STATS 2107
Statistical Modelling and Inference II

Workshop 6: The `rstatix` package

Matt Ryan

School of Mathematical Sciences, University of Adelaide

Semester 2 2022

The `rstatix` package

What is it

The `rstatix` package provides a pipe-friendly (%>% this thing) interface for performing hypothesis test, that fits nicely into the `tidyverse` framework and philosophy.

This package allows for a single framework to do many different hypothesis tests, a full list of which can be found in here

Installing the package

This is available on CRAN, so you can install with

```r
install.packages("rstatix") # Install the package
library(rstatix) # Load the package
```

One more package

```r
install.packages("ggpubr") # Install the package
library(ggpubr) # Load the package
```

Some data to play with

For todays workshop, we will use the classic `iris` dataset by Fisher. Load this dataset and convert it into a tibble with the following code

```
data(iris)
iris <- as_tibble(iris)
```

What is the `iris` dataset

The `iris` dataset contains the following observations on the 50 flowers from 3 different species

variable	units
Sepal.Length	centimeters
Sepal.Width	centimeters
Petal.Length	centimeters
Petal.Width	centimeters
Species	Setosa, Veriscolor, Virginica

First taste of `rstatix`

We can use `rstatix` to easily generate summary statistics for our variables:

```
iris %>%
  get_summary_stats() %>%
  select(variable, min, q1,
         median, q3, max) # Just so it all fits
```

```
## # A tibble: 4 x 6
##   variable       min    q1 median    q3   max
##   <chr>        <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1 Petal.Length   1     1.6   4.35   5.1   6.9
## 2 Petal.Width    0.1   0.3   1.3    1.8   2.5
## 3 Sepal.Length   4.3   5.1   5.8    6.4   7.9
## 4 Sepal.Width    2     2.8   3      3.3   4.4
```

1 sample t-test

What is wrong with base R

When performing hypothesis testing in normal R, it is functional and pretty to look at, but hard to use.

For example, consider the following hypothesis test. Let μ_{pl} be the true mean petal length of iris flowers in centimeters, and consider the hypothesis

$$H_0 : \mu_{pl} = 5 \quad \text{vs} \quad H_a : \mu_{pl} \neq 5\,.$$

What is wrong with base R

```r
(petal_t_test <- t.test(iris$Petal.Length, mu = 5))
```

```
##
##  One Sample t-test
##
## data:  iris$Petal.Length
## t = -8.6169, df = 149, p-value = 9.235e-15
## alternative hypothesis: true mean is not equal to 5
## 95 percent confidence interval:
##  3.473185 4.042815
## sample estimates:
## mean of x
##     3.758
```

Getting things from base R

If we want to access the information about the t-test, we need to use $ notation to get it, i.e.

```
petal_t_test$statistic
```

```
##         t
## -8.616862
```

```
petal_t_test$p.value
```

```
## [1] 9.235041e-15
```

`rstatix` brings dataframes!

The beauty of `rstatix` is that it takes all of this information and bundles it into an easy to use data frame (good for things like plotting). We perform the t-test as follows:

```
petal_t_test_rstatix <- iris %>%
  t_test(Petal.Length ~ 1, mu = 5)
petal_t_test_rstatix
```

```
## # A tibble: 1 x 7
##   .y.          group1 group2         n statistic    df        p
## * <chr>        <chr>  <chr>      <int>     <dbl> <dbl>    <dbl>
## 1 Petal.Length 1      null model   150     -8.62   149 9.24e-15
```

Your turn

What to do

1. Obtain summary statistics for each Species of iris in the `iris` dataset. Which species has the largest average petal length? Which species has the smallest average petal width?
2. Let μ_{pw} be the true mean petal width of iris flowers in centimeters, and consider the hypothesis

$$H_0 : \mu_{pw} = 1.3 \quad \text{vs} \quad H_a : \mu_{pl} \neq 1.3\,.$$

   Test this hypothesis at the $\alpha = 0.05$ level of significance using `rstatix`. Conclude your hypothesis in context.
3. Perform the hypothesis test from Q2 for each species of iris flower in the `iris` dataset. What do you notice?

Two sample t-tests

Let's narrow our scope

We will consider the two groups of flowers from the versicolor species and the virginica species. We can subset our data as

```
iris_sub <- iris %>%
  filter(Species != "setosa") %>%
  mutate(Species = fct_drop(Species))
head(iris_sub)
```

```
## # A tibble: 6 x 5
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
##          <dbl>       <dbl>        <dbl>       <dbl> <fct>
## 1          7           3.2          4.7         1.4 versicolor
## 2          6.4         3.2          4.5         1.5 versicolor
## 3          6.9         3.1          4.9         1.5 versicolor
## 4          5.5         2.3          4           1.3 versicolor
## 5          6.5         2.8          4.6         1.5 versicolor
## 6          5.7         2.8          4.5         1.3 versicolor
```

The hypothesis to test

Let $\mu_{versi-pl}$ and $\mu_{virgin-pl}$ be the true mean petal length of versicolor and virginica iris flowers in centimeters, respectively. We will test the hypothesis

$$H_0 : \mu_{versi-pl} = \mu_{virgin-pl} \quad \text{vs} \quad H_a : \mu_{versi-pl} \neq \mu_{virgin-pl} \,.$$

at the $\alpha = 0.05$ level.

The how-to in rstatix

This is simple in `rstatix` in the following way:

```
two_sample_petal_length <- iris_sub %>%
  t_test(Petal.Length ~ Species)
two_sample_petal_length
```

```
## # A tibble: 1 x 8
##   .y.          group1     group2        n1    n2 statistic    df        p
## * <chr>        <chr>      <chr>      <int> <int>     <dbl> <dbl>    <dbl>
## 1 Petal.Length versicolor virginica     50    50     -12.6  95.6 4.9e-22
```

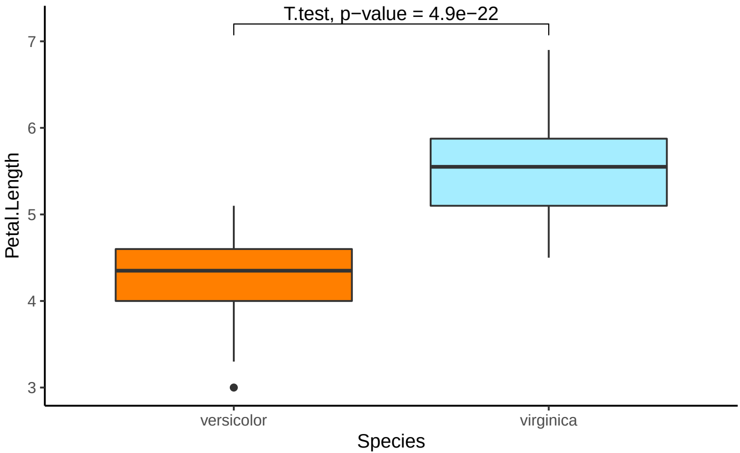
Conclusion?

Since the p-value is MUCH smaller than 0.05, we reject the hypothesis and conclude that there is sufficient evidence to suggest the true mean petal length between versicolor and virginica is different.

Let's visualise this! (the code)

```
p <- iris_sub %>%
  ggplot(aes(x = Species, y = Petal.Length)) +
  geom_boxplot(aes(fill = Species), show.legend = FALSE) +
  scale_fill_manual(values = cols) + # This line of code won't run for you
  theme_classic() +
  stat_pvalue_manual(two_sample_petal_length, # This is where ggpubr comes in!
                     label = "T.test, p-value = {p}",
                     y.position = 7.2)
```

Let's visualise this! (the plot)



Your turn

What to do

1. Let $\mu_{versi-pw}$ and $\mu_{setosa-pw}$ be the true mean petal width of versicolor and setosa iris flowers in centimeters, respectively. Test the hypothesis

   $$H_0 : \mu_{versi-pw} = \mu_{setosa-pw} \quad \text{vs} \quad H_a : \mu_{versi-pw} \neq \mu_{setosa-pw} \, .$$

   at the $\alpha = 0.05$ level. Conclude your hypothesis test in context.
2. Visualise your results from Q1 on a boxplot.
3. **Challenge**: Perform pairwise t-tests on Sepal Length for each species of iris flower, and visualise this on a boxplot. Brownie points for using (and remembering) adjusted p-values.