Chapter 6 Textbook exercises

Solutions to even-numbered questions
Statistics and statistical programming
Northwestern University
MTS 525

Aaron Shaw

October 22, 2020

All exercises taken from the *OpenIntro Statistics* textbook, 4^{th} edition, Chapter 6.

6.10 Marijuana legalization, Part I

(a) It is a sample statistic (the sample mean), because it comes from a sample. The population parameter (the population mean) is unknown in this case, but can be estimated from the sample statistic.

(b) As given in the textbook, confidence intervals for proportions are equal to:

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p} \times (1 - \hat{p})}{n}}$$

Where we calculate z^* from the Z distribution table. For a 95% confidence interval, $z^* = 1.96$, so we can plug in the values $\hat{p}$ and n given in the problem and calculate the interval like this:

```
lower = .61 - 1.96 * sqrt(.61 * .39 / 1578)
upper = .61 + 1.96 * sqrt(.61 * .39 / 1578)

ci = c(lower, upper)
print(ci)
```

```
## [1] 0.5859342 0.6340658
```

This means that we are 95% confident that the true proportion of Americans who support legalizing marijuana is between ~58.6% and ~63.4%.

(c) We should believe that the distribution of the sample proportion would be approximately normal because we have a large enough sample (collected randomly) in which enough responses are drawn from each potential outcome to assume (1) that the observations are independent and (2) that the distribution of the sample proportion is approximately normal.

(d) Yes, the statement is justified, since our confidence interval is entirely above 50%.

6.16 Marijuana legalization, Part II

We can use the point estimate of the poll to estimate how large a sample we would need to have a confidence interval of a given width.

In this case, we want the margin or error to be 2%. Using the same formula from above, this translates to:

$$1.96 \times \sqrt{\frac{.61 \times .39}{n}} \leq .02$$

Rearrange to solve for n:

$$\begin{aligned} \sqrt{\frac{.61 \times .39}{n}} &\leq \frac{.02}{1.96} \\ \frac{.61 \times .39}{n} &\leq \left(\frac{.02}{1.96}\right)^2 \\ \frac{(.61 \times .39)}{\left(\frac{.02}{1.96}\right)^2} &\leq n \end{aligned}$$

Let R solve this:

```
(.61 * .39)/(.02/1.96)^2
```

```
## [1] 2284.792
```

So, we need a sample of at least 2,285 people (since we can't survey fractions of a person).

6.22 Sleepless on the west coast

Before we march ahead and plug numbers into formulas, it's probably good to review that the conditions for calculating a valid confidence interval for a difference of proportions are met. Those conditions (and the corresponding situation here) are:

1. *Indepdendence*: Both samples are random and less than 10% of the total population in the case of each state. The observations within each state are therefore likely independent. In addition, the two samples are independent of each other (the individuals sampled in Oregon do not (likely) have any dependence on the individuals sampled in California).
2. *Success-failure*: The number of "successes" in each state is greater than 10 (you can multiply the respective sample sizes by the proportions in the "success" and "failure" outcomes to calculate this directly).

With that settled, we can move on to the calculation. The first equation here is the formula for a confidence interval for a difference of proportions. Note that the subscripts CA and OR indicate parameters for the observed proportions ($\hat{p}$) reporting sleep deprivation from each of the two states.

$$\hat{p}_{CA} - \hat{p}_{OR} \pm z^* \sqrt{\frac{\hat{p}_{CA}(1-\hat{p}_{CA})}{n_{CA}} + \frac{\hat{p}_{OR}(1-\hat{p}_{OR})}{n_{OR}}}$$

Plug values in and recall that $z^* = 1.96$ for a 95% confidence interval:

$$0.8 - 0.088 \pm 1.96 \sqrt{\frac{0.08 \times 0.92}{11545} + \frac{0.088 \times 0.912}{4691}}$$

Let's let R take it from there:

```
var.ca <- (0.08*0.92 ) / 11545
var.or <- (0.088*0.912) / 4691
se.diff <- 1.96 * sqrt(var.or + var.ca)
upper <- 0.08-0.088 + se.diff
lower <- 0.08-0.088 - se.diff
print(c(lower, upper))
```

```
## [1] -0.017498128  0.001498128
```

The data suggests that the 95% confidence interval for the difference between the proportion of California residents and Oregon residents reporting sleep deprivation is between −1.75% and 0.1%. In other words, we can be 95% confident that the true difference between the two proportions falls within that range.

6.30 Apples, doctors, and informal experiments on children

tl;dr answer: No. Constructing the test implied by the question is not possible without violating the assumptions that define the estimation procedure, and thereby invalidating the estimate.

longer answer: The question the teacher wants to answer is whether there has been a meaningful change in a proportion across two data collection points (the students pre- and post-class). While the tools we have learned could allow you to answer that question for *two independent groups*, the responses are not independent in this case because they come from the same students. You could go ahead and calculate a statistical test for difference in pooled proportions (after all, it's just plugging values into an equation!) and explain how the data violates a core assumption of the test. However, since the dependence between observations violates that core assumption, the baseline expectations necessary to construct the null distribution against which the observed test statistic can be evaluated are not met. The results of the hypothesis test under these conditions may or may not mean what you might expect (the test has nothing to say about that).

6.40 Website experiment

(a) The question gives us the total sample size and the proportions cross-tabulated for treatment condition (position) and outcome (download or not). I'll use R to work out the answers here.

```
props <- data.frame(
  "position" = c("pos1", "pos2", "pos3"),
  "download" = c(.138, .146, .121),
  "no_download" = c(.183, .185, .227)
)
props
```

```
##   position download no_download
## 1     pos1    0.138       0.183
## 2     pos2    0.146       0.185
## 3     pos3    0.121       0.227
```

Now multiply those values by the sample size to get the counts:

```
counts <- data.frame(
  "position" = props$position,
  "download" = round(props$download*701, 0),
  "no_download" =  round(props$no_download*701, 0)
)

counts
```

```
##   position download no_download
## 1     pos1       97         128
## 2     pos2      102         130
## 3     pos3       85         159
```

(b) This set up is leading towards a χ^2 test for goodness of fit to evaluate balance in a one-way table (revisit the section of the chapter dealing with this test for more details). We can construct and conduct the

test using the textbook's (slightly cumbersome, but delightfully thorough and transparent) "prepare-check-calculate-conclude" algorithm for hypothesis testing. Let's walk through that:

Prepare: The first thing to consider is the actual values the question is actually asking us to compare: the total number of study participants in each condition. We can do that using the table from part (a) above:

```
counts$total <- counts$download + counts$no_download
counts$total
```

```
## [1] 225 232 244
```

So the idea here is to figure out whether these counts are less balanced than might be expected. (And this is maybe a good time to point out that you might eyeball these values and notice that they're all pretty close together.)

Here are the hypotheses stated more formally:
H_0: The chance of a site visitor being in any of the three groups is equal.
H_A: The chance of a site visitor being in one group or another is not equal.

Check: Now we can check the assumptions for the test. If H_0 were true, we might expect 1/3 of the 701 visitors (233.67 visitors) to be in each group. This expected (and observed) count is greater than 5 for all three groups, satisfying the *sample size /distribution condition*. Because the visitors were assigned into the groups randomly and only appear in their respective group once, the *indepdendence condition* is also satisfied. That's both of the conditions for this test, so we can go ahead and conduct it.

Calculate: For a χ^2 test, we need to calculate a test statistic as well as the number of degrees of freedom. Here we go, in that order.

First up, let's set up the test statistic given some number of cells (k) in the one-way table:

$$\begin{aligned}\chi^2 &= \sum_{n=1}^{k} \frac{(Observed_k - Expected_k)^2}{Expected_k} \\ &= \frac{(225-233.67)^2}{233.67} + \frac{(232-233.67)^2}{233.67} + \frac{(244-233.67)^2}{233.67} \\ &= 0.79\end{aligned}$$

Now the degrees of freedom:

$$df = k - 1 = 2$$

You can look up the results in the tables at the end of the book or calculate it in R using the `pchisq()` function. Note that the `pchisq()` function returns "lower tail" area values from the χ^2 distribution. However, for these tests, we usually want the corresponding "upper tail" area, which can be found by subtracting the results of a call to `pchisq()` from 1.

```
1-pchisq(.79, df=2)
```

```
## [1] 0.67368
```

Conclude: Because this p-value is *larger* than 0.05, we cannot reject H_0. That is, we do not find evidence that randomization of site visitors to the groups is imbalanced.

(c) I said you did *not* need to do this one, but I'll walk through the setup and solution anyway because it's useful to have an example. We're doing a χ^2 test again, but this time for independence in a two-way table. Because the underlying setup is pretty similar, my solution here is a bit more concise.

Prepare: Create the null and alternative hypotheses (in words here, but we could do this in notation too).
H_0: No difference in download rate across the experiment groups.
H_A: Some difference in download rate across the groups.

Check: Each visitor was randomly assigned to a group and only counted once in the table, so the observations are independent. The expected counts can also be computed by following the procedure described on p.241

of the textbook to get the expected counts under H_0. Those expected counts in this case are (reading down the first column then down the second): 91.2, 94.0, 98.9, 133.8, 138.0, 145.2. All of these expected counts are at least 5 (which, let's be honest, you might have been able to infer/guess just by looking at the observed counts). Therefore we can use the χ^2 test.

Calculate: the test statistic and corresponding degrees of freedom. For the test statistic

$$\begin{aligned}
\chi^2 &= \sum_{n=1}^{k} \frac{(Observed_k - Expected_k)^2}{Expected_k} \\
&= \frac{(97-91.2)^2}{91.2} + \ ... \ + \frac{(159-145.2)^2}{145.2} \\
&= 5.04 \\
df &= 3 - 1 = 2
\end{aligned}$$

Once again, I'll let R calculate the p-value:

```
1-pchisq(5.04, 2)
```

```
## [1] 0.08045961
```

Conclude: The p-value is (just a little bit!) greater than 0.05, so assuming a typical hypothesis testing framework, we would be unable to reject H_0 that there is no difference in the download rates. In other words, we do not find compelling evidence that the position of the link led to any difference in download rates. That said, given that the p-value is quite close to the conventional threshold, you might also note that it's possible that there's a small effect that our study design was insufficiently sensitive to detect.