Live data analysis demonstration

Owen Petchey

2/12/2019

Live data analysis demonstration

Introduction

I thought it would be good to get hands on as early as possible, and to do so for something directly relevant to yourselves. So, we're going to attempt something quite ambitious – in the next two hours or so, we'll go through a whole data analysis from start to end. Lets get started.

Meta-task

Write things that you don't understand, and need to know about in subsequent classes. We will cover these things.

The question

What should be our question? As always, there are some influences and some constraints. We should ask a question of interest to us, and of some importance. And we should be able to collect the data, within our current constraints, necessary to answer the question.

The question we will address is "do male and female reaction times of students at the University of Zurich differ?".

More specifically "how different is the average reaction time of a man compared to the average reaction time of a woman"? And is that difference large or small relative to the variation among men, and among women?

Why this question? Reaction times are important, safety, sport...

Expectation

We can have a look on the internet, and pretty easily find lots of studies of reaction times and gender (e.g., A comparative study of visual and auditory reaction times on the basis of gender and physical activity levels of medical first year students (by the way, we will later in the course critique this paper – it has some pretty poor features).

Generally, we see that males tend to have faster reaction times than females. So we expect that to be the same for students at the University of Zurich.

Given that you know this pattern, and you are the subjects, its interesting to see if you women can buck the trend, perhaps by trying especially hard. Though know the men know you might do this, it probably won't work!

How are we going to present the results?

Thinking backwards from how we present the final results, can often be quite useful.

I think a nice box and whisker plot will work here (Owen will sketch this). We will have two groups of reaction times. Put another way, we will have one explanatory variable (gender) and one response variable (reaction time). The explanatory variable gender is a categorical / discrete variable. The response variable is continuous.

We expect the distribution of male reaction times to be have a lower mean than the distribution of female reaction times.

We will look at this graph, and answer our question (wow, without any statistical test – YES!).

What statistical test will we use?

Reaction times (the response variable) we expect to be quite normally distributed, though cannot be negative. Gender will be categorical with two levels (male and female). We don't expect greater or less variation in reaction times among males compared to among females.

Based on these expectations, we will use a linear model, which assumes normally distributed residuals and equal variances among groups. The traditional name for the test is the T-test.

Based on convention, and little else, we will say there is a significant difference between male and female reaction times if the observed difference has a p-value of less than 0.05.

Selection of subjects

We usually need to very carefully select the subjects of our study. Ideally, as we're interested in students at the University of Zurich (see the question), we would select a cross section of such students. Instead, you are going to be the subjects, and you are not representative of all students. You're relatively young, on average, you're studying natural sciences, etc. So we will have to be very cautious if we make statements about students at the University in general, and perhaps one might even now conclude that we can't really answer the question.

Perhaps we need the question "do male and female reaction times of biology and biomedicine students, in their first year, at the University of Zurich differ?"

Ethical clearance and considerations

If we aimed to publish these results, or in some other way disseminate them, we would need ethical clearance for research involving humans. We're not, so we don't have ethical clearance.

However, please do not include any personal information in any of the data you contribute to the exercise.

Data collection

Create for yourself a unique ID code, so that if we want to collect more data about you, we could related the reaction time data to that. Write this code down somewhere safe, keep it.

Go to the Human Benchmark website.

Do the reaction time test and write down your reaction time in fractions of a second.

While you're there, also please do the other three tests, and write down your score. We don't need these for our current question about reaction times, but we might look at this data later in the course.

Now go to the web page with the form to enter your information and enter your ID, gender, and four scores. Please be careful!

Go ahead and do all that.

Look at the data!

Here is the link to the datasheet containing all the data you just recorded (online only)

Lets have a look at it, see what you've done. (I'm scared! I know how difficult it is to enter data without making mistakes! And I have some experience of how different people are!)

Lets get the data into our data analysis software of choice (R, via RStudio)

First note that Owen has gone to the responses googlesheet, and in the "File" menu, clicked on "Publish to web..." and chosen to publish as "Comma separate values".

Now we can read that web page of comma separated values into R:

```
## First we load a required package (we need to install this if we haven't already)
library(readr)
```

```
## Warning: package 'readr' was built under R version 4.1.2
```

```
## also we will use some other packages
library(dplyr)
library(ggplot2)

## Now read in the data, using the read_csv() function. We give it the URL of the published version of
the_URL <- "https://docs.google.com/spreadsheets/d/e/2PACX-1vQDI5oZ54MD4Qm_WDydUAWgRYX1PRWMoOWJCSFN5mZJ
class_RTs <- read_csv(the_URL)

## FS22 csv link: https://docs.google.com/spreadsheets/d/e/2PACX-1vQDI5oZ54MD4Qm_WDydUAWgRYX1PRWMoOWJCS



## Have a look at the data in R
#View(class_RTs)
## or just do
class_RTs
```

```
## # A tibble: 173 x 16
##    Timestamp `Please enter t~ `What was your ~ `Please enter y~ `Are you right ~
##    <chr>     <chr>            <chr>                      <dbl> <chr>
##  1 16/02/20~ catmanPH22       Male                          67 Right handed
##  2 21/02/20~ fir              Female                        45 Right handed
##  3 21/02/20~ Nišama           Female                        65 Right handed
##  4 21/02/20~ Why              Female                        60 Right handed
##  5 21/02/20~ sprit            Female                        49 Right handed
##  6 21/02/20~ Samaden          Male                          65 Right handed
##  7 21/02/20~ id123            Female                        56 Right handed
##  8 21/02/20~ 2SW0LE           Female                        61 Right handed
##  9 21/02/20~ Nase2016         Male                          84 Right handed
## 10 21/02/20~ 123456789        Male                          80 Right handed
## # ... with 163 more rows, and 11 more variables: ...
```

Now we need to do some data wrangling (cleaning and tidying)

Clean up the column / variable names:

```
## Must be very careful to get the next line right!!! Really important!!!
names(class_RTs) <- c("Timestamp", "ID", "Gender", "Weight",
                      "Handedness", "Pref_Reaction_time_1",
                      "Pref_Reaction_time_2", "Pref_Reaction_time_3",
                      "Pref_Reaction_time_4", "Pref_Reaction_time_5",
                      "Pref_Reaction_time",
                      "Nonpref_Reaction_time_ave",
                      "Verbal_memory_score", "Number_memory_score",
                      "Visual_memory_score", "Random_number")
class_RTs
```

```
## # A tibble: 173 x 16
##    Timestamp           ID    Gender Weight Handedness Pref_Reaction_t~ Pref_Reaction_t~
##    <chr>               <chr> <chr>   <dbl> <chr>                 <dbl>            <dbl>
##  1 16/02/2022 14:43:37 catm~ Male       67 Right han~              272              285
##  2 21/02/2022 14:14:21 fir   Female     45 Right han~              382              353
##  3 21/02/2022 14:14:22 Niša~ Female     65 Right han~              210              211
##  4 21/02/2022 14:14:59 Why   Female     60 Right han~              409              989
##  5 21/02/2022 14:16:38 sprit Female     49 Right han~              329              249
##  6 21/02/2022 14:24:57 Sama~ Male       65 Right han~              284              273
##  7 21/02/2022 14:21:36 id123 Female     56 Right han~              265              299
##  8 21/02/2022 14:17:21 2SW0~ Female     61 Right han~              315              343
##  9 21/02/2022 14:17:42 Nase~ Male       84 Right han~              376              294
## 10 21/02/2022 14:19:00 1234~ Male       80 Right han~              300              300
## # ... with 163 more rows, and 9 more variables: Pref_Reaction_time_3 <dbl>,
## #   Pref_Reaction_time_4 <dbl>, Pref_Reaction_time_5 <dbl>,
## #   Pref_Reaction_time <dbl>, Nonpref_Reaction_time_ave <dbl>,
## #   Verbal_memory_score <dbl>, Number_memory_score <dbl>,
## #   Visual_memory_score <dbl>, Random_number <dbl>
```

Check the variable types are correct.

- Timestamp should be a character
- ID should be a character
- Gender should be a character
- The remaining four variables should be numeric (if fractional, if whole numbers).

```
class_RTs
```

```
## # A tibble: 173 x 16
##    Timestamp           ID    Gender Weight Handedness Pref_Reaction_t~ Pref_Reaction_t~
##    <chr>               <chr> <chr>   <dbl> <chr>                 <dbl>            <dbl>
##  1 16/02/2022 14:43:37 catm~ Male       67 Right han~              272              285
##  2 21/02/2022 14:14:21 fir   Female     45 Right han~              382              353
##  3 21/02/2022 14:14:22 Niša~ Female     65 Right han~              210              211
##  4 21/02/2022 14:14:59 Why   Female     60 Right han~              409              989
##  5 21/02/2022 14:16:38 sprit Female     49 Right han~              329              249
```

```
##  6 21/02/2022 14:24:57 Sama~ Male       65 Right han~              284              273
##  7 21/02/2022 14:21:36 id123 Female     56 Right han~              265              299
##  8 21/02/2022 14:17:21 2SW0~ Female     61 Right han~              315              343
##  9 21/02/2022 14:17:42 Nase~ Male       84 Right han~              376              294
## 10 21/02/2022 14:19:00 1234~ Male       80 Right han~              300              300
## # ... with 163 more rows, and 9 more variables: Pref_Reaction_time_3 <dbl>,
## #   Pref_Reaction_time_4 <dbl>, Pref_Reaction_time_5 <dbl>,
## #   Pref_Reaction_time <dbl>, Nonpref_Reaction_time_ave <dbl>,
## #   Verbal_memory_score <dbl>, Number_memory_score <dbl>,
## #   Visual_memory_score <dbl>, Random_number <dbl>
```

Correct or exclude problematic data

This section should not be necessary, as the google form allows only numeric entries in fields that should have numbers.

If we have problems here, with variables of the wrong type, it probably means some of the data entry is a bit messed up.

```
## Have to do this live!!!
## e.g. to exclude observations with character entries in Reaction_time variable
class_RTs <- filter(class_RTs, !is.na(as.numeric(Pref_Reaction_time)))
```

Once fixed, we need to make the variable have the correct type

```
## try using type_convert() from readr package.
class_RTs <- type_convert(class_RTs)
```

Check numbers of data points in each gender

```
table(class_RTs$Gender)
```

```
##
## Female   Male
##    110     63
```

Check the number of observations

Should be the same as we saw in the datasheet, which should be number of you in this classroom.

The number of observations and variables is given by R in the first line of output when we type the name of the data object:

```
class_RTs
```

```
## # A tibble: 173 x 16
##    Timestamp           ID    Gender Weight Handedness Pref_Reaction_t~ Pref_Reaction_t~
##    <chr>               <chr> <chr>   <dbl> <chr>                 <dbl>            <dbl>
##  1 16/02/2022 14:43:37 catm~ Male       67 Right han~              272              285
##  2 21/02/2022 14:14:21 fir   Female     45 Right han~              382              353
```

```
##  3 21/02/2022 14:14:22 Niša~ Female     65 Right han~              210              211
##  4 21/02/2022 14:14:59 Why   Female     60 Right han~              409              989
##  5 21/02/2022 14:16:38 sprit Female     49 Right han~              329              249
##  6 21/02/2022 14:24:57 Sama~ Male       65 Right han~              284              273
##  7 21/02/2022 14:21:36 id123 Female     56 Right han~              265              299
##  8 21/02/2022 14:17:21 2SW0~ Female     61 Right han~              315              343
##  9 21/02/2022 14:17:42 Nase~ Male       84 Right han~              376              294
## 10 21/02/2022 14:19:00 1234~ Male       80 Right han~              300              300
## # ... with 163 more rows, and 9 more variables: Pref_Reaction_time_3 <dbl>,
## #   Pref_Reaction_time_4 <dbl>, Pref_Reaction_time_5 <dbl>,
## #   Pref_Reaction_time <dbl>, Nonpref_Reaction_time_ave <dbl>,
## #   Verbal_memory_score <dbl>, Number_memory_score <dbl>,
## #   Visual_memory_score <dbl>, Random_number <dbl>
```

Visualise the data

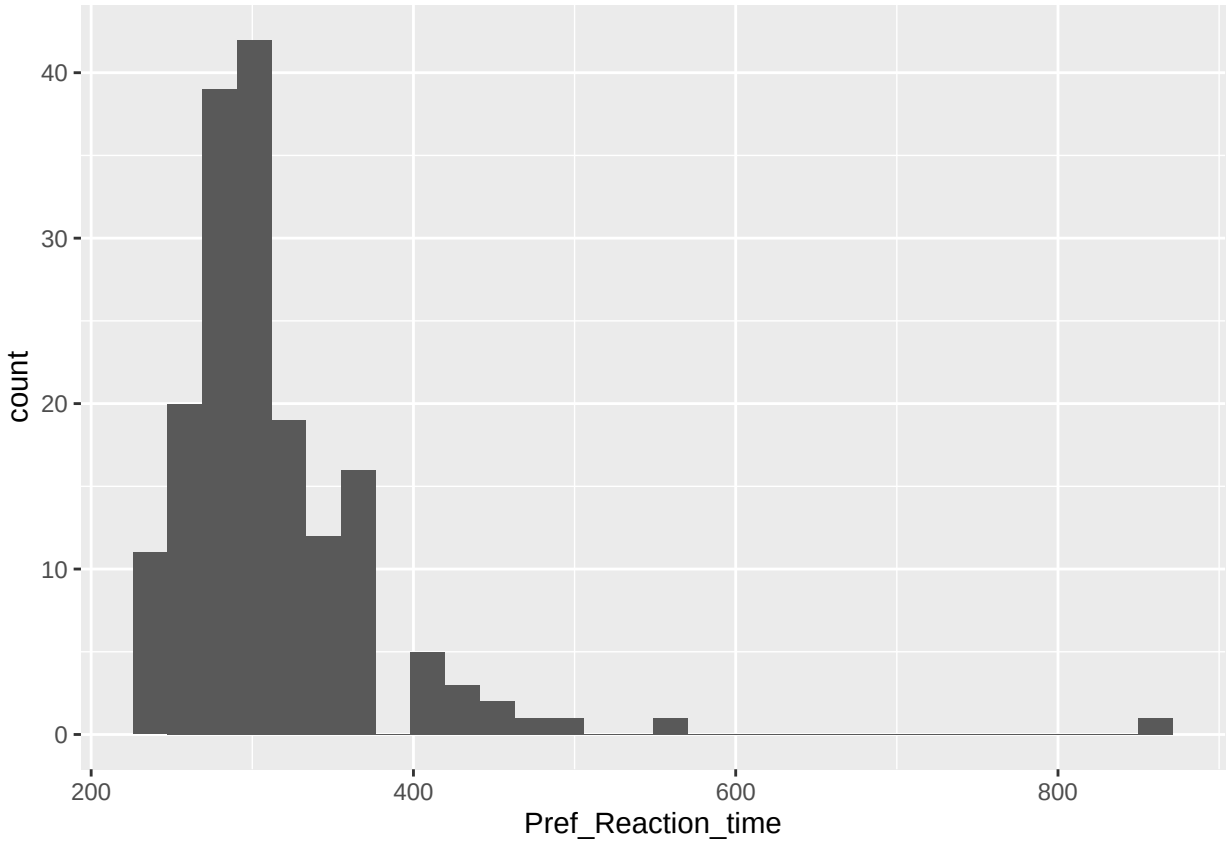
When we visualise the data, we're trying to do at least three things, and are not trying to do at least one.

We're not trying to make the most beautiful graph in the world, so we can put it in our report / presentation etc. We just want to clearly see the data.

We are trying to 1) do further checks for possible errors in the data, 2) making some initial assessments of how the data is distributed, 3) see what we think is the answer to our question.

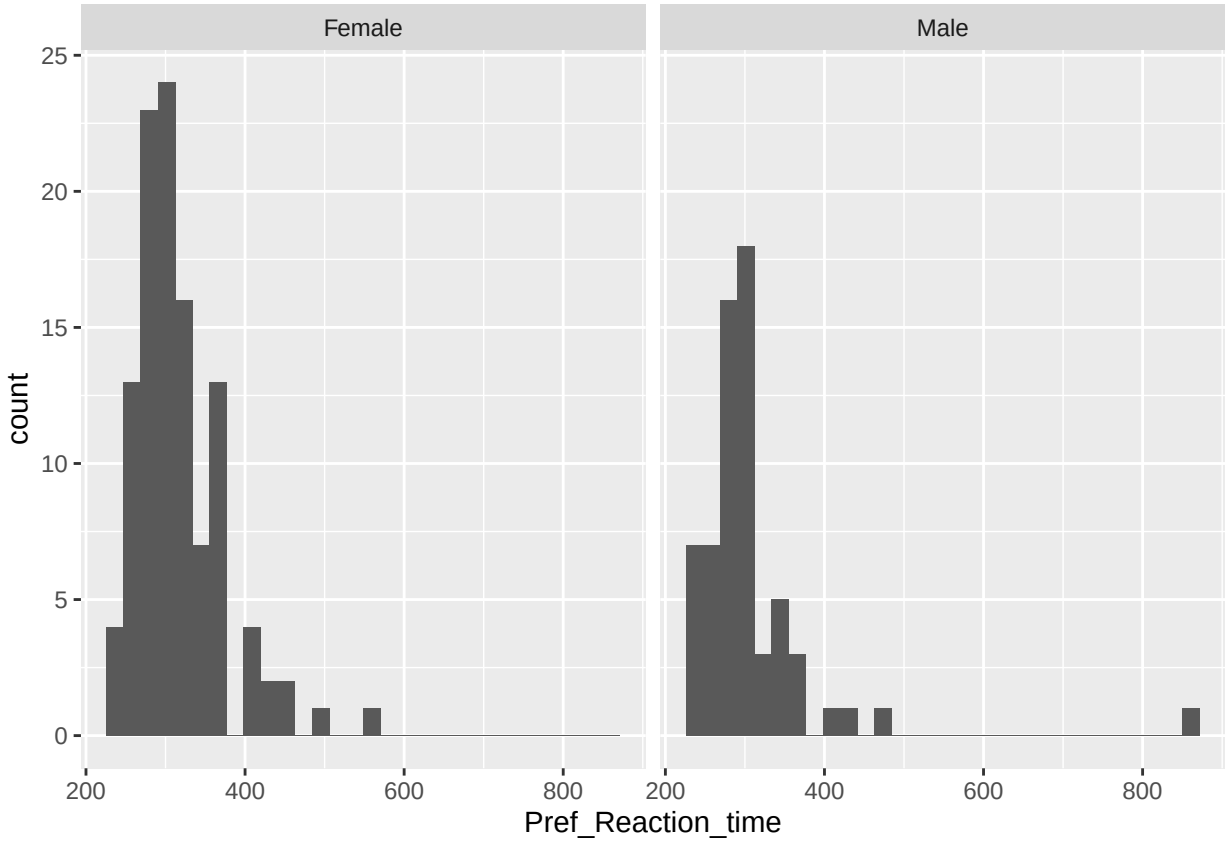
A histogram of all the data:

```
ggplot(class_RTs, aes(x=Pref_Reaction_time)) +
  geom_histogram()
```



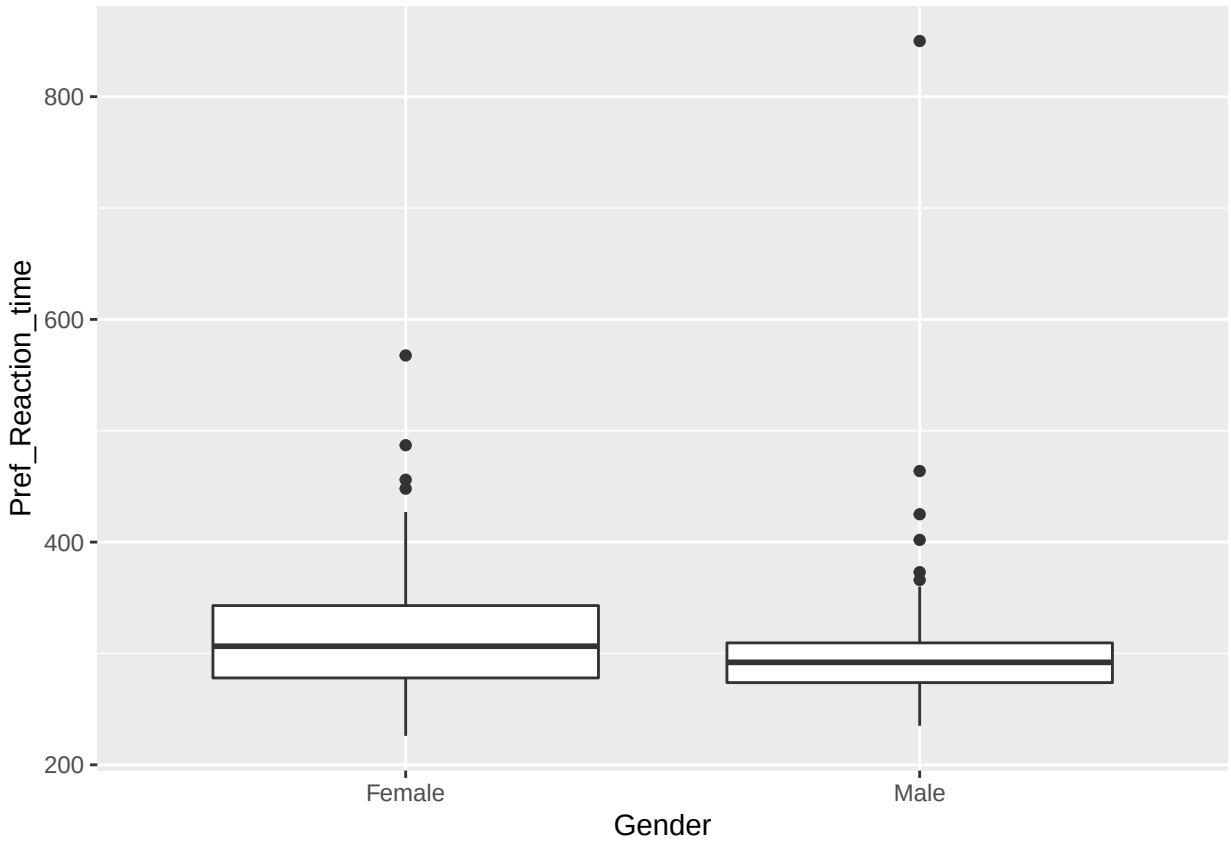
Separate histograms for each gender:

```
ggplot(class_RTs, aes(x=Pref_Reaction_time)) +
  geom_histogram() +
  facet_wrap(~ Gender)
```



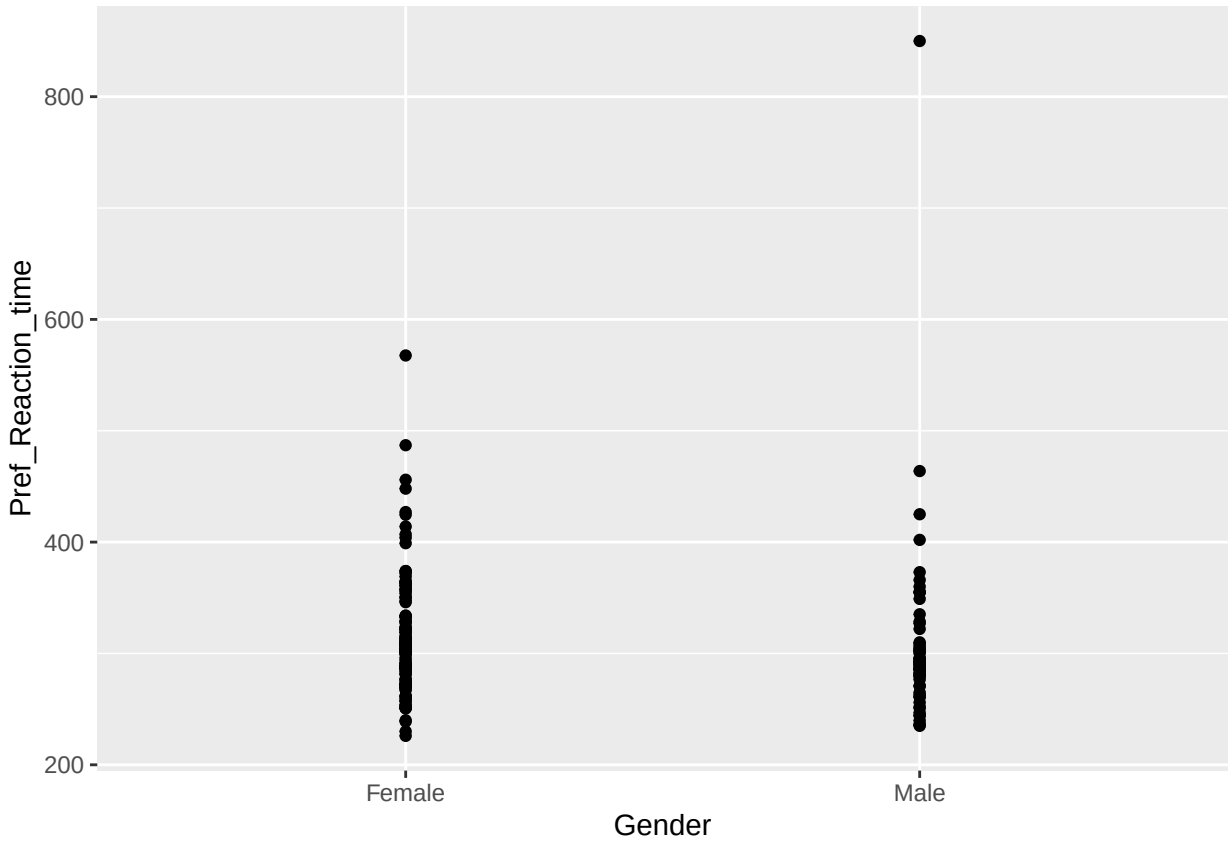
A box and whisker plot:

```
ggplot(class_RTs, aes(x=Gender, y=Pref_Reaction_time)) +
  geom_boxplot()
```



Or just the data points (with some jitter, to separate overlapping points):

```
ggplot(class_RTs, aes(x=Gender, y=Pref_Reaction_time)) +
  geom_point()
```



What do we think about the three things? Any likely errors? How is the data distributed (within and between groups)? Does it look like there is a difference in reaction times (if so, by how much on average, and which group is faster)?

Get the means

```
class_RTs %>% group_by(Gender) %>%
  summarise(mean_RT=mean(Pref_Reaction_time),
            sd_RT=sd(Pref_Reaction_time))
```

```
## # A tibble: 2 x 3
##   Gender mean_RT sd_RT
##   <chr>    <dbl> <dbl>
## 1 Female    316.  55.3
## 2 Male      307.  82.6
```

Effect size and practical importance?

Does the difference between the means (i.e. the effect size) seem of practical importance? How does that size of difference correspond to the difference, for example, between the reaction time of an elite athlete, and a random person? How does it correspond to the difference between the reaction time of a human and a fly?

Assess assumptions

Before we even start to think about running a statistical test, we must check if the specific test we intend to run is justified. That is, we must check if the assumptions of the test are likely to be met.

Independence

The t-test assumes that observations are independent.

How was the data collected (hopefully not one gender on one day, and the other on another day, or something similarly confounding)?

Do we have more than one observation per subject? Not in this case, because you typed in the average. But we could have. Then the observations from the same individual would not be independent of each other. They would share in common the person they originated from. This would make the statistical test unreliable. Its something we'll look more closely at later in the course.

Normally distributed residuals

We can get a good idea about this, in this case, by looking at the distribution of the two groups of reaction times (see above). Obviously we need to have in mind some idea of what the normal distribution looks like, and how close the data have to look like one. There are quantitative tests for normality, we may look at them later.

Equal variance

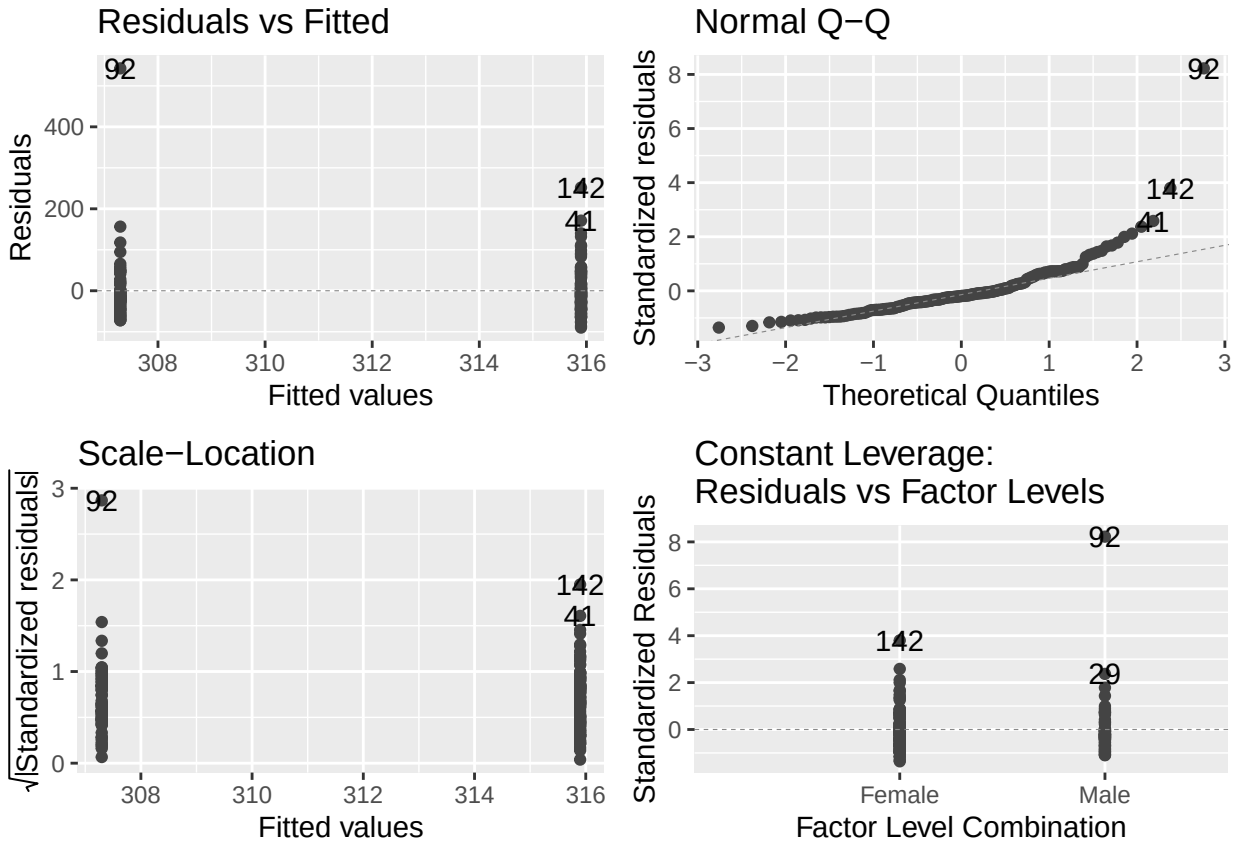
The spread of the reaction times for men, and the spread for women, should be about the same.

We can get a good idea about this, in this case, by looking at the distribution of the two groups of reaction times (see above). Again, we need to have in mind how similar the variance can be, without invalidating this assumption the data. There are quantitative tests for equal variance, we may look at them later.

```
m1 <- lm(Pref_Reaction_time ~ Gender, data=class_RTs)
library(ggfortify)
autoplot(m1, smooth.colour = NA)
```

```
## Warning: Removed 173 row(s) containing missing values (geom_path).
```

```
## Warning: Removed 173 row(s) containing missing values (geom_path).
```



Do the statistical test

We have to do a test, or more generally, some statistics, to give some kind of assessment of certainty / uncertainty in our answer. Traditionally, this is done with a p-value, and if its lower than 0.05 we say the result is significant (i.e. the results are very consistent with no difference). If its higher than 0.05 we accept the null hypothesis that there is no difference.

Another way to quantify uncertainty, is to give the difference in the means of the two groups, and a measure of certainty in this difference. If the difference between the means close to zero, and the uncertainty overlaps zero, then we conclude there is no strong difference.

We'll do this with a T-test, as we already planned. Before we go on, there is something very important we should figure out, and we should do this every time before we run a statistical test. Figure out the degrees of freedom.

There will be learning about this later in the course. For now, know that for a t-test the degrees of freedom are the number of observations minus two. Here that is [r nrow(class_RTs)] - 2 (= [r nrow(class_RTs)-2]). This is really important to figure out in advance, as its a great way to check that R is doing the test we think we're telling it to do.

```
my_ttest <- t.test(Pref_Reaction_time ~ Gender, data=class_RTs, var.equal=TRUE)
my_ttest
```

```
##
##  Two Sample t-test
##
## data:  Pref_Reaction_time by Gender
```

```
## t = 0.81846, df = 171, p-value = 0.4142
## alternative hypothesis: true difference in means between group Female and group Male is not equal to
## 95 percent confidence interval:
##  -12.14427  29.34880
## sample estimates:
## mean in group Female   mean in group Male
##             315.8991             307.2968
```

Lots of information there. We will teach you how to read this in later lectures.

For now, we can find the p-value: [r round(my_ttest$p.value, 5)]. And the difference between the means : [r round(diff(my_ttest estimate),3)] and the lower ([r round(my_ttest$conf.int[1], 3)]) and upper ([r round(my_ttest conf.int[2],3)]) 95% confidence limits on that difference.

Critical thinking

- How might the work be flawed?
- How might the analysis be flawed (assumptions violated)?
- Is the difference (i.e. effect size) small, medium, large, relative to differences caused by other factors?
- How general might be the finding?
- How do the qualitative and quantitative findings compare to those in previous studies?
- What could have been done better?
- What are the implications of the findings?

Practical significance

100 km per hour is 100 x 10^3 m per hour = 10^5 m per hour = 10^5 / 3600 m per sec 10^5 / 3600 = 27 m per sec

How many individual women are faster than the average of men? And the other way around?

Report and communicate the results

The results as a sentence

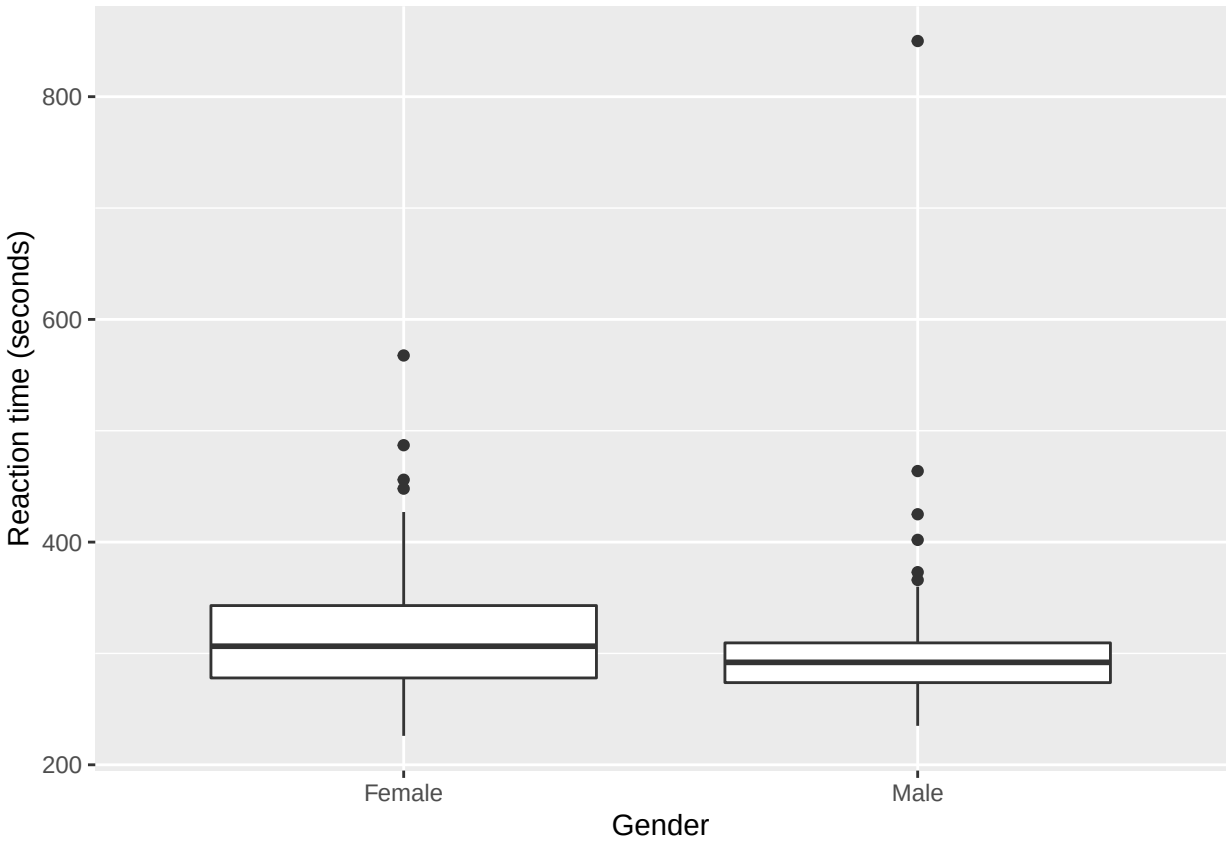
We should write a sentence that gives the direction and extent of difference, and a measure of certainty / uncertainty in that finding. It is totally unnaceptable, though common, to just write "there was a significant difference". If we want to give a p-value (and most people tend to expect to see one), we should remind about the statistical test used (remind because we may have already mentioned it) and give the degrees of freedom, the value of the test statistic, and the p-value.

Insert sentence here, once we have the results.

The results graphically

The aim here is to make a beautiful graph that very clearly communicates the findings! This doesn't mean "fancy" and or "complex". Often simpler is better. Getting the basic right is essential, of course.

```
ggplot(class_RTs, aes(x=Gender, y=Pref_Reaction_time)) +
  geom_boxplot() +
  ylab("Reaction time (seconds)")
```



Wow! That was easy.

Do not use a table

Here, a table is not necessary. The results are in the sentence and in the graph.