Baby Birth Weight Analysis

JoJo Summersett, Klaire Pham, Jason Phillip

Introduction:

We are interested in the relationship between certain factors (including smoking habits, mother's weight gain, etc) have on a baby's birth weight. Generally, mothers are advised not to smoke tabacco during their pregnancy, so we want to find out whether or not that advise is statistically sound. Different studies have also shown the advantages of abstinence from cigarette smoking during pregnancy, thus, we also want to reproduce evidence to those claims.

Is there a difference in average weights of babies birthed by smoking mothers and non-smoking mothers? Is there a difference in proportion of low-weight babies birthed by smoking mothers and non-smoking mothers? Does a mother's smoking habits have an effect on the baby's birth weight?

As we explore the answers to these questions, we decided to study the variability of the difference of the proportions of low weight babies with smoking moms and low weight babies for non-smoking moms through both simulation- and theory-based inference methods.

Data:

We sourced our dataset from the R OpenIntro package, BIRTHS14. The US Government runs a continous observational study recording information about the weight, demographic, sex, and other attributes of US babies and births. It is observational because there is no random assignment or treatment groups. Because of this, findings from this dataset can be generalized to the general population because it uses random sampling methods; however, it is not an experiment so the data cannot be used to establish a causal link.

The US released a large dataset containing information on births recorded within the country in 2014, and the dataset cases represent 1,000 birth records, randomly sampled from the US Department of Health. We are interested in the birth weight classification (low or not low) of babies (nominal, categorical) and if other variables, like smoking habits (nominal, categorical) and mother's age (discrete, numerical), have a positive or negative effect on it. We will at times also consider other variables such as the weight gained by mother (continuous, numerical).

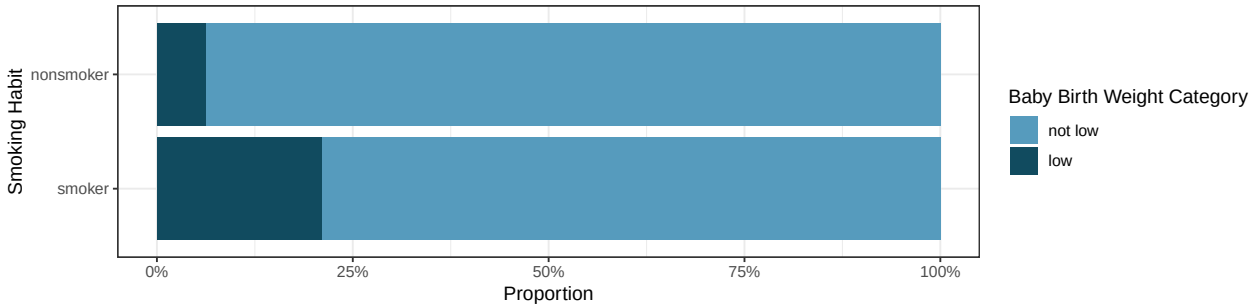
Citation:

United States Department of Health and Human Services. Centers for Disease Control and Prevention. National Center for Health Statistics. Natality Detail File, 2014 United States. Inter-university Consortium for Political and Social Research, 2016-10-07. doi: 10.3886/ICPSR36461.v1 .

Exploratory Data Analysis:

```
# summary stats for numerical baby birth weight
births_cleaned %>%
  group_by(habit) %>%
  summarise(iqr    = IQR(weight),
            median = median(weight),
            n      = n())
```

```
## # A tibble: 2 x 4
##   habit       iqr median     n
##   <chr>     <dbl>  <dbl> <int>
## 1 nonsmoker  1.49   7.35   867
## 2 smoker     1.85   7.03   114
```



Initial Observations:

The summary statistics verify that the median baby weight in the nonsmoker category is 0.29 lbs higher than than the smoker category and the IQR is 0.355 lbs larger in the smoker category than the nonsmoker category. This may suggest that the two variables are not independent, and there may exist a relationship between a mother's smoking habit and their baby's birth weight.

The bar chart above shows the proportion of smoking vs. nonsmoking mothers whose baby was categorized as low weight. We can see a larger proportion of the smoking mothers had babies with low birth weights. This supports the summary statistics that the median and IQR for baby's weight for nonsmoker mothers is slightly higher than those of smoker mothers, and suggests that the two categorical variables are not independent.

Inference:

```
## # A tibble: 4 x 3
## # Groups:   habit [2]
##   habit     lowbirthweight     n
##   <chr>     <chr>          <int>
## 1 nonsmoker low               54
## 2 nonsmoker not low          813
## 3 smoker    low               24
## 4 smoker    not low           90
```

From our sample, we see that the proportion of low weight babies in each habit type are nonsmokers $\hat{P}_{nonsmokers} = 54/867 = 0.06228374$ and smokers $\hat{P}_{smokers} = 24/114 = 0.2105263$. The difference of these two proportions, nonsmoker-smoker = $\hat{P}_{nonsmokers} - \hat{P}_{smokers} = -0.1482426$

To test the difference in proportions, we define the following hypotheses:

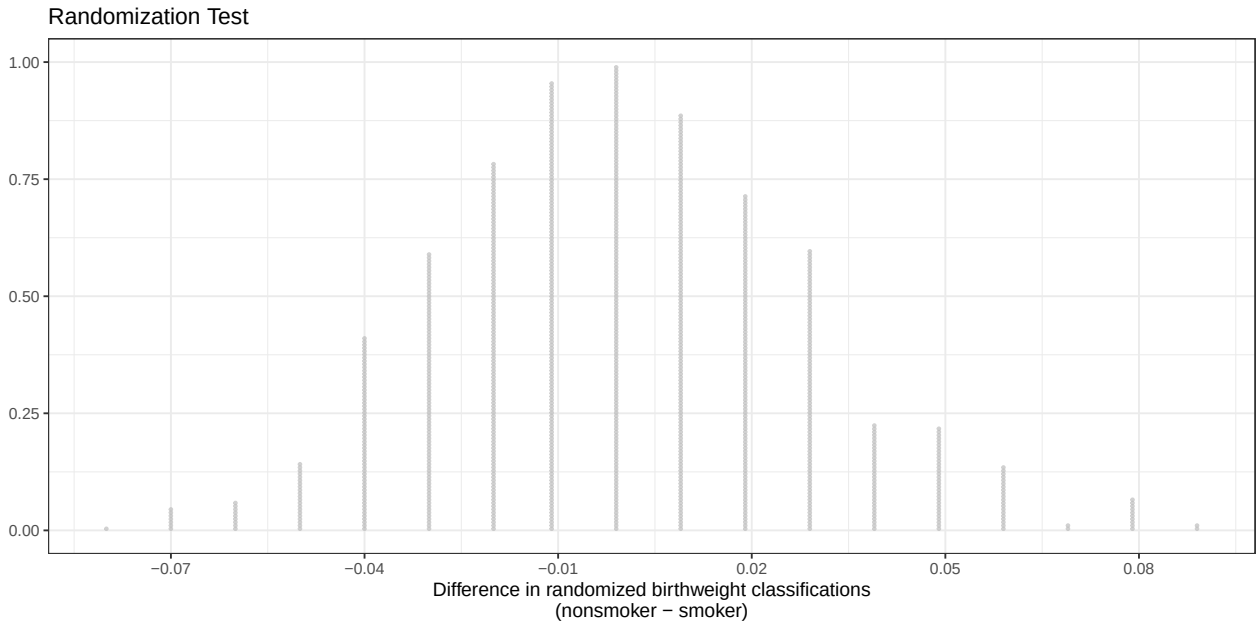
Null Hypothesis = $\hat{P}_{nonsmokers} - \hat{P}_{smokers} = 0$, Alternative Hypothesis = $\hat{P}_{nonsmokers} - \hat{P}_{smokers} \neq 0$, and $\alpha = 0.05$.

Checking the conditions for Central Limit Theorem:
- Independence: The samples are independent as they were randomly sampled from all the US births recorded in 2014.
- Success-failure: The dataset satisfies the success-failure condition as both categories have at least 10 successes (non low birth weight) and 10 failures (low birth weight).

Randomization/Simulation-Based Hypothesis Testing

```
# randomization test
set.seed(132)
```

```
births_cleaned %>%
  specify(response = lowbirthweight, explanatory = habit, success = "not low") %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in props", order = c("nonsmoker", "smoker")) %>%
mutate(stat = round(stat, 3)) %>%
  ggplot(aes(x = stat)) +
  geom_dotplot(binwidth = 0.001, dotsize = .5) +
  labs(
    title = "Randomization Test",
    y = NULL,
    x = "Difference in randomized birthweight classifications\n(nonsmoker - smoker)"
  ) +
theme(axis.title.y = element_blank(),
        axis.text.y  = element_blank(),
        axis.ticks.y = element_blank()) +
  scale_x_continuous(breaks=seq(-0.1,0.1,0.03)) +
  gghighlight(stat >= 0.1482426 |
                stat <= -0.1482426) +
  theme_bw()
```

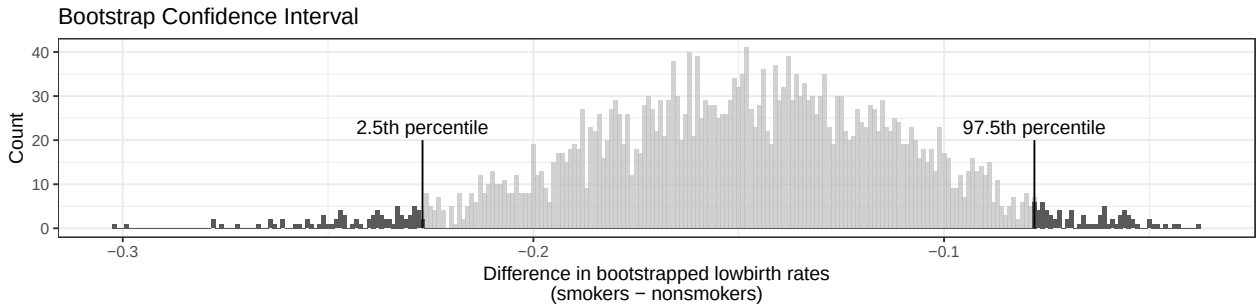


The random distribution generated above does not include the observed difference in proportion, -0.14822426, so the p-value = 0. Because p (0) < alpha (0.05), we have evidence to reject the null hypothesis. In other words, the observed difference in proportions between the smoking habit of the mothers and whether they birth low-weight babies is not likely due to random chance. It supplies evidence that there may be a relationship between the smoking habit of the mothers and whether they birth low-weight babies.

Bootstrap Confidence Interval

```
# bootstrap sample
set.seed(162)

births_boot_dist <- births_cleaned %>%
  specify(response = lowbirthweight, explanatory = habit, success = "not low") %>%
```

```
  generate(reps = 3000, type = "bootstrap") %>%
  calculate(stat = "diff in props", order = c("smoker", "nonsmoker"))

births_boot_95 <- births_boot_dist %>%
  summarise(
    lower = round(quantile(stat, 0.025), 3),
    upper = round(quantile(stat, 0.975), 3)
  )

births_boot_95_lower_perc <- births_boot_95 %>% pull(lower)
births_boot_95_upper_perc <- births_boot_95 %>% pull(upper)

ggplot(births_boot_dist, aes(x = stat)) +
  geom_histogram(binwidth = 0.001) +
  labs(
    title = "Bootstrap Confidence Interval",
    x = "Difference in bootstrapped lowbirth rates\n(smokers - nonsmokers)",
    y = "Count"
  ) +
  gghighlight(stat <= births_boot_95_lower_perc | stat >= births_boot_95_upper_perc) +
  annotate(
    "segment",
    x = c(births_boot_95_lower_perc, births_boot_95_upper_perc),
    xend = c(births_boot_95_lower_perc, births_boot_95_upper_perc),
    y = 0, yend = 20
  ) +
  annotate("text", x = births_boot_95_lower_perc, y = 23, label = "2.5th percentile") +
  annotate("text", x = births_boot_95_upper_perc, y = 23, label = "97.5th percentile") +
  theme_bw()
```




```
births_boot_95_lower_perc
```

```
##   2.5%
## -0.227
```

```
births_boot_95_upper_perc
```

```
##  97.5%
## -0.078
```

Confidence Interval = [-0.225, -0.076]. Our Null Hypothesis = $\hat{P}_{nonsmokers} - \hat{P}_{smokers} = 0$ is not within the interval, so we can say the null hypothesis claim is NOT supported by the confidence interval. The Alternative Hypothesis on the other hand, $\hat{P}_{nonsmokers} - \hat{P}_{smokers} \neq 0$, and $\hat{P}_{nonsmokers} - \hat{P}_{smokers} = -0.1482426$ is supported by the confidence interval. This means we are 95% confident that the true difference between the proportions of low-weight babies birthed by nonsmoking and smoking mothers is between -0.225 and -0.076.

Mathematical Model: Theory-Based Confidence Interval

```
p_1_hat <- 54 / (54 + 813)
p_2_hat <- 24 / (24 + 90)
point_est <- p_1_hat - p_2_hat
se <- sqrt((p_1_hat * (1 - p_1_hat))/(54 + 813)
           + (p_2_hat* (1 - p_2_hat))/(24 + 90))
births_math_95_lower_perc <- point_est - qnorm(0.95) * se
births_math_95_upper_perc <- point_est + qnorm(0.95) * se
births_math_95_lower_perc
```

```
## [1] -0.2124825
```

```
births_math_95_upper_perc
```

```
## [1] -0.08400262
```

Through theory-based confidence interval, we are 95% confident that the true difference between the proportions of low-weight babies birthed by nonsmoking and smoking mothers is between -0.212 and -0.084. 0 is not included in that interval, so again, we have evidence to reject the Null Hypothesis.

Conclusion:

In conclusion, through the simulation-based randomization test and the simulation-based as well as theory-based confidence intervals, we have rejected the Null Hypothesis, which claims that the there is no significant difference in the proportion of low eight babies born to smoking vs nonsmoking mothers, each time. Consequently, we find that the smoking habit of the mother has an effect on the classification of her baby's weight, and, more specifically, smoking mothers give birth to a higher proportion of low weight babies. However, the data also suggests that the potential effect isn't very large. Since this is an observational study, there is no causal link established – there might be other reasons acting as confounding variables that reduce the baby's weight.

Looking forward, we would be interested in using linear regression between two variables to predict the weight of a newborn baby based on variables like smoking habit and weight gained during a mother's pregnancy.