Inference for numerical data

The data

Every two years, the Centers for Disease Control and Prevention conduct the Youth Risk Behavior Surveillance System (YRBSS) survey, where it takes data from high schoolers (9th through 12th grade), to analyze health patterns.

```
data(yrbss)
```

There are observations on 13 different variables, some categorical and some numerical. The meaning of each variable can be found by bringing up the help file:

```
?yrbss     # A data frame with 13583 observations on the following 13 variables
View(yrbss)
glimpse(yrbss)
```

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

13585 students in ages 12-18, although mostly older than 14

The count() and table() functions also are useful to get an overview of the data

```
yrbss %>% count(race)
```

```
## # A tibble: 6 x 2
##   race                                     n
##   <chr>                                <int>
## 1 American Indian or Alaska Native       323
## 2 Asian                                  552
## 3 Black or African American             3229
## 4 Native Hawaiian or Other Pacific Islander  258
## 5 White                                 6416
## 6 <NA>                                  2805
```

```
yrbss %>% count(text_while_driving_30d)
```

```
## # A tibble: 9 x 2
##   text_while_driving_30d     n
##   <chr>                  <int>
## 1 0                       4792
## 2 1-2                      925
## 3 10-19                    373
## 4 20-29                    298
## 5 3-5                      493
```

```
## 6 30                       827
## 7 6-9                      311
## 8 did not drive           4646
## 9 <NA>                     918
```

```
table(yrbss[c("gender","age")])
```

```
##         age
## gender    12   13   14   15   16   17   18
##   female   9    6  720 1575 1543 1700 1041
##   male    17   12  648 1522 1659 1771 1278
```

Exploratory data analysis

Using visualization and summary statistics, describe the distribution of weights.

Fairly normal

There are some extremes on both tails. (1 KG = 2.2 pounds)

and to explore that i used quantile() to see where the bottom and top 5% are.

```
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
```

```
quantile(yrbss$weight,.95, na.rm=TRUE)
```

```
##    95%
## 100.25
```

```
quantile(yrbss$weight,.05, na.rm=TRUE)
```

```
##    5%
## 46.72
```

1. How many observations are we missing weights from?

1004

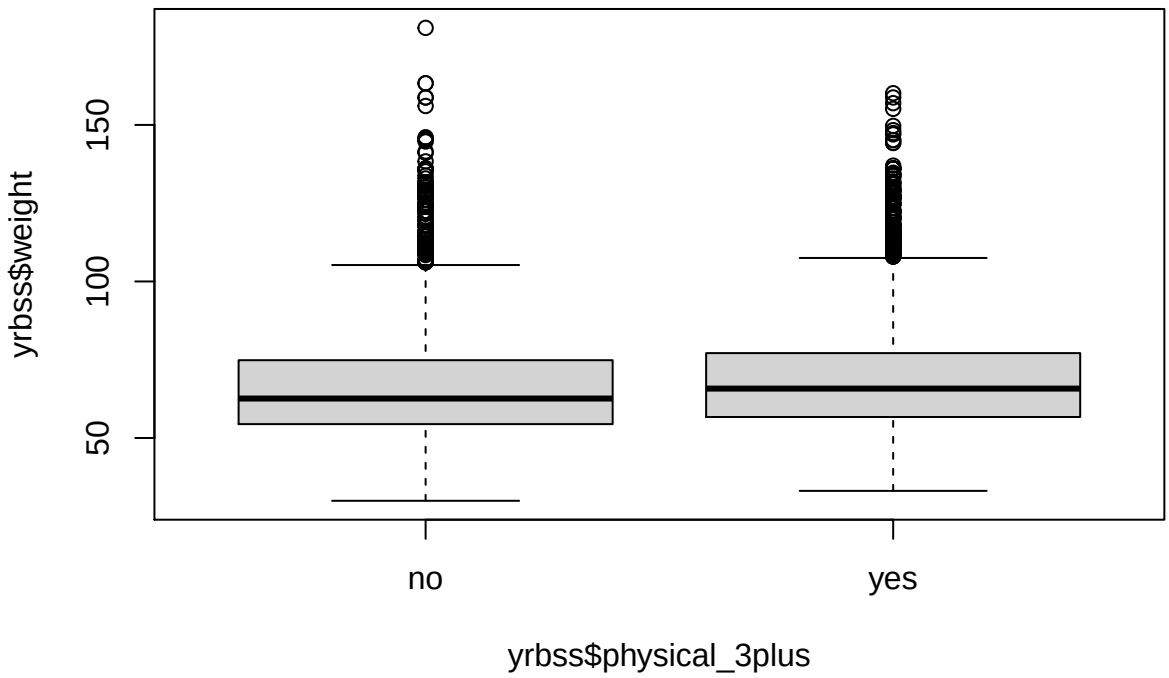
Next, consider the possible relationship between a high schooler's weight and their physical activity. Plotting the data is a useful first step because it helps us quickly visualize trends, identify strong associations, and develop research questions.

First, let's create a new variable `physical_3plus`, which will be coded as either "yes" if they are physically active for at least 3 days a week, and "no" if not.

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

# glimpse(yrbss)
```

1. 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?

```
boxplot(yrbss$weight ~ yrbss$physical_3plus)
```



Those who exercise seem to weigh a little more.

A little surprising but not totally. The exercise may result in some weight loss. But Weight Lifters may weigh more.

Im thinking if the same study was done on a different age group, the results might be different.

The box plots show how the medians of the two distributions compare, but we can also compare the means of the distributions using the following to first group the data by the `physical_3plus` variable, and then calculate the mean `weight` in these groups using the `mean` function while ignoring missing values by setting the `na.rm` argument to `TRUE`.

```
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
```

There is an observed difference, but is this difference statistically significant? In order to answer this question we will conduct a hypothesis test.

Inference

1. 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(observations = n(),mean_weight = mean(weight, na.rm = TRUE))
```

```
## # A tibble: 3 x 3
##   physical_3plus observations mean_weight
##   <chr>                 <int>       <dbl>
## 1 no                     4404        66.7
## 2 yes                    8906        68.4
## 3 <NA>                    273        69.9
```

When dealing with a comparison between groups, the normality check should include the independence *extended* check.

We might consider whether organized sports creates an important influence on the yes group but I doubt it effects our normality requirement.

Sample sizes are over 30 with no extreme outliers.

1. 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.

$$H_0 : \bar{w}_{yes} = \bar{w}_{no}$$

$$H_A : \bar{w}_{yes} <> \bar{w}_{no}$$

Next, we will introduce a new function, `hypothesize`, that falls into the `infer` workflow. You will use this method for conducting hypothesis tests.

But first, we need to initialize the test, which we will save as `obs_diff`.

```
obs_diff <- yrbss %>%
  specify(weight ~ physical_3plus) %>%    # creates a 2 column table of quantile of weight and category
  calculate(stat = "diff in means", order = c("yes", "no"))

# obs_diff = 1.77 which we knew
```

Notice how you can use the functions `specify` and `calculate` again like you did for calculating confidence intervals. Here, though, the statistic you are searching for is the difference in means, with the order being `yes - no != 0`.

After you have initialized the test, you need to simulate the test on the null distribution, which we will save as `null`.

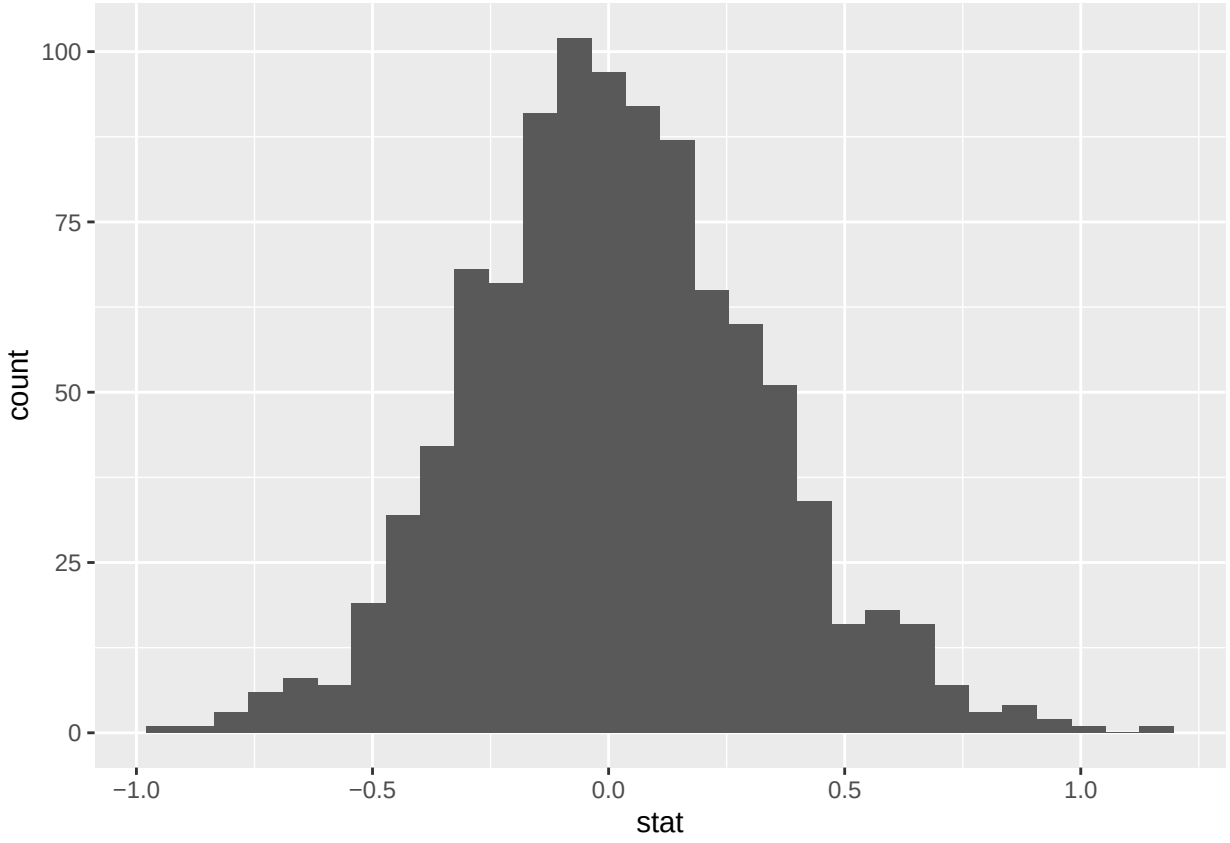
```r
null_dist <- yrbss %>%
  specify(weight ~ physical_3plus) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("yes", "no"))

# quantile(yrbss$weight, c(0.25, 0.75), na.rm = TRUE)
# View(null_dist)
# vignette("infer")
```

Here, `hypothesize` is used to set the null hypothesis as a test for independence. In one sample cases, the `null` argument can be set to "point" to test a hypothesis relative to a point estimate.

Also, note that the `type` argument within `generate` is set to `permute`, which is the argument when generating a null distribution for a hypothesis test.

We can visualize this null distribution with the following code:

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



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

```
gt_obs_stat<-nrow(subset(null_dist, abs(stat)>obs_diff))

sprintf("The generated distribution has %d differences greater than %.4f", gt_obs_stat,obs_diff)
```

```
## [1] "The generated distribution has 0 differences greater than 1.7746"
```

Apparently the null distribution ranges from -1.04 to 1.02 so we would reject the null hypothesis.

Im not clear how the infer code knew what the standard deviation should be since the actual range of weights is much higher.

Now that the test is initialized and the null distribution formed, you can calculate the p-value for your hypothesis test using the function `get_p_value`.

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

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

This the standard workflow for performing hypothesis tests.

1. 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.

```
s_1<-subset(yrbss,physical_3plus=="yes")["weight"]
s_2<-subset(yrbss,physical_3plus=="no")["weight"]

sd1<-sd(s_1$weight, na.rm=TRUE)
sd2<-sd(s_2$weight, na.rm=TRUE)
n1<-nrow(s_1)
n2<-nrow(s_2)

se<-sqrt(sd1^2/n1 + sd2^2/n2)

# validating...
# mean(s_1$weight, na.rm=TRUE) - mean(s_2$weight, na.rm=TRUE)

z_score<-qnorm(.95)        #  The z score of 90% (1.90/2) is 1.645
moe<-z_score*se

sprintf("The 90%s Confidence Interval is between %.2f and %.2f ", '%', obs_diff - moe, obs_diff+moe)
```

```
## [1] "The 90% Confidence Interval is between 1.25 and 2.30 "
```

If we were to continually perform similar studies, and respecting the first study as our best data.

then 90 percent of those studies would result between -0.06 and 3.61 (where -0.06 means the non exercisers would weigh more).

More Practice

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

```
sigma<-sd(yrbss$height,na.rm = TRUE)
mu<-mean(yrbss$height,na.rm = TRUE)

n <- yrbss %>%
  drop_na(height) %>%
    summarise(n())


se<-sigma/sqrt(n)

z_score<-qnorm(.975)        #  The z score of 95 (1.95/2) is 1.96

moe<-z_score*se

sprintf("The 95%s Confidence Interval is between %.6f and %.6f ", '%', mu - moe, mu+moe)
```

```
## [1] "The 95% Confidence Interval is between 1.689411 and 1.693071 "
```

With 12000 observations there is not much standard error.

2. 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.

```
z_score<-qnorm(.95)        #  The z score of 90% (1.9/2) is 1.645

moe<-z_score*se

sprintf("The 90%s Confidence Interval is between %.6f and %.6f ", '%', mu - moe, mu+moe)
```

```
## [1] "The 90% Confidence Interval is between 1.689705 and 1.692776 "
```

By being less confident we are restricting the interval of possible values.

3. 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.

```
s_1<-subset(yrbss,physical_3plus=="yes")["height"]
s_2<-subset(yrbss,physical_3plus=="no")["height"]

mu1<-mean(s_1$height, na.rm=TRUE)
mu2<-mean(s_2$height, na.rm=TRUE)

obs_diff<-mu1-mu2
h0_diff<-0

sd1<-sd(s_1$height, na.rm=TRUE)
```

```
sd2<-sd(s_2$height, na.rm=TRUE)
n1<-nrow(s_1)
n2<-nrow(s_2)

se<-sqrt(sd1^2/n1 + sd2^2/n2)

t_score<-obs_diff-h0_diff/se
df=min(n1-1,n2-1)

p_value<-pt(t_score, df=df, lower.tail=FALSE)


sprintf("The p-value is %.2f ",  p_value)
```

```
## [1] "The p-value is 0.48 "
```

We can not reject the null hypothesis.

4. 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.

```
yrbss %>%
  group_by(hours_tv_per_school_day) %>%
    summarise(n())
```

```
## # A tibble: 8 x 2
##   hours_tv_per_school_day `n()`
##   <chr>                   <int>
## 1 <1                       2168
## 2 1                        1750
## 3 2                        2705
## 4 3                        2139
## 5 4                        1048
## 6 5+                       1595
## 7 do not watch             1840
## 8 <NA>                      338
```

5. 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 α level, and conclude in context.

\textcolor{blue}{Does sleep effect either height or weight? Well use 95% Confidence Level.}

```
new_yrbss <- yrbss %>%
  drop_na(height) %>%
  drop_na(weight) %>%
  drop_na(school_night_hours_sleep)



res.aov <- aov(weight ~ school_night_hours_sleep, data = new_yrbss)
summary(res.aov)
```

```
##                              Df  Sum Sq Mean Sq F value  Pr(>F)
## school_night_hours_sleep      6   11333    1889   6.581 5.9e-07 ***
## Residuals                 11474 3293032     287
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
res.aov <- aov(height ~ school_night_hours_sleep, data = new_yrbss)
summary(res.aov)
```

```
##                              Df Sum Sq Mean Sq F value Pr(>F)
## school_night_hours_sleep      6   0.17 0.02776   2.538 0.0186 *
## Residuals                 11474 125.49 0.01094
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
new_yrbss %>%
  group_by(school_night_hours_sleep) %>%
    summarise(obs=n(),mu_h=mean(height),mu_w=mean(weight))
```

```
## # A tibble: 7 x 4
##   school_night_hours_sleep   obs  mu_h  mu_w
##   <chr>                    <int> <dbl> <dbl>
## 1 <5                         859  1.69  70.3
## 2 10+                        255  1.68  69.3
## 3 5                         1378  1.68  68.4
## 4 6                         2496  1.69  68.3
## 5 7                         3283  1.69  67.4
## 6 8                         2505  1.69  67.5
## 7 9                          705  1.69  65.6
```

```
mu_yrbss <- select(new_yrbss, c("school_night_hours_sleep","height","weight")) %>%
  group_by(school_night_hours_sleep) %>%
    summarise(obs=n(),mu_h=mean(height),mu_w=mean(weight))

new_mu_yrbss <- select(mu_yrbss, c("mu_h","mu_w"))

chisq.test(new_mu_yrbss)
```

```
##
##  Pearson's Chi-squared test
##
## data:  new_mu_yrbss
## X-squared = 0.0053893, df = 6, p-value = 1
```

The aov function is designed to return a pvalue that is high if there is variance between the means between groups.

It displayed a p-value for heights as .0186 and for weights, esentially 0.

Neither exceeds my alpha of .05 so I conclude that sleep does not effect height or weight.

The chi squared function is designed to analyze if the row/column distribution equals the row/column value.

Im not sure if chisq.test worked as expected. I would think the p-value would be close to 0 not 1.

Clearly average height doesnt change across sleep groups. Casual observation of the deviation in weight is that its small as well.
