Test 3 Info

Stat 145

Test 3 Information

Coverage

The test is cumulative, but will emphasize the material since the last test.

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

Format

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.

Content

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:

1. Understand, use and explain the statistical terminology.
    a. Be sure to focus on important distinctions being made by terms like case vs. variable, categorical vs. quantitative, explanatory variable vs. response variable, statistic vs. parameter, sample vs. population, sample vs. sampling distribution, sampling distribution vs. bootstrap distribution, etc.
    b. 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
2. Understand the issues involved in collecting good data and the design of studies, including the distinctions between observational studies and experiments.
3. Understand how confidence intervals are computed
    a. how to get R to generate a bootstrap distribution.
    b. using a bootstrap distribution to compute a confidence interval
    c. using standard error formulas to compute a confidence interval
    d. using summary information from a linear regression model
    e. how to determine good sample sizes for a desired margin of error.
4. Understand what a confidence interval tells you
    a. meaning of confidence level
    b. recognizing incorrect ways to interpret a confidence interval and what is wrong with them.
    c. relationship between p-values and confidence intervals

5. Use the 4-step process for conducting a hypothesis test, including
    a. expressing null and alternative hypotheses
    b. computing an appropriate test statistic
    c. how to get R to generate a randomization distribution
    d. determining a p-value from a randomization distribution
    e. determining a p-value from a using formulas (SE, Chi-squared, degrees of freedom)
    f. expressing the logic of a p-value in words (in the context of a particular example).
    g. the difference between 1-sided and 2-sided tests
    h. why we use upper tails for Chi-squared tests.
6. Perform and interpret Chi-squared tests a Chi-squared goodness of fit vs. Chi-squared for two-way tables
    b. How to compute expected counts
    c. Chi-squared test statistic
    d. degrees of freedom
7. Perform and interpret 1-way ANOVA
    a. null and alternative hypotheses
    b. using `lm()` to fit the model
    c. computation of F statistic (ANOVA table, degrees of freedom, SS, MS, etc.)
    d. $R^2 = \frac{SSM}{SST}$ and what it tells us
    e. Tukey's Honest Significant Differences (`TukeyHSD()`) and why we use it.
    f. checking assumptions (normality, equal standard deviation)
    g. residuals and residuals plots
8. Perform and interpret simple linear regression
    a. linear relationships and equations for lines (slope, intercept, etc.)
    b. hat notation ($\hat{y}$, $\hat{\beta}_1$, etc.)
    c. using `lm()` to fit the model
    d. computation of F statistic (ANOVA table, degrees of freedom, SS, MS, etc.)
    e. $R^2 = \frac{SSM}{SST}$ and what it tells us
    f. correlation coefficient (R)
    g. checking assumptions (LINE)
    h. residuals and residuals plots
9. Check conditions/rules of thumb to see whether approximations (normal, t, Chi-squared) are good enough for our purposes.
10. Important functions to review include
    a. `gf_histogram()`, `gf_boxplot()`, `gf_point()`
    b. `df_stats()`, `tally()`, `mean()`, `prop()`, `diffmean()`, `diffprop()`
    c. `pnorm()`, `pt()`, `pchisq()`, `qnorm()`, `qt()`, `qchisq()`
    d. `rbind()`
    e. `do()`, `resample()`, `shuffle()`
    f. `chisq.test`, `xchisq.test()`, `t.test()`, `prop.test()`
    g. `lm()`, `msummary()` `mplot()`, `anova()`

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 included on the test:

parameter type	one group	two groups
proportion	$SE = \sqrt{\frac{p(1-p)}{n}}$	$SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$
mean	$SE = \frac{\sigma}{\sqrt{n}}$	$SE = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

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

Example Problems

These problems are from past tests I have given.

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 **THREE** 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 (aka 1 mean), 2-sample t (aka 2 means), paired t, Chi-squared goodness of fit, Chi-squared 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.

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

   How much of the ANOVA table do you need in order to be able to fill out the rest?

3. Below are some numerical summaries from the study of Atlanta commuters.

```
df_stats( ~ Time,  data = CommuteAtlanta ) %>% pander()
```

response	min	Q1	median	Q3	max	mean	sd	n	missing
Time	1	15	25	40	181	29.11	20.72	500	0

```
df_stats( Time ~ Sex, data = CommuteAtlanta ) %>% pander()
```

response	Sex	min	Q1	median	Q3	max	mean	sd	n	missing
Time	F	1	15	25	35	120	26.8	17.26	246	0
Time	M	1	15	30	40	181	31.34	23.41	254	0

   a. Compute a 95% confidence interval for the difference between the mean commute time for men and for women based on a sample of Atlanta commuters.
   b. Is there enough evidence to conclude that men and women have different mean commute times? Explain.

4. The following code can be used to test the null hypothesis that smoking rates are the same for men and women in a population of students.

```
library(Lock5withR)
prop.test(Smoke ~ Sex, data = StudentSurvey)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  tally(Smoke ~ Sex)
## X-squared = 1.3548, df = 1, p-value = 0.2444
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.02623048  0.11667411
## sample estimates:
##    prop 1    prop 2
## 0.9053254 0.8601036
```

```
prop.test(Smoke ~ Sex, data = StudentSurvey, correct = FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  tally(Smoke ~ Sex)
## X-squared = 1.7603, df = 1, p-value = 0.1846
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.02068122  0.11112486
## sample estimates:
##    prop 1    prop 2
## 0.9053254 0.8601036
```

   a. The code above does not check to see whether the sample sizes are large enough for the normal approximations being used by `prop.test()`. Use the following output to decide whether the normal approximation can be used in this situation:

```
tally( Smoke ~ Sex, data = StudentSurvey, format = "count" ) %>% pander()
```

	Female	Male
No	153	166
Yes	16	27

   b. We could do the hypothesis test part of this another way. How?

5. You are interested in knowing whether moderate exercise for one month reduces cholesterol.

   a. Outline a randomized comparative experiment that attempts to find out. Be brief, but answer the following:

   b. What are your **explanatory** and **response** variables?

   c. For each variable, identify whether it is **categorical** or **quantitative**.

   d. What makes your study an **experiment** and not an **observational study**?

   e. What advantage does your study have over an observational study designed to answer the same question?

f. As part of your preliminary analysis of the data, you are asked to present a graphical summary of the data. What type of plot will you make?

g. Write down the R command to make this plot.

h. Sketch what it might look like if exercise has a large effect on reducing cholesterol levels.

6. A certain IQ test is standardized in such a way that the mean score is 100 and the standard deviation is 10. For parts b), c) and d) please sketch the distribution and shade the appropriate region.

   a. What Z-score is associated with a score of 117?
   b. Aproximately what percentage of people have IQs between 90 and 110?
   c. Approximately what percentage of people have IQs above 120?
   d. Approximately what percentage of people have IQs between 90 and 120?

7.