Dealing with Data II Homework 6 Solutions

Joshua D. Ingram

2022-12-02

Reading

a) Chapter 7: Section 7.1 and 7.2

On The Central Limit Theorem for Proportions and Means. In particular, check out the yellow boxes.

b) Chapter 10: Section 10.1 and 10.2

On comparing two populations (or groups) in terms of their proportions or means. Only focus on those parts that talk about a confidence interval for the difference in proportions, or a confidence interval for a difference in means. You can ignore the details about hypothesis testing.

c) Chapter 11: Section 11.1 and 11.2

On the Chi-squared Test for Independence

Textbook Exercises Chapter 7

7.7 - Delays at airport

During the month of January 2017, a total of 29,544 flights took off from the Atlanta International Airport. Of all these flights, 23.9% had a departure delay of more than 10 minutes. If we were to randomly sample just 100 of these flights.

a) What sample proportion should we expect to see in such a sample, and how much should we expect the proportio to vary from sample to sample in samples of size 100?

Answer:

We would expect to see a sample proportion close to 0.239 (the population proportion). We would expect the proportion to vary from sample to sample by about $\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.239(1-0.239)}{100}} = 0.0426$.

b) What is the approximate shape of the sampling distribution of the sample proportion of flights delayed by more than 10 minutes in samples of size 100? Why?

Answer:

The shape of the sampling distribution would be normal, with the mean equal to the population proportion and variance $\frac{p(1-p)}{n} = 0.0018$. This is because of the central limit theorem.

c) Would it be unusual for us to observe a sample proportion of 40% or higher of delayed flights? Why or why not?

Answer:

Yes. Observing a sample with a proportion of 40% or higher would be extremely unlikely, as this is more than 3 standard deviations from the center of the sampling distribution.

d) Of all the flights scheduled to leave the airport in January 2017, 2% got canceled. If we were to randomly sample just 100 of all flights scheduled to leave the airport, would the sampling distribution of the sample proportion of canceled flights have an approximate normal shape? Why or why not?

Answer:

No. For the central limit theorem to apply with proportions, we must have $np > 15$ and $n(1-p) > 15$, but here $np = 0.02 * 100 = 2$.

7.21 - Shared family phone plan

A recent personalized information sheet from your wireless phone carrier claims that the mean duration of all your phone calls was $\mu = 2.8$ minutes with a standard deviation of $\theta = 2.1$ minutes.

a) Is the population distribution of the duration of your phone calls likely to bell shaped, right skewed, or left skewed?

Answer:

The population distribution would likely be right skewed. This is because of the lower limit on call lengths of zero minutes.

b) You are on a shared wireless plan with your parents, who are statisticians. They look at some of your recent monthly statements that list each call and its duration and randomly sample 45 calls form the thousands listed there. They construct a histogram of the duration to look at the data distribution. Is this distribution likely to be bell shaped, right skewed, or left skewed, or is it impossible to tell?

Answer:

It would most likely look right skewed, like the population distribution, but will vary from sample to sample.

c) From the sample of n = 45 calls , your parents compute the mean duration. Is the sampling distribution of the sample mean likely to be bell shaped, right skewed, or left skewed, or is it impossible to tell? Explain.

Answer:

The sampling distribution will be bell shaped. This is because of the central limit theorem.

7.22 - Dropped from plan

The previous exercise mentions that the duration of your phone calls follows a distribution with mean $\mu = 2.8$ minutes and standard deviation $\theta = 2.1$ minutes. From a random sample of $n = 45$ calls, your parents computed a sample mean of $\bar{x} = 3.4$ minutes and a sample standard deviation of $s = 2.9$ minutes.

a) Give the likely shape of the population distribution of your phone calls. What are its mean and standard deviation?

Answer:

The population distribution would likely be right skewed. This is because of the lower limit on call lengths of zero minutes. The mean and standard deviation are $\mu = 2.8$ minutes and $\theta = 2.1$ minutes.

b) What are the mean and standard deviation of the data distribution?

Answer:

The sample mean is $\bar{x} = 3.4$ minutes and the sample standard deviation is $s = 2.9$ minutes.

c) Find the mean and standard deviation of the sampling distribution of the sample mean.

Answer:

The mean of the sampling distribution is the population mean $\mu = 2.8$ minutes, and its standard deviation (standard error) is $\frac{\sigma}{\sqrt{(n)}} = \frac{2.1}{\sqrt{45}} = 0.313$ minutes.

d) Is the sample mean of 3.4 minutes unusually high? Find its z-score and comment.

Answer:

The z-score is $z = \frac{3.4-2.6}{0.313} = 1.916933$. It is high, with only 2.8% of sample mean call times expected to be as long or longer.

e) Your parents told you that they will kick you off the plan when they find a sample mean larger than 3.5 minutes. How likely is this to happen?

Answer:

This extremely unlikely, with only a 1.27% chance of a sample mean call time being as long or longer.

Textbook Exercises Chapter 10

10.5 - Do you believe in miracles

Let p_1 and p_2 denote the population proportions of males and females in the United States who answer, yes, definitely when asked whether they believe in miracles. Based on results from the 2008 GSS, 277 males (out of 603 responding) and 461 females (out of 730 responding) indicated "Yes, definitely."

a) Report point estimates of p_1 and p_2

Answer:

$\hat{p}_1 = \frac{277}{603} = 0.4593698$, $\hat{p}_2 = \frac{461}{730} = 0.6315068$

b) Construct a 95% confidence interval for $(p_1 - p_2)$, specifying the assumptions you make to use this method. Interpret.

Answer:

$(\hat{p}_1 - \hat{p}_2) = 0.4593698 - 0.6315068 = -0.172137$

```
prop.test(x = c(277, 461), n = c(603, 730), correct = FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(277, 461) out of c(603, 730)
## X-squared = 39.595, df = 1, p-value = 3.124e-10
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.2251152 -0.1191588
```

```
## sample estimates:
##    prop 1    prop 2
## 0.4593698 0.6315068
```

We assume that the our samples are random and independent. The 95% confidence interval is (-0.2251152, -0.1191588). We are 95% confident that the true proportion of males that "yes, definitely" believe in miracles in the United States is between 22.5 and 11.9 percentage points lower than females.

c) Based on the interval in part b, explain why the proportion believing in miracles may have been quite a bit larger for females, or it might have been only moderately larger.

Answer:

The difference is negative, and the lower bound has a considerable distance from 0.

10.12 - Obama A/B testing

To increase Barack Obama's visibility and to raise money for the campaign leading up to the 2008 presidential election, Obama's analytics team conducted an A/B test with his website. In the original version, the button to join the campaign said "Sign Up". In an alternative version, it read "Learn More". Of 77,858 visitors to the original version, 5,851 clicked the button. Of 77,729 visitors to the alternative version, 6,927 clicked the button. Is there evidence that one version was more successful than the other in recruiting campaign members? The following shows computer output from statistical software, but you can also enter the data into the *Compare Two Proportions* app.

(don't do parts a, b and c, but use the data from this exercise and R to find the confidence interval, interpret it in context, and then mention, how based on the confidence interval, you can tell whether the proportions are significantly different.)

Answer:

```
# sign up - learn more
prop.test(x = c(5851, 6927), n = c( 77858, 77729), correct = FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(5851, 6927) out of c(77858, 77729)
## X-squared = 100.67, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.01669550 -0.01123987
## sample estimates:
##     prop 1     prop 2
## 0.07514963 0.08911732
```

The 95% confidence interval is (-0.01669550, -0.01123987). We are 95% confident that the true proportion of visitors that signed up with the "sign up" button is between 1.67 and 1.12 percentage points lower than visitors that signed up with the "learn more" button. The difference is statistically significant, as the interval does not contain 0.

10.28, parts a, b, c - Student survey

Refer to the FL Student Survey data file on the book's website. Use the number of times reading a newspaper as the response variables and gender as the explanatory variable. The observations are as follows:

Females: 5, 3, 6, 3, 7, 1, 1, 3, 0, 4, 7, 2, 2, 7, 3, 0, 5, 0, 4, 4, 5, 14, 3, 1, 2, 1, 7, 2, 5, 3, 7

Males: 0, 3, 7, 4, 3, 2, 1, 12, 1, 6, 2, 2, 7, 7, 5, 3, 14, 3, 7, 6, 5, 5, 2, 3, 5, 5, 2, 3, 3

Using the *Compare Two Means* app or other software,

a) Construct and interpret an appropriate plot comparing responses by females and males.

Answer:

```
library(readr)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5      v dplyr   1.0.10
## v tibble  3.1.8      v stringr 1.4.0
## v tidyr   1.2.0      v forcats 0.5.1
## v purrr   0.3.4
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
survey <- read_csv("/Users/joshuaingram/Main/Projects/masters_coursework/teaching_assistant/dealing_wit
```

```
## Rows: 60 Columns: 18
```

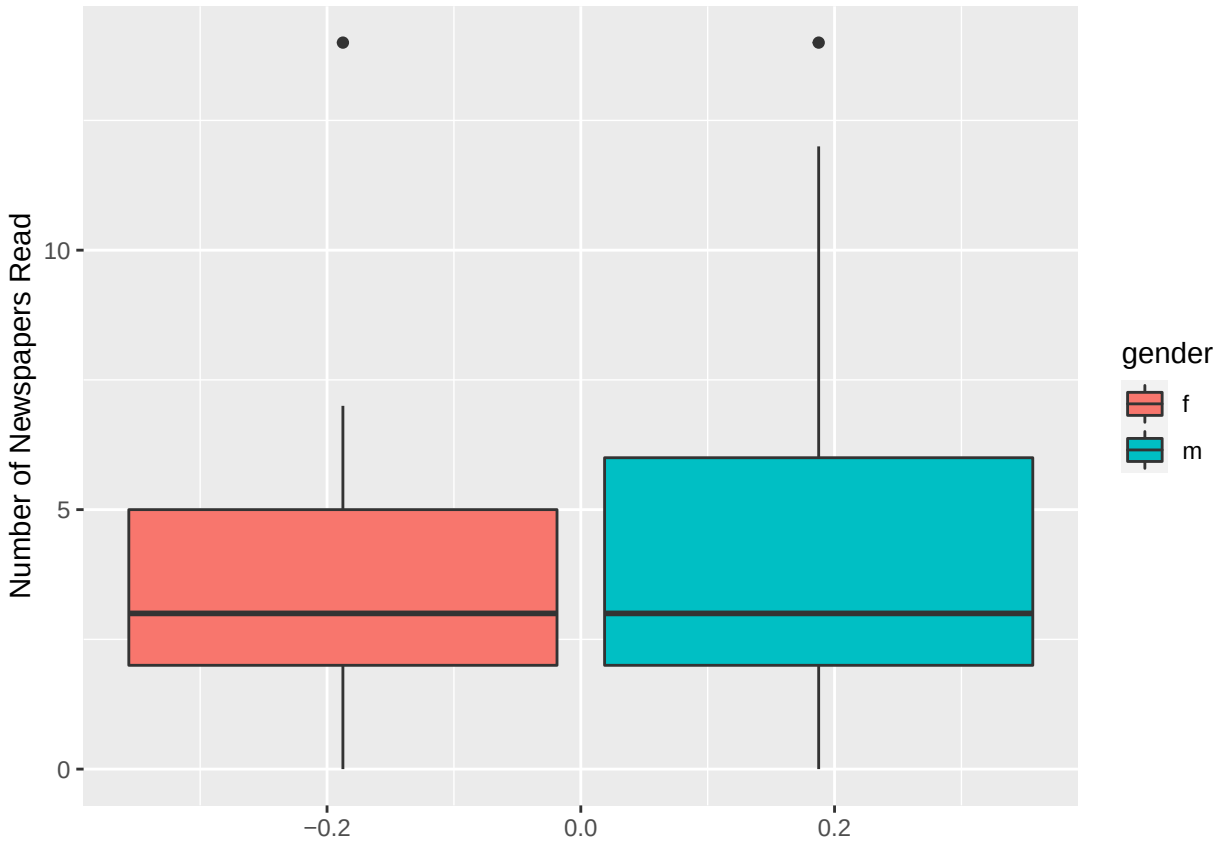
```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (7): gender, newspapers, vegetarian, political_affiliation, abortion_le...
## dbl (11): subject, age, high_sch_GPA, college_GPA, distance_home, distance_r...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
ggplot(survey, aes(y = as.numeric(newspapers), fill = gender)) +
  geom_boxplot() +
  labs(y = "Number of Newspapers Read")
```

```
## Warning in FUN(X[[i]], ...): NAs introduced by coercion
```

```
## Warning in FUN(X[[i]], ...): NAs introduced by coercion
```

```
## Warning: Removed 2 rows containing non-finite values (stat_boxplot).
```



b) Find and interpret a 95% confidence interval comparing population means for females and males.

Answer:

```
t.test(as.numeric(newspapers) ~ gender, data = survey)
```

```
## Warning in eval(predvars, data, env): NAs introduced by coercion
```

```
##
##  Welch Two Sample t-test
##
## data:  as.numeric(newspapers) by gender
## t = -0.81836, df = 55.067, p-value = 0.4167
## alternative hypothesis: true difference in means between group f and group m is not equal to 0
## 95 percent confidence interval:
##  -2.2745639  0.9555163
## sample estimates:
## mean in group f mean in group m
##        3.733333        4.392857
```

The confidence interval contains 0, so there is not a significant difference.

c) State the assumptions of the t test.

Answer:

1. Sample must be random
2. Sample size must be greater than 30
3. The data must be normally distributed

4. The variance of both samples must be the same.

Textbook Exercises Chapter 11

11.11 - Marital happiness and income

The contingency table for relative family income and marital status in 2012 from Exercise 11.5 is as follows:

(enter the data as a matrix into R, as shown on the handout, and obtain the Chi-squared test in R, confirming the results as given in the exercise. Also, include a side-by-side barchart to visualize the data.)

Answer:

$$H_o : \text{Marital happiness is independent of family income.}$$

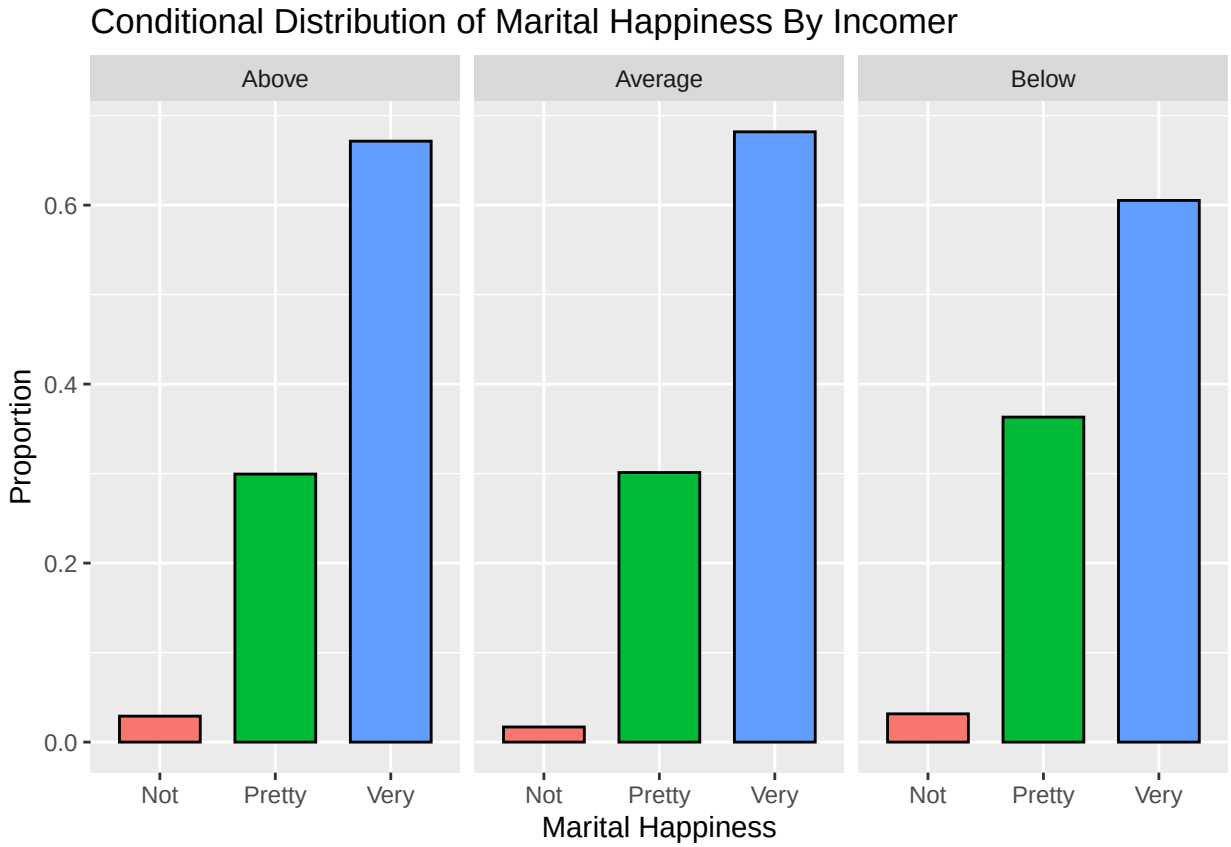
$$H_a : \text{There is a relationship between marital happiness and family income.}$$

```
table <- as.table(
matrix(c(6, 62, 139, 7, 125, 283, 6, 69, 115),ncol=3,byrow=TRUE))
dimnames(table) <- list("Income" = c("Above", "Average", "Below"), "Marital Happiness" = c("Not", "Pret
```

```
table_cond <- prop.table(table,1)
table_cond <- as.data.frame(table_cond)
table_cond
```

```
##    Income Marital.Happiness       Freq
## 1   Above               Not 0.02898551
## 2 Average               Not 0.01686747
## 3   Below               Not 0.03157895
## 4   Above            Pretty 0.29951691
## 5 Average            Pretty 0.30120482
## 6   Below            Pretty 0.36315789
## 7   Above              Very 0.67149758
## 8 Average              Very 0.68192771
## 9   Below              Very 0.60526316
```

```
ggplot(data=table_cond, aes(x=str_wrap(Marital.Happiness,width=10), y=Freq, fill=Marital.Happiness)) +
  geom_bar(stat="identity", width=0.7, position="dodge",color="black",show.legend=FALSE,alpha=1) +
  labs(y="Proportion", x="Marital Happiness", title="Conditional Distribution of Marital Happiness By I
  facet_wrap(~Income, nrow=1)
```




```
expected <- chisq.test(table)$expected
```

```
## Warning in chisq.test(table): Chi-squared approximation may be incorrect
```

```
expected
```

```
##          Marital Happiness
## Income         Not    Pretty     Very
##   Above   4.843596  65.26108 136.8953
##   Average 9.710591 130.83744 274.4520
##   Below   4.445813  59.90148 125.6527
```

```
chisq.test(table)
```

```
## Warning in chisq.test(table): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  table
## X-squared = 4.5831, df = 4, p-value = 0.3328
```

With a chi-squared statistic of 4.58 and p-value of 0.3328, we do not have sufficient evidence to reject the null hypothesis, that income and marital happiness are independent.