Exercises worked on in class on Thursday, Apr. 16

Before carrying out these exercises, some packages are needed:

```
> require(mosaic)
> require("Lock5withR")    # quotation marks aren't necessary, but are OK
```

While the work is largely that submitted by students (attributed by name), there is some editorializing from Professor Scofield in most (all?) of the answers.

1. (This answer supplied by team consisting of deHaan, LaCroix and Coria)

   Our data here is univariate, with the single variable (`HeartRate`) being quantitative. Thus, we will use a 1-sample t procedure (as in Sections 6.4–6.6).

   The goal of a CI here is to find the range of the mean heart rates that is likely to contain the mean heart rate within the population of people whose current health status warrants admission to the Intensive Care Unit (ICU).

   In Chapter 3 we learned to generate confidence intervals using bootstrapping. If we do so here, our first step would be to generate a bootstrap distribution, calling for RStudio commands such as these:

   ```
   > x = do(5000) * (mean(~HeartRate, data=resample(ICUAdmissions)))
   > histogram(~result, data=x)
   ```

   As our goal here is to practice the methods of Chapter 6, the commands above serve merely to provide some evidence that the sampling distribution of the sample mean $\overline{X}$ appears (to the naked eye) normal. We might also decide this is so simply on the basis of sample size, as this data set contains $n = 200$ cases, which is much larger than 30:

   ```
   > dim(ICUAdmissions)

   [1] 200  42
   ```

   So, we compute the sample mean, the sample standard deviation, and the appropriate t^*-value (for $df = n - 1 = 199$)

   ```
   > mean(~HeartRate, data=ICUAdmissions)

   [1] 98.925

   > qt(.975, df=199)

   [1] 1.971957

   > sd(~HeartRate, data=ICUAdmissions)

   [1] 26.82962
   ```

   and put these together in the usual way

   $$(\text{point estimate}) \pm t^*\text{SE}.$$

   ```
   > 98.925 + c(-1,1) * 1.97 * 26.83/sqrt(200)

   [1]  95.18758 102.66242
   ```

   Our 95% CI is: $[95.17, 102.68]$

2. (Work here is typical of that submitted by team of Cobb and Ivancich)

Our data here is bivariate; the two variables ("does the subject have dyslexia?", and "has the subject experienced gene disruption?") are binary categorical. Thus, we will use a 2-proportion procedure (as in Sections 6.7–6.9). The sorts of questions one might ask include

- Is there a connection between a disrupted DYXC1 gene and Dyslexia? (hypothesis test)
- Is the porportion of people who have this disrupted gene and have Dyslexia greater than the proportion of non-sufferers who have this disrupted gene? What is a likely range for the difference? (CI question)

Letting p_D and p_N stand for the proportion of people with gene disruption from those with and without dyslexia, respectively, we might use these hypotheses for an hypothesis test:

$$\mathrm{H}_0\colon\ p_D - p_N = 0, \qquad \mathrm{H}_\mathrm{a}\colon\ p_D - p_N > 0.$$

The methods of Chapter 6 are not so appropriate to use on the original data, as the rules of thumb for normality are not met. For instance,

$$n_N\hat{p}_N = 195 \cdot \frac{5}{195} = 5,$$

which is smaller than 10. However, on the modified data,

$$\begin{aligned} n_N\hat{p}_N &= 390 \cdot \frac{10}{390} = 10, & n_N(1-\hat{p}_N) &= 390 \cdot \frac{380}{390} = 380, \\ n_D\hat{p}_D &= 20, & n_D(1-\hat{p}_D) &= 198 \end{aligned}$$

so the rules of thumb are now met. (Note: For the purpose of practicing Chapter 6 methods, we have modified our data; this is **not** the sort of practice one should do with data in general!) While our point estimate (test statistic) is

$$\hat{p}_D - \hat{p}_N = \frac{20}{218} - \frac{10}{390} \doteq 0.0661,$$

in Chapter 6, the general procedure is to *standardize* this value (turn it into a z- or t-score). We are dealing with proportions, so ours is a z-score:

$$z = \frac{(\text{test statistic}) - (\text{hypothesized value})}{\text{SE}}.$$

To find SE, we recall that, under the null hypothesis, the two groups are considered the *same*—that is, they experience gene disruption as one population, not two, with a **pooled sample proportion** rate of

$$\hat{p} = \frac{20+10}{218+390} = \frac{30}{608} \doteq 0.0493.$$

We use this to calculate the standard error

$$\text{SE} = \sqrt{p(1-p)\left(\frac{1}{n_D} + \frac{1}{n_N}\right)}$$

```
>  sqrt((30/608)*(1 - 30/608) * (1/218 + 1/390))

[1] 0.01831522
```

So, our z-score is approximately

```
> (.0661 - 0) / .0183

[1] 3.612022
```

and we obtain our P-value

```
>  1 - pnorm(3.612)

[1] 0.0001519223
```

This P-value is small enough (even at the 1% significance level) to reject the null hypothesis and conclude that there is some link between disruption of the DYXC1 gene and instances of dyslexia.

Were we to construct a confidence interval, we would not have pooled together our data to obtain a single proportion. In that instance, we would have

$$\text{SE} = \sqrt{\frac{(p_D(1-p_D)}{n_D} + \frac{p_N(1-p_N)}{n_N}} = \sqrt{\frac{(0.0917)(0.9083)}{218} + \frac{(0.0256)(0.9744)}{390}} \doteq 0.0211.$$

3. (This answer partly supplied by team consisting of Cochran and Moentmann)

   A natural question would be: Do women and men exercise, on average, the same number of hours per week?

   One way to state hypotheses here is

   $$\text{H}_0\text{: } \mu_f = \mu_m, \qquad \text{H}_\text{a}\text{: } \mu_f \neq \mu_m.$$

   We obtain the sample means, sds for the two groups:

```
> favstats(Exercise ~ Gender, data=StudentSurvey)

  .group min Q1 median Q3 max     mean       sd   n missing
1      F   0  4      7 12  27 8.110119 5.198579 168       1
2      M   0  5     10 14  40 9.875648 6.068625 193       0
```

   The standardized test statistic (t, since we are dealing with means) is

   $$t \;=\; \frac{(\overline{x}_m - \overline{x}_f) - 0}{\text{SE}} \;=\; \frac{9.876 - 8.110}{\sqrt{\frac{5.199^2}{168} + \frac{6.069^2}{193}}} \;\doteq\; 2.978.$$

   Using a t-distribution with $df = 167$, we get P-value

```
> 2 * (1 - pt(2.978, df=167))

[1] 0.003333505
```

   We may reject the null hypothesis and conclude that the average number of hours exercising per week is not the same for women and men.

4. (This answer supplied by team consisting of Clark and Peplinski)

   The data here is paired; the population is wife–husband pairs. The variable of interest is the *difference in age* at marriage. We compute that variable on the couples sampled:

```
> ageDiffs = MarriageAges$Husband-MarriageAges$Wife
```

   This variable is quantitative and, as is usual in paired situations, a paired t-test (really just 1-sample t) is called for. (This is the content of Section 6.13.)

   Perhaps the natural "skeptic's view" is that, on average, these differences are 0. We express this in null and alternative hypotheses:

   $$\text{H}_0\text{: } \mu_d = 0, \qquad \text{H}_0\text{: } \mu_d \neq 0.$$

   We calculate the mean and standard deviation for these differences in the sampled couples

```
> mean(ageDiffs)

[1] 2.828571

> sd(ageDiffs)

[1] 4.995107
```

The standardized test statistic (a t-value) is

```
> (2.826-0)/(4.995/(sqrt(105)))

[1] 5.797374
```

We compute the area to the right of this t-score, then double it (because it is a 2-sided alternative hypothesis). There are 105 couples, so $df = 104$:

```
> (1-pt(5.802708, df=104))*2

[1] 7.114905e-08
```

This is quite a small P-value, leading us to reject the null hypothesis and conclude couples have a mean difference in age at the time of marriage that is nonzero.

5. (This answer supplied by team consisting of DeBoer and Witte)

   The data here is univariate categorical (i.e., for each *case*, what was recorded was the server who collected the tip.) We need to use 1-proportion procedures (Sections 6.1–6.3).

   The hypotheses:

$$\mathrm{H}_0\colon\ p_B = \frac{1}{3}, \qquad \mathrm{H}_\mathrm{a}\colon\ p_B > \frac{1}{3}$$

   To carry out the analysis, we note that every tip falls into *two* categories, those collected by Server B, and those not collected by Server B. There are 157 tips in all, so we have sample statistic ($\hat{p}$)

```
> p.hat = 65/157
> p.hat

[1] 0.4140127
```

   The standard error

```
> se = sqrt((1/3)*(2/3)/157)
> se

[1] 0.03762218
```

   so the P-value is

```
> 1 - pnorm(p.hat, 1/3, se)

[1] 0.01599786
```

   Or, following our usual approach (when normality is in play), we find the (z) standardized value

```
> z = (p.hat - 1/3) / se
> z

[1] 2.144464
```

   and obtain the P-value with

```
> 1 - pnorm(z)

[1] 0.01599786
```

At the 5% significance level we can reject the null hypothesis and say that there is evidence that Server B gets more than 1/3 of tips.