

Instructions

For the first problem, an exercise from the textbook (OpenIntro Statistics) has been worked, with any requisite R (or RStudio) commands provided. Use it as a reference when answering the similar exercise specified. The inputs that will require new values are highlighted in yellow. All input and output will be in the RStudio console. For the second problem an R script file has been provided to generate all requisite output. Further, it is recommended that sections 6.2 and 6.6 on pp. 280 - 286; 307 - 311 be read, or at least reviewed, prior to completing this lab assignment.

1. Exercise 6.29 on page 319

A 2010 survey asked 827 randomly sampled registered voters in California “Do you support? ... Or do you oppose? ... drilling for oil and natural gas ... off the Coast of California? Or do you not know enough to say?” Below is the distribution of responses, separated based on whether or not the respondent graduated from college.¹

¹ Survey USA, [Election Poll #16804](#), data collected July 8-11, 2010.

	<u>College Grad</u>	
	Yes	No
Support	154	132
Oppose	180	126
Do not know	104	131
Total	438	389

- (a) What percent of college graduates and what percent of the non-college graduates in this sample do not know enough to have an opinion on drilling for oil and natural gas off the Coast of California?

If we let $\hat{p}_{Yes}$ denote the percent of *college* graduates and $\hat{p}_{No}$ denote the percent of the *non-college* graduates in this sample that *do not know* enough to have an opinion on drilling for oil and natural gas off the Coast of California, their respective values are:

$$\hat{p}_{Yes} = \frac{104}{438} \approx 23.74\% \qquad \hat{p}_{No} = \frac{131}{389} \approx 33.68\%$$

Note: these are the point estimates for the true proportion that *do not know* enough to have an opinion on drilling for oil and natural gas off the Coast of California from the respective college graduate and non-college graduate populations.

- (b) Conduct a hypothesis test to determine if the data provide strong evidence that the proportion of college graduates who do not have an opinion on this issue is different than that of non-college graduates.

- State hypotheses.

$$\begin{array}{llll} H_0: p_{Yes} = p_{No} & \text{or} & H_0: p_{Yes} - p_{No} = 0 & \text{or} & H_0: p_{No} - p_{Yes} = 0 \\ H_A: p_{Yes} \neq p_{No} & \text{or} & H_A: p_{Yes} - p_{No} \neq 0 & \text{or} & H_A: p_{No} - p_{Yes} \neq 0 \end{array}$$

H_0 : There is no difference in the proportion that *do not know* enough to have an opinion on drilling for oil and natural gas off the Coast of California between college and non-college graduates.

H_A : There is a difference in the proportion that *do not know* enough to have an opinion on drilling for oil and natural gas off the Coast of California between college and non-college graduates.

- *Check relevant conditions.*

As stated on page 280, “the difference $\hat{p}_1 - \hat{p}_2$ tends to follow a normal model when (1) each [sample] proportion follows a normal model, and (2) the two samples are independent of each other.” In other words, we need to verify the **conditions for the sampling distribution of $\hat{p}$ being nearly normal** (see p. 275) for both the college graduate and non-college graduate populations and their respective sample must be independent.

(1.a) the sample observations are independent

This assumption is reasonable as both the college graduate and non-college graduate samples were randomly selected (we shall presume using simple random sampling) from all registered voters in California. Further, their respective sample sizes, 438 and 389 are much less than 10% of the college graduate and non-college graduate populations. Alternatively, we can reasonably assume there are more than 4,380 college graduates and 3,890 non college graduates among the registered voters in California.

(1.b) success-failure condition

IMPORTANT: If we were constructing a confidence interval, we would check this condition for each sample. Otherwise, we would use the pooled proportion, as defined on pp. 282 - 283, for testing a hypothesis; both approaches are shown here.

CONFIDENCE INTERVAL

For the sample of college graduates, 104 do not know enough to have an opinion (success), while $438 - 104 = 334$ either support or oppose (failure); both are at least 10.

For the sample of non-college graduates, 131 do not know enough to have an opinion (success), while $389 - 131 = 258$ either support or oppose (failure); both are at least 10.

The success-failure condition is satisfied since all four values are at least 10.

HYPOTHESIS TEST (Pooled Proportion)

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{104 + 131}{438 + 389} = \frac{235}{827} \approx 28.42\%$$

Under the assumption that there is no difference between the two proportions, i.e., H_0 : the two proportions are equal is true, we expect to find...

For the sample of college graduates, $438(0.2842) \approx 124.4619$ do not know enough to have an opinion (success), while $438 - 124.4619 = 313.5381$ either support or oppose (failure); both are at least 10.

For the sample of non-college graduates, $389(0.2842) \approx 110.5318$ do not know enough to have an opinion (success), while $389 - 110.5318 = 278.4619$ either support or oppose (failure); both are at least 10.

The success-failure condition is satisfied since all four values are at least 10.

(2) two samples are independent (of each other)

Since the 438 college graduates and 389 non-college graduates were each randomly selected from the population of registered voters in California, the two samples are independent; that is the selection of the college graduates did not determine/influence the selection of the non-college graduates (and vice-versa) - the two groups are unrelated.

- *Interpret results in context of the data.*

As discussed on Example 6.14 on pages 283 - 284, the point estimate of the difference in the proportion that *do not know* enough to have an opinion on drilling for oil and natural gas off the Coast of California between the college graduates and non-college graduates among California's registered voters is

$$\hat{p}_{Yes} - \hat{p}_{No} = \frac{104}{438} - \frac{131}{389} \approx -0.099318 = -9.93\%$$

with a standard error of (assuming no difference, i.e., H_0 is true)

$$SE = \sqrt{\hat{p}(1 - \hat{p}) \left(\frac{1}{n_{Yes}} + \frac{1}{n_{No}} \right)} = \sqrt{\frac{235}{827} \left(1 - \frac{235}{827} \right) \left(\frac{1}{438} + \frac{1}{389} \right)} \approx 0.03142174 \approx 3.14\%$$

The test statistic, Z , is then,

$$Z = \frac{(\hat{p}_{Yes} - \hat{p}_{No}) - \text{null value}}{SE} \approx \frac{(-0.099318) - 0}{0.03142174} \approx -3.160806$$

with a corresponding p-value of

$$P(Z < -3.160806 \cup Z > 3.160806) = P(|Z| > 3.160806) \approx 0.00157333$$

These calculations can be done altogether by typing (or copying and pasting) the following R commands into the console in RStudio:

```
# inputs
x1 <- 104; n1 <- 438; x2 <- 131; n2 <- 389

# calculations
p.pool <- (x1 + x2)/(n1 + n2)
p.diff <- x1/n1 - x2/n2
se <- sqrt(p.pool*(1-p.pool)*(1/n1+1/n2))
z <- p.diff/se
p.value <- 2*pnorm(abs(z), lower.tail=F)

# print results
p.pool
p.diff
se
z
p.value
```

which yields the resulting [edited] console output:

```
> p.pool
[1] 0.2841596

> p.diff
[1] -0.099318

> se
[1] 0.03142174

> z
[1] -3.160806

> p.value
[1] 0.00157333
```

We could have also obtained the value of the test statistic (albeit as a chi-squared statistic with 1 degree of freedom, which is equivalent to Z^2 - see section 6.3.2 on pp. 288 - 289) and corresponding p-value by using the more convenient `prop.test` command. Here we only need to provide the values x and n for each sample. To group the two values of each together, i.e., both x -values and both sample sizes, we use the `c()` command, which defines a vector (or array) containing the specified values. Further, to match the results of the “by hand” calculation performed earlier using the normal model, we will need to “turn off” the continuity correction in the `prop.test` procedure by including the `correct=F` input. Copy and paste the command below (into the Rstudio console) to obtain the corresponding [formatted] output.

COMMAND

```
prop.test(x=c(104,131), n=c(438,389), correct=F)
```

CONSOLE OUTPUT

```
2-sample test for equality of proportions without continuity correction
data:  c(104, 131) out of c(438, 389)
X-squared = 9.9907, df = 1, p-value = 0.001573
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.16091078 -0.03772522
sample estimates:
   prop 1    prop 2 
0.2374429 0.3367609
```

Notice that the p-values are identical (to the millionth) and the test statistic, X-squared, is $(-3.160806)^2 \approx 9.990695$.

Decision:

Since the p-value (0.00157333) < α (0.05), we reject H_0 .

Conclusion:

At the 5% significance level, we have sufficient (statistically significant) evidence to conclude there is a difference in the proportion that *do not know* enough to have an opinion on drilling for oil and natural gas off the Coast of California between college and non-college graduates.

- (c) Although not explicitly requested in this particular problem, we shall attempt to estimate the difference in the proportion that *do not know* enough to have an opinion on drilling for oil and natural gas off the Coast of California between college and non-college graduates at the 95% confidence level.

If we refer to the output from the `prop.test` command discussed in (b), we see that a 95% confidence interval estimate for $p_{Yes} - p_{No}$ has already been computed; specifically (-0.16091078, -0.03772522) or (-16.1%, -3.8%). For completeness, similar to hypothesis test just performed, the following R commands effectively implement the “by hand” method presented in section 6.2.2 on pages 280 - 281. Copy and paste them into the RStudio console to obtain the resulting output.

COMMANDS

```
# inputs
x1 <- 104; n1 <- 438; x2 <- 131; n2 <- 389; conf.level <- 0.95

# calculations
p1 <- x1/n1; p2 <- x2/n2
p.diff <- p1-p2
se <- sqrt(p1*(1-p1)/n1+p2*(1-p2)/n2)
z <- qnorm(0.5+conf.level/2)

# limits of CI
p.diff + z*se*c(-1,1)
```

CONSOLE OUTPUT

```
> p.diff + z*se*c(-1,1)
[1] -0.16091078 -0.03772522
```

Notice the limits for the two approaches (in RStudio) are identical.

Interpretation:

We are 95% confident the proportion college graduates (among all registered voters in California) that *do not know* enough to have an opinion on drilling for oil and natural gas off the Coast of California is between 3.8% and 16.1% lower than that of non-college graduates.

On Your Own

Exercise 6.36 on page 321; while (b) asks that you “test for independence of incidences of diabetes and employment states”, you are effectively testing if the proportion of employed Americans with diabetes is different from the proportion of unemployed Americans with diabetes. Also, please add a part (d) and compute/interpret a 95% confidence interval estimate of the difference.

2. *Adopted from Exercise 6.56 on page 330.* An experiment conducted by the MythBusters, a science entertainment TV program on the Discovery Channel, tested if a person can be subconsciously influenced into yawning if another person near them yawns. 50 people were randomly assigned to two groups: 34 to a group where a person near them yawned (treatment) and 16 to a group where there wasn't a person yawning near them (control). The following table shows the results of this experiment.

		Group		Total
		Treatment	Control	
Result	Yawn	10	4	14
	Not Yawn	24	12	36
	Total	34	16	50

Let p_{trmt} and p_{ctrl} represent the proportion of all persons that yawn under the conditions for the *treatment* and *control* groups, respectively.

- (a) What are the hypotheses for testing if yawning is contagious, i.e. whether it is **more likely** for someone to yawn if they see someone else yawning?
- (b) Compute the **pooled proportion** estimate, $\hat{p}$, under the assumption the null hypothesis is true and use it to explain why we could not reliably use the normal model for the sampling distribution of the test statistic. *See example 6.13, page 283.*
- (c) In light of the violation of the requisite **success-failure condition** to use the normal model for the sampling distribution of $\hat{p}_{trmt} - \hat{p}_{ctrl}$, we turn to a simulation strategy called the **randomization test** (see pp. 307-311) to estimate the p-value.

While we will make use of RStudio to expedite the process, a less sophisticated approach for simulating the distribution of the test statistic under the assumption of independence, i.e. having someone yawn near another person has no influence on if the other person will yawn, is discussed below:

A researcher writes *yawn* on 14 index cards and *not yawn* on 36 index cards to indicate whether or not a person yawned. He then shuffles the cards and deals them into two groups of size 34 and 16 for *treatment* and *control*, respectively. Finally, he counts the number of index cards in each group with *yawn* written on them and then calculates the difference between the respective proportions: $\hat{p}_{trmt(sim)} - \hat{p}_{ctrl(sim)}$. This process was repeated to obtain 10,000 differences that are due to chance alone.

We now implement the algorithm (procedure) outlined above in RStudio with the following script:

```
N_SIM <- 10000 # number of simulations
x <- c(10, 4) # number of successes in each group
n <- c(34, 16) # size of each group

# initialize vector to store simulated differences
p.diff <- rep(0, N_SIM)

# intermediate calculations
x.total <- sum(x)
n.total <- sum(n)

# generate simulated differences
for (i in 1:N_SIM)
{
  sim.data <- sample(c(rep(1, x.total), rep(0, n.total-x.total)))
  p.diff[i] <- mean(sim.data[1:n[1]]) - mean(sim.data[(n[1]+1):n.total])
}

# construct histogram
hist(
  x = p.diff,
  breaks = seq(-0.6, 0.6, 0.1),
  labels = T,
  right = F,
  ylim = c(0, 3000),
  xlab = "Simulated Difference (Treatment - Control)",
  ylab = "Number of Simulations",
  main = "Histogram of Difference in Sample Proportions (10000 Simulations)",
)

# observed difference
p.diff.obs <- x[1]/n[1]-x[2]/n[2]; p.diff.obs

# p-value (right-tailed)
p.value <- mean(p.diff >= p.diff.obs); p.value
```

Run this script in RStudio [either copy/paste the above code, or load the .R file that is included with this document on Blackboard]. Capture the console output and any plots generated and include them in your submission (for this part) as you will use it to answer the questions that follow in (d) and (e).

- (d) Determine the observed difference between the yawning rates for the two groups and the proportion of the 10,000 simulated differences that are as extreme or more extreme than the observed sample difference. Explain how this proportion can be found using the histogram.

(e) Use the p-value to evaluate the hypotheses. State your **decision** and **conclusion** in the context of the study.