Chapter 7 Textbook exercises

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

Aaron Shaw

October 28, 2020

All exercises taken from the *OpenIntro Statistics* textbook, 4^{th} edition, Chapter 7.

7.12 Lead exposure

(a) The hypotheses can be written:

$$H_0 \ : \ \mu = 35$$
$$H_A \ : \ \mu \neq 35$$

(b) The conditions we need to evaluate are *independence* of observations and *normality* of the distribution. More about each below:
Indepdendence: If the 52 officers represent a random sample, then independence would be satisfied. Unfortunately, we cannot check this and the assumption seems like a bit of a stretch.
Normality: There's no plot or summary information about the distribution, so it's hard to check whether this condition holds or not either. That said, with $n \geq 30$ the distribution would need to be quite skewed for the t-test procedure to be biased or invalid, so this is probably not all that crucial/concerning.

(c) The test statistic, degrees of freedom, and p-value can be calculated from the information provided:

$$T = \frac{124.32 - 35}{\frac{37.74}{\sqrt{52}}} \approx 17.07$$
$$df = 52 - 1 = 51$$
$$p = 2 \times P(T_{51} > 17.07) < 0.001$$

With the test statistic and the degrees of freedom, we could do that last bit in R:

```
pt(17.07, 51, lower.tail=FALSE)
```

```
## [1] 4.914893e-23
```

That's quite a small p-value! The hypothesis test suggests that we can reject H_0. Given that the observed difference of means ($124.32 - 35 = 89.32$) is large relative to the range and standard deviation of the distribution for the officers (and presumably even larger in reference to the distribution of the suburbanites or the pooled distribution of both groups), the data provides compelling evidence that the police officers have a higher lead concentration in their blood than the suburbanites. Further inferences, such as whether or not this difference can be attributed to the effect of the additional exposure experienced by the officers, would require additional data and some way to disentangle the causal effects of traffic enforcement from any other observed or unobserved differences between the officers and the suburbanites.

7.24 Diamonds, Part I

We want to test the following hypotheses:

$$H_0 \ : \ \mu_{0.99} = \mu_1$$

$$H_A \ : \ \mu_{0.99} \neq \mu_1$$

To do so, we can use a two-sample t-test to compare the two sample means. The conditions we'd like to satisfy are independence and normality. Re: independence, the samples are random and not exhaustive of the populations (presumably less than 10% of all the diamonds of each carat rating on earth), so we should be good to go. Re: normality, visual inspection of the histograms presented in the textbook suggests that what skew may be present in either distribution is not extreme.

Given that the conditions are met, here's how you could construct the test statistic T:

$$T \ = \ \frac{Point \ estimate \ - \ Null}{Standard \ error_{difference}}$$

Plugging in formulas from the textbook this looks like:

$$T \ = \ \frac{(\overline{x}_1 - \overline{x}_2) - (0)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Now, plug in values from the table provided in the question text:

$$T \ = \ \frac{(44.51 - 56.81) - (0)}{\sqrt{\frac{13.32^2}{23} + \frac{16.13^2}{23}}}$$

Work that out and you should have $T \ = \ -2.82$. The degrees of freedom are estimated by the smaller of $n_1 - 1$ or $n_2 - 1$ (which are equal in this case), so $df \ = \ 22$. Consulting the table of T-statistics from the back of the book, we find:

$$p_{value} = P(T_{22} > 2.82) \approx 0.01$$

Or, you might calculate that in R:

```
pt(-2.82, 22) ## lower.tail == TRUE since t* is negative
```

```
## [1] 0.004985866
```

Assuming we're okay with a false positive rate of $p \leq 0.05$, this provides support for the alternative hypothesis and we can reject the null of no difference between the average standardized prices of 0.99 and 1 carat diamonds.

7.26 Diamonds, Part II

To construct the confidence interval for the difference of means, I need to calculate the following:

$$(\overline{x_1} - \overline{x_2}) \ \pm \ t^*_{df} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

All of this is available from the table except the critical value t^*_{df}. To find that, I can either use the table in the textbook (and find that it is about 2.07) or calculate it directly in R using the `qt()` function. The degrees of freedom are approximated by the smaller value $n - 1$ from either sample (in this case, both yield the same number: 22).

```r
t.star <- qt(0.025, df=22, lower.tail=FALSE)

diff.means <- 56.81-44.51
se <- sqrt( ((16.13^2)/23) + ((13.32^2)/23) )

diff.means - (t.star*se) ## lower
```

```
## [1] 3.254
```

```r
diff.means + (t.star*se) ## upper
```

```
## [1] 21.346
```

In words, I am 95% confident that the average difference between the standardized prices of 1 carat diamond and a 0.99 carat diamond falls between \$3.27 and \$21.33 (the 1 carat diamond costs more).