Selected Nonparametric Methods

Lecture 11

1 Outline

- Nonparametric Statistics
- Test for Normality
- One-Sample Wilcoxon Test
- Paired-Sample Test
- Two-Sample Test
- Kruskal-Wallis Test
- Spearman Correlation

2 Nonparametric Statistics

2.1 What is Non-Parametric Statistics?

Most of the inferential statistical procedures we have discussed so far can be classified as *parametric methods*.

> **Note**
>
> Parametric methods are based on statistical parameters like the mean, proportion, or variance.
> An examle of a test that did not use any statistical parameter assumption are the chi-square tests.

2.2 Nonparametric Tests

2.2.1 Advantages

Unknown form of sampled population

Nonparametric tests allow us to develop tests that do not involve parameters. Hence, they may be used when the form of the sampled population is unknown.

Small Sample Size

Nonparametric tests can be applied to small data sets and are generally easy to do computationally.

2.2.2 Disadvantages

Lower power

Nonparametric tests may have less power compared to some parametric tests.

Economics of Scale

Nonparametric tests may be tedious to apply to large data sets.

Ties

For rank-based tests, some methods assume that there are no "ties" in ranking, hence adjustments need to be performed if ties are present.

3 Tests for Normality

3.1 Formal Tests for Normality

There are nonparametric tests that can be used to test if a variable measured in a data set follow a Gaussian distribution.

> **Note**
>
> The Shapiro-Wilks test is a common test for normality.
> The Kolgomorov-Smirnov (K-S),Anderson-Darling and Lilliefors tests are also used.

3.1.1 Notes about K-S

The K-S test can be used for any given reference probability distribution or to compare if two samples came from the same distribution.

3.1.2 Lilliefors Test

The Lilliefors test is an improvement on the K-S test for the normality assumption.

3.2 Statistical Hypotheses

> **Important**
>
> For all these tests, the null hypothesis is that the data follow the normal distribution.
> The alternative hypothesis is that the data does not follow the normal distribution.
> Rejecting the null hypothesis means we have sufficient evidence of a deviation from the normal distribution.

3.3 `R` implementation

The Kolgomorov-Smirnov test can be implemented using the `ks.test()` function in `R`.

4 One-Sample tests

4.1 Wilcoxon Signed-Rank Test for Location

The Wilcoxon Signed-Rank Test for Location is used when we want to test a mean, but neither z or t are appropriate.

Assumptions

The Wilcoxon test assumes the following about the data:

- The sample is random
- The random variable is continuous
- The population is symmetric about its mean.
- The measurement scale is at least interval.

4.2 Statistical Hypotheses

Unlike the one-sample t-test or z-test that hypothesize about the mean, the hypothesis of the one-sample Wilcoxon test involves the median (assuming the symmetry assumption holds).

The null hypothesis is that the median of the data set, $\tilde{\mu}$ is equal to a null value $\tilde{\mu}_0$.

$$H_0 : \tilde{\mu} = \tilde{\mu}_0$$

The alternative can be one-sided ($\tilde{\mu} > \tilde{\mu}_0$ or $\tilde{\mu} < \tilde{\mu}_0$) or two-sided ($\tilde{\mu} \neq \tilde{\mu}_0$).

4.3 How it Works

- The difference between the data points and the null value are calculated
- These differences are then ranked from highest to lowest. If the difference is zero, it is ignored.
- Reapply the original signs to these ranks
- Sum the positive and negative ranks to obtain the test statistic.
- The test statistic is the lower value between the absolute value of the sums of the positive and negative ranks.

4.4 `R` Implementation

The `wilcox.test()` function can be used to perform a one-sample Wilcoxon-test in `R`. For a vector `x` in a data frame `df` and a null value `m0`, an example code of a two-sided test can be written as

```
wilcox.test(df$x,mu=mu0,alternative="two.sided")
```

4.5 Example

Consider the sleep health data set in `SleepHealthData.csv`.

```
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   4.0.0     v tibble    3.3.0
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.1.0
-- 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 becor
```

```
sleep <- read.csv("SleepHealthData.csv")
glimpse(sleep)
```

```
Rows: 374
Columns: 13
$ person_id               <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14,~
$ gender                  <chr> "Male", "Male", "Male", "Male", "Male", "Male"~
$ age                     <int> 27, 28, 28, 28, 28, 28, 29, 29, 29, 29, 29, 29~
$ occupation              <chr> "Software Engineer", "Doctor", "Doctor", "Sale~
$ sleep_duration          <dbl> 6.1, 6.2, 6.2, 5.9, 5.9, 5.9, 6.3, 7.8, 7.8, 7~
$ quality_of_sleep        <int> 6, 6, 6, 4, 4, 4, 6, 7, 7, 7, 6, 7, 6, 6, 6, 6~
$ physical_activity_level <int> 42, 60, 60, 30, 30, 30, 40, 75, 75, 75, 30, 75~
$ stress_level            <int> 6, 8, 8, 8, 8, 8, 7, 6, 6, 6, 8, 6, 8, 8, 8, 8~
$ bmi_category            <chr> "Overweight", "Normal", "Normal", "Obese", "Ob~
$ blood_pressure          <chr> "126/83", "125/80", "125/80", "140/90", "140/9~
$ heart_rate              <int> 77, 75, 75, 85, 85, 85, 82, 70, 70, 70, 70, 70~
$ daily_steps             <int> 4200, 10000, 10000, 3000, 3000, 3000, 3500, 80~
$ sleep_disorder          <chr> "None", "None", "None", "Sleep Apnea", "Sleep ~
```

4.5.1 Question

Use the one-sample Wilcoxon test to test whether the median stress level for the sampled population is higher than 4.5.

4.5.2 Answer

The statistical hypothesis can be written as:

$$H_0 : \tilde{\mu} = 4.5; H_a : \tilde{\mu} > 4.5$$

```
wtest <- wilcox.test(sleep$stress_level,mu=4.5,alternative = "greater")
wtest
```

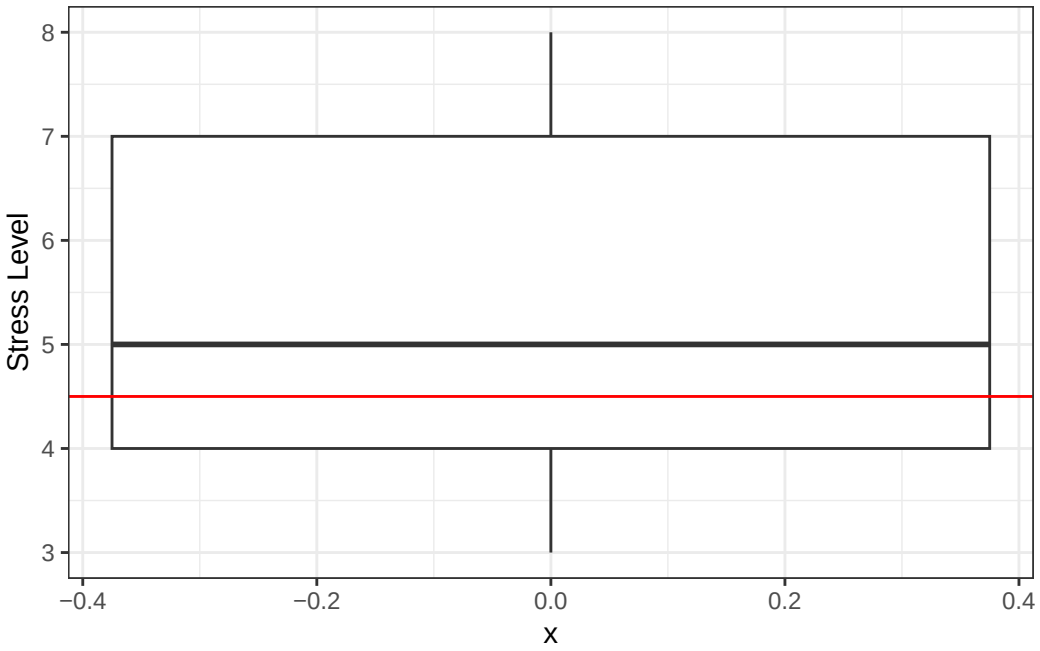
```
    Wilcoxon signed rank test with continuity correction

data:  sleep$stress_level
V = 51379, p-value = 1.556e-15
alternative hypothesis: true location is greater than 4.5
```

The p-value is $1.5555249 \times 10^{-15}$, which is less than 0.05. We reject the null hypothesis. We have sufficient evidence that the median stress level of the sampled population is greater than 4.5.

We can also visually check if this makes sense using a box plot

```
ggplot(data=sleep,aes(x=NULL,y=stress_level)) +
  geom_boxplot() +
  geom_hline(aes(yintercept=4.5),color="red")+
  theme_bw() +
  labs(y="Stress Level")
```



4.6 Notes

If the assumption about the symmetry of the data set is not satisfied, the Wilcoxon test might have faulty results about the inference involving medians.

4.6.1 Warning

> **Warning**
>
> Note that the test statistic is based on the ranked differences and that zero differences are ignored. This means that if the data is asymmetric and the assumed median is part of the data set, it might lead to a significant result even if the medians are the same.
> If we are interested in a pure test of medians for a skewed sample, the **sign test** might be a better alternative.

4.6.2 Sign test

The `BSDA` package includes the `SIGN.test()` function that can be used to run the sign test.

4.7 Exercise

Consider the sleep health data set in `SleepHealthData.csv`.

```
library(tidyverse)
sleep <- read.csv("SleepHealthData.csv")
glimpse(sleep)
```

```
Rows: 374
Columns: 13
$ person_id               <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14,~
$ gender                  <chr> "Male", "Male", "Male", "Male", "Male", "Male"~
$ age                     <int> 27, 28, 28, 28, 28, 28, 29, 29, 29, 29, 29, 29~
$ occupation              <chr> "Software Engineer", "Doctor", "Doctor", "Sale~
$ sleep_duration          <dbl> 6.1, 6.2, 6.2, 5.9, 5.9, 5.9, 6.3, 7.8, 7.8, 7~
$ quality_of_sleep        <int> 6, 6, 6, 4, 4, 4, 6, 7, 7, 7, 6, 7, 6, 6, 6, 6~
$ physical_activity_level <int> 42, 60, 60, 30, 30, 30, 40, 75, 75, 75, 30, 75~
$ stress_level            <int> 6, 8, 8, 8, 8, 8, 7, 6, 6, 6, 8, 6, 8, 8, 8, 8~
$ bmi_category            <chr> "Overweight", "Normal", "Normal", "Obese", "Ob~
$ blood_pressure          <chr> "126/83", "125/80", "125/80", "140/90", "140/9~
$ heart_rate              <int> 77, 75, 75, 85, 85, 85, 82, 70, 70, 70, 70, 70~
$ daily_steps             <int> 4200, 10000, 10000, 3000, 3000, 3000, 3500, 80~
$ sleep_disorder          <chr> "None", "None", "None", "Sleep Apnea", "Sleep ~
```

4.7.1 Question

Use the one-sample Wilcoxon test to test whether the median physical activity level for the sampled population is not equal to 60.

4.7.2 Answer

The statistical hypothesis can be written as:

$$H_0 : \tilde{\mu} = 60; H_a : \tilde{\mu} \neq 60$$

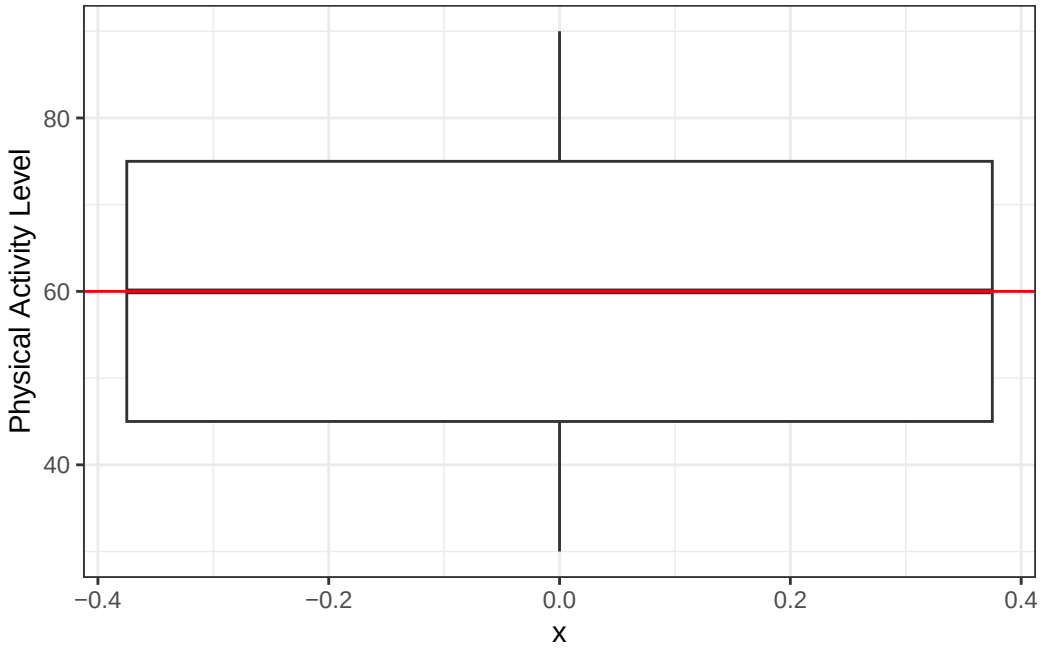
```
wtest <- wilcox.test(sleep$physical_activity_level,mu=60,alternative = "two.sided")
wtest
```

```
    Wilcoxon signed rank test with continuity correction

data:  sleep$physical_activity_level
V = 22196, p-value = 0.5121
alternative hypothesis: true location is not equal to 60
```

We can also visually check if this makes sense using a box plot

```
ggplot(data=sleep,aes(x=NULL,y=physical_activity_level)) +
  geom_boxplot() +
  geom_hline(aes(yintercept=60),color="red")+
  theme_bw() +
  labs(y="Physical Activity Level")
```



5 Paired-Samples Test

5.1 Paired-Samples Test

the Wilcoxon signed-rank test can be used to test for a difference between paired samples.

> **Note**
>
> This is the non-parametric equivalent of a paired t-test.

5.2 Statistical Hypotheses

The hypothesis of the paired-sample Wilcoxon signed rank test involves the median of the paired difference.

The null hypothesis is that the median of the data set, $\tilde{\mu}_d$ is equal to a null value $\tilde{\mu}_{d,0}$.

$$H_0 : \tilde{\mu}_d = \tilde{\mu}_{d,0}$$

The alternative can be one-sided ($\tilde{\mu}_d > \tilde{\mu}_{d,0}$ or $\tilde{\mu}_d < \tilde{\mu}_{d,0}$) or two-sided ($\tilde{\mu}_d \neq \tilde{\mu}_{d,0}$).

> **Tip**
>
> The null value commonly used is 0 (null hypothesis of equality).

5.3 `R` Implementation

The `wilcox.test()` can also be used for paired data. For a data frame `df` with paired data `x1` and `x2` and a one-sided "less" alternative hypothesis , the sample code would look like:

```
wilcox.test(x=df$x1,y=df$x2,paired=T,alternative="less")
```

5.4 Example

The following data set `example` shows data from research that investigated the possible beneficial effects of singing on wellbeing during a single singing lesson. One of the variables of interest was the change in cortisol as a result of the singing lesson.

```
example <- data.frame(
  before=c(214,362,202,158,403,219,307,331),
  after=c(232,276,224,412,562,203,340,313)
)
glimpse(example)
```

```
Rows: 8
Columns: 2
$ before <dbl> 214, 362, 202, 158, 403, 219, 307, 331
$ after  <dbl> 232, 276, 224, 412, 562, 203, 340, 313
```

5.4.1 Question

Use the Wilcoxon Signed-Rank Test to test whether the cortisol levels increased after singing.

5.4.2 Answer

The alternative hypothesis is one-sided because of the direction assumed in the problem. We assume that $\tilde{\mu}_d$ is calculated as `before-after` Hence, the hypotheses can be written as:

$$H_0 : \tilde{\mu}_d = 0; H_0 : \tilde{\mu}_d > 0$$

```
wtest <- wilcox.test(x=example$before,y=example$after,paired=T,alternative="less",correct=TRU
```

```
Warning in wilcox.test.default(x = example$before, y = example$after, paired =
T, : cannot compute exact p-value with ties
```

```
wtest
```

```
    Wilcoxon signed rank test with continuity correction

data:  example$before and example$after
V = 9.5, p-value = 0.131
alternative hypothesis: true location shift is less than 0
```

The resulting Wilcoxon test p-value is 0.1310164. We fail to reject the null hypothesis at the significance level of 0.05. We have insufficient evidence to claim that the cortisol levels increased after singing.

5.5 Exercise

The data set `anorexia` from the package `MASS` contains weight change data for young female anorexia patients.

```
library(MASS)
```

```
Attaching package: 'MASS'

The following object is masked from 'package:dplyr':

    select
```

```
library(tidyverse)
glimpse(anorexia)
```

```
Rows: 72
Columns: 3
$ Treat  <fct> Cont, Cont, Cont, Cont, Cont, Cont, Cont, Cont, Cont, Cont, Con~
$ Prewt  <dbl> 80.7, 89.4, 91.8, 74.0, 78.1, 88.3, 87.3, 75.1, 80.6, 78.4, 77.~
$ Postwt <dbl> 80.2, 80.1, 86.4, 86.3, 76.1, 78.1, 75.1, 86.7, 73.5, 84.6, 77.~
```

5.5.1 Question

Perform a paired-sample non-parametric test to test whether there is a difference between the preweight `prewt` and `postwt` for all patients, disregarding the type of treatment. Use a significance level of 0.05.

5.5.2 Answer

The alternative hypothesis is two-sided. We assume that $\tilde{\mu}_d$ is calculated as `Postwt-Prewt` Hence, the hypotheses can be written as:

$$H_0 : \tilde{\mu}_d = 0; H_0 : \tilde{\mu}_d \neq 0$$

```
wtest <- wilcox.test(x=anorexia$Postwt,y=anorexia$Prewt,paired=T,alternative="two.sided")
wtest
```

```
    Wilcoxon signed rank test with continuity correction

data:  anorexia$Postwt and anorexia$Prewt
V = 1724.5, p-value = 0.0106
alternative hypothesis: true location shift is not equal to 0
```

The p-value is 0.0106022. We reject the null hypothesis at the 0.05 level of significance. There is sufficient evidence to conclude that there is a difference in post-weight and pre-weight scores across the sample.

6 Two-Sample Test

6.1 Mann-Whitney U Test

The Mann-Whitney U Test is equivalent to the two-sample Wilcoxon test.

> **Assumptions**
>
> The Mann-Whitney test assumes the following:
>
> - The two samples are independent and randomly drawn from their respective populations
> - The measurement scale is at least ordinal
> - The variable of interest is continuous
> - If the populations differ at all, they differ only with respect to their medians.

6.2 Statistical Hypotheses

The null hypothesis for the Mann-Whitney U test is that there is no difference between the distribution of the two independent groups.

> **Note**
>
> The distribution can be measured using the median or sometimes, the mean of the ranks.

The alternative hypotheses can be:

- One-sided: One group tends to have higher or lower values than the other
- Two-sided: There is a difference between the distributions of the two groups.

6.3 The U Statistic

To calculate the test statistic:

- Assign ranks to the values from the two groups pooled together in order from smallest to largest.
- Sum the ranks of each group.

- Calculate the U_j value as:

$$U_1 = n_1 n_2 + n_1(n_1 + 1)/2 - \sum_i R_{i1}$$

$$U_2 = n_1 n_2 + n_2(n_2 + 1)/2 - \sum_i R_{i2}$$

- The test statistic is the lower value between U_1 and U_2.

6.4 R implementation

The `wilcox.test()` function can implement the Mann-Whitney U test for independent groups. The syntax is similar to the paired-samples test but specifying that `paired=FALSE`.

```
wilcox.test(x=df$x1,y=df$x2,alternative="less")
```

If the grouping variable and the response variables are defined as two distinct columns, we can use the formula notation as well.

```
wilcox.test(response~grouping,data=df,alternative="less")
```

6.5 Example

Consider the data set `SleepHealthData.csv`.

```
sleep <- read.csv("SleepHealthData.csv")
glimpse(sleep)
```

```
Rows: 374
Columns: 13
$ person_id               <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14,~
$ gender                  <chr> "Male", "Male", "Male", "Male", "Male", "Male"~
$ age                     <int> 27, 28, 28, 28, 28, 28, 29, 29, 29, 29, 29, 29~
$ occupation              <chr> "Software Engineer", "Doctor", "Doctor", "Sale~
$ sleep_duration          <dbl> 6.1, 6.2, 6.2, 5.9, 5.9, 5.9, 6.3, 7.8, 7.8, 7~
$ quality_of_sleep        <int> 6, 6, 6, 4, 4, 4, 6, 7, 7, 7, 6, 7, 6, 6, 6, 6~
$ physical_activity_level <int> 42, 60, 60, 30, 30, 30, 40, 75, 75, 75, 30, 75~
$ stress_level            <int> 6, 8, 8, 8, 8, 8, 7, 6, 6, 6, 8, 6, 8, 8, 8, 8~
$ bmi_category            <chr> "Overweight", "Normal", "Normal", "Obese", "Ob~
```

```
$ blood_pressure          <chr> "126/83", "125/80", "125/80", "140/90", "140/9~
$ heart_rate              <int> 77, 75, 75, 85, 85, 85, 82, 70, 70, 70, 70, 70~
$ daily_steps             <int> 4200, 10000, 10000, 3000, 3000, 3000, 3500, 80~
$ sleep_disorder          <chr> "None", "None", "None", "Sleep Apnea", "Sleep ~
```

6.5.1 Question

Use the Mann-Whitney U test to determine if there is a difference in quality of sleep between males and females.

6.5.2 Answer

The alternative hypothesis should be two-sided because of the lack of directionality specified in the difference that we want to detect.

The null hypothesis is that there is no difference in quality of sleep between males and females, while the alternative hypothesis is that there is a difference in quality of sleep between males and females.

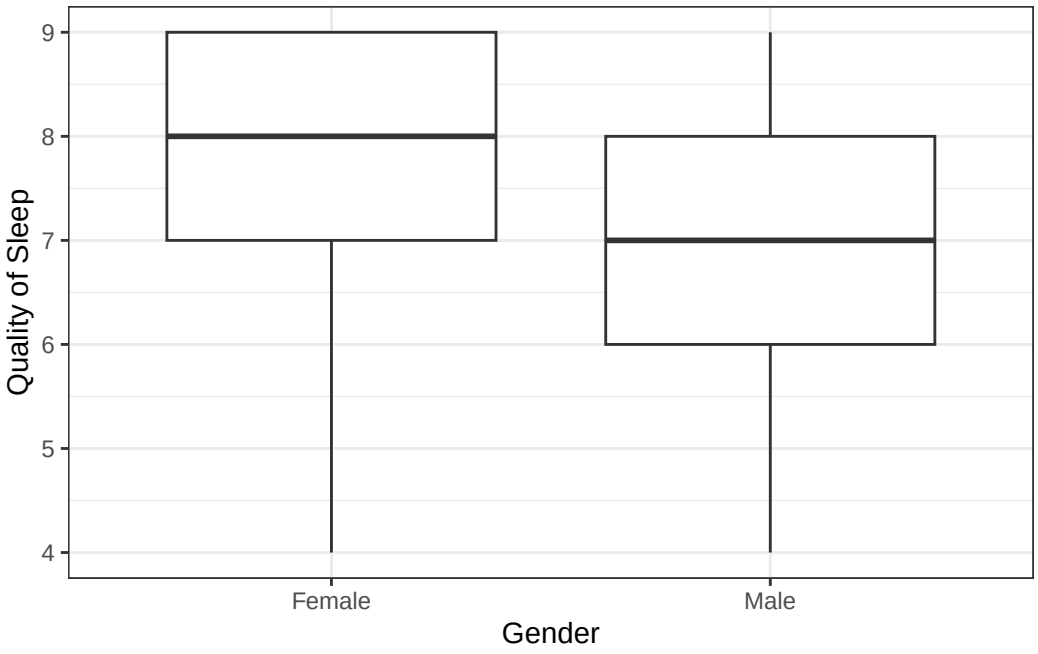
```
wtest<- wilcox.test(quality_of_sleep~gender,data=sleep,alternative="two.sided")
wtest
```

```
    Wilcoxon rank sum test with continuity correction

data:  quality_of_sleep by gender
W = 23369, p-value = 6.048e-09
alternative hypothesis: true location shift is not equal to 0
```

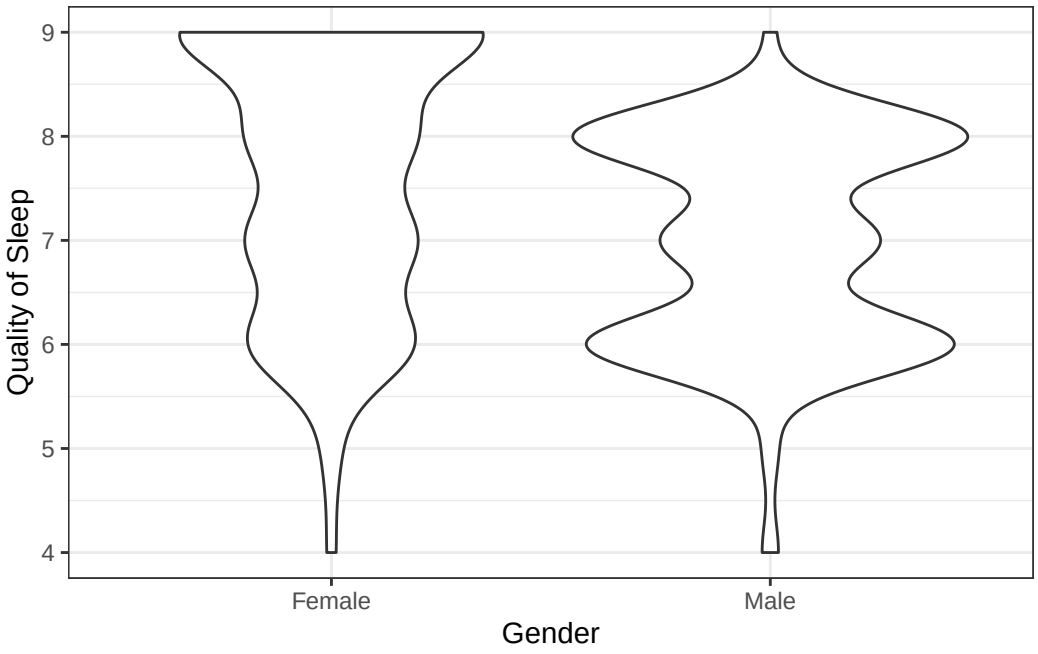
The resulting p-value is 6.0476062×10^{-9}. We reject the null hypothesis. We have sufficient evidence to conclude that there is a difference in quality of sleep between males and females.

```
ggplot(data=sleep,aes(x=gender,y=quality_of_sleep))+
  geom_boxplot() +
  theme_bw() +
  labs(y="Quality of Sleep",x="Gender")
```



6.5.3 Violin Plot

```
ggplot(data=sleep,aes(x=gender,y=quality_of_sleep))+
  geom_violin() +
  theme_bw() +
  labs(y="Quality of Sleep",x="Gender")
```



6.6 Exercise

Consider the sleep health data set `SleepHealthData.csv`.

```
sleep <- read.csv("SleepHealthData.csv")
glimpse(sleep)
```

```
Rows: 374
Columns: 13
$ person_id               <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14,~
$ gender                  <chr> "Male", "Male", "Male", "Male", "Male", "Male"~
$ age                     <int> 27, 28, 28, 28, 28, 28, 29, 29, 29, 29, 29, 29~
$ occupation              <chr> "Software Engineer", "Doctor", "Doctor", "Sale~
$ sleep_duration          <dbl> 6.1, 6.2, 6.2, 5.9, 5.9, 5.9, 6.3, 7.8, 7.8, 7~
$ quality_of_sleep        <int> 6, 6, 6, 4, 4, 4, 6, 7, 7, 7, 6, 7, 6, 6, 6, 6~
$ physical_activity_level <int> 42, 60, 60, 30, 30, 30, 40, 75, 75, 75, 30, 75~
$ stress_level            <int> 6, 8, 8, 8, 8, 8, 7, 6, 6, 6, 8, 6, 8, 8, 8, 8~
$ bmi_category            <chr> "Overweight", "Normal", "Normal", "Obese", "Ob~
$ blood_pressure          <chr> "126/83", "125/80", "125/80", "140/90", "140/9~
$ heart_rate              <int> 77, 75, 75, 85, 85, 85, 82, 70, 70, 70, 70, 70~
$ daily_steps             <int> 4200, 10000, 10000, 3000, 3000, 3000, 3500, 80~
$ sleep_disorder          <chr> "None", "None", "None", "Sleep Apnea", "Sleep ~
```

6.6.1 Question

Use the Mann-Whitney U test to determine if there is a difference in daily steps between males and females.

6.6.2 Answer

The alternative hypothesis should be two-sided because of the lack of directionality specified in the difference that we want to detect.

The null hypothesis is that there is no difference in daily number of steps between males and females, while the alternative hypothesis is that there is a difference in daily number of steps between males and females.

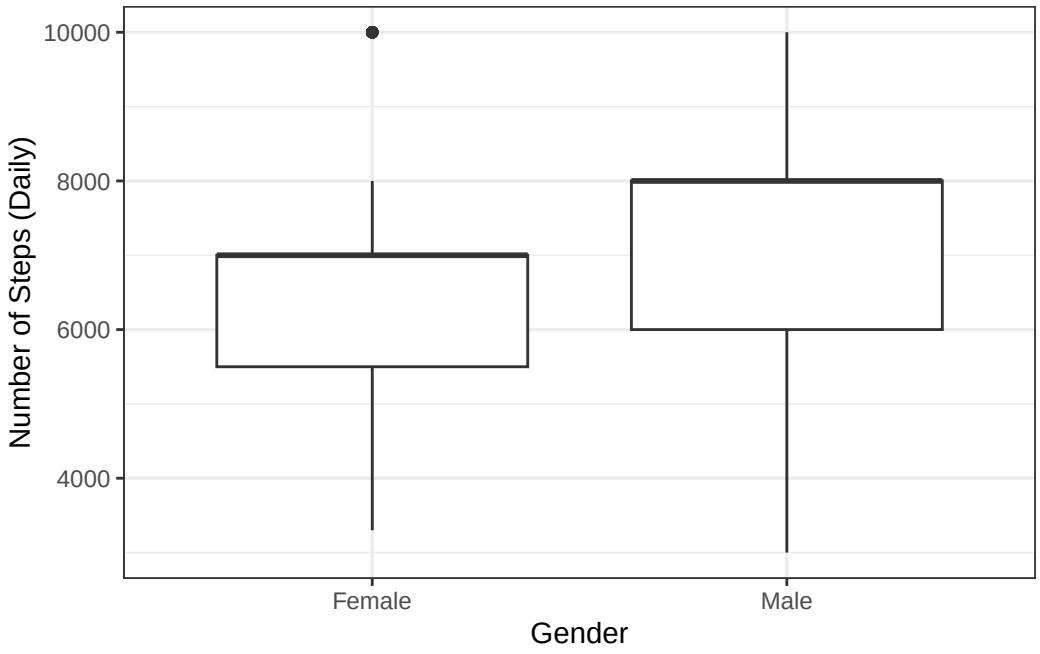
```
wtest<- wilcox.test(daily_steps~gender,data=sleep,alternative="two.sided")
wtest
```

```
    Wilcoxon rank sum test with continuity correction

data:  daily_steps by gender
W = 15890, p-value = 0.1203
alternative hypothesis: true location shift is not equal to 0
```

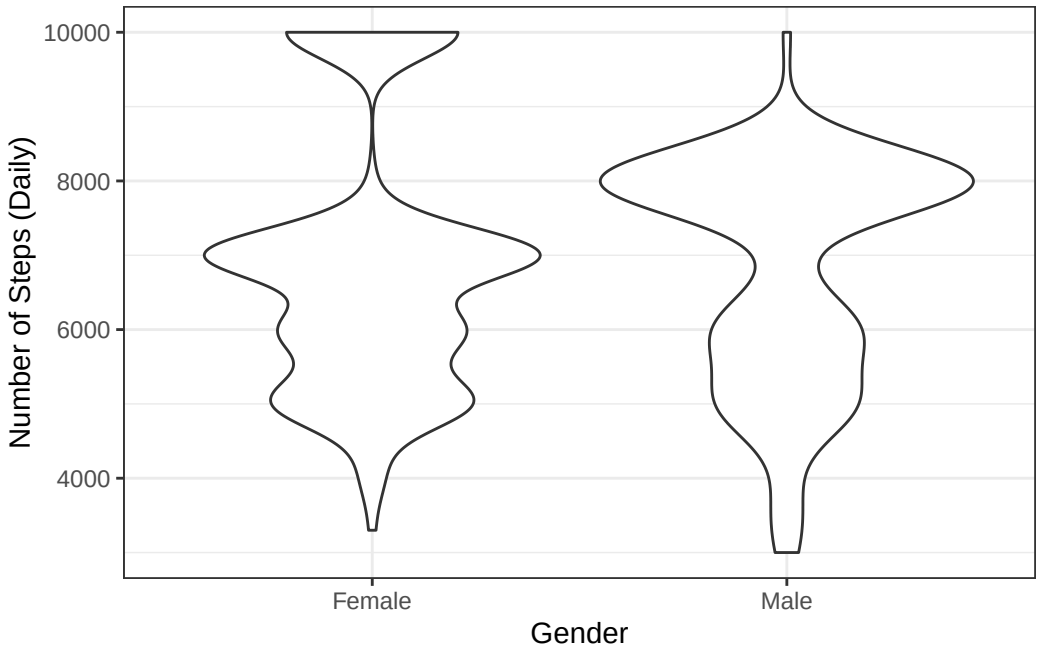
The resulting p-value is 0.1203365. We fail to reject the null hypothesis at the 0.05 level of significance. We have insufficient evidence to conclude that there is a difference in number of daily steps between males and females.

```
ggplot(data=sleep,aes(x=gender,y=daily_steps))+
  geom_boxplot() +
  theme_bw() +
  labs(y="Number of Steps (Daily)",x="Gender")
```



6.6.3 Violin Plot

```
ggplot(data=sleep,aes(x=gender,y=daily_steps))+
  geom_violin() +
  theme_bw() +
  labs(y="Number of Steps (Daily)",x="Gender")
```



7 ANOVA-like Tests

7.1 ANOVA-like tests

There are non-parametric tests that compare the distributions of more than two groups.

> **!** Important
>
> The Kruskal-Wallis Rank Sum test works similarly to the one-way ANOVA, while the Friedman's test works similarly to the two-way ANOVA. Both use assumptions that are the same as the Mann-Whitney test for two independent samples.

7.2 Statistical Hypotheses: Kruskal-Wallis Test

The null hypothesis of the Kruskal-Wallis Rank Sum Test states that there is no difference in the distribution of the response variable across the different groups.

The alternative hypothesis states that there is at least one group that has a different distribution of the response variable across the different groups.

> **Note**
>
> This can be stated in terms of medians, similar to how ANOVA was stated in terms of means.

7.3 R implementation

The `kruskal.test()` function implements the Kruskal-Wallis test. The function works best with the formula notation.

```
kruskal.test(response~group,data=df)
```

7.4 Example

Consider the `penguins` data set in R.

```
library(tidyverse)
glimpse(penguins)
```

```
Rows: 344
Columns: 8
$ species     <fct> Adelie, Adelie, Adelie, Adelie, Adelie, Adelie, Adelie, Ad~
$ island      <fct> Torgersen, Torgersen, Torgersen, Torgersen, Torgersen, Tor~
$ bill_len    <dbl> 39.1, 39.5, 40.3, NA, 36.7, 39.3, 38.9, 39.2, 34.1, 42.0, ~
$ bill_dep    <dbl> 18.7, 17.4, 18.0, NA, 19.3, 20.6, 17.8, 19.6, 18.1, 20.2, ~
$ flipper_len <int> 181, 186, 195, NA, 193, 190, 181, 195, 193, 190, 186, 180,~
$ body_mass   <int> 3750, 3800, 3250, NA, 3450, 3650, 3625, 4675, 3475, 4250, ~
$ sex         <fct> male, female, female, NA, female, male, female, male, NA, ~
$ year        <int> 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007~
```

7.4.1 Question

Use the Kruskal-Wallis test to test for a difference between the body mass across the different years (2007,2008,2009).

7.4.2 Answer

The null hypothesis is that there is no difference in the distribution of body mass of the penguins across the different years. The alternative is that at least one year is different from the others.

```
ktest <- kruskal.test(body_mass~as.factor(year),data=penguins)
ktest
```

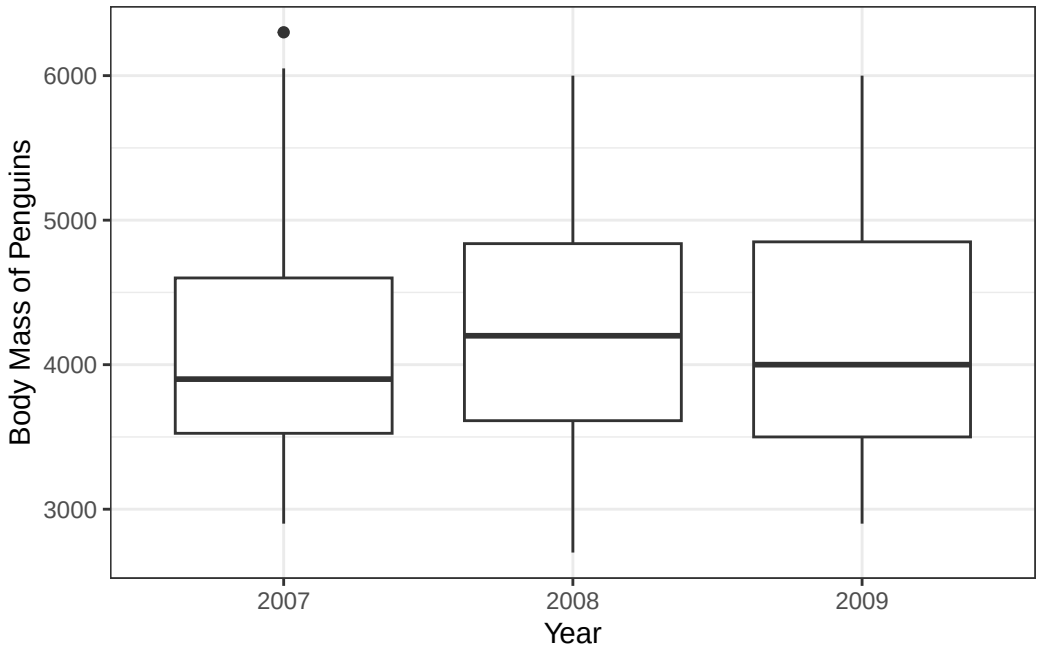
```
    Kruskal-Wallis rank sum test

data:  body_mass by as.factor(year)
Kruskal-Wallis chi-squared = 2.5523, df = 2, p-value = 0.2791
```

The resulting p-value is 0.279111. At 0.05 level of significance, we fail to reject the null hypothesis. We have insufficient evidence that there is at least one year with a different distribution of body mass compared to the others.

```
ggplot(data=penguins,aes(x=as.factor(year),y=body_mass)) +
  geom_boxplot() +
  theme_bw() +
  labs(x="Year",y="Body Mass of Penguins")
```

```
Warning: Removed 2 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```

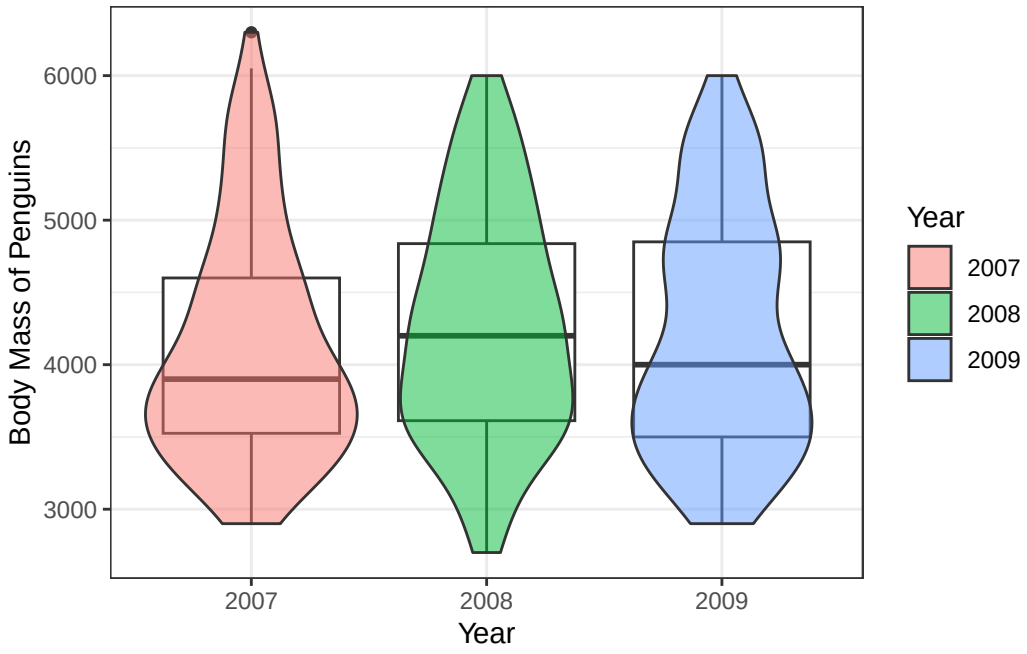


7.4.3 Violin Plot

```
ggplot(data=penguins,aes(x=as.factor(year),y=body_mass)) +
  geom_boxplot() +
  geom_violin(aes(fill=as.factor(year)),alpha=0.5) +
  theme_bw() +
  labs(x="Year",y="Body Mass of Penguins",fill="Year")
```

```
Warning: Removed 2 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```

```
Warning: Removed 2 rows containing non-finite outside the scale range
(`stat_ydensity()`).
```



7.5 Exercise

Consider the `iris` data set in `R`.

```
glimpse(iris)
```

```
Rows: 150
Columns: 5
$ Sepal.Length <dbl> 5.1, 4.9, 4.7, 4.6, 5.0, 5.4, 4.6, 5.0, 4.4, 4.9, 5.4, 4.~
$ Sepal.Width  <dbl> 3.5, 3.0, 3.2, 3.1, 3.6, 3.9, 3.4, 3.4, 2.9, 3.1, 3.7, 3.~
$ Petal.Length <dbl> 1.4, 1.4, 1.3, 1.5, 1.4, 1.7, 1.4, 1.5, 1.4, 1.5, 1.5, 1.~
$ Petal.Width  <dbl> 0.2, 0.2, 0.2, 0.2, 0.2, 0.4, 0.3, 0.2, 0.2, 0.1, 0.2, 0.~
$ Species      <fct> setosa, setosa, setosa, setosa, setosa, setosa, setosa, s~
```

7.5.1 Question

Use the Kruskal-Wallis test to test for a difference between the petal widths of the species of iris.

7.5.2 Answer

The null hypothesis is that there is no difference in petal widths of the species of iris. the alternative hypothesis is that there is at least one species that has a different distribution of petal widths compared to the others.

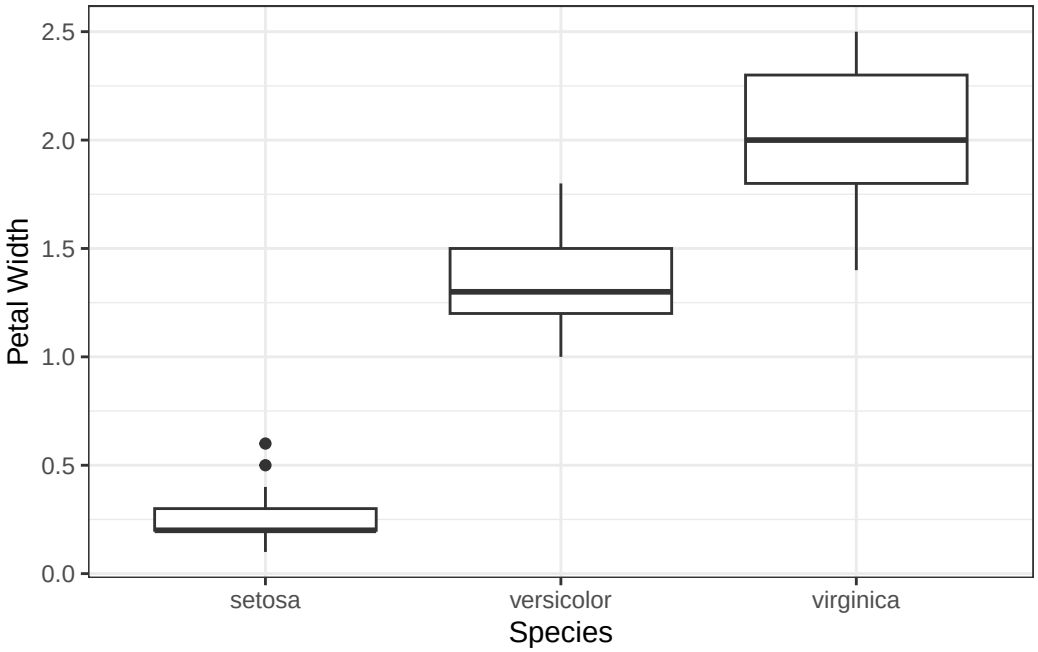
```
ktest <- kruskal.test(Petal.Width~Species,data=iris)
ktest
```

```
    Kruskal-Wallis rank sum test

data:  Petal.Width by Species
Kruskal-Wallis chi-squared = 131.19, df = 2, p-value < 2.2e-16
```

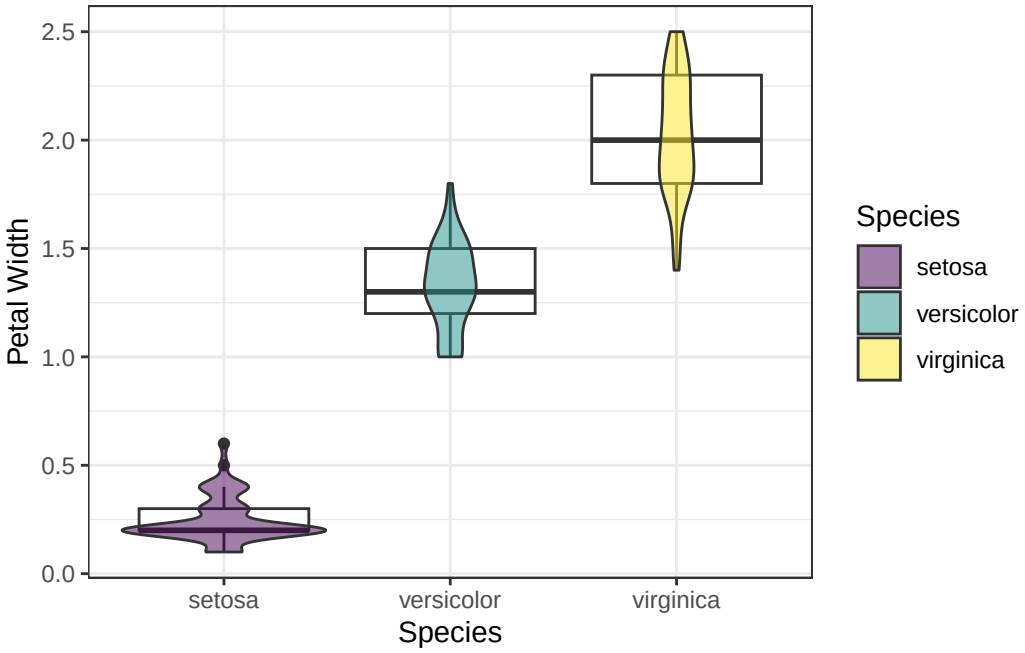
The p-value is $3.2617956 \times 10^{-29}$. At the 0.05 level of significance, we reject the null hypothesis. We have sufficient evidence that at least one species has a different distribution of petal widths compared to the others.

```
ggplot(data=iris,aes(x=Species,y=Petal.Width)) +
  geom_boxplot() +
  theme_bw() +
  labs(x="Species",y="Petal Width")
```



7.5.3 Violin Plot

```
ggplot(data=iris,aes(x=Species,y=Petal.Width)) +
  geom_boxplot() +
  geom_violin(aes(fill=Species),alpha=0.5) +
  scale_fill_discrete(palette="viridis") +
  theme_bw() +
  labs(x="Species",y="Petal Width",fill="Species")
```



8 Non-Parametric Correlation Tests

8.1 Spearman Correlation

Recall Lecture 9: The Spearman correlation coefficient is a measure of association that measures monotonic correlation between the two variables.

> **!** Important
>
> The Spearman correlation coefficient is based on the data ranks, and can be used if the Pearson correlation assumptions are not met.

8.2 Statistical Hypotheses

The null hypothesis is that there is no association between the two variables. The alternative hypothesis can be one-sided (negative or positive association) or two-sided (there is an association).

8.3 R Implementation

The function `cor.test()` can perform a test for the Spearman correlation, as long as the method is specified.

> **Warning**
>
> If the method (`method="spearman"`) is unspecified, `cor.test()` defaults to the Pearson correlation coefficient.

```
cor.test(df$x,df$y,alternative="two.sided",method="spearman")
```

8.4 Example

Consider the sleep health data set `SleepHealthData.csv`.

```
library(tidyverse)
sleep <- read.csv("SleepHealthData.csv")
glimpse(sleep)
```

```
Rows: 374
Columns: 13
$ person_id               <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14,~
$ gender                  <chr> "Male", "Male", "Male", "Male", "Male", "Male"~
$ age                     <int> 27, 28, 28, 28, 28, 28, 29, 29, 29, 29, 29, 29~
$ occupation              <chr> "Software Engineer", "Doctor", "Doctor", "Sale~
$ sleep_duration          <dbl> 6.1, 6.2, 6.2, 5.9, 5.9, 5.9, 6.3, 7.8, 7.8, 7~
$ quality_of_sleep        <int> 6, 6, 6, 4, 4, 4, 6, 7, 7, 7, 6, 7, 6, 6, 6, 6~
$ physical_activity_level <int> 42, 60, 60, 30, 30, 30, 40, 75, 75, 75, 30, 75~
$ stress_level            <int> 6, 8, 8, 8, 8, 8, 7, 6, 6, 6, 8, 6, 8, 8, 8, 8~
$ bmi_category            <chr> "Overweight", "Normal", "Normal", "Obese", "Ob~
$ blood_pressure          <chr> "126/83", "125/80", "125/80", "140/90", "140/9~
$ heart_rate              <int> 77, 75, 75, 85, 85, 85, 82, 70, 70, 70, 70, 70~
$ daily_steps             <int> 4200, 10000, 10000, 3000, 3000, 3000, 3500, 80~
$ sleep_disorder          <chr> "None", "None", "None", "Sleep Apnea", "Sleep ~
```

8.4.1 Question

Calculate the Spearman correlation coefficient between the daily steps and sleep duration. Test for an association at the 0.01 level of significance.

8.4.2 Answer

The null hypothesis is that there is no association between daily steps and sleep duration, while the alternative hypothesis is that there is an association between daily steps and sleep duration.

```
ctest <- cor.test(sleep$daily_steps,sleep$sleep_duration,alternative="two.sided",method="spea
```

```
Warning in cor.test.default(sleep$daily_steps, sleep$sleep_duration,
alternative = "two.sided", : Cannot compute exact p-value with ties
```

```
ctest
```

```
    Spearman's rank correlation rho

data:  sleep$daily_steps and sleep$sleep_duration
S = 8630931, p-value = 0.8458
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.01008658
```

The estimate of the Spearman correlation coefficient is 0.0101, corresponding to a p-value of 0.8458492. We fail to reject the null hypothesis. There is insufficient evidence of an association between the daily number of steps and sleep duration.

8.5 Exercise

Consider the sleep health data set `SleepHealthData.csv`.

```
library(tidyverse)
sleep <- read.csv("SleepHealthData.csv")
glimpse(sleep)
```

```
Rows: 374
Columns: 13
$ person_id               <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14,~
$ gender                  <chr> "Male", "Male", "Male", "Male", "Male", "Male"~
$ age                     <int> 27, 28, 28, 28, 28, 28, 29, 29, 29, 29, 29, 29~
$ occupation              <chr> "Software Engineer", "Doctor", "Doctor", "Sale~
$ sleep_duration          <dbl> 6.1, 6.2, 6.2, 5.9, 5.9, 5.9, 6.3, 7.8, 7.8, 7~
$ quality_of_sleep        <int> 6, 6, 6, 4, 4, 4, 6, 7, 7, 7, 6, 7, 6, 6, 6, 6~
$ physical_activity_level <int> 42, 60, 60, 30, 30, 30, 40, 75, 75, 75, 30, 75~
$ stress_level            <int> 6, 8, 8, 8, 8, 8, 7, 6, 6, 6, 8, 6, 8, 8, 8, 8~
$ bmi_category            <chr> "Overweight", "Normal", "Normal", "Obese", "Ob~
$ blood_pressure          <chr> "126/83", "125/80", "125/80", "140/90", "140/9~
$ heart_rate              <int> 77, 75, 75, 85, 85, 85, 82, 70, 70, 70, 70, 70~
$ daily_steps             <int> 4200, 10000, 10000, 3000, 3000, 3000, 3500, 80~
$ sleep_disorder          <chr> "None", "None", "None", "Sleep Apnea", "Sleep ~
```

8.5.1 Question

Calculate the Spearman correlation coefficient between the daily steps and quality of sleep. Test for an association at the 0.01 level of significance.

8.5.2 Answer

The null hypothesis is that there is no association between daily steps and quality of sleep, while the alternative hypothesis is that there is an association between daily steps and quality of sleep

```
ctest <- cor.test(sleep$daily_steps,sleep$quality_of_sleep,alternative="two.sided",method="sp
```

```
Warning in cor.test.default(sleep$daily_steps, sleep$quality_of_sleep,
alternative = "two.sided", : Cannot compute exact p-value with ties
```

```
ctest
```

```

    Spearman's rank correlation rho

data:  sleep$daily_steps and sleep$quality_of_sleep
S = 8520266, p-value = 0.6606
alternative hypothesis: true rho is not equal to 0
```

```
sample estimates:
       rho
0.02277921
```

The estimate of the Spearman correlation coefficient is 0.0228, corresponding to a p-value of 0.6605808. We fail to reject the null hypothesis. There is insufficient evidence of an association between the daily number of steps and quality of sleep.