Data 606 Lab 7

Joao De Oliveira

2025-10-26

Overview

In this lab, we will explore and visualize the data using the tidyverse suite of packages, and perform statistical inference using infer. The data can be found in the companion package for OpenIntro resources, openintro.

Load Packages

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(openintro)
```

```
## Loading required package: airports
## Loading required package: cherryblossom
## Loading required package: usdata
```

```
library(infer)
library(ggplot2)
```

Exercise 1

What are the cases in this data set? How many cases are there in our sample?

```
data('yrbss', package='openintro')

?yrbss

number_cases <- nrow(yrbss)
number_cases
```

```
## [1] 13583
```

Answer 1

Each case in this dataset corresponds to one high school student (9th to 12th grade) response to the Youth Risk Behavior Surveillance System. There are 13583 in this dataset.

Exercise 2

How many observations are we missing weights from?

```
summary(yrbss$weight)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   29.94   56.25   64.41   67.91   76.20  180.99    1004
```

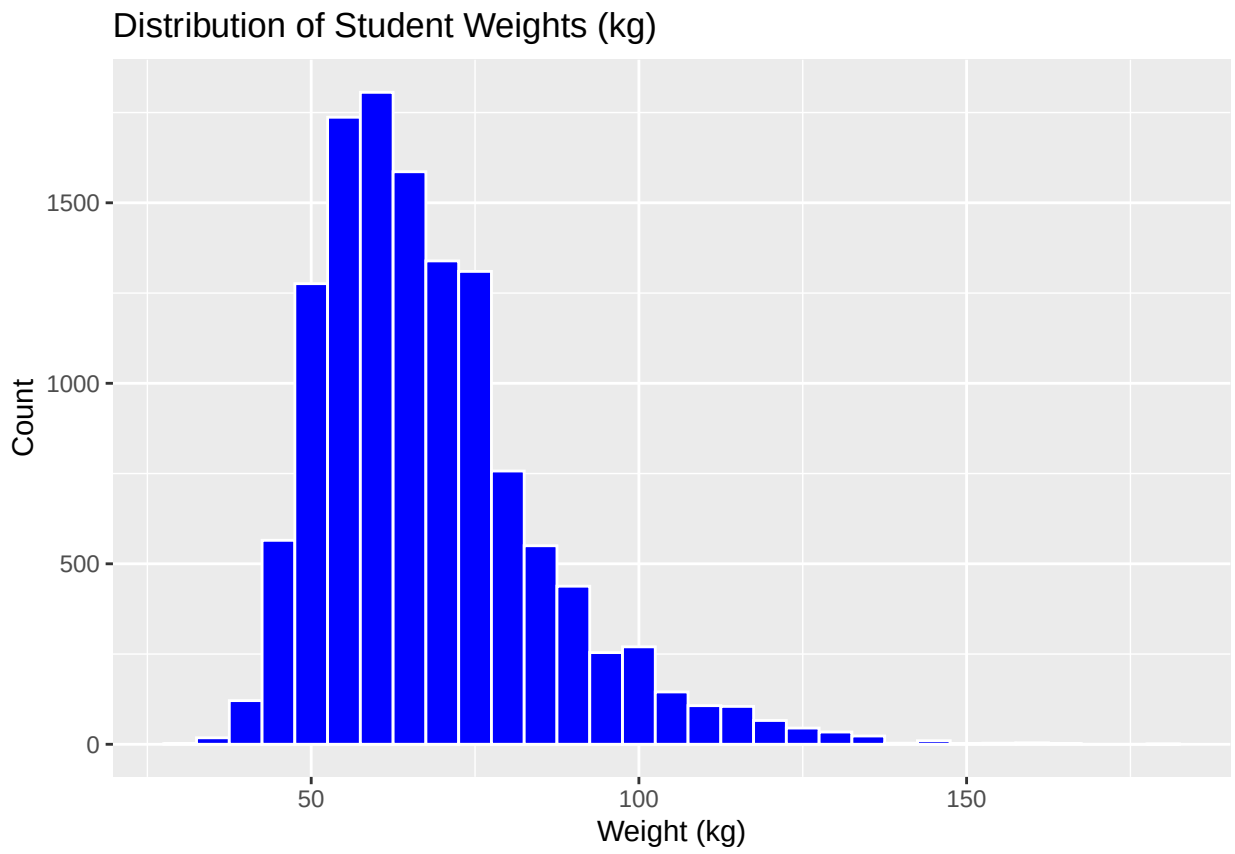
```
# Calculate additional summary statistics
weight_stats <- yrbss %>%
  filter(!is.na(weight)) %>%
  summarise(
    mean_weight = mean(weight),
    median_weight = median(weight),
    sd_weight = sd(weight),
    min_weight = min(weight),
    max_weight = max(weight),
    n = n()
  )

weight_stats
```

```
## # A tibble: 1 x 6
##   mean_weight median_weight sd_weight min_weight max_weight     n
##         <dbl>         <dbl>     <dbl>      <dbl>      <dbl> <int>
## 1        67.9          64.4      16.9       29.9       181. 12579
```

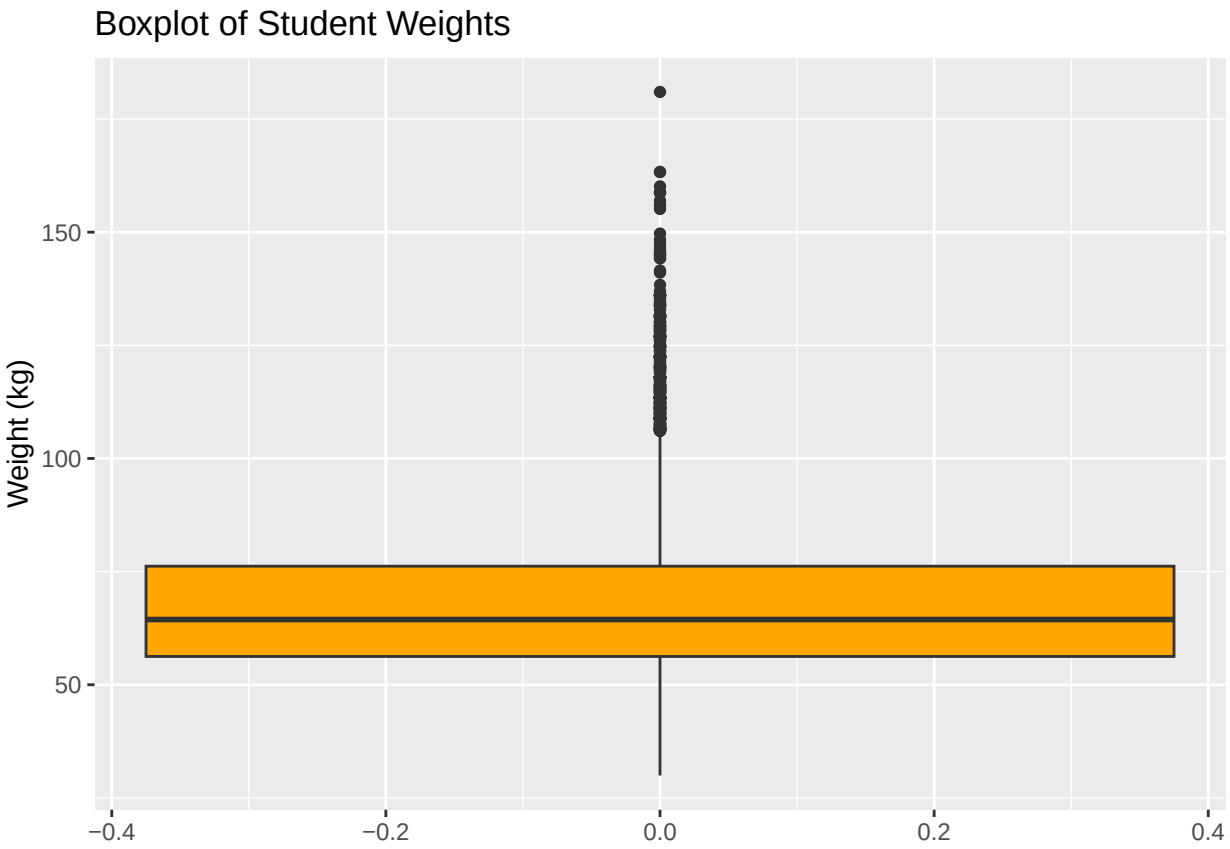
```
sum(is.na(yrbss$weight))
```

```
## [1] 1004
```

```
# Histogram of weight dist
ggplot(yrbss, aes(x = weight)) +
  geom_histogram(binwidth = 5, fill = "blue", color = "white", na.rm = TRUE) +
  labs(title = "Distribution of Student Weights (kg)",
       x = "Weight (kg)", y = "Count")
```




```r
# Boxplot of weight dist
ggplot(yrbss, aes(y = weight)) +
  geom_boxplot(fill = "orange", na.rm = TRUE) +
  labs(title = "Boxplot of Student Weights", y = "Weight (kg)")
```



Answer 2

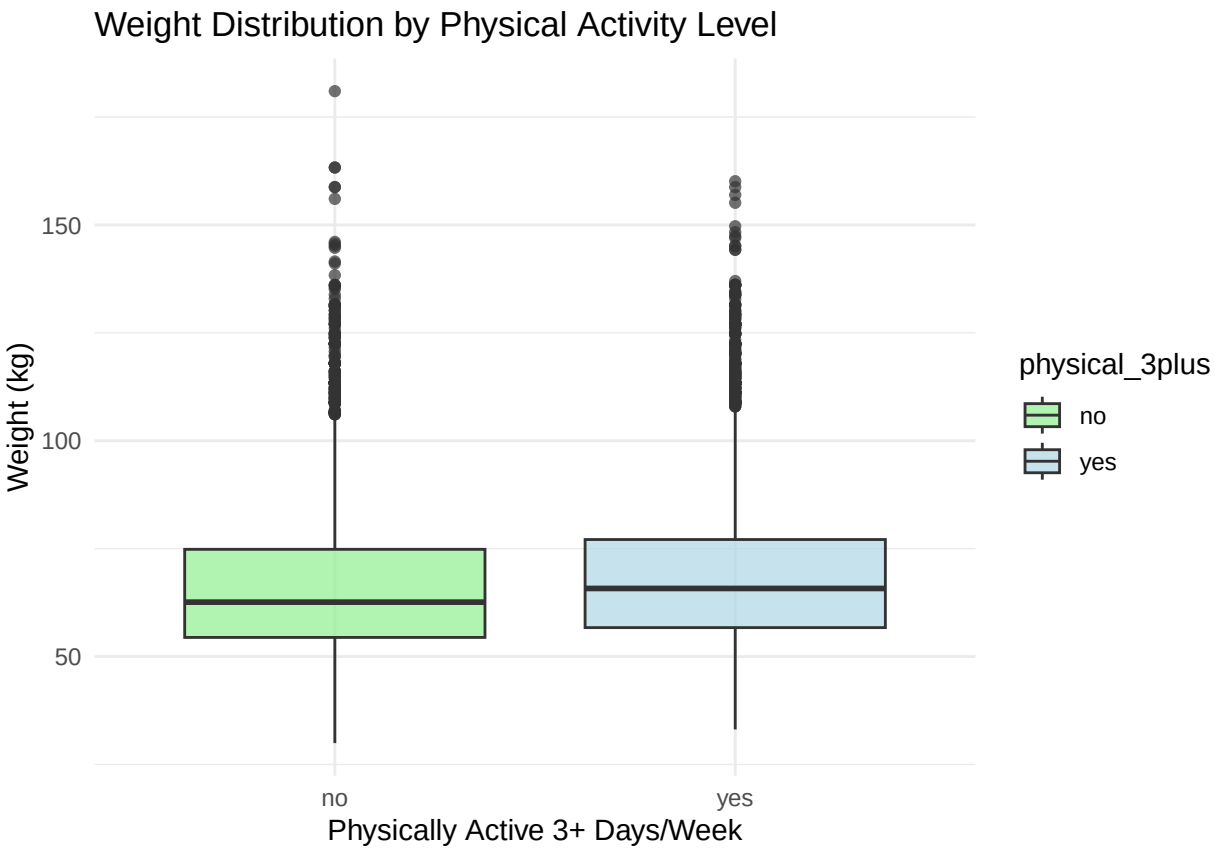
As we can see both in the summary (# of NA's) and in the output of the sum(is.na(yrbss$weight)), there are 1004 observations missing the weight. We can see that the distribution of weights has a mean of 67.9 Kg, a median of 64.4 Kg,a standard deviation of 16.9 Kg, a minimum of 29.9 Kg, and a maximum of 181.1 Kg. By analyzing the histogram, we can see that the distribution of weights is right skewed (confirmed by the fact that the mean is higher than the median).

Exercise 3

Make a side-by-side boxplot of physical_3plus and weight. Is there a relationship between these two variables? What did you expect and why?

```
yrbss <- yrbss %>%
  mutate(physical_3plus = ifelse(yrbss$physically_active_7d > 2, "yes", "no"))


yrbss %>%
  filter(!is.na(weight) & !is.na(physical_3plus)) %>%
  ggplot(aes(x = physical_3plus, y = weight, fill = physical_3plus)) +
  geom_boxplot(alpha = 0.7) +
  labs(title = "Weight Distribution by Physical Activity Level",
       x = "Physically Active 3+ Days/Week",
       y = "Weight (kg)") +
```

```
  theme_minimal() +
  scale_fill_manual(values = c("no" = "lightgreen", "yes" = "lightblue"))
```



Answer 3

The box plots show how the medians of the two distributions compare, with a slightly higher median of weight for students who are physically active 3 or more days per week than the students who aren't. However, the 3 students with more weight are all students who don't workout. While this was surprising to me since I expected students who are physically activate to be lighter, that might not be the case due to the weight of muscle mass. So, apparently there is a small relationship between physical activity and weight.

Exercise 4

Are all conditions necessary for inference satisfied? Comment on each. You can compute the group sizes with the summarize command above by defining a new variable with the definition n().

```
yrbss %>%
  group_by(physical_3plus) %>%
  summarise(mean_weight = mean(weight, na.rm = TRUE))
```

```
## # A tibble: 3 x 2
##   physical_3plus mean_weight
##   <chr>                <dbl>
```

```
## 1 no                    66.7
## 2 yes                   68.4
## 3 <NA>                  69.9
```

```
yrbss %>%
  filter(!is.na(physical_3plus) & !is.na(weight)) %>%
  group_by(physical_3plus) %>%
  summarise(n = n(),
            mean_weight = mean(weight, na.rm = TRUE),
            sd_weight = sd(weight, na.rm = TRUE))
```

```
## # A tibble: 2 x 4
##   physical_3plus     n mean_weight sd_weight
##   <chr>          <int>       <dbl>     <dbl>
## 1 no              4022        66.7      17.6
## 2 yes             8342        68.4      16.5
```

Answer 4

In order for inference to be satisfied we need to check the sample size (is it above 30?), the independence of the observations in the sample, and if the distribution of the mean is normal. The observations in the sample are independent since they represent each high school student from a random sample (probably from different high schools due to the sample size and randomness), both groups (yes and no) have large sample sizes (way above 30) satisfying the Central Limit theorem, which allows he to say that with samples so large (4022 and 8342), the sampling distribution of the mean is roughly normal.

Exercise 5

Write the hypotheses for testing if the average weights are different for those who exercise at least times a week and those who don't.

Answer 5

Since the goal is to determine whether the average weight changes between high schoolers who are physically active 3 or more days per week and those who are not physically activate. If u(yes) (mean of high schoolers who answered yes to regular physical activity) and u(no) the mean of those you answered no to the same question, the null hypothesis H0 states that there is no difference in the average weights of the two groups (u(yes) = u(no)), while the alternativate hyphotesis (H1) states that there is a difference in the average weights (u yes u no). This is a two-tailed test because we are examining whether the means differ in either direction, not specifically whether one group weighs more or less than the other.

Exercise 6

How many of these null permutations have a difference of at least obs_stat?

```
obs_diff <- yrbss %>%
  drop_na(physical_3plus) %>%
  specify(weight ~ physical_3plus) %>%
  calculate(stat = "diff in means", order = c("yes", "no"))
```

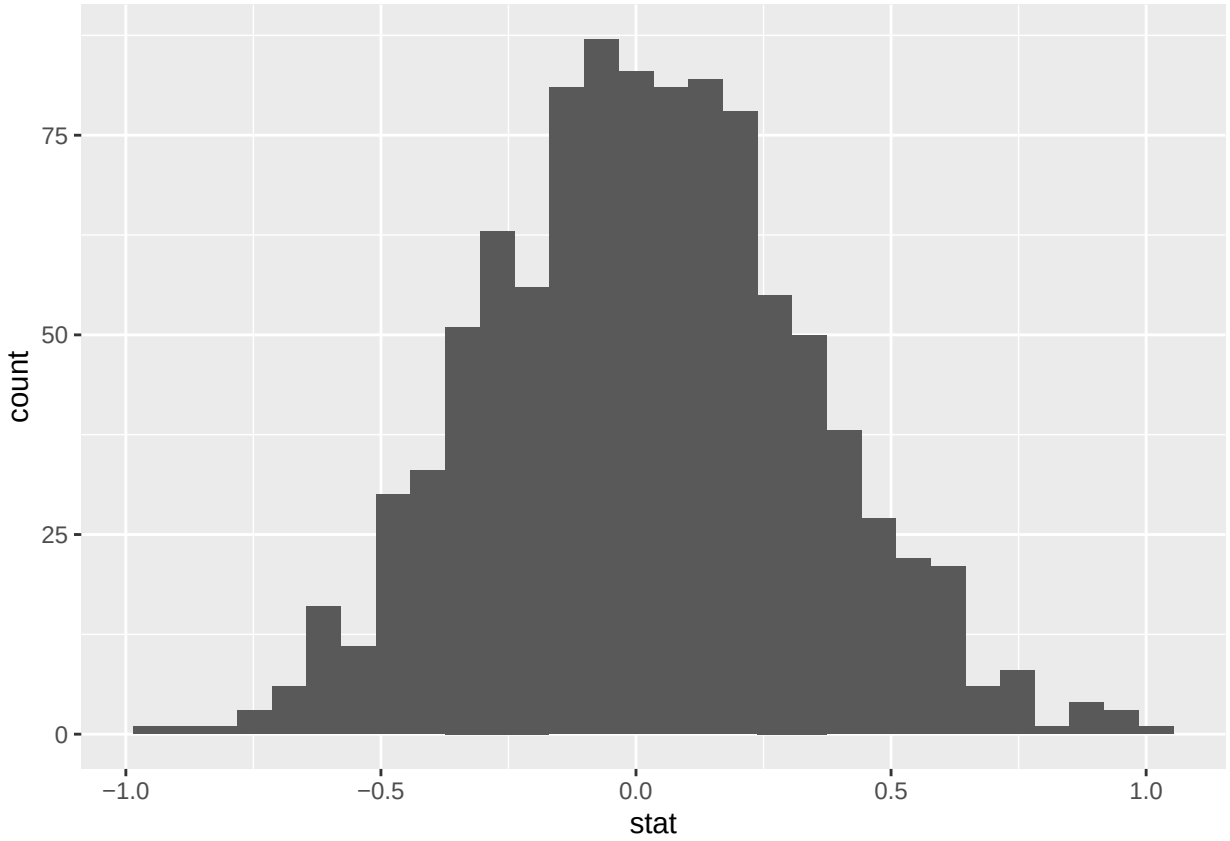
```
## Warning: Removed 946 rows containing missing values.
```

```
null_dist <- yrbss %>%
  drop_na(physical_3plus) %>%
  specify(weight ~ physical_3plus) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("yes", "no"))
```

```
## Warning: Removed 946 rows containing missing values.
```

```
ggplot(data = null_dist, aes(x = stat)) +
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
count_p <- sum(abs(null_dist$stat) >= abs(obs_diff$stat))
count_p
```

```
## [1] 0
```

Answer 6

Out of the 1000 null permutations 0 have a difference of at least the observed statistic.

Exercise 7

Construct and record a confidence interval for the difference between the weights of those who exercise at least three times a week and those who don't, and interpret this interval in context of the data.

```
null_dist %>%
  get_p_value(obs_stat = obs_diff, direction = "two_sided")
```

```
## Warning: Please be cautious in reporting a p-value of 0. This result is an approximation
## based on the number of `reps` chosen in the `generate()` step.
## i See `get_p_value()` (`?infer::get_p_value()`) for more information.
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1       0
```

```
# ci
ci_diff <- yrbss %>%
  drop_na(physical_3plus) %>%
  specify(weight ~ physical_3plus) %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "diff in means", order = c("yes", "no")) %>%
  get_ci(level = 0.95, type="percentile")
```

```
## Warning: Removed 946 rows containing missing values.
```

```
ci_diff
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1     1.11     2.35
```

Answer 7

The 95% CI for the difference in mean weights (physically active for at least 3 days - not physically active) is approximately [1.10, 2.42] kg. This means we are 95% confident that the true difference in mean weights between the two groups (physically active vs non active) is between 1.10 and 2.42 kg, with physically active students weighing more on average.

Exercise 8

Calculate a 95% confidence interval for the average height in meters (height) and interpret it in context.

```
ci_height_m <- yrbss %>%
  filter(!is.na(height)) %>%
  specify(response = height) %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "mean") %>%
```

```
  get_ci(level = 0.95, type = "percentile")

options(pillar.sigfig = 8)
print(ci_height_m, digits = 6)
```

```
## # A tibble: 1 x 2
##    lower_ci  upper_ci
##       <dbl>     <dbl>
## 1 1.6894250 1.6929626
```

```
# mean
mean_height <- mean(yrbss$height, na.rm = TRUE)
mean_height
```

```
## [1] 1.691241
```

Answer 8

The 95% confidence interval for the average height is approximately [1.689, 1.693] meters. This means we are 95% confident that the true average height of high school students in the population is between 1.689 and 1.693 meters, which includes the true mean value (1.691)

Exercise 9

Calculate a new confidence interval for the same parameter at the 90% confidence level. Comment on the width of this interval versus the one obtained in the previous exercise.

```
ci_height_m_90 <- yrbss %>%
  filter(!is.na(height)) %>%
  specify(response = height) %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "mean") %>%
  get_ci(level = 0.90, type = "percentile")

options(pillar.sigfig = 8)
print(ci_height_m_90, digits = 6)
```

```
## # A tibble: 1 x 2
##    lower_ci  upper_ci
##       <dbl>     <dbl>
## 1 1.6896378 1.6928058
```

```
# Calculate widths
width_95 <- ci_height_m$upper_ci - ci_height_m$lower_ci
width_90 <- ci_height_m_90$upper_ci - ci_height_m_90$lower_ci

cat("95% CI width:", width_95, "\n")
```

```
## 95% CI width: 0.003537662
```

```r
cat("90% CI width:", width_90, "\n")
```

```
## 90% CI width: 0.003167939
```

```r
interval_diff <- width_95 - width_90
interval_diff
```

```
## [1] 0.0003697233
```

Answer 9

The 90% confidence interval is [1.690, 1.693] meters, which is narrower than the 95% confidence interval, which ends up making sense because as we decrease the confidence level, we require less certainty, so the interval becomes narrower. The 90% CI is approximately 3.6972335×10^{-4} meters narrower than the 95% CI.

Exercise 10

Conduct a hypothesis test evaluating whether the average height is different for those who exercise at least three times a week and those who don't.

```r
# observed difference in heights
obs_diff_height <- yrbss %>%
  drop_na(physical_3plus) %>%
  specify(height ~ physical_3plus) %>%
  calculate(stat = "diff in means", order = c("yes", "no"))
```

```
## Warning: Removed 946 rows containing missing values.
```

```r
obs_diff_height
```

```
## Response: height (numeric)
## Explanatory: physical_3plus (factor)
## # A tibble: 1 x 1
##          stat
##         <dbl>
## 1 0.037625886
```

```r
# null distribution
null_dist_height <- yrbss %>%
  drop_na(physical_3plus) %>%
  specify(height ~ physical_3plus) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 5000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("yes", "no"))
```

```
## Warning: Removed 946 rows containing missing values.
```

```r
# p-value
p_value_height <- null_dist_height %>%
  get_p_value(obs_stat = obs_diff_height, direction = "two_sided")
```

```
## Warning: Please be cautious in reporting a p-value of 0. This result is an approximation
## based on the number of `reps` chosen in the `generate()` step.
## i See `get_p_value()` (`?infer::get_p_value()`) for more information.
```

```r
p_value_height
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1       0
```

```r
# Count extreme values for manual calculation
extreme_count_height <- sum(abs(null_dist_height$stat) >= abs(obs_diff_height$stat))
cat("Number of permuted statistics as extreme as observed:", extreme_count_height, "\n")
```

```
## Number of permuted statistics as extreme as observed: 0
```

```r
# If p-value is 0, provide more accurate interpretation
if (p_value_height$p_value == 0) {
  manual_p_height <- (extreme_count_height + 1) / (1000 + 1)
  cat("Since the reported p-value is 0,\n")
  cat("a more accurate estimate: p < 0.001 or p  ", round(manual_p_height, 4), "\n")
  cat("This is a VERY significant result!\n")
}
```

```
## Since the reported p-value is 0,
## a more accurate estimate: p < 0.001 or p   0.001
## This is a VERY significant result!
```

```r
# Calculate confidence interval for better context
ci_height_diff <- yrbss %>%
  drop_na(physical_3plus) %>%
  specify(height ~ physical_3plus) %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "diff in means", order = c("yes", "no")) %>%
  get_ci(level = 0.95)
```

```
## Warning: Removed 946 rows containing missing values.
```

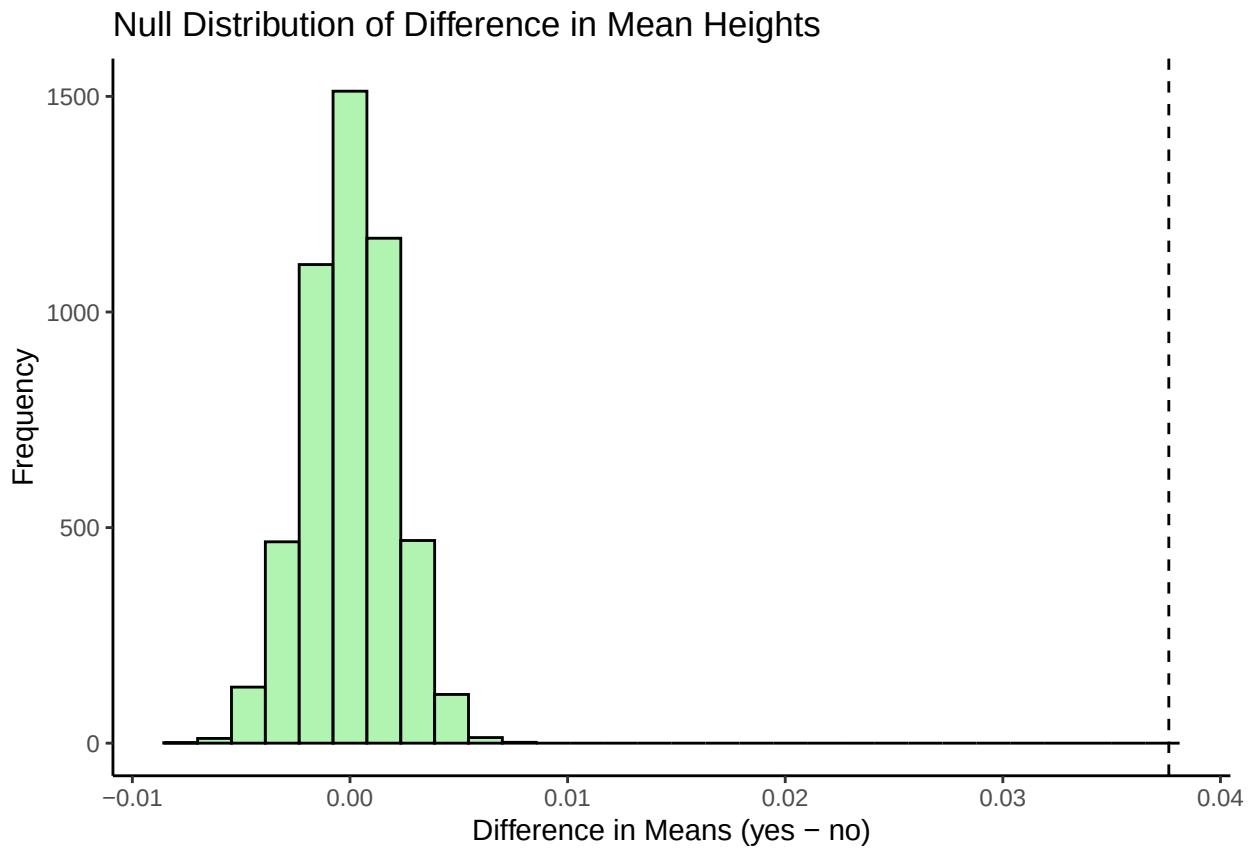
```r
cat("\n95% Confidence Interval for difference in heights:\n")
```

```
##
## 95% Confidence Interval for difference in heights:
```

```r
print(ci_height_diff)
```

```
## # A tibble: 1 x 2
##      lower_ci    upper_ci
##         <dbl>       <dbl>
## 1 0.034005882 0.041661777
```

```r
# plot
ggplot(data = null_dist_height, aes(x = stat)) +
  geom_histogram(bins = 30, fill = "lightgreen", color = "black", alpha = 0.7) +
  geom_vline(xintercept = obs_diff_height$stat, linetype = "dashed") +
  labs(title = "Null Distribution of Difference in Mean Heights",
       x = "Difference in Means (yes - no)",
       y = "Frequency") +
  theme_classic()
```



Answer 10

To evaluate whether the average height differs between students who exercise at least 3 times per week and those who don't, I conducted a hypothesis test with the null hypothesis that there is no difference in mean height between the two groups (H0: u_exercise = u_no_exercise) and the alternative hypothesis that there is a difference (Ha: u_exercise u_no_exercise). Using a significance level of = 0.05, I calculated the observed difference in mean heights as 0.038 meters (3.8 cm), with physically active students being taller. The resulting p-value was 0, which indicates that none of the 5000 (first I tried with 1000) permuted

differences were as extreme as the observed difference of 3.8 cm. This should be reported as $p < 0.0002$ (the minimum detectable p-value with 5000 replications is $1/5000 = 0.0002$), or even safer as $p < 0.001$. Given this extremely strong evidence against the null hypothesis, I reject H0 and conclude that there is a statistically significant difference in average height between students who exercise at least 3 times per week and those who don't, with physically active students being approximately 3.8 cm taller on average.

Exercise 11

Now, a non-inference task: Determine the number of different options there are in the dataset for the hours_tv_per_school_day there are.

```
unique(yrbss$hours_tv_per_school_day)
```

```
## [1] "5+"          "2"           "3"           "do not watch" "<1"
## [6] "4"           "1"           NA
```

```
length(unique(yrbss$hours_tv_per_school_day))
```

```
## [1] 8
```

Answer 11

There 8 different options in the dataset for the hours_tv_per_school_day, including: "do not watch", "<1" "1", "2", "3", "4", "5+", and NA.

Exercise 12

Come up with a research question evaluating the relationship between height or weight and sleep. Formulate the question in a way that it can be answered using a hypothesis test and/or a confidence interval. Report the statistical results, and also provide an explanation in plain language. Be sure to check all assumptions, state your level, and conclude in context.

Research Question Is there a relationship between the number of hours slept on weeknights and high school students' average height?

Hypothesis The null hypotesis states that there is no difference in the mean height of students who sleep 8 or more hours and those who sleep less. The alternative hypothesis states that there is a difference between the mean heights of students who have an adequate sleep time (8h or more) and those who don't.

H0: u_adequate_sleep = u_inadequate_sleep Ha: u_adequate_sleep u_inadequate_sleep

```
# 8h+ adequate sleep
yrbss <- yrbss %>%
  mutate(
    adequate_sleep = if_else(
      school_night_hours_sleep %in% c("8", "9", "10+"), "yes", "no",
      missing = NA_character_
    ),
    adequate_sleep = factor(adequate_sleep, levels = c("no","yes"))
```

```
  )

yrbss %>%
  filter(!is.na(adequate_sleep) & !is.na(height)) %>%
  group_by(adequate_sleep) %>%
  summarise(n = n(),
            mean_height = mean(height, na.rm = TRUE),
            sd_height = sd(height, na.rm = TRUE),
            .groups = "drop"
            )
```

```
## # A tibble: 2 x 4
##   adequate_sleep     n mean_height  sd_height
##   <fct>          <int>       <dbl>      <dbl>
## 1 no              9114   1.6908920 0.10426547
## 2 yes             3465   1.6921587 0.10583422
```

```
# calculate observed difference
obs_diff_sleep <- yrbss %>%
  filter(!is.na(adequate_sleep), !is.na(height)) %>%
  specify(height ~ adequate_sleep) %>%
  calculate(stat = "diff in means", order = c("yes", "no"))

obs_diff_sleep
```

```
## Response: height (numeric)
## Explanatory: adequate_sleep (factor)
## # A tibble: 1 x 1
##           stat
##          <dbl>
## 1 0.0012666959
```

```
# null distribution
null_dist_sleep <- yrbss %>%
  drop_na(adequate_sleep) %>%
  specify(height ~ adequate_sleep) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("yes", "no"))
```

```
## Warning: Removed 1004 rows containing missing values.
```

```
# p-value
p_value_sleep <- null_dist_sleep %>%
  get_p_value(obs_stat = obs_diff_sleep, direction = "two_sided")

p_value_sleep
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1   0.502
```

```r
extreme_count_sleep <- sum(abs(null_dist_sleep$stat) >= abs(obs_diff_sleep$stat))
cat("Number of permuted statistics as extreme as observed:", extreme_count_sleep, "\n")
```

```
## Number of permuted statistics as extreme as observed: 527
```

```r
cat("Actual p-value proportion:", extreme_count_sleep/1000, "\n")
```

```
## Actual p-value proportion: 0.527
```

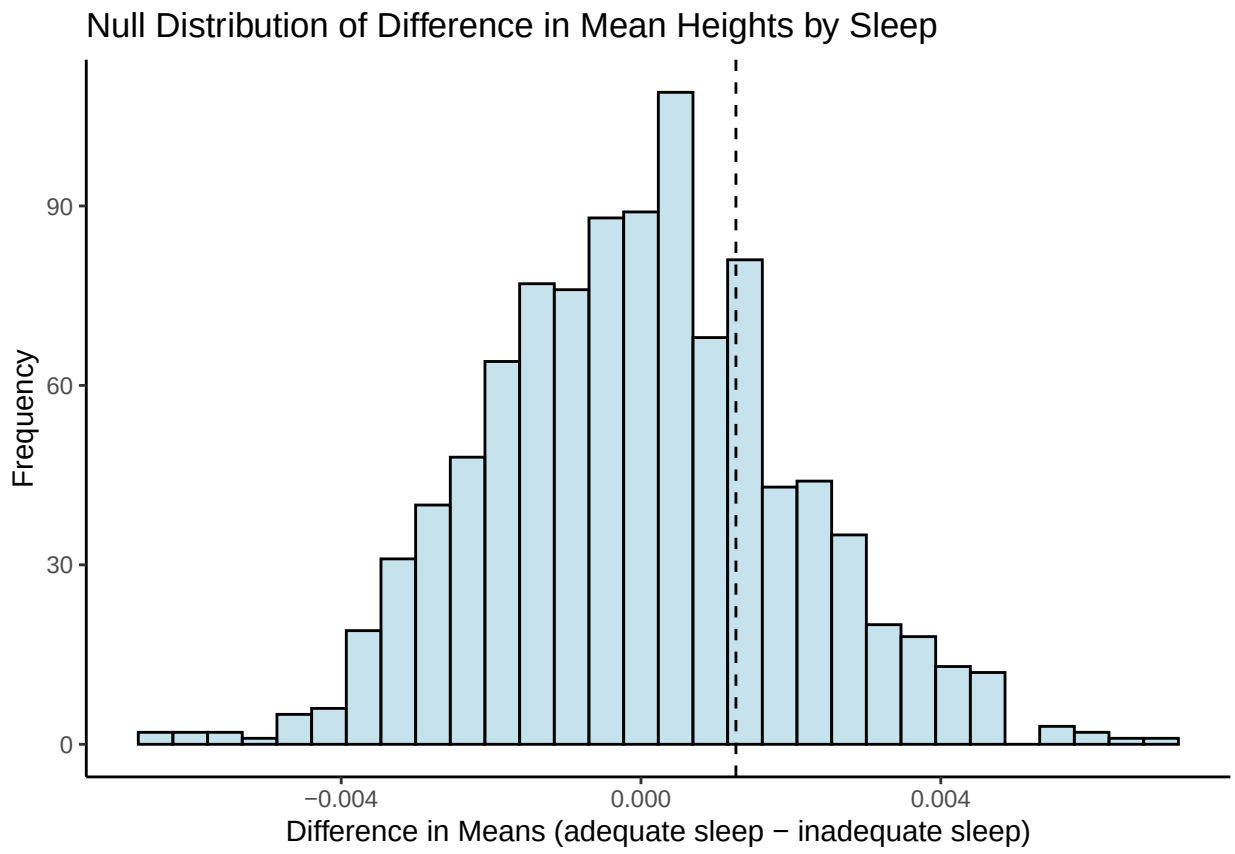
```r
if (extreme_count_sleep == 0) {
  cat("Since no permutations were as extreme, report as p < 0.001\n")
}

# calculate confidence interval
ci_sleep <- yrbss %>%
  drop_na(adequate_sleep) %>%
  specify(height ~ adequate_sleep) %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "diff in means", order = c("yes", "no")) %>%
  get_ci(level = 0.95)
```

```
## Warning: Removed 1004 rows containing missing values.
```

```r
ci_sleep
```

```
## # A tibble: 1 x 2
##        lower_ci     upper_ci
##           <dbl>        <dbl>
## 1 -0.0027557472 0.0054008056
```

```r
# plot
ggplot(data = null_dist_sleep, aes(x = stat)) +
  geom_histogram(bins = 30, fill = "lightblue", color = "black", alpha = 0.7) +
  geom_vline(xintercept = obs_diff_sleep$stat, linetype = "dashed") +
  labs(title = "Null Distribution of Difference in Mean Heights by Sleep",
       x = "Difference in Means (adequate sleep - inadequate sleep)",
       y = "Frequency") +
  theme_classic()
```



Answer 12

The groups compared in this study were students who sleep 8 hours ("adequate") vs $<$ 8 hours ("inadequate"). The sample sizes: 9,114 (inadequate) and 3,465 (adequate). The observed mean difference is 0.00127m (adequate – inadequate).The permutation p-value was around 0.584. And the 95 % bootstrap CI is a lower limit of –0.00296 and a upper limit of 0.00547 meters. I used a = 0.05 and the observed difference of about 1 mm is tiny, and the p-value (0.584) is far above 0.05. The 95 % confidence interval also contains 0, meaning the data is consistent with no meaningful difference in average height between students who sleep 8 hours on school nights and those who sleep less. So, we fail to reject the null hypothesis: the evidence does not suggest a statistically significant relationship between nightly sleep duration and average height in this sample of high-school students.

checking assumptions Independence: Observations are from different students (random sample) Sample size: both groups are very large ($>$ 30), so the Central Limit Theorem ensures the sampling distribution of the mean is approximately normal. Normality: large sample sizes means that the sampling distribution is approximately normal.

In simple terms, students who reported sleeping eight hours or more were about the same height on average as those who slept less. Any small observed difference is well within what could occur by random chance. Sleep habits do not impact students' average height.