R Assignment 6 - Differences in Proportions

Data Set

To apply our knowledge comparing two proportions, we are going to work with data from a 2004 North Carolina observational study on births in the state. The purpose for collecting this data was for researchers to understand the relationship between a pregnant mothers habits and their children. We are interested in the relationship between a mothers smoking status and whether their baby was born with low birth weight. Let p_1 be the proportion of babies born with low birth weight for mothers who are non-smokers. Let p_2 be the proportion of babies born with low birth weight for mothers who are smokers.

```
# load data from same folder
load("ncbirths.rda")

# create additional categorical variables to code low birth weight and smoking stat
us as 0 and 1 for easier handling
ncbirths$lowbirthweight0 <- factor(ncbirths$lowbirthweight,
                                   levels = c("not low", "low"),
                                   labels = c(0,1))
ncbirths$habit0 <- factor(ncbirths$habit,
                          levels = c("nonsmoker", "smoker"),
                          labels = c(0, 1))
```

Questions

1. Based on the above, what is our parameter of interest? What would be a point estimate of this parameter of interest?

```
p1_hat = length(which(ncbirths$habit0 == 0 & ncbirths$lowbirthweight0 == 1)) / leng
th(which(ncbirths$habit0 == 0))

p2_hat = length(which(ncbirths$habit0 == 1 & ncbirths$lowbirthweight0 == 1)) / leng
th(which(ncbirths$habit0 == 1))

p1_p2_hat = p1_hat - p2_hat
```

The parameter of interest is p_1 - p_2. The point estimate form the given data is calculated in the variable `p1_p2_hat` as $\hat{p_1} - \hat{p_2} = -0.03747$.

2. Using the data, compute the following:

a. The sample proportion of babies born with low birth weight among nonsmoking women ($\hat{p_1}$)
b. The sample proportion of babies born with low birth weight among smoking women ($\hat{p_2}$)
c. The point estimate for $p_1 - p_1$, the difference in population proportions of babies born with low birth weight between smoking and non-smoking women.
d. The z* needed for a 90% confidence interval.

```
# z_90 = qnorm(0.9, p1_p2_hat, se_p1_p2)
z_90 = qnorm(0.95)
```

As seen in question 1, the sample proportion of babies born with low birth weight among nonsmoking women ($\hat{p_1}$) is 0.1054 and the sample proportion of babies born with low birth weight among smoking women ($\hat{p_2}$) is 0.1429. The point estimate for $p_1 - p_1$ is $\hat{p_1} - \hat{p_2} = -0.03747$.

The z* for a 90% confidence interval is 1.28155.

3. Check the assumptions for the sampling distribution of $\hat{p_1} - \hat{p_2}$ to be normal. In other words, check the conditions necessary to construct a confidence interval for $p_1 - p_1$. Recall, these conditions are (1) independence within groups, (2) independence between groups, and (3) success-failure condition in BOTH groups.

```
n1 = length(which(ncbirths$habit0 == 0))  # number of nonsmokers in data, sample si
ze 1
n2 = length(which(ncbirths$habit0 == 1))  # number of smokers in data, sample size
2

# check success-failure condition:
n1*p1_hat >= 10 & n1*(1-p1_hat) >= 10
```

```
## [1] TRUE
```

```
n2*p2_hat >= 10 & n2*(1-p2_hat) >= 10
```

```
## [1] TRUE
```

We assume that the data is from a random sample and that the subjects within and between each group are independent. As seen above, the success-failure condition is given. Therefor, we can assume that the sampling distribution for $\hat{p_1} - \hat{p_2}$ is normal.

4. Calculate the standard error for the sampling distribution of $\hat{p_1} - \hat{p_2}$. Then, compute the 90% confidence interval for $p_1 - p_2$.

```
se_p1_p2 = sqrt(((p1_hat*(1-p1_hat))/n1) + ((p2_hat*(1-p2_hat))/n2))

lowerBd = p1_p2_hat - z_90 * se_p1_p2
upperBd = p1_p2_hat + z_90 * se_p1_p2
```

The 90% confidence interval for the difference in proportion of babies born with low birth weight between non-smoking mothers and smoking mothers is (-0.079586, 0.004639).

5. Interpret the confidence interval you computed in Question 5 given the context of the data.

I am 90% confident that the difference of babies born with low birth weight among nonsmoking women and babies born with low birth weight among smoking women is between -0.079586 and 0.004639.

6. State the null and alternative hypotheses, if we are interested in comparing the proportion of babies born with low birth weight between non-smoking and smoking mothers.

a. State the hypotheses in words and with statistical notation.
b. Why is the null rather than the alternative hypothesis a statement of equality?

a. The null hypothesis is that there is no difference in the proportion of babies born with low birth weight among nonsmoking women (p_1) and babies born with low birth weight among smoking women (p_2), in other words, p_1 and p_2 are equal. The alternative hypothesis is, that there is a difference between p_1 and p_2, that they are not equal.

$H_0 : p_1 - p_2 = 0$ or $p_1 = p_2$

$H_a : p_1 - p_2 \neq 0$ or $p_1 \neq p_2$

b. The null hypothesis is always a very specific statement that we want to disprove. A statement of equality is easily falsifiable, which is what we are trying to do here.

7. Compute the pooled proportion of babies born with low birth weight between non-smoking and smoking mothers. Explain why we use a pooled proportion.

We assume that our null hypothesis is correct, so we assume that there is no difference between both proportions p_1 and p_2. Since we assume no difference between both proportions, we use a common *pooled* proportion.

```
p_pooled = (length(which(ncbirths$habit0 == 0 & ncbirths$lowbirthweight0 == 1)) + l
ength(which(ncbirths$habit0 == 1 & ncbirths$lowbirthweight0 == 1))) / (n1 + n2)
```

8. Using the pooled proportion computed in Question 7, check the conditions necessary to use the normal distribution to perform a hypothesis test. Show all your work.

The condition of independence stands as stated above. The success-failure condition needs to be checked for the pooled proportion.

```
n1*p_pooled >= 10 & n1*(1 - p_pooled) >= 10
```

```
## [1] TRUE
```

```
n2*p_pooled >= 10 & n2*(1 - p_pooled) >= 10
```

```
## [1] TRUE
```

The success-failure condition is also met, so a normal distribution van be used to perform a hypothesis test.

9.

a. Compute the standard error using the pooled proportion computed in Question 7.
b. Calculate your Z-statistic/test statistic.
c. Compute the associated p-value.
d. Report your conclusion from the hypothesis test based on the given significance level above and include the confidence interval and p-value. State your conclusion in the context of the data.
e. Define what the p-value means in context.

```
se_pooled = sqrt(((p_pooled * (1 - p_pooled))/ n1) + ((p_pooled * (1 - p_pooled))/
n2))
z_statistic = (p1_p2_hat - 0) / se_pooled
p_value = 2*pnorm(z_statistic, mean = 0, sd = 1)
```

The standard error from the pooled proportions is 0.0298, the Z-statistic is -1.256 qith an associated p-value of 0.209.

With a significance level of 0.1, we fail to reject our null hypothesis and conclude that there is no evidence in our data that there is a difference between the proportion of babies born with low birth weight to non-smoking and smoking mothers (p-value = 0.209, 90%-CI = (-0.079586, 0.004639)).

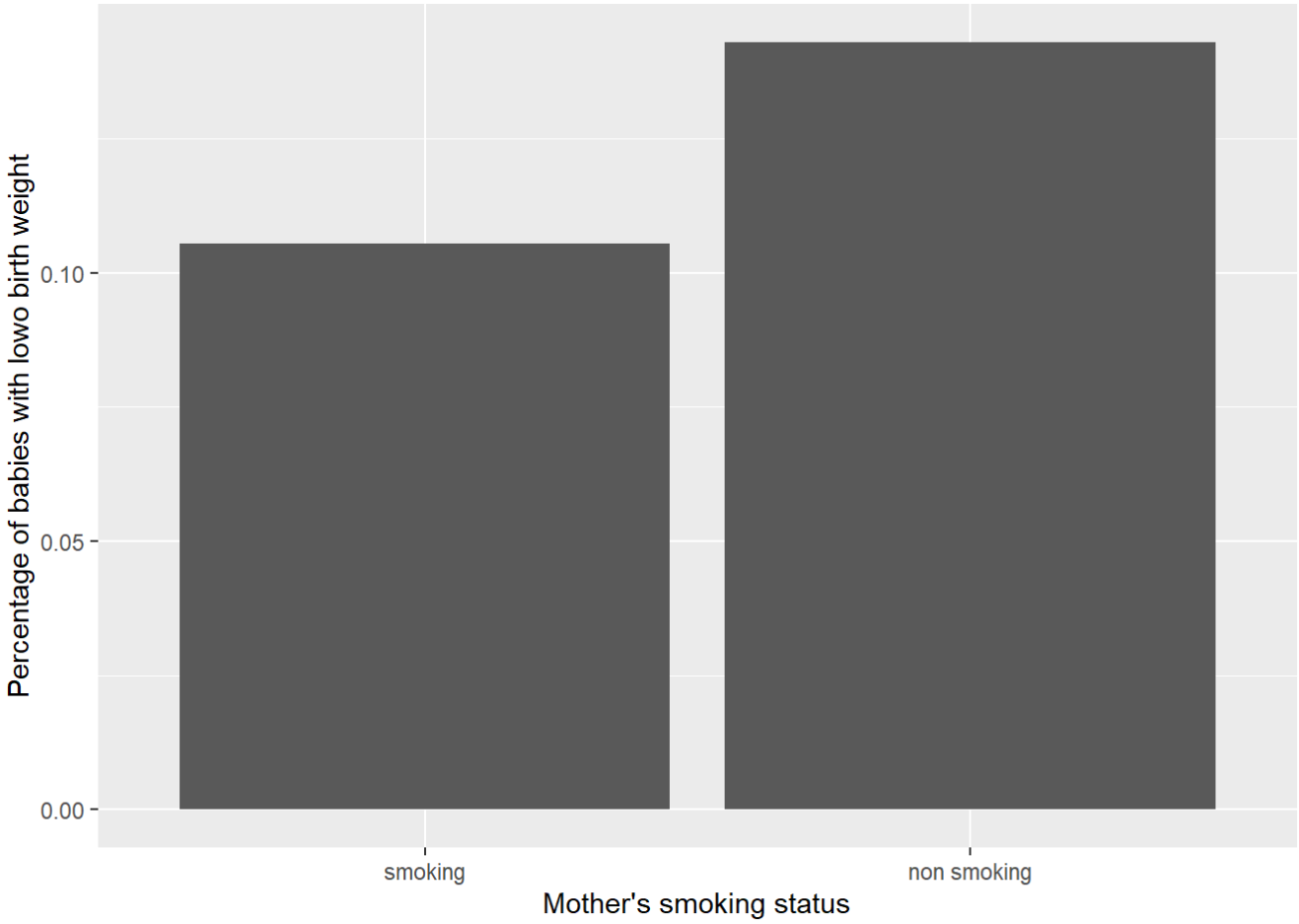
The p-value in this context means, that if we live in a world, where the null hypothesis is true, it would not be rare (20.91%) to get the results from this data or more extreme, i.e. if there is no difference between the proportion of babies born with low birth weight to non-smoking and smoking mothers, 20.91% of all samples

would be at least as extreme as the observed data in this study.

10. Provide an appropriate visualization for your data. (Look at the Week 2 slides). EXTRA CREDIT (2 points): Use the ggplot2() or plot_ly R packages to create visualizations. You will need to look up how to do this (you may refer to the R demo posted in the Week 3 module).

```
library(ggplot2)

df = data.frame(
  mothers = factor(c("smoking", "non smoking"), levels=c("smoking", "non smokin
g")),
  low_birth_weight = c(p1_hat, p2_hat)
)
ggplot(data = df, aes(x=mothers, y=low_birth_weight)) +
  geom_bar(stat = "identity") +
  xlab("Mother's smoking status") +
  ylab("Percentage of babies with lowo birth weight")
```



11. Exercise 6.19 in the OpenIntro 4rth edition textbook (page 225): A study asked 1,924 male and 3,666 female undergraduate college students their favorite color. A 95% confidence interval for the difference between the proportions of males and females whose favorite color is black ($p_{male} - p_{female}$) was calculated to be (0.02, 0.06). Based on this information, determine if the following statements are true or false, and explain your reasoning for each statement you identify as false.

    a. We are 95% confident that the true proportion of males whose favorite color is black is 2% lower to 6% higher than the true proportion of females whose favorite color is black.
    b. We are 95% confident that the true proportion of males whose favorite color is black is 2% to 6% higher than the true proportion of females whose favorite color is black.
    c. 95% of random samples will produce 95% confidence intervals that include the true difference between the population proportions of males and females whose favorite color is black.
    d. We can conclude that there is a significant difference between the proportions of males and females whose favorite color is black and that the difference between the two sample proportions is too large to plausibly be due to chance.
    e. The 95% confidence interval for $p_{female} - p_{male}$ cannot be calculated with only the information given in this exercise.

a. False, the range for the 95% confidence interval only includes positive values, so the true proportion of males whose favorite color is black cannot be higher or lower, than the the true proportion of females whose favorite color is black.

b. true

c. true

d. True. If the decision of whether or not the difference is too large to be due to chance is made based on the confidence interval alone, it is possible to conclude, that there is a significant difference. The conclusion should be stated with a given α and p-value.

e. False. $p_{female} - p_{male}$ is the negated value of $p_{male} - p_{female}$ and can be calculated by negating the given confidence interval.