Test 3 Information

Time, Location, Coverage

Test 3 will be given in class on Tuesday, December 2. The test is cumulative, but will emphasize the material from Chapters 7 – 9 (Chi-Squared, ANOVA, and Regression). You are responsible for material covered in the text, in the problem sets, and in class.

Format

Test questions will be designed to try to see how well you understand the material, not how well you can perform various procedures mindlessly.

A variety of question formats may be used. You may be required to compute numerical statistics; produce graphs by hand or explain how to get a computer to produce them; or to analyze data or numerical or graphical summaries of data. Some items may be be tested using ”short answers” (a couple sentences to a paragraph), multiple choice, or true/false.

You should bring your laptop to the test, but the only software you may use is RStudio. You will also be allowed to use a basic calculator to do arithmetic (but, of course, RStudio can do this sort of arithmetic as well), but no other phone, etc., may be used during the exam.

Instructions

Read through these prior to coming to the test and follow them when you take your test.

1. Always show your work and explain your reasoning. Answers without work or reasoning will not receive full credit.
   - Use mathematical notation (especially the equals sign) correctly.
   - Don’t be afraid to use words in your explanations.
   - If you get an unreasonable answer, be sure to say so. Give a brief explanation about how you know your answer is wrong (for example, “the mean I calculated is less than 10, but I can see from the graph that it should be at least 20”). Then go on to other problems and come back and try to fix the error if you have time at the end of the test period.
   - Even if you cannot do a problem completely, show me what you do know.

2. You may use your calculator during the exam, but for each number you write on the exam, it must be clear where it came from. For example, if you got .25 by multiplying .5 by .5, I want to see .5 · .5 = .25 on your paper (or words indicating the same).

3. Short answer questions will be graded based on truth, accuracy, clarity, significance, and brevity. In short, I’m looking for high quality answers. (Example: If you are asked to give an example of something, pick the best example you can think of, one that makes the issue especially clear.)

4. Test restrictions.
   - The test is closed book No notes are allowed.
   - Do not write in purple on the exam. (The exam will be graded in purple.)
Here is a list of things you should be sure you know how to do. It is not intended to be an exhaustive list, but it is an important list.

You should be able to:

- Understand, use and explain the statistical vocabulary/terminology.
  - Be sure to focus on important distinctions being made by terms like case vs. variable, statistic vs. parameter, sample vs. population, categorical vs. quantitative, sample vs. sampling distribution, sampling distribution vs. bootstrap distribution, explanatory variable vs. response variable, etc.
  - Some other important terms: significance level, confidence level, margin of error, statistically significant, type I error, type II error, critical value, paired design, blinding, residual, correlation

- Work with normal, t, and chi-squared distributions. (z-scores, 68-95-99.7 Rule, \texttt{pnorm()}, \texttt{qnorm()}, \texttt{pt()}, \texttt{qt()}, \texttt{pchisq()}, \texttt{qchisq}(), etc.)

- Understand the issues involved in collecting good data and the design of studies, including the distinctions between observational studies and experiments.

- Understand how confidence intervals are computed from a bootstrap distribution (percentile method and standard error method) and using standard error formulas, including
  - what the resulting interval tells you (the meaning of the confidence level, etc.)
  - recognizing incorrect ways to interpret a confidence interval and what is wrong with them.
  - how to get \texttt{RStudio} to generate a bootstrap distribution.
  - how to determine good sample sizes for a desired margin of error.

- Use the 4-step process for conducting a hypothesis test, including
  - expressing null and alternative hypotheses
  - determining a p-value from a randomization distribution or using formulas (SE, Chi-squared, F, degrees of freedom)
  - expressing the logic of a p-value in words (in the context of a particular example).
  - how to get \texttt{RStudio} to generate a randomization distribution
  - the difference between 1-sided and 2-sided tests
  - why we use upper tails for Chi-squared and F tests.

- Use R’s built-in functions for doing tests and confidence intervals: \texttt{binom.test()}, \texttt{prop.test()}, \texttt{t.test()}, \texttt{chisq.test()}, \texttt{xchisq.test()}, \texttt{lm()}, \texttt{anova()}, \texttt{summary}().

- Interpret output from the above functions.

- Check rules of thumb to see whether approximations (normal, t, Chi-squared) are good enough for our purposes.
• Use R to compute numerical summaries, make plots, compute probabilities, create bootstrap and randomization distributions, and compute p-values and confidence intervals. Important functions to review include `histogram()`, `bwplot()`, `mplot()`, `favstats()`, `do()`, `mean()`, `prop()`, `pnorm()`, `pt()`, `qnorm()`, `qt()`, `rbind()`, `tally()`, `resample()`, `shuffle()`.

• Use the relationship between confidence intervals and p-values.

• Perform and interpret Chi-squared tests
  – Chi-squared goodness of fit vs. Chi-squared for two-way tables
  – How to compute expected counts
  – Chi-squared test statistic
  – degrees of freedom

• Perform and interpret 1-way ANOVA
  – null and alternative hypotheses
  – using `lm()` to fit the model
  – computation of F statistic (ANOVA table, degrees of freedom, SS, MS, etc.)
  – \( R^2 = \frac{SSM}{SST} \) and what it tells us
  – Comparing two means with \( SE = \sqrt{MSE \left( \frac{1}{n_1} + \frac{1}{n_2} \right)} \)
  – Tukey’s Honest Significant Differences (TukeyHSD()) and why we use it.
  – checking assumptions (normality, homoskedasticity)
  – residuals and residuals plots

• Perform and interpret simple linear regression
  – linear relationships and equations for lines (slope, intercept, etc.)
  – hat notation (\( \hat{y}, \hat{\beta}_1 \), etc.)
  – using `lm()` to fit the model
  – computation of F statistic (ANOVA table, degrees of freedom, SS, MS, etc.)
  – \( R^2 = \frac{SSM}{SST} \) and what it tells us
  – correlation coefficient \( (R) \)
  – checking assumptions (linear shape, normality, homoskedasticity)
  – residuals and residuals plots

Note that the test will be a sample from the possible topics; it is not possible to cover everything on the test.
The following formulas will be printed on the test:

\[ SE = \sqrt{\frac{p(1-p)}{n}} \quad SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \]

\[ SE = \frac{\sigma}{\sqrt{n}} \quad SE = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \]

You will need to know how to adjust these for use with confidence intervals and p-values and how to determine the correct degrees of freedom for t-distributions.

Example Problems

1 What do I do?

In each of the following situations, pretend you want to know some information and you are designing a statistical study to find out about it. Give the following THREEM pieces of information for each:

(i) what variables you would need to have in your data set (ii) whether they are categorical or quantitative, and (iii) what statistical procedure you would use to analyze the results.

Select your procedures from the following list: 1-proportion, 2-proportion, 1-sample t, 2-sample t, Paired t, Chi-square goodness of fit, Chi-square for 2-way table, 1-way ANOVA, linear regression, none of these.

a) You want to know if boys or girls score better on reading tests in Kent County grade schools.

(See exercises C.31–38 and D.19–27 for more example scenarios.)

2 Complete the ANOVA table (from a partially filled in ANOVA table)