W203: Exploring and Analyzing Data
Summer 2015
Lab 2

Instructions

Please construct an R markdown report that addresses Part 3

You will receive a web-based link from bCourses. This link will include all questions from Part 1 as well as questions (where applicable) from Parts 2 and 3. Do not leave any answers empty on bcourses.

As a suggestion, complete your report first, and include your answers for all parts in the report. Then access the web-based link from bCourses and answer **all of the questions** found there. Finally, be sure to upload your report to bCourses (you will be able to do so at the end of the bCourses quiz).

Part 1: Multiple Choice (40 points)

1) Martha administers a one-time survey and wants to clearly display the religious affiliations of her respondents. What kind of display(s) should she consider using?

a) Bar graphs
b) 3 dimensional, animated pie charts with a fancy 3-point font
c) Stem-and-leaf plots
d) Box plots
e) Line charts
f) a and e

2) Your friend claims that, after consuming the memory-boosting drug Beiberol-5, Berkeley students can recall the names of more than 10 Justin Bieber songs, even without any opportunity to practice them beforehand. What would be an appropriate set of hypotheses to test this claim (where H_0 is the null hypothesis and H_a is the alternative hypothesis)?

a) H_0: $\mu > \mu_0$; H_a: $\mu = \mu_0$

b) H_0: $\mu = \mu_0$; H_a: $\mu \neq \mu_0$

c) H_0: $\mu = \mu_0$; H_a: $\mu > \mu_0$

d) H_0: $\mu \neq \mu_0$; H_a: $\mu > \mu_0$

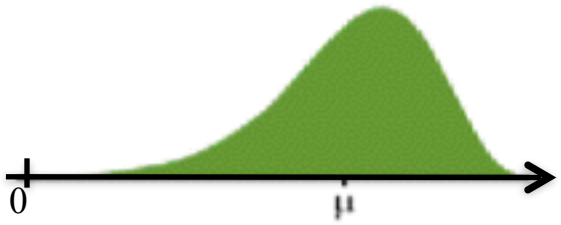
3) We conduct a study using a national survey and get p-value of .01 for a statistical test that is of interest to us. Based on the p-value alone, which of the following must be true?

a) Our alpha is less than .10.
b) The effect size of the test is medium or large.
c) The effect size of the test is small
d) a and b
e) a and c
f) none of the above.

4) A medical researcher replicates one of her earlier studies precisely (same method, same sample size, same analysis), and calculates the same type of test statistic for a one-tailed test in each study. In the original study she used an alpha value of .10 as her decision rule for rejecting the null hypothesis. However, in the second study she uses an alpha of .05 for rejecting the null hypothesis. As a result of this one change, what can we say about the second study compared to the first?

a) The possibility of a Type II error will go up in the second study.
b) The possibility of a Type II error will go down in the second study.
c) The statistical power will go up in the second study.
d) The statistical power will go down in the second study.
e) a and d
f) a and c
g) b and c

5) Based on the following picture, what is the most effective way to make this variable's distribution less skewed?



a) Use the inverse of the variable
b) Divide the variable by its standard deviation
c) Compute the variable minus its mean
d) Take the square-root of the variable

e) Raise the variable to a power greater than 1

f) Take a log transform

6) Which of the following could serve as the null hypothesis of a statistical test?

a) The mean age of Berkeley students lies between 20 and 25.

b) The standard deviation of Berkeley student ages is 2 years.

c) The mean age of Berkeley students is different than the national average.

d) The distribution of ages at Berkeley is not normal.

7. Which of the following questions would you answer with a frequentist statistical test?

a) What is the probability that our model is the correct one?

b) Does our measurement instrument have construct validity?

c) Is the relationship between two variables causal?

d) What is the probability of the data we observe, assuming that the null hypothesis is true?

8. You know that the number of times an American smiles in a day is normally distributed with a mean of 50 and a standard deviation of 8. You conduct a study of 20 Berkeley Students to see if Berkeley students smile more or less than the national average. You perform a z-test on your data and arrive at a p-value of .04. Which of the following do you know to be true?

a) You have absolutely disproved the null hypothesis (that Berkeley students have the same distribution as the nation at large).

b) You have found the probability that the null hypothesis is true.

c) Assuming your null hypothesis is actually false, your p-value is likely to decrease as you increase your sample size.

d) You have absolutely proved your alternate hypothesis (that Berkeley students smile a different amount than the national average).

e) You can deduce the probability that your alternative hypothesis is true.

f) Assuming your null hypothesis is actually true, and you were to repeat the experiment a large number of times, you would expect a type 1 error 4% of the time.

9. An online web-based experiment is administered to all 120 Berkeley students in a single dormitory. Participants are given 60 seconds to memorize the ingredients of Poultronix-brand chicken-substitute nuggets and immediately earn gift cards for correct

recall. Students may complete the experiment at any time during a three-month period. Which of the following assumptions should you be most concerned about when assessing this design?

a) Normal sampling distribution of the test statistic.
b) Homogeneity of variance.
c) Interval data.
d) Independence of observations.
e) None of the above.

10. Steve is convinced that the color red incites aggression in men. To test his hypothesis, Steve conducts an experiment on undergraduate students taking his "Human Aggression" seminar, placing subjects in either a red room or a blue one. He is able to reject the null hypothesis (that room color has no effect on aggression) and produces a p-value of .007. Intrigued, Martha replicates Steve's procedure on the students in her "World Religions" course, but cannot reject the null hypothesis (p=.34). Based on this information, which of the following must be true?

a) Steve committed a Type II error.
b) Martha committed a Type II error.
c) Steve's study had more power than Martha's.
d) Steve's results will generalize to the US population.
e) Steve committed a Type I error.
f) None of the above.

Part 2: Test Selection (10 points)

Data.

Every other year, the General Social Survey collects responses to thousands of questions, covering a wide variety of topics. You will be using a subset of data from 1993, including a small number of variables. This may be found in the file, GSS.Rdata.

Like any survey, GSS data creates additional concerns that would normally go into a statistical analysis. Surveys are usually weighted in order to compensate for over- or under-representation of subgroups. For this lab, however, you will be using unweighted data, which limits how well your findings generalize to the U.S. population.

For the following problems, select the most appropriate statistical test to answer the question from the given choices.

Note: You do **NOT** need to execute the test.

11. Is the variance of the age variable the same for men and for women?
a) Shapiro-Wilk test
b) Levene's test
c) z-test

12. Is the age variable normally distributed?
a) Shapiro-Wilk test
b) Levene's test
c) z-test

Part 3: Data Analysis and Short Answer (50 points)

Write a well-commented R markdown report to perform each of the following tasks, then answer the provided questions in your bcourses submission.

The answers on bcourses should be sufficient. I will look at your report only if I want to see the mechanics are correct. Include all code, graph(s), and output in the report.

13. Data Import and Error Checking

    a. Examine the "agewed" variable (age when married). What are the value(s) of agewed, if any, that do not meaningfully correspond to ages?

    b. Recode any value(s) that do not correspond to age as NA. What is the mean of the agewed variable?

14. Checking assumptions

    a. Produce a QQ plot for the agewed variable. Using this plot information, is agewed normal and how do you know?

    b. Perform a Shapiro-Wilk test on the agewed variable.

        i. What is the null and alternative hypothesis for your test?

        ii. What is your p-value, and what is your specific conclusion?

c. What is the variance of agewed for men? What is the variance of agewed for women?

d. Perform a Levene's test for the agewed variable grouped by men and women.

    i. What is the null and alternative hypothesis for this test?

    ii. What is your p-value, and what is your specific conclusion?

15. More hypothesis testing

a. Suppose we have a hypothesis that the age of marriage (agewed) in the population has a mean of exactly 23, with a standard deviation of 5 years (you should assume this value is correct rather than estimating the standard deviation from the data). Perform a z-test to analyze this hypothesis.

    i. What is the null and alternative hypothesis for this test?

    ii. What is your p-value, and what is your specific conclusion?