Statistics 302 Midterm 1

Larget, Spring 2014

Solutions

---

1. True/False Problems. **3 points each, 15 points total**. Write *very brief explanations*.

   (a) Circle either TRUE or FALSE (and explain/correct if FALSE):

   When data is strongly skewed to the left, the mean is typically smaller than the median.

   Solution: True.

   (b) Circle either TRUE or FALSE (and explain/correct if FALSE):

   Based on the same sample data, a 90% confidence interval will be wider than a 95% confidence interval.

   Solution: A 90% confidence interval will be narrower.

   (c) Circle either TRUE or FALSE (and explain/correct if FALSE):

   A plot of variables X and Y shows that the points lie tightly around a line with slope 2. Therefore, the correlation coefficient r between these variables will be just a bit smaller than 2.

   Solution: False. Correlation coefficient cannot be more than one.

   (d) Circle either TRUE or FALSE (and explain/correct if FALSE):

   A 99% confidence interval for a population mean μ based on a random sample of size ten is 55.5 to 71.9. It follows that the p-value from the hypothesis test $H_0\colon \mu = 70$ versus $H_A\colon \mu \neq 70$ is significant at the $\alpha = 0.05$ level.

   Solution: False. 70 is in the interval, so the hypothesis is not significant at the $\alpha = 0.01$ (not $\alpha = 0.05$) level.

   (e) Circle either TRUE or FALSE (and explain/correct if FALSE):

   Random samples from two populations result in sample proportions $\hat{p}_1 = 0.55$ and $\hat{p}_2 = 0.47$. To estimate the difference $p_1 - p_2$ with a confidence interval, the scientist produces a bootstrap distribution of differences between sample proportions. This distribution is centered at zero.

   Solution: False. The bootstrap distribution is centered at $0.55 - 0.47 = 0.08$.

2. **9 parts, 5 points each, 45 points total.**

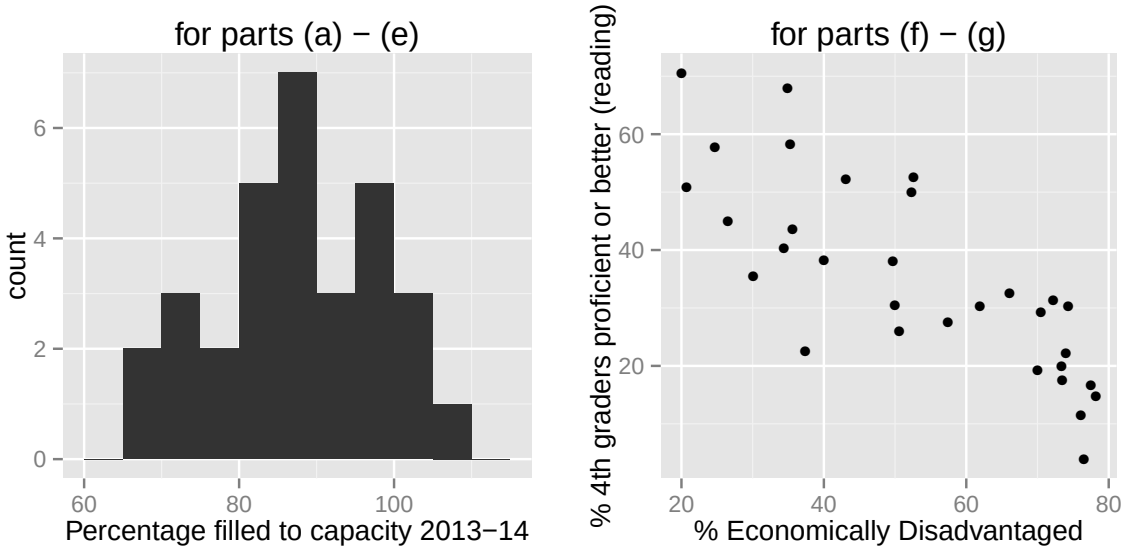
The February 23, 2014 *Wisconsin State Journal* published data on all schools in the Madison Public School System, including 31 elementary schools. Here is a subset of the data.

```
##      school highSchool cap.2013 cap.2018 disadvantaged proficient
## 1     Allis LaFollette       66       70          76.5        3.9
## 2    Chavez   Memorial       95      114          26.5       44.9
## 3 Crestwood   Memorial       84       82          43.1       52.3
## 4  Elvehjem LaFollette       93       81          34.4       40.3
## 5      Falk   Memorial       76       68          73.4       20.0
## 6  Franklin       West       91       85          20.7       50.8
```

The variables are:

- `school`, which is the name of the elementary school;
- `highSchool`, which is the name of the high school that the elementary school eventually feeds into, one of East, La Follette, Memorial and West;
- `cap.2013`, which is the percentage of students enrolled relative to the capacity of the school in the 2013–2014 school year, as a percentage (which can exceed 100% if the school is overcrowded);
- `cap.2018`, which is the projected percentage of students enrolled relative to the capacity of the school in the 2018–2019 school year, as a percentage;
- `disadvantaged`, which is the percentage of students at the school that are economically disadvantaged; and
- `proficient`, which is the percentage of 4th graders at the school who read at the proficient or advanced levels for their age group.

Here are two graphs that display some aspects of the data.



(a) How many elementary schools are overcrowded (have capacity greater than 100%) in 2013–14?

Solution: From the histogram, there are two bars to the right of 100 with a total of four schools.

(b) Which number is closest to the mean 2013–14 capacity percentage? Circle your response.

65.7 79.4 87.3 99.3

Solution: The mean is 87.3. All other answers are too far from the center.

(c) Which number is closest to the standard deviation of the 2013–14 capacity percentage? Circle your response.

3.2 10.6 18.2 101.2

Solution: The standard deviation is 10.6. 87.3 ± 18.2 extends from 69.1 to 105.5 which encompasses far more than about two thirds of the data, so 18.2 is too big and 87.3 ± 3.2 is the interval from 84.1 to 90.5 which is far less than about two thirds of the data.

(d) The capacity of Muir elementary school in 2013–14 is 79 percent. Using your previous two answers, write an expression for and calculate the z-score of this observation. Round your answer to two digits after the decimal point.

Solution:

$$z = \frac{79 - 87.3}{10.6} \doteq -0.78$$

(e) Interpret what the z-score computed in the previous part means by either filling in missing information or circling the best choice of words.

The 2013–14 capacity of Muir elementary school is _____

[means | standard deviations | percentage points] below the

[mean | standard deviation | standard error] 2013–14 capacity of all

Madison elementary schools.

Solution: 79 is 0.78 standard deviations below the mean.

(f) Which number is closest to the correlation coefficient between `disadvantaged` and `proficient`?

$$-1.21 \qquad -0.79 \qquad -0.08 \qquad 0.27 \qquad 0.82$$

Solution: The correlation coefficient is closest to -0.79. There is a fairly strong negative relationship, -1.21 is impossible, and -0.08 represents a very weak linear relationship.

(g) If the variable `disadvantaged` had been expressed as a proportion instead of a percentage (by dividing each value by 100), briefly explain how this would affect the calculation of the correlation coefficient.

Solution: There would be no change in the correlation coefficient.

Here is why the previous sentence is true. The plot of the points would have been the same, just with one axis relabeled. The correlation coefficient is unitless. The formula for the correlation coefficient uses the data only as the z-scores of x and y, and changes by adding any number or multiplying by a positive number have no effect on the z-score.

There is no plot of the relationship between the two variables for parts (h) and (i).

(h) The correlation coefficient between `cap.2013` and `proficient` is 0.27. A bootstrap procedure samples 31 of the points with replacement many times and calculates the correlation coefficient for each set of sampled points. The standard deviation of these bootstrap correlation coefficients is 0.18 and the following quantiles describe the distribution.

```
 0.5%  2.5%    5%   95% 97.5% 99.5%
-0.27 -0.13 -0.06  0.54  0.58  0.66
```

Use this information to find the end points of a 90% confidence interval for a population correlation ρ. *Do not interpret this interval.*

Solution: We use the 0.05 and 0.95 quantiles of the bootstrap distribution as the confidence interval, so the interval is from -0.06 to 0.54.

(i) Based on this sample of 31 Madison public elementary schools, we are 95% confident that the correlation between the percentage of students with 4th grade reading at the proficient or higher levels and the percentage of overcrowding in the school for all public elementary schools in the state of Wisconsin is between -0.13 and 0.58.

Circle the label of the most appropriate comment about the above conclusion.

A. The statement is an accurate inference based on principles of statistics.

B. The confidence interval is inaccurate because it did not use the estimate plus or minus twice the standard error.

C. The statement should contain a 90% confidence interval.

D. The inference is not justified because the sample of Madison schools is not a random or representative sample from the population of schools in the entire state of Wisconsin.

E. The official should have made an inference about all United States schools.

Solution: The most appropriate comment is D.

The 31 elementary schools in Madison are not a random sample of all public elementary schools in Wisconsin or in the USA. and these schools should not be considered as representative from a larger population of schools. There may be important differences between Madison elementary schools and schools elsewhere on these two variables. Inference to a larger population would be biased due to the sampling method.

3. **5 parts, 8 points each, 40 points total.**

Scientists studied a sample of college students to examine attitudes and habits. Below is a table that summarizes the stress levels of females and males in the study.

	Stress Level		
Gender	High	Low	Total
Female	38	113	151
Male	18	84	102
Total	56	197	253

(a) Define parameters and state null and alternative hypotheses to examine if the proportions of female and male college students that are experiencing high levels of stress are different.

Solution: p_F is the proportion of female college students that experience high stress.
p_M is the proportion of male college students that experience high stress.
$H_0 : p_F = p_M$ versus $H_A : p_F \neq p_M$

(b) Define a test statistic and compute its value from the sample data.

Solution: For a test that compares two population proportions, examine a difference between sample proportions.

$$\hat{p}_F - \hat{p}_M = \frac{38}{151} - \frac{18}{102} \doteq 0.0752$$

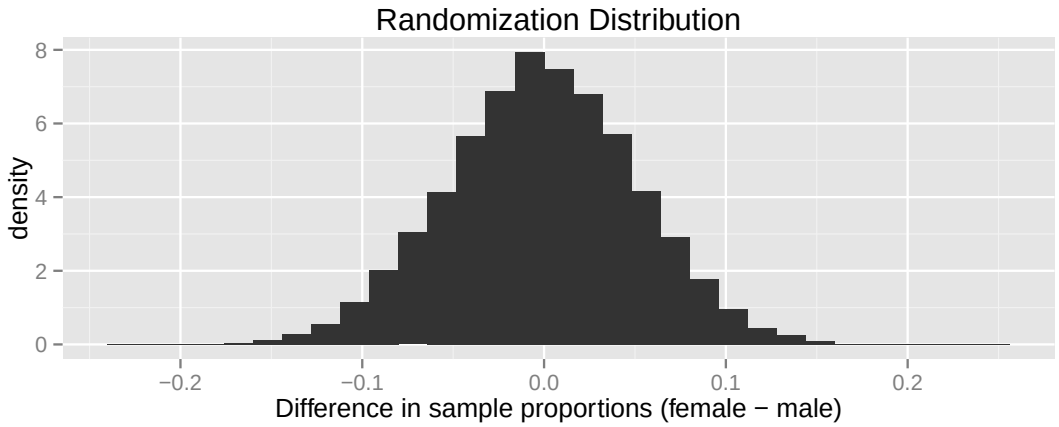
(c) Fill in the blanks to complete the steps in constructing the randomization distribution for the differences in proportions and calculating a p-value.

1. Create an array of numbers with 56 ones and **<u>197</u>** zeros.

2. For a large number of times, put these values into a random order.
3. For each random order, calculate the proportion of 1s among the first **151** numbers and the proportion of 1s among the last **102** numbers, and take the difference between these.
4. Find the proportion of the sampled differences **greater than** or equal to **0.0752** and multiply by **2** to calculate the p-value.

(d) Use the following summaries of the randomization distribution to select the value closest to the p-value: circle it.

0.01 0.05 0.10 0.20 0.50




```
##      5%     10%     25%     40%     60%     75%     90%     95%
## -0.089 -0.073 -0.040 -0.007  0.009  0.042  0.075  0.092
```

Solution: The p-value is close to 0.20.

The test statistic 0.0752 is very close to the 0.90 quantile, so about ten percent of the randomization distribution is at or exceeds 0.0752. This was a two-sided test, so we double the p-value and get 0.20.

(e) Summarize the results of the hypothesis test in the context of the problem.

Solution: There is very little evidence from the data to suggest a difference in the proportions of female and male college students that are experiencing high stress ($p = 0.20$, randomization test for difference in proportions).

Rant on interpretation of hypothesis tests in context.— Based on what many of you wrote in your exams, I must not have given sufficient examples in lecture on how to

properly summarize the results of a hypothesis test in context. When doing so, I want you to avoid using technical terms such as *reject*, *accept*, *hypothesis*, *null*, *alternative*, *significant*, α, and *Type I* or *Type II errors* and just say how strong the evidence is in favor of the alternative in context. The figure on page 260 can be a guide for language choice: notice that there are no strict boundaries attached to any significance levels. If a textbook question asks you to test a hypothesis at a specific significance level, it is okay to say that the test is either rejected or not rejected (but never accepted!) at a given level, or better, to say the result is statistically significant at the given level. *But then follow up with a jargon-free statement that says what this means in context.*

When making inferences, we are rarely making decisions, but rather are trying to explain to some audience what the statistical evidence is about a question. There is a historical importance in knowing the language of hypothesis testing with regard to decision theory, but rejecting or not rejecting a hypothesis at a fixed level of significance is not a proper way to address specific questions of inference. (Many statisticians will take issue with this last sentence, I imagine, but I believe it nonetheless.)

I reread the homework solutions from Katherine and looked at examples in the textbook where hypothesis tests are summarized in context and did not find any bad examples. It must be the case that many of you learned some bad habits in a previous exposure to statistics; we will work hard this semester to eliminate these bad habits. Often, especially if the problem asked you to test something at a fixed significance level, a summary in context may be something like:

> There is sufficient evidence at the $\alpha = 0.05$ significance level to reject the null hypothesis, so (some reasonable words that explains what this means without jargon).

If the question does not specify a significance level, I much prefer a format where you first say what the results mean in context without jargon and follow with a parenthetic expression that summarizes the statistical evidence on which you based the conclusion, namely what the p-value is and the name of the method you used to compute it. Then, the audience knows the essential statistical evidence if they care to, and they know how you are choosing to interpret what this evidence means. The audience can make their own decision as to whether or not they agree with you.

One important thing to recall is that we *never* accept the null hypothesis. We might *fail to reject* the null hypothesis. A lack of sufficient evidence to conclude that the alternative hypothesis is true is *quite distinct from* having ample evidence that the null hypothesis is true. When I read a student say they accept the null hypothesis, I become enraged and disappointed. When I am enraged and disappointed and holding a red pen, I take off many points. So, if you do not want to lose many points, then never again accept a null hypothesis.

Finally, recognize that in most situations, the null hypothesis, which is a specific statement that some parameter or difference in parameters which may have a continuum of possible values is exactly equal to a specific number. In the absence of a theory based on some strong symmetry, the null hypothesis is almost assuredly false. The hypothesis test is merely an exercise to see if we have sufficient data to confirm what must be true. However, this is not to say that everything is important. Often, the truth may be close enough to the null hypothesis that acting as if the null hypothesis were true is not a bad way to go. The

distinction between these viewpoints is best seen by making inferences with *estimation with a quantitative measure of uncertainty*, for example with a confidence interval. We may be highly confident that the true parameter value is so small that it does not matter in context. We may need an interval so wide for a high level of confidence that we merely confirm that the data had very limited information on the question at hand. Or, we may be confident that an effect is important because we are confident that all of the plausible values of a parameter are worth carrying about.

So, know the language and jargon associated with hypothesis testing, but estimate first when asked to do inference and be cautious when assessing the meaning of the results of a formal hypothesis test.