HW5

Caitlin Jagla

03/28/2022

1. pwr

Using the pwr package, estimate the power of a two-sided, two-sample t-test to detect the difference between the mean relative vascularized areas for WT cells versus mutants. Use a significance level (alpha) of 0.05 and a pooled standard deviation of .036.

Your pilot study indicates the following:

> WT: ybar = 0.106, n = 5
>
> Mutant: ybar = 0.161, n = 7

A. Calculate the effect size. Is this effect size small (~.3), medium (~.5), or large (~.8)?

```
#Effect size (Cohen's d) - difference between the means divided by the pooled standard deviation
d <- (0.161 - 0.106)/0.036

d
```

```
## [1] 1.527778
```

> The effect size for a two-sample t-test is the difference in means divided by the pooled SD, which in this case is 1.528, which is large.

B. Calculate the power and interpret your findings

```
library(pwr)

pwr <- pwr.t2n.test(n1 = 5, n2 = 7, d = d, sig.level = 0.05, alternative="two.sided")
pwr
```

```
##
##      t test power calculation
##
##              n1 = 5
##              n2 = 7
##               d = 1.527778
##       sig.level = 0.05
##           power = 0.6532558
##     alternative = two.sided
```

The power is 0.6532558, meaning there is a 65% probability of rejecting H_0 if H_0 is false, given the observed effect size.

C. What sample sizes would the experimenter need to reach power of 0.90? Try out some different n1 and n2 values and explain the effect of unbalanced sample sizes on power.

```
# calculate sample sizes needed to reach power = 0.90 (with equal n)
pwr90 <- pwr.t.test(d = d, power = 0.90, sig.level = 0.05, alternative="two.sided")
pwr90
```

```
##
##      Two-sample t test power calculation
##
##               n = 10.06853
##               d = 1.527778
##       sig.level = 0.05
##           power = 0.9
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

```
# see what n2 is required to maintain 0.90 power if n1 is less than required when n's are equal
pwr_n1low <- pwr.t2n.test(n1 = 8, d = d, power = 0.90, sig.level = 0.05, alternative="two.sided")
pwr_n1low
```

```
##
##      t test power calculation
##
##              n1 = 8
##              n2 = 13.32031
##               d = 1.527778
##       sig.level = 0.05
##           power = 0.9
##     alternative = two.sided
```

```
# see what n2 is required to maintain 0.90 power if n1 is greater than required when n's are equal
pwr_n1high <- pwr.t2n.test(n1 = 18, d = d, power = 0.90, sig.level = 0.05, alternative="two.sided")
pwr_n1high
```

```
##
##      t test power calculation
##
##              n1 = 18
##              n2 = 6.763743
##               d = 1.527778
##       sig.level = 0.05
##           power = 0.9
##     alternative = two.sided
```

If n for both groups is equal, then 11 samples are needed per group to reach $power = 0.90$.

If n_1 is less than the n required to achieve $power = 0.90$ when $n_1 = n_2$, then n_2 must be higher to compensate. In this example, $n_1 = 8$ results in a requirement of $n_2 = 14$ to achieve $power = 0.90$.

Conversely, if n_1 is greater than the n required to achieve $power = 0.90$ when n of both groups are equal, then n_2 can be lower. In this example, $n_1 = 18$ results in a requirement of $n_2 = 7$ to achieve $power = 0.90$.

2. nhanes

Use the nhanes.csv dataset to assess the difference in Testosterone values in adult males with health insurance and those without. Make sure to follow all the steps of hypothesis testing and clearly state your conclusions about the test assumptions to validate your choice of test.

```r
library(tidyverse)
nh <- read_csv("nhanes.csv")
```

A. Filter the dataset so that only males at or above age 18 are included. Then filter out participants who are missing values on the Testosterone or on the Insured variable. Print your dataset (not the whole thing, please!) to illustrate that the filtering step worked

```r
nh <- nh |> filter(Gender == "male" & Age >= 18 & !is.na(Testosterone) & Insured != "No")

nh
```

```
## # A tibble: 48 x 32
##       id Gender   Age Race  Education   MaritalStatus RelationshipStatus Insured
##    <dbl> <chr>  <dbl> <chr> <chr>       <chr>         <chr>              <chr>
##  1 65162 male      32 White High School NeverMarried  Single             Yes
##  2 64151 male      21 White Some Colle~ NeverMarried  Single             Yes
##  3 64964 male      80 White High School Married       Committed          Yes
##  4 67828 male      67 Black 8th Grade   Married       Committed          Yes
##  5 62757 male      54 White High School Divorced      Single             Yes
##  6 68058 male      50 White Some Colle~ Married       Committed          Yes
##  7 69419 male      57 Black College Gr~ Married       Committed          Yes
##  8 68271 male      43 White Some Colle~ NeverMarried  Single             Yes
##  9 69896 male      49 White High School NeverMarried  Single             Yes
## 10 66678 male      21 White Some Colle~ NeverMarried  Single             Yes
## # i 38 more rows
## # i 24 more variables: Income <dbl>, Poverty <dbl>, HomeRooms <dbl>,
## #   HomeOwn <chr>, Work <chr>, Weight <dbl>, Height <dbl>, BMI <dbl>,
## #   Pulse <dbl>, BPSys <dbl>, BPDia <dbl>, Testosterone <dbl>, HDLChol <dbl>,
## #   TotChol <dbl>, Diabetes <chr>, DiabetesAge <dbl>, nPregnancies <dbl>,
## #   nBabies <dbl>, SleepHrsNight <dbl>, PhysActive <chr>, PhysActiveDays <dbl>,
## #   AlcoholDay <dbl>, AlcoholYear <dbl>, SmokingStatus <chr>
```

B. Clearly state your hypotheses

I hypothesize that adult men with diabetes have lower testosterone levels than adult men without diabetes.

Put in a more formal way,

H_0: mean testosterone levels do not differ between adult men with and without diabetes.

H_1: mean testosterone levels differ between adult men with and without diabetes.

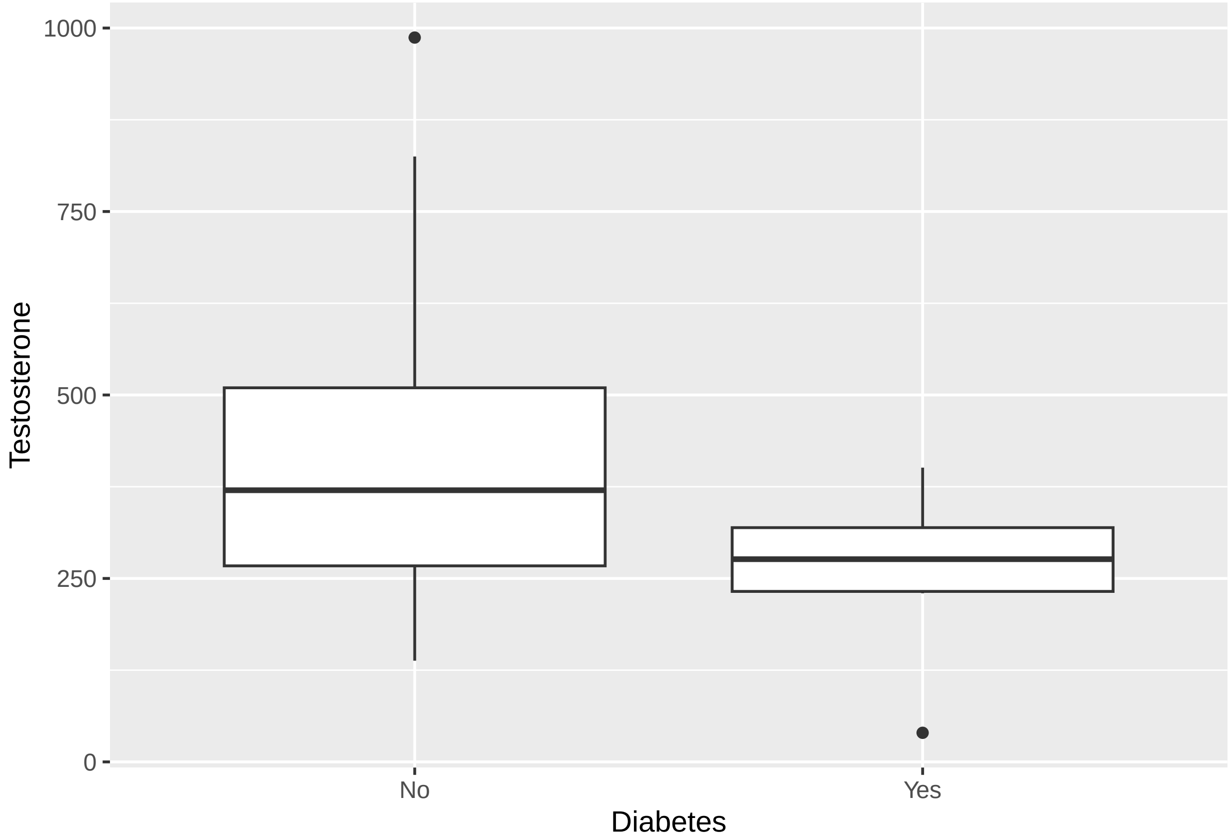
$\alpha = 0.05$

C. Show an exploratory plot

```
nh |> group_by(Diabetes) |> summarise(n = n(),
                                      mean = mean(Testosterone),
                                      sd = sd(Testosterone))
```

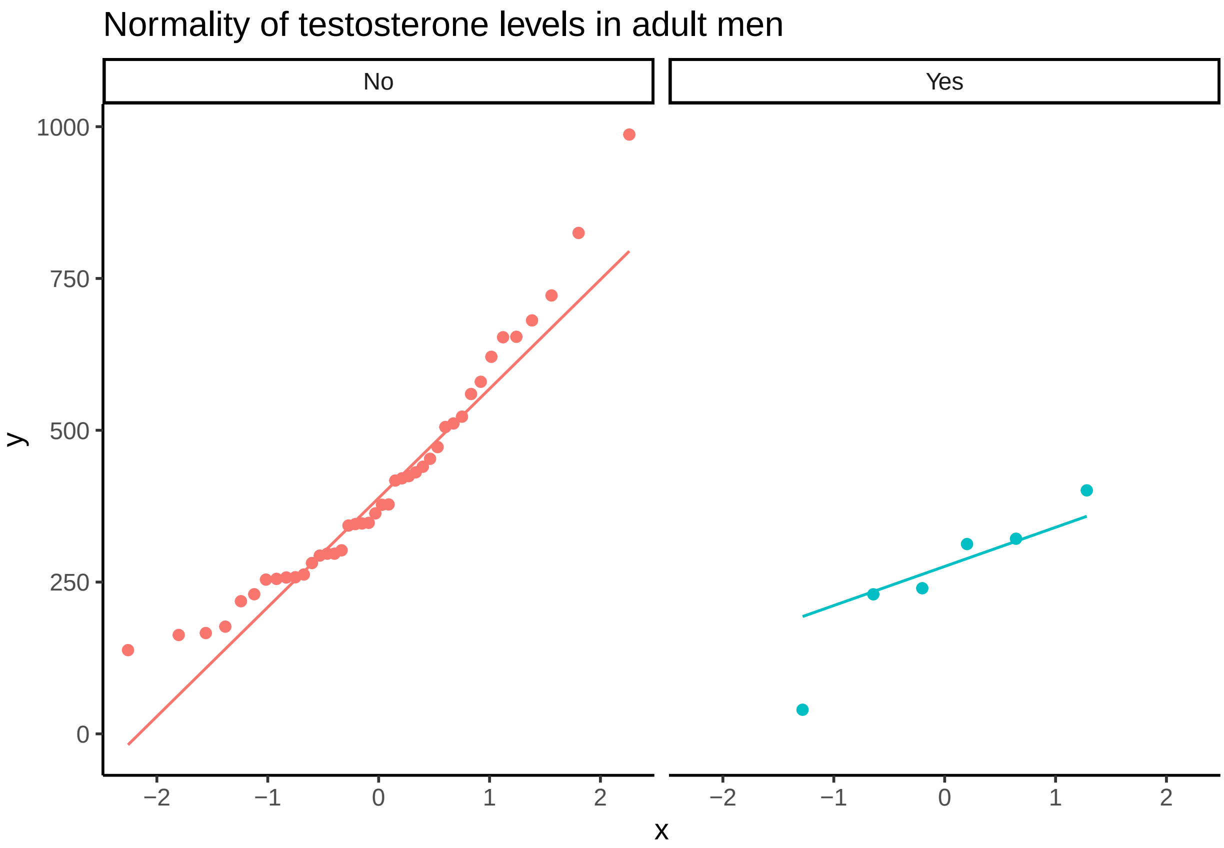
```
## # A tibble: 2 x 4
##   Diabetes     n  mean    sd
##   <chr>    <int> <dbl> <dbl>
## 1 No          42  410.  188.
## 2 Yes          6  257.  123.
```

```
nh |>
  ggplot(aes(x = Diabetes, y = Testosterone)) + geom_boxplot()
```

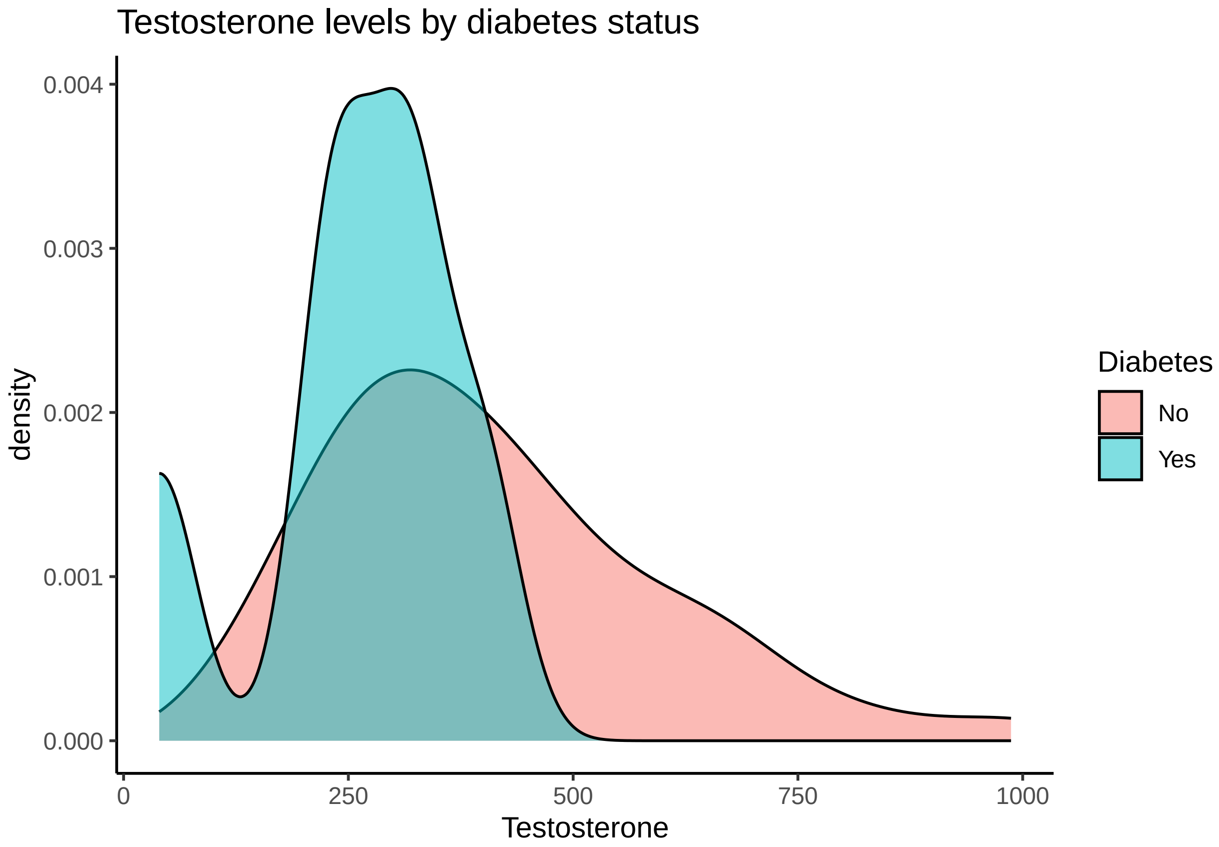


D. Check the assumptions of a two-sample t-test and clearly explain your logic of which hypothesis test to conduct based on your assessment of the assumptions.

```
# check normality
p_qq <- nh |>
          ggplot(aes(sample = Testosterone, color = Diabetes)) +
          geom_qq(show.legend=FALSE) + geom_qq_line(show.legend=FALSE) +
          facet_wrap(~Diabetes) +
          labs(title = "Normality of testosterone levels in adult men") +
          theme_classic()
p_qq
```



```
# check within-group variance / distribution with density plot
p_dens <- nh |>
            ggplot(aes(x = Testosterone, fill = Diabetes)) +
              geom_density(alpha = 0.5) +
              labs(title = "Testosterone levels by diabetes status") +
              theme_classic()
p_dens
```



```r
# check within-group variance with descriptive statistics
summary_stats <- nh |>
                    group_by(Diabetes) |>
                    summarize(n = n(),
                              median = median(Testosterone),
                              mean = mean(Testosterone),
                              stdev = sd(Testosterone),
                              variance = var(Testosterone))

summary_stats
```

```
## # A tibble: 2 x 6
##   Diabetes     n median  mean stdev variance
##   <chr>    <int>  <dbl> <dbl> <dbl>    <dbl>
## 1 No          42   370.  410.  188.   35430.
## 2 Yes          6   276.  257.  123.   15248.
```

Because each datapoint represents a separate man, I can be sure that these data are independently sampled.

Based on the density plot and the descriptive statistics, the two groups do not have equal variance.

Based on the QQ plot and density plot, these data seem somewhat normally distributed. However, there are long tails among the men without diabetes, particularly the upper tail, and a single outlier on the low end for the men with diabetes. Because of this, I would rather not assume normality, just to be safe.

For these reasons, I will run a Wilcoxon-Mann-Whitney U-test. This will slightly alter the hypotheses such that:

H_0: men with and without diabetes have the same distribution of testosterone levels.

H_1: men with and without diabetes have different distributions of testosterone levels.

$\alpha = 0.05$

E. Run the test and interpret the results

```r
wt <- wilcox.test(Testosterone ~ Diabetes, data = nh)
wt
```

```
##
##  Wilcoxon rank sum exact test
##
## data:  Testosterone by Diabetes
## W = 186, p-value = 0.06255
## alternative hypothesis: true location shift is not equal to 0
```

There is no difference in the distributions of testosterone levels in men with vs. without diabetes ($p = 0.063$, which is greater than $\alpha = 0.05$).

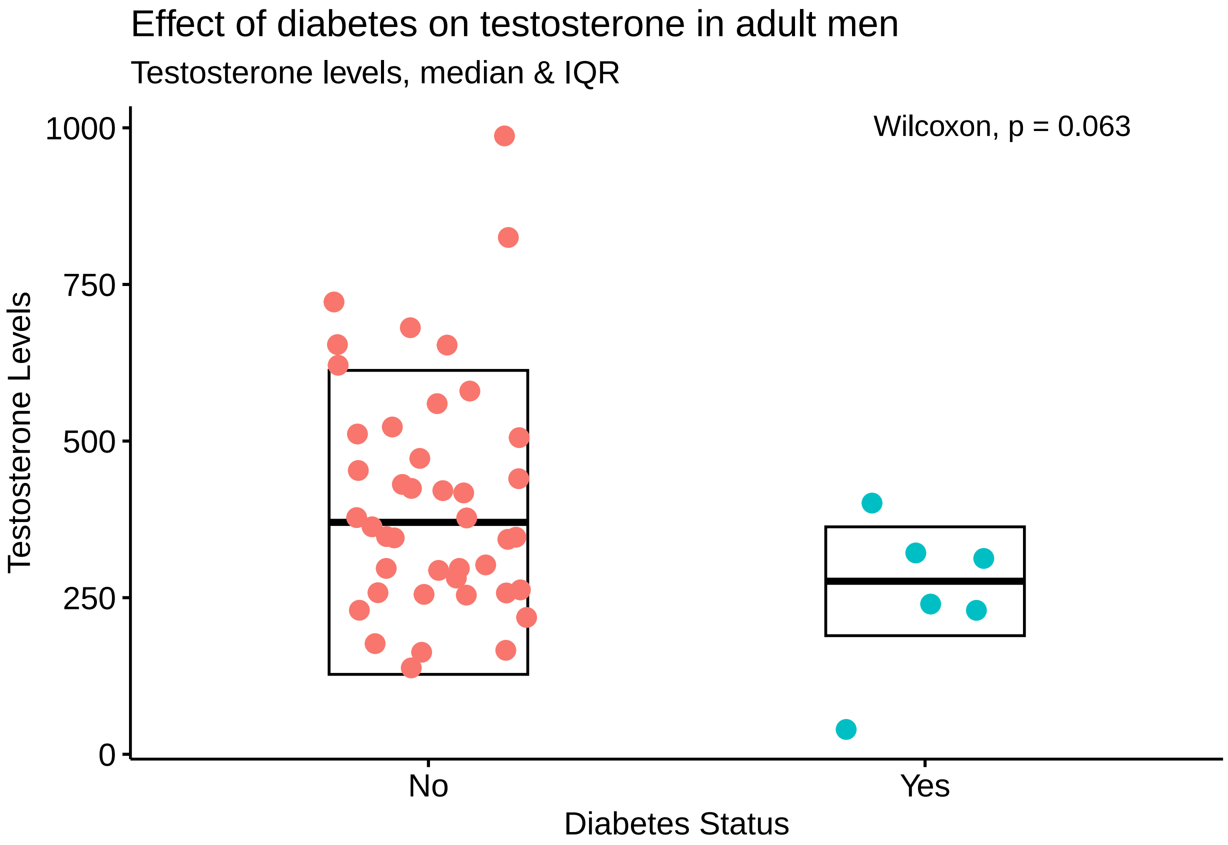
F. Which descriptive statistics would you use to describe the results (mean/sd or median/IQR)?

Median and IQR should be used to describe the results, because the Wilcoxon-Mann-Whitney U-test compares the distributions of two groups, not the means.

```r
library(ggpubr)

# use ggpubr to make pretty plot
# display individual data points, group median and IQR, and the t-test result
p <- nh |>
  ggstripchart(
  # assign aes() parameters
      x = "Diabetes", y = "Testosterone",
      color = "Diabetes", size = 3,
  # plot median and IQR overlaid on stripchart of individual datapoints
      add = "median_iqr",
  # format plot visualization of median and IQR
      add.params = list(color = "black", width = 0.4), error.plot = "crossbar",
  # set plot labels
      title = "Effect of diabetes on testosterone in adult men",
      subtitle = "Testosterone levels, median & IQR",
      xlab = "Diabetes Status",
      ylab = "Testosterone Levels") +
  # hide legend
      rremove("legend") +
  # annotate plot with pvalue from unpooled, unpaired t-test
      stat_compare_means(method = "wilcox",
                         method.args = list(alternative = "two.sided", paired = FALSE),
                         label.x.npc = "right")

p
```



3. penguins

Use the penguins dataset from the palmerpenguins package to assess the difference in flipper length (flipper_length_mm) for Adelie and Chinstrap penguins. Make sure to follow all the steps of hypothesis testing and clearly state your conclusions about the test assumptions to validate your choice of test.

```r
library(palmerpenguins)
penguins
```

```
## # A tibble: 344 x 8
##    species island    bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
##    <fct>   <fct>              <dbl>         <dbl>             <int>       <int>
##  1 Adelie  Torgersen           39.1          18.7               181        3750
##  2 Adelie  Torgersen           39.5          17.4               186        3800
##  3 Adelie  Torgersen           40.3          18                 195        3250
##  4 Adelie  Torgersen           NA            NA                  NA          NA
##  5 Adelie  Torgersen           36.7          19.3               193        3450
##  6 Adelie  Torgersen           39.3          20.6               190        3650
##  7 Adelie  Torgersen           38.9          17.8               181        3625
##  8 Adelie  Torgersen           39.2          19.6               195        4675
##  9 Adelie  Torgersen           34.1          18.1               193        3475
## 10 Adelie  Torgersen           42            20.2               190        4250
## # i 334 more rows
## # i 2 more variables: sex <fct>, year <int>
```

A. Use the filter() function to subset the penguins dataset so that only males from the Adelie and Chinstrap species are included (no females and no Gentoo species). Save the result intoan object and show the resulting object in the report. The resulting dataset should have dimensions 107 x 8

```r
df <- penguins |>
        filter(sex == "male" & species != "Gentoo")
glimpse(df)
```

```
## Rows: 107
## Columns: 8
## $ species           <fct> Adelie, Adelie, Adelie, Adelie, Adelie, Adelie, Adel~
## $ island            <fct> Torgersen, Torgersen, Torgersen, Torgersen, Torgerse~
## $ bill_length_mm    <dbl> 39.1, 39.3, 39.2, 38.6, 34.6, 42.5, 46.0, 37.7, 38.2~
## $ bill_depth_mm     <dbl> 18.7, 20.6, 19.6, 21.2, 21.1, 20.7, 21.5, 18.7, 18.1~
## $ flipper_length_mm <int> 181, 190, 195, 191, 198, 197, 194, 180, 185, 180, 18~
## $ body_mass_g       <int> 3750, 3650, 4675, 3800, 4400, 4500, 4200, 3600, 3950~
## $ sex               <fct> male, male, male, male, male, male, male, male, male~
## $ year              <int> 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007~
```

B. Clearly state your hypotheses

I hypothesize that Chinstrap penguins have longer bills than Adelie penguins.

Put in a more formal way,

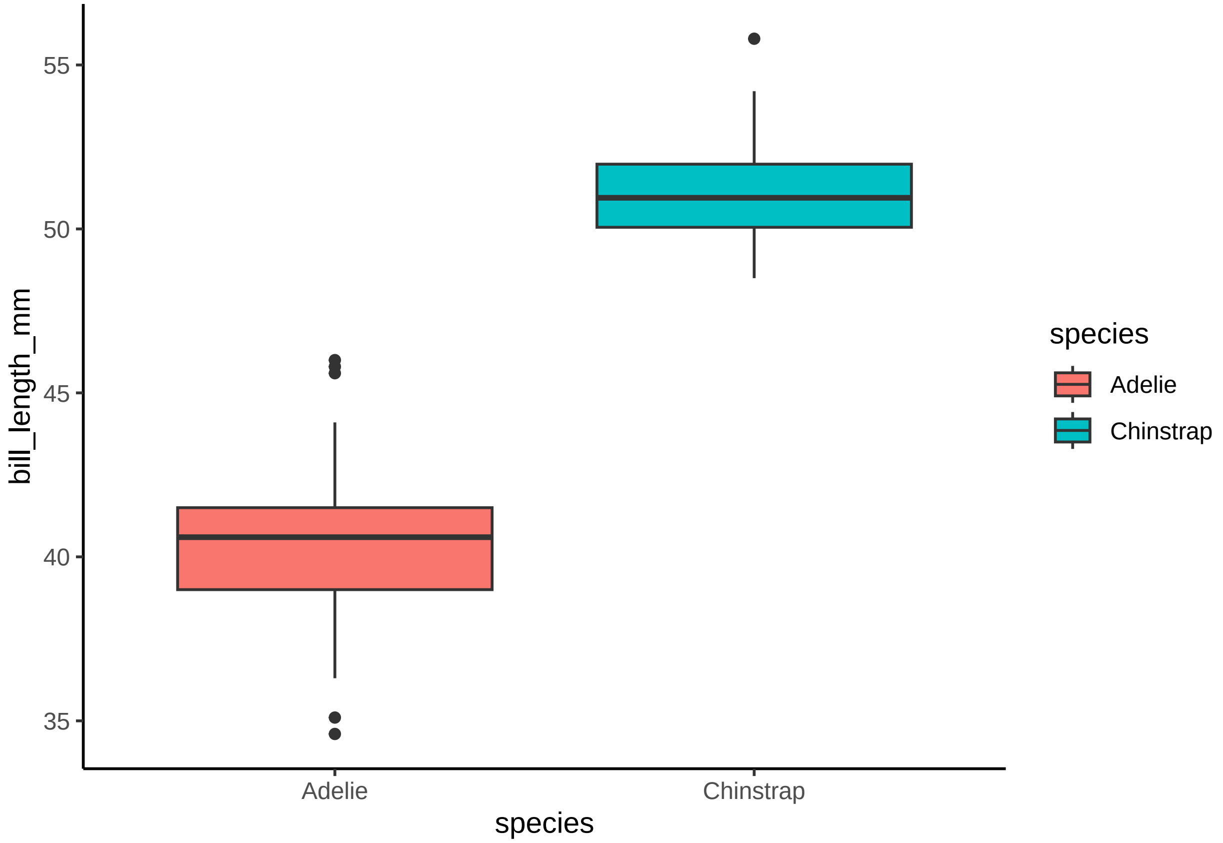
H_0: bill lengths do not differ between penguin species.

H_1: mean bill length is longer in Chinstrap penguins than Adelie penguins.

$\alpha = 0.05$

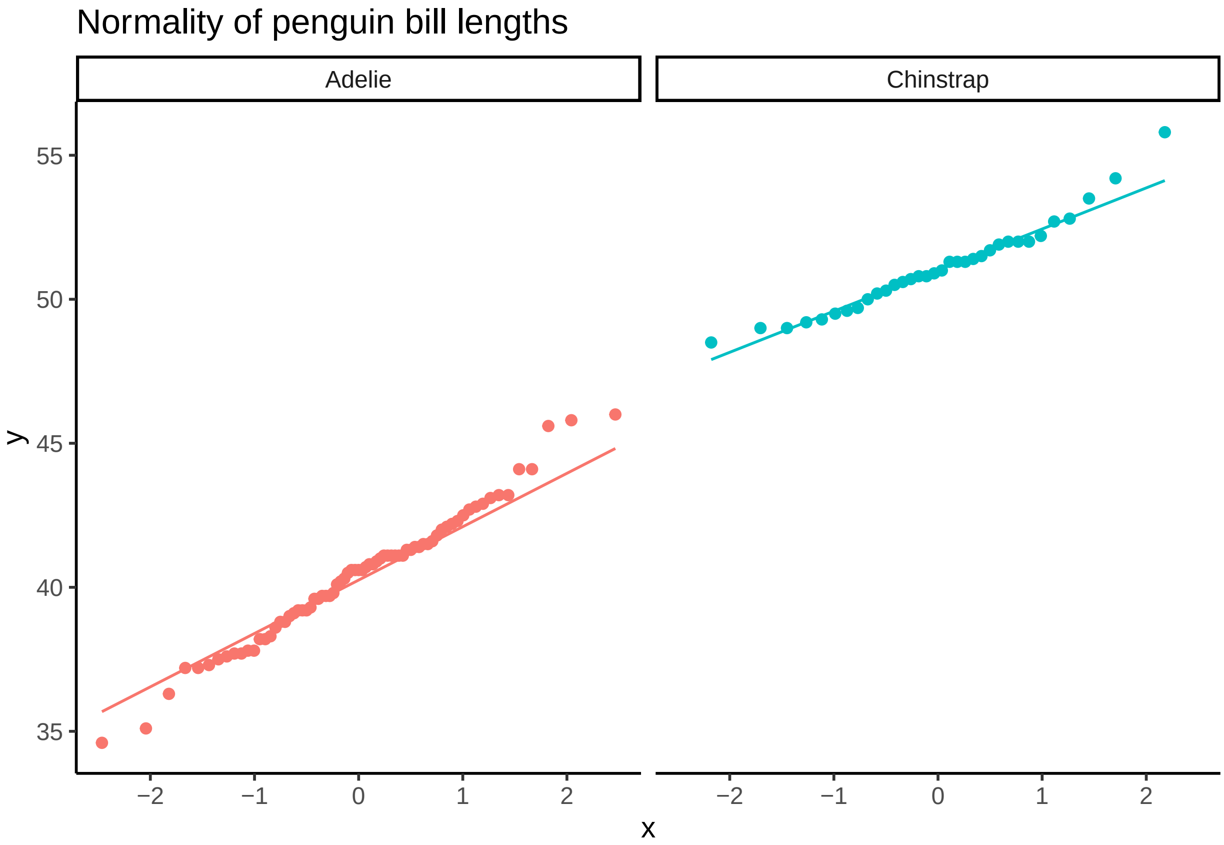
C. Show an exploratory plot

```
df |>
  ggplot(aes(x = species, y = bill_length_mm, fill = species)) +
  geom_boxplot() + theme_classic()
```

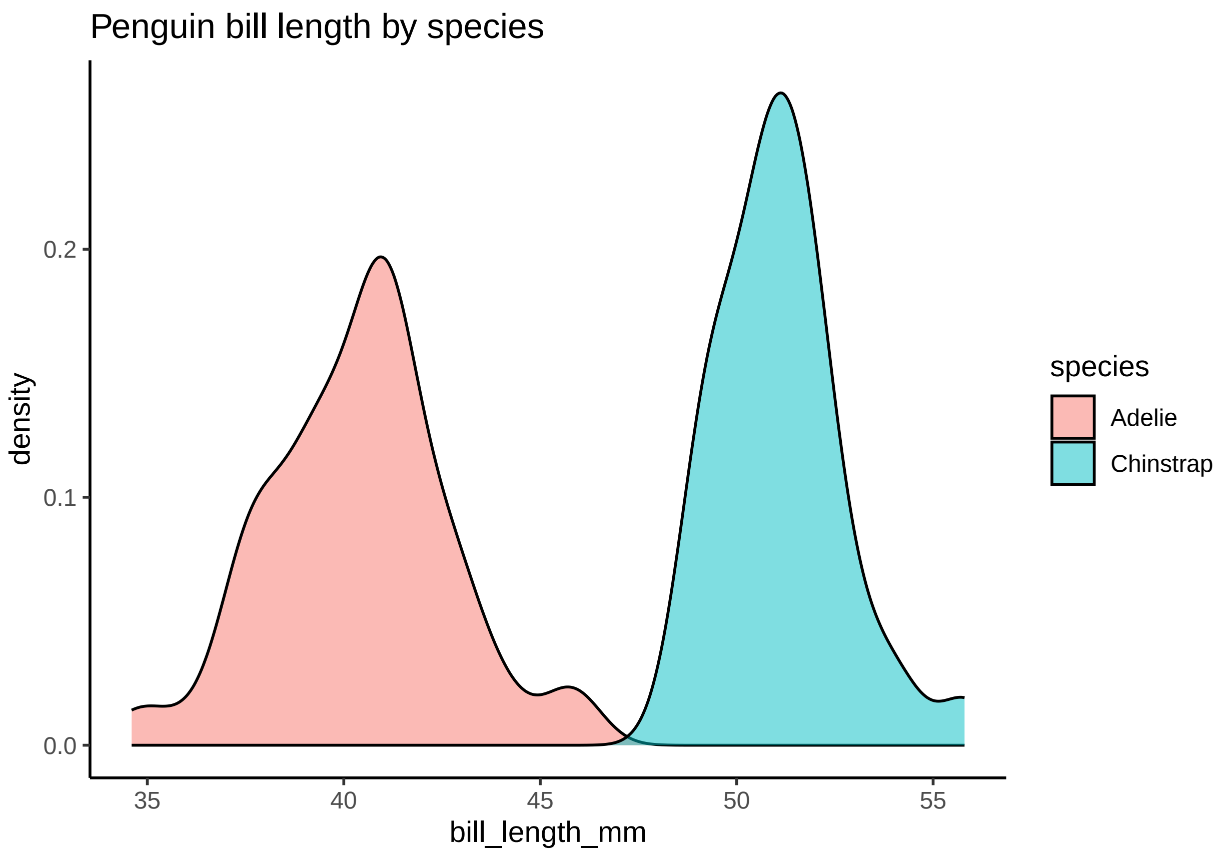


D. Check the assumptions of a two-sample t-test and clearly explain your logic of which hypothesis test to conduct based on your assessment of the assumptions.

```
# check normality
p2_qq <- df |>
          ggplot(aes(sample = bill_length_mm, color = species)) +
          geom_qq(show.legend=FALSE) + geom_qq_line(show.legend=FALSE) +
          facet_wrap(~species) +
          labs(title = "Normality of penguin bill lengths") +
          theme_classic()
p2_qq
```



```r
# check within-group variance / distribution with density plot
p2_dens <- df |>
            ggplot(aes(x = bill_length_mm, fill = species)) +
              geom_density(alpha = 0.5) +
              labs(title = "Penguin bill length by species") +
              theme_classic()
p2_dens
```



```
# check within-group variance with descriptive statistics
summary_stats2 <- df |>
                    group_by(species) |>
                    summarize(n = n(),
                              median = median(bill_length_mm),
                              mean = mean(bill_length_mm),
                              stdev = sd(bill_length_mm),
                              variance = var(bill_length_mm))

summary_stats2
```

```
## # A tibble: 2 x 6
##   species       n median  mean stdev variance
##   <fct>     <int>  <dbl> <dbl> <dbl>    <dbl>
## 1 Adelie       73   40.6  40.4  2.28     5.19
## 2 Chinstrap    34   51.0  51.1  1.56     2.45
```

I am fairly certain these data are independently sampled, though I can not be sure because I did not collect the dataset. Theoretically, if the penguins were tagged and the same penguins tracked from year-to-year, then the measurements would not be independent. However, there is no ‘penguin ID‘ variable in the dataset, which would be necessary for tracking that type of data. So I feel comfortable with the assumption that these are independent samples.

Based on the density plot and the descriptive statistics, the two groups have equal variance.

Based on the QQ plot, these data are normally distributed.

For these reasons, I will run a pooled t-test with $\alpha = 0.05$.

E. Run the test and interpret the results

```
penguin_t <- t.test(bill_length_mm ~ species, data = df, var.equal = TRUE)
penguin_t
```

```
##
##  Two Sample t-test
##
## data:  bill_length_mm by species
## t = -24.789, df = 105, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group Adelie and group Chinstrap is not equa
## 95 percent confidence interval:
##  -11.559886  -9.847527
## sample estimates:
##    mean in group Adelie mean in group Chinstrap
##                40.39041                51.09412
```

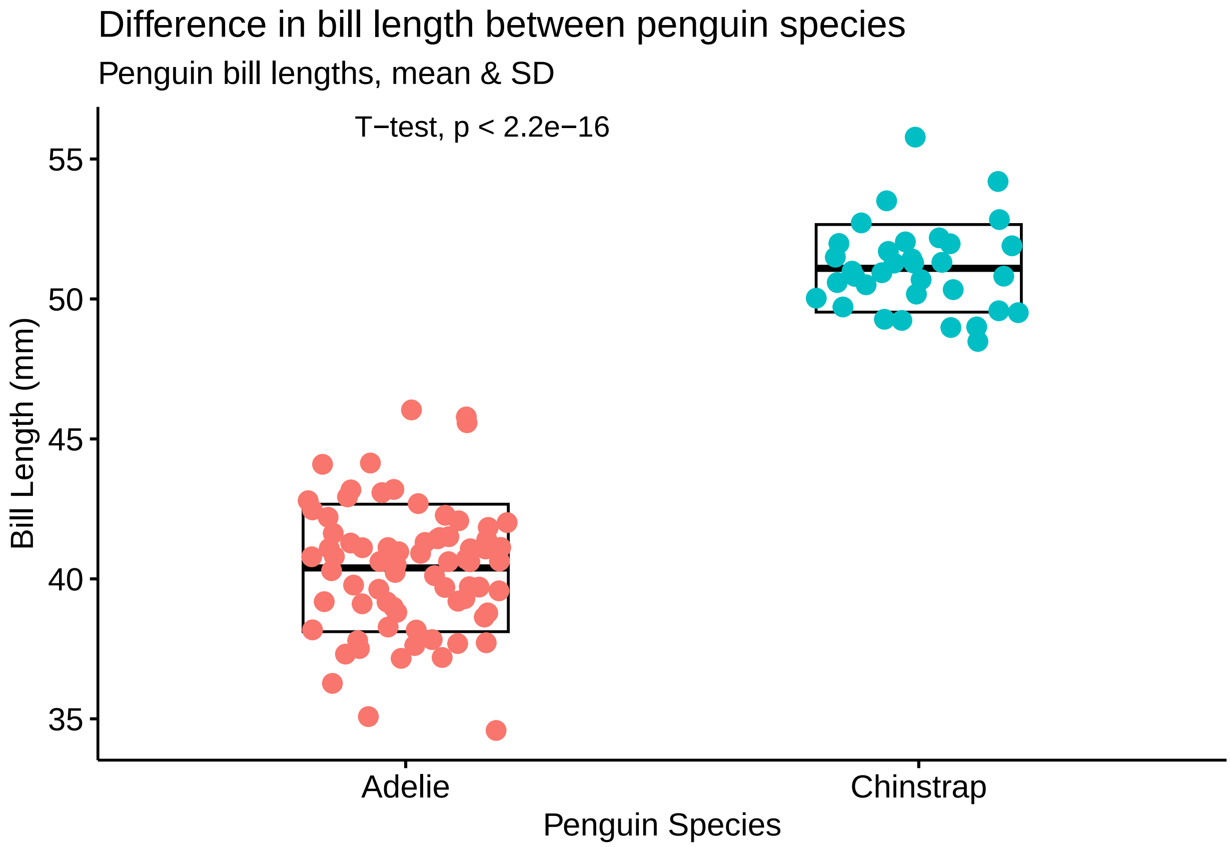
Mean bill length is significantly lower in Adelie penguins than in Chinstrap penguins ($p = 1.1041861 \times 10^{-45}$, which is less than $\alpha = 0.05$).

F. Which descriptive statistics would you use to describe the results (mean/sd or median/IQR)?

Mean and standard deviation should be used to describe these results, because the t-test compares the means of two groups. Median and IQR describe distributions of data, so they aren't appropriate to combine with the results of the t-test.

```
# use ggpubr to make pretty plot
# display individual data points, group means & stdev, and the t-test result
p2 <- df |>
  ggstripchart(
  # assign aes() parameters
      x = "species", y = "bill_length_mm",
      color = "species", size = 3,
  # plot mean & stdev overlaid on stripchart of individual datapoints
      add = "mean_sd",
  # format plot visualization of mean & stdev
      add.params = list(color = "black", width = 0.4), error.plot = "crossbar",
  # set plot labels
      title = "Difference in bill length between penguin species",
      subtitle = "Penguin bill lengths, mean & SD",
      xlab = "Penguin Species",
      ylab = "Bill Length (mm)") +
  # hide legend
      rremove("legend") +
  # annotate plot with pvalue from pooled, unpaired t-test
      stat_compare_means(method = "t.test",
                         method.args = list(var.equal = TRUE, paired = FALSE),
                         label.x.npc = "left")

p2
```


Difference in bill length between penguin species
Penguin bill lengths, mean & SD
T−test, p < 2.2e−16
Bill Length (mm)
55
50
45
40
35
Adelie
Chinstrap
Penguin Species