Answers (Midterm 2 practice questions)

Online S520

1. *Wilder and Rypstra (2004) tested the effect of praying mantis excrement on the behavior of wolf spiders. They put 12 wolf spiders in individual containers. Each container had two semicircles of filter paper: one semicircle that had been smeared with praying mantis excrement, and one without excrement. They observed each spider for one hour, and measured its walking speed while it was on each half of the container. They used a t-test at level $\alpha = 0.05$ to see if, on average, there was a difference between walking speed on the paper with excrement and on the paper without excrement.*

   (a) *What is the experimental unit? What measurements are taken on the experimental units?*

   The experimental unit is a wolf spider. Two measurements are taken: Walking speed on filter paper with excrement, and walking speed on filter paper without excrement.

   There are two nominal variables, filter paper type (with or without excrement) and the individual spider, and one measurement variable (walking speed). Different spiders may have different overall walking speed, so a paired analysis is appropriate to test whether the presence of praying mantis excrement changes the walking speed of a spider.

   (b) *Give null and alternative hypotheses for an appropriate two-tailed t-test. If the null hypothesis is true, what is the distribution of the t-statistic?*

   Let μ be the average difference in the walking speed of a wolf spider on filter paper with excrement compared to filter paper without excrement.

$$
\begin{aligned}
H_0 &: \mu = 0 \\
H_1 &: \mu \neq 0.
\end{aligned}
$$

   If the population is normal, then under the null, the t-statistic has a t-distribution with 11 degrees of freedom. If it's not normal then we don't know the distribution of the t-statistic.

   (c) *The P-value (significance probability) was calculated to be 0.053, so the null hypothesis was not rejected. From this and the other information given, is it correct to conclude that we are sure that wolf spiders' walking speed is not affected by praying mantis excrement? Explain.*

   This conclusion is too strong. We are right on the border of statistical significance, and the sample size is only 12.

2. *In a controversial 2011 paper in the Journal of Personality and Social Psychology, a researcher claimed to have found evidence of precognition. In one experiment, Cornell students sat in front of computer screens with two windows. They were asked to click the window that they thought had a picture behind it; after clicking, a picture would then randomly appear behind one of the two windows. In all, the students correctly clicked the window with the picture 1844 out of 3600 times. However, when the pictures were erotic, the students correctly clicked the window with the picture 828 out of 1560 times.*

   Note: Recall that the standard error of a proportion is $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

   (a) *Using the Central Limit Theorem, calculate a 95% confidence interval for the proportion of the time students click correctly for all pictures.*

   The sample proportion is $\hat{p} = 1844/3600 \approx 0.5122$. We estimate the standard error of this proportion as

   $$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \approx 0.0083$$

   Using 1.96 standard errors, the 95% confidence interval is

   $$0.5122 \pm 1.96 \times 0.0083$$

   49.6% to 52.9%.

   (b) *Using the Central Limit Theorem, calculate a 95% confidence interval for the proportion of the time students click correctly for erotic pictures.*

   The sample proportion is $\hat{p} = 828/1560 \approx 0.5308$. We estimate the standard error of this proportion as

   $$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \approx 0.013$$

   Using 1.96 standard errors, the 95% confidence interval is

   $$0.5308 \pm 1.96 \times 0.013$$

   50.6% to 55.6%.

   (c) *Another team of researchers wanted to repeat the experiment for erotic pictures, but with a larger sample size, so that the width of a 95% confidence interval for the proportion would be 0.02. How large a sample would they need?*

   The width of the interval is $2 \times 1.96 \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.02$. Solving for n gives

   $$n = \frac{1.96^2 \hat{p}(1-\hat{p})}{0.01^2}$$

   We don't know what $\hat{p}$ will be in the new study, but if we think that it will be the same as the previous study, we guess it'll be the same as in the old study. This gives $n = 9568$. If we think that quantum erotophysics does not exist, we might instead guess $\hat{p} = 0.5$, which would give $n = 9604$. In either case, a sample of size 10,000 should suffice. (This sounds like a lot, but we can get e.g. 50 students to do 200 pictures each.)

3. *The file* `ozone.txt` *contains 116 measurements of the concentration of ozone (in parts per billion) at Roosevelt Island, New York. After saving the file to your computer, you can load it into R by entering the command:*

```
ozone = scan(file.choose())
```

and then selecting the file.

   (a) *Does the data look like it comes from a normally distributed population? Include a graph to support your answer.*

   ```
   qqnorm(ozone)
   ```

   The normal QQ plot curves upward. It's difficult to believe the data comes from a normal population.

   (b) *Treating the data as an iid sample, find a 95% confidence interval for the mean ozone concentration.*

   ```
   mean(ozone)-1.96*sd(ozone)/sqrt(116)
   mean(ozone)+1.96*sd(ozone)/sqrt(116)
   ```

   A 95% normal interval is 36.1 to 48.1 parts per billion. (A t-interval is 36.1 to 48.2.)

4. *The Breakfast Club believes the ages of its members are approximately normally distributed. It sends out a survey to a random sample of 10 of its members. The ages of the sample are:*

$$18, 13, 18, 16, 12, 16, 17, 17, 20, 18.$$

The following R output will be useful for this question:

```
> ages = c(18, 13, 18, 16, 12, 16, 17, 17, 20, 18)
> sd(ages)
[1] 2.415229
> pt(0.655, df=9)
[1] 0.7355744
> qt(0.975, df=9)
[1] 2.262157
```

   (a) *Test the null hypothesis that the mean age of all members of the Breakfast Club is no more than 16, finding a P-value and giving a conclusion.*

   A quick glance at the numbers show no obvious outliers or skewness, so the normal assumption seems OK. The sample has mean 16.5 and SD 2.42. The t-statistic is $(16.5 - 16)/(2.42/\sqrt{10}) = 0.6547$. The P-value is `1 - pt(0.6547, df=9)`, or about 0.26. There's no evidence against the null hypothesis.

(b) *Find a 95% confidence interval for the mean age of all members of the Breakfast Club.*

The confidence interval is $16.5 \pm$ `qt(0.975, df=9)` $\times 2.42/\sqrt{10}$, or 14.8 to 18.2 years.

(c) *The Deadbeat Club sends out a survey to a random sample of 10 of their members. The ages of the sample are:*

$$13, 20, 26, 41, 67, 13, 79, 15, 20, 15.$$

Explain why a t-test would not be the best choice to test the hypothesis that the mean age of all members of the Deadbeat Club is no greater than 30.

This sample is skewed and so clearly non-normal, so we should avoid the standard normal and t methods, especially with such a small sample.

5. *In a study of a wave power generator, experiments were carried out on scale models in a wave tank to establish how two different choices of mooring method affected the bending stress produced in part of the device. The wave tank could simulate a wide range of sea states and the model system was subjected to the same sample of sea states with each of two mooring methods, one of which was considerably cheaper than the other. The question of interest is whether bending stress differs for the two mooring methods.*

   The data frame `waves` *contains the following:*

```
> waves
method1 method2
1      2.23    1.82
2      2.55    2.42
3      7.99    8.26
4      4.09    3.46
5      9.62    9.77
6      1.59    1.40
7      8.98    8.88
8      0.82    0.87
9     10.83   11.20
10     1.54    1.33
11    10.75   10.32
12     5.79    5.87
13     5.91    6.44
14     5.79    5.87
15     5.50    5.30
16     9.96    9.82
17     1.92    1.69
18     7.38    7.41
```

   Here `method1` *and* `method2` *are the stresses (in Newton meters) for the two different mooring methods. Some summary statistics follow:*

```
> summary(waves$method1)
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.820   2.310   5.790   5.736   8.732  10.830
> summary(waves$method2)
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.870   1.970   5.870   5.674   8.725  11.200
> summary(waves$method2 - waves$method1)
Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-0.63000 -0.20750 -0.11500 -0.06167  0.08000  0.53000

> var(waves$method1)
[1] 11.82893
> var(waves$method2)
[1] 12.54725
> var(waves$method2 - waves$method1)
[1] 0.08414412
```

The data appears close to normal.

(a) *We want to test for an average difference in stress between the methods. Explain what test we should use and why.*

Each pair of measurements is taken under the same conditions, so they are strongly dependent. Thus even though we have sample data for both "methods", we should reduce to a one-sample problem by taking differences.

(b) *Write down the null and alternative hypotheses for a one-sample t-test, and calculate an appropriate t-statistic.*

Let μ be the expected value of the difference (method 2 minus method 1.) Since there's no direction given in the question,

$$\begin{aligned} H_0 &: \mu = 0 \\ H_1 &: \mu \neq 0 \\ t &= \frac{5.674 - 5.736}{\sqrt{\frac{0.08414412}{18}}} \\ &= -0.907 \end{aligned}$$

(c) *Fine the P-value and a 95% confidence interval for the average difference (Method 2 minus Method 1). What do you conclude about the two methods?*

The P-value for the test is:
`2*pt(-0.907, 17)` $= 0.377$.
The 95% confidence interval for the average difference (Method 2 minus Method 1):
$(5.674 - 5.736) \pm \texttt{qt(0.025, 17)} * \sqrt{\frac{0.08414412}{18}} = (-0.21, 0.08)$
There's no evidence of a difference, so unless very small differences matter, they might as well choose the considerably cheaper method.

6. *In 2014 in the US, there were 4,010,532 recorded births in total. Of these, 327,680 were in June. June has 30 days and 2014 had 365 days. For this question, treat the births in 2014 as an IID sample from a larger population of births (this is not literally true but is a sufficient approximation.)*

(a) *Suppose we wish to test the hypothesis that the probability of being born in June is proportional to the number of days in June. Write down mathematical null and alternative hypotheses for this test.*

$$\begin{aligned} H_0 &: p = 30/365 \approx 0.0822 \\ H_1 &: p \neq 30/365 \end{aligned}$$

(b) *Using the Central Limit Theorem, calculate an approximate 95% confidence interval for the probability a child is born in June.*

$\hat{p} = 327680/4010532 = 0.08170$. The CI is $\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, which is $0.08170 \pm 1.96 \times 0.00014$, which is 0.0814 to 0.0820.

(c) *Find the P-value for the test of your hypotheses in (a) using the Binomial probability and normal approximation. Using both this and your confidence interval, explain what you can conclude about the probability of being born in June.*

Binomial probability:

`2 * pbinom(327680, 4010532, 30/365)` $= 0.0003818962 \approx 0.0004$

Normal approximation: z-statistic $= \frac{0.08170-0.0822}{0.00014} = 3.571429$

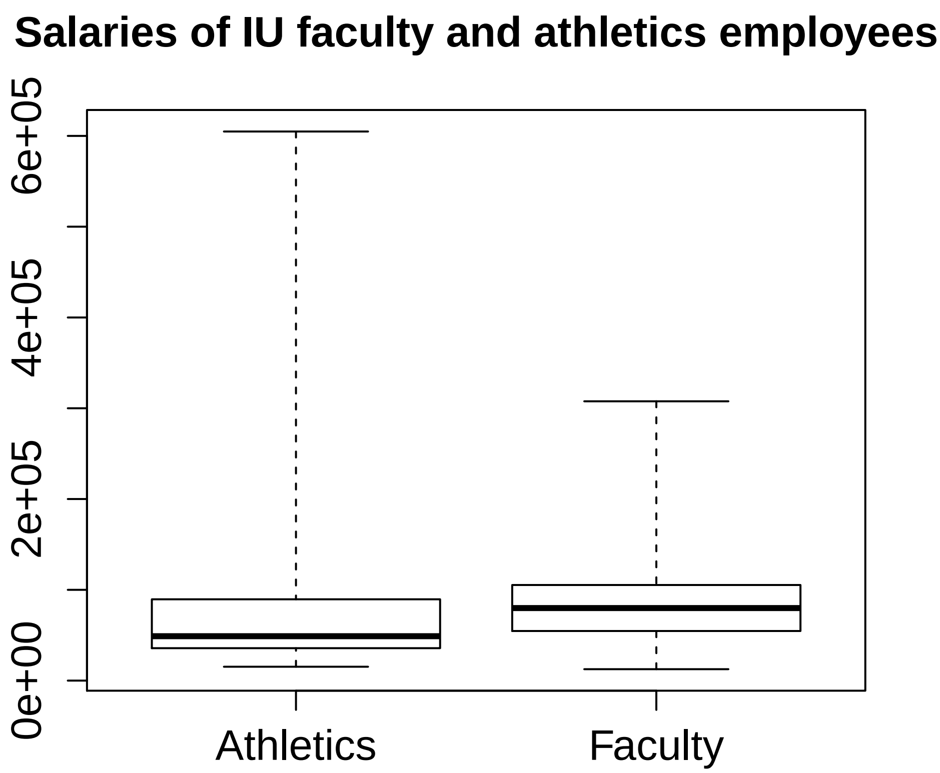
`2 * pnorm(-3.571429)` $= 0.0003550388$
Based on the p-value and the 95% confidence interval, we have evidence that the probability of being born in June is not proportional to the number of days in June. It's a little less than 30/365, but only a little.

7. *The file* `IUSalaries.txt` *contains the salaries of a random sample of 50 IU Bloomington faculty, along with the salaries of a random sample of 50 IU athletics employees. Download the data from Canvas and read it into R, e.g. by using the following code:*

```
salaries = read.table(file.choose(), header=TRUE)
Faculty = salaries$Salary[salaries$Job == "Faculty"]
Athletics = salaries$Salary[salaries$Job == "Athletics"]
```

(a) *Draw side-by-side boxplots of the two sets of salaries, and describe what you see in a sentence.*

```
boxplot(Salary~Job, data=salaries, range=0,
main="Salaries of IU faculty and athletics employees")
```



The striking visual difference is that the athletics distribution is more more strongly skewed than the faculty distribution.

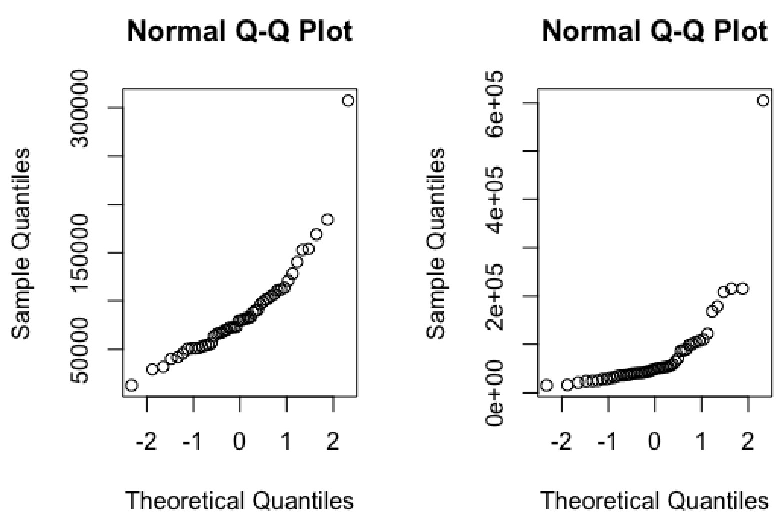
(b) *Suppose we wish to test for a difference between faculty salaries and athletics salaries. I performed Welch's two sample t-test to test whether the population means for faculty and athletics salaries were equal. I got the following output:*

```
> t.test(Faculty, Athletics)

Welch Two Sample t-test

data:  Faculty and Athletics
t = 0.6388, df = 74.075, p-value = 0.5249
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-19826.38  38539.46
sample estimates:
mean of x mean of y
87608.22  78251.68
```

The P-value produced by this test is 0.5249. Explain why we might not trust this P-value. Include graphs to support your answer.

The t-test assumes data from normal distributed populations. The boxplots alone are sufficient to show this is not the case, but for further evidence, look at the normal QQ plots.

```
par(mfrow=c(1,2))
qqnorm(Faculty)
qqnorm(Athletics)
```



Both plots have systematic bends — they're not remotely close to straight lines, and our sample sizes aren't large enough to totally ignore this.

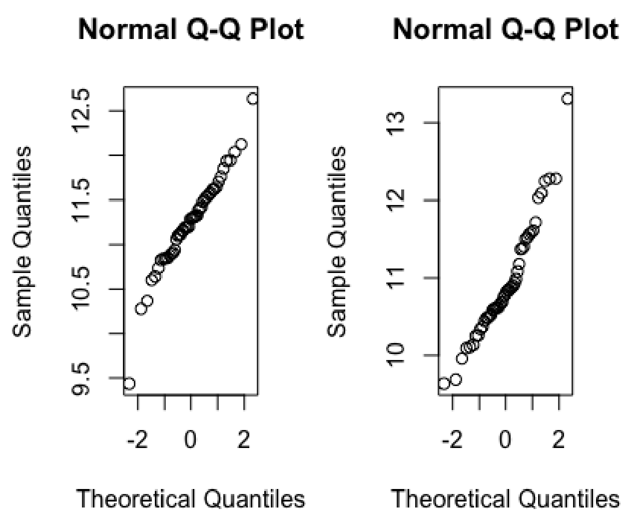
(c) *Suppose we wish to test the hypothesis that the mean of the logged faculty salaries is equal to mean of the logged athletics salaries. Should we use Welch's two-sample t-test, Student's two-sample t-test, or something else? Justify your choice.*

Welch's two-sample t-test is the better choice, because the sample variances are not close:

```
> var(log(Faculty))
[1] 0.2743252
> var(log(Athletics))
[1] 0.5406859
```

Although it won't make that much difference, it's safer not to make the additional equal variance assumption of Student's test.
We should also check normality on the log scale:

```
qqnorm(log(Faculty))
qqnorm(log(Athletics))
```



These look pretty close to straight lines, so the normality assumption is close to true.

(d) *Perform the test you chose in part (c).*

```
> log.faculty = log(Faculty)
> log.athletics = log(Athletics)
> se = sqrt(var(log.faculty)/50 + var(log.athletics)/50)
> t = (mean(log.faculty) - mean(log.athletics)) / se
> nu = (var(log.faculty)/50 + var(log.athletics)/50)^2 /
+ ((var(log.faculty)/50)^2/49 + (var(log.athletics)/50)^2/49)
> 2 * (1 - pt(abs(t), df=nu))
[1] 0.01648609
```

Double-check:

```
> t.test(log(Faculty), log(Athletics))

Welch Two Sample t-test

data:  log(Faculty) and log(Athletics)
t = 2.4445, df = 88.543, p-value = 0.01649
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
0.05839702 0.56579749
sample estimates:
mean of x mean of y
11.2526   10.9405
```

The P-value is 0.016, indicating some incompatibility with the null hypothesis of equal means (on the log scale.)

(e) *Do IU Bloomington faculty and IU Athletics employees have the same distribution of salaries? Explain.*

Pretty clearly they don't. The means might actually be similar, but the boxplots alone show the distributions aren't the same. Looking deeper, we can interpret the t-test on the log scale as saying the population *medians* are different: The sample median for faculty is around \$80,000 while the sample median for athletics is around \$49,000, and it's hard to believe this big difference just happened because of random sampling. The reasons the means are similar while the median are different is the extreme right skew of the athletics salaries: Some athletics employees earn a boatload.