampling distribution model, with P-value shown; state your final result in words and infer an appropriate conclusion.

Common errors

* Assumptions not verified -- or incorrectly, incompletely or vaguely verified.
* Doing a confidence interval, when question asks for hypothesis test, or vice versa.
* Confidence interval errors: incorrect z* or t* value; incorrect values in SE calculation; missing conclusion.
* Hypothesis test errors: incorrect hypotheses; missing sketch &/or indication of type of model; forgetting to "pool" when needed; incorrect lookup for P-value; garbled/confused conclusions.

My recommendation: Work through the "preliminary inventory" & setup for several practice exercises, even if you don't solve the whole problem. E.g: Pg. 534-540: 3, 6, 10, 17, 21, 23, 30, 37, 44.
Pg. 689-700: 5, 10, 12, 14, 18, 24, 50.
**Pointers on parts I & IV topics**

Emphasis will be on proficiency with basic concepts and graphical displays. Here are some key things you must know:

**Chapters 1-5**

* How to write a descriptive summary of quantitative variables
  - in context
  - covering shape, center, spread
  - with numerical details and graphics
* How to make key graphical displays from given data
  - histograms, boxplots, stemplots, segmented bar graphs
  - must clearly label every axis & use units when appropriate
  - must use reasonable scale for all graphs
* How to read & interpret data from graphical displays.
* How to compute numerical summaries for given data
  - must know the effect of rescaling data; effect of skew
* How to compute marginal & conditional distributions for categorical variables.

**Chapters 6-7**

Emphasis on basic concepts. Here are some things that I don't plan to put on the test:

- How to construct a regression equation
- How to compute correlation coefficients

Here are some basic things you must know:

- The purpose & objectives of a regression model
- How to identify explanatory & response variable
- How to read & interpret scatterplots
- How to interpret correlation coefficient
- How to give "in context" interpretation of the slope & intercept of a regression equation
- Simple interpretation of R-squared

**Chapters 9-13**

Expectations for these chapters are the same as what we had for Test 2. Please see the Test2-review for details.
The quality of fruit juice produced is frequently monitored at juice manufacturing plants. One of the qualities that is closely monitored is a quantitative measure of sweetness (the "sweetness index"). A company wants to determine whether the amount of the chemical pectin in the fruit juice can be used to predict the sweetness index. They collected data from 24 production runs and created the scatterplot shown below. They want to use simple linear regression to model the relationship.

(a) Identify the explanatory variable and the response variable.
(b) Discuss whether linear regression is appropriate.
(c) Find equation of regression line from the given information.
(d) Interpret the slope.
(e) The regression line has $R^2=22.8\%$. Explain, in statistical terms, what this means.

**Solution outline:**

(c) $\text{sweetness index} = mx + b$ [equation has form: $y = mx + b$]

Slope: $m = \frac{r \cdot S_y}{S_x}$. Here: $r = -0.477$, $S_y = 0.2394$, $S_x = 49.5834$

Therefore: $\text{sweetness index} = -0.0023 \cdot \text{pectin} + b$.

Plugin means & solve for $b$: $5.658 + 0.0023 \times 257.083 = b = 6.25$.

Final regression line: $\text{sweetness index} = -0.0023 \cdot \text{pectin} + 6.25$

(d) For each 1 part per million increase in pectin level, the sweetness index decreases by 0.0023.

(e) 22.8% of the variance in sweetness index is accounted for by the variance in pectin levels.
Example: Assume that 13% people are left-handed. If we pick 5 people at random, find the probability of each of the following outcomes: (a) The first lefty is the 5th person chosen. (b) There are some lefties in the 5 people chosen. (c) The first lefty is the 2nd or 3rd person chosen.

Solution: Summarize key pieces of useful information:

\[ P(L) = 0.13, \quad P(\text{not } L) = 0.87 \quad [L=\text{Left-handed}] \]

Each of the random 5 persons is independent, so we can use simple multiplication rule for "AND" type events.

(a) First L is 5th person.

Only 1 way this can happen: (not L) AND (not L) AND . . . AND (L)

\[ P(1\text{st }L \text{ is 5th person}) = 0.87 \times 0.87 \times 0.87 \times 0.87 \times 0.13 = 0.0745 \]

(b) Some L among the 5.

Many different ways this can happen. Best to use complement.

\[ P(\text{No } L \text{ among the 5}) = 0.87^5 = 0.4984 \]

Probability of at least some L = 1 - 0.4984 = 0.5016

(c) First L is 2nd or 3rd person.

Very similar to (a), except we have to add probabilities of 2 disjoint events: (1) First L is 2nd person, (2) First L is 3rd.

\[ P(\text{First }L \text{ is 2nd}) = 0.87 \times 0.13 = 0.1131, \quad P(\text{First }L \text{ is 3rd}) = 0.0984 \]

\[ P(\text{First }L \text{ is 2nd or 3rd person}) = 0.1131 + 0.0984 = 0.2115 \]

[P.395, Ex.21] Tree diagram works well here.

\[ \text{Prof}_S: \frac{70}{120} \quad \text{Prof}_S \text{ and Pass} = \frac{46}{120} = 0.3833 \]

\[ \text{Fail}: \frac{2}{20} \quad \text{Prof}_S \text{ and Fail} = \frac{11}{120} = 0.0917 \]

\[ \text{Prof}_K: \frac{50}{120} \quad \text{Prof}_K \text{ and Pass} = \frac{25}{120} = 0.2083 \]

\[ \text{Fail}: \frac{4}{20} \quad \text{Prof}_K \text{ and Fail} = \frac{16}{120} = 0.1333 \]

(a) \[ P(\text{Pass}) = 0.4667 + 0.25 = 0.7167 \quad \text{[Note that } P(\text{Fail}) = 1-0.7167]\]

(b) \[ P(\text{Prof}_K | \text{Fail}) = \frac{P(\text{Prof}_K \text{ and Fail})}{P(\text{Fail})} = \frac{0.1667}{1-0.7167} \]

Thus, \[ P(\text{Prof}_K | \text{Fail}) = 0.5884 \]
Jail inmates in a certain region can be classified according to the type of crime committed. Random samples of 500 male inmates and 500 female inmates are selected and classified as in the table.

<table>
<thead>
<tr>
<th>Type of crime</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>violent</td>
<td>117</td>
<td>66</td>
</tr>
<tr>
<td>property</td>
<td>150</td>
<td>160</td>
</tr>
<tr>
<td>drug</td>
<td>109</td>
<td>168</td>
</tr>
<tr>
<td>public-order</td>
<td>124</td>
<td>106</td>
</tr>
</tbody>
</table>

Answer the questions below based on these data. Indicate the calculations needed. Give the proper notation and/or terminology.

A. What is the probability that an inmate chosen at random has committed a crime against property?

B. Are "female" and "drug crime" independent events?

C. Are "female" and "drug crime" mutually exclusive events?

D. What is the conditional probability that a male inmate chosen at random has committed a violent crime?
Some more questions from previous final exams

(I) You want to estimate the mean price of textbooks at your college bookstore. A random sample of 25 textbooks has mean price of $72.24 and standard deviation of $8.60. Assuming the conditions necessary for inference apply, answer true/false (with reason) for the following statements:

(a) The best sampling distribution model here is the normal model N(72.24, 1.72).
(b) Suppose the 95% confidence interval is ($68.69, $75.79). Then the margin of error is $3.55.
(c) If we increase the sample size to 36, the margin of error would drop below $3.0. Assume the sample mean and standard deviation are the same ($72.24 and $8.60, respectively).
(d) We can decrease the margin of error further by choosing a higher confidence level.

(II) The distribution of SAT scores of incoming freshmen at a large university is known to be unimodal with strong right skew. The mean and standard deviation are 1180 and 50 respectively. Describe the sampling distribution model for mean SAT scores of random samples of 100 freshmen.

(III) Given the probabilities P(U)=0.3, P(V)=0.6, find the probability P(U and V) for each of the following situations: (a) U and V are disjoint.
(b) U and V are independent.

(IV) A popular TV talk show asked viewers to call in with their opinion regarding the question: "Should the United Nations continue to have its headquarters in the U.S.?" 186,000 callers responded, with 67% saying "No." A nationwide random sample of 500 adults found that 72% answered "Yes" to the same question. Which estimate is likely to more accurately reflect the opinion of all adults in the U.S.? Does sample size affect your conclusion? Explain.
Suppose you have 20 data values in a distribution for which you have calculated the mean, standard deviation, median and IQR. If you subtract 10 from each data value, how would it affect each of the summary statistics you had calculated previously?

Stores that use checkout scanners estimate that 1% of scanner sales are overcharges. In a study of a random sample of stores, a consumer group found 20 overcharges out of a total of 1234 scanned items. Carry out a hypothesis test to assess the stores' claim. Use a 5% significance level \( \alpha = 0.05 \). Remember to follow all the key steps for hypothesis testing.

Final word

* This is a very bare-bones, minimum review -- even if it runs into several pages. Please make sure your preparation efforts match your grade expectations!

* Test questions will not be designed to be tricky or hard. Instead, my primary goal will be to test understanding of key concepts, and your ability to apply them for problem solving.

It has been my pleasure, and privilege, to have been your instructor! I wish each of you the very best on the final & in your academic career!