Learning Objectives: Comparison Tests

- **Perform a paired t-test on a data set(s)**
- **Perform an independent t-test on a data set(s)**
- **Perform an ANOVA test on a data set(s)**

definition

Assumptions

- Learners are comfortable creating vectors and data frames with numerical elements.
- Learners are comfortable applying statistical functions on vectors and data frames.

Limitations

- This section will cover only the most commonly used comparison tests.

Paired T-Test

Before we begin, let’s open up the `comparison.r` file within RStudio. See instructions below:

info

Open the `comparison.r` file

Within RStudio, open the `comparison.r` file by selecting: `File` —> `Open File...` —> `code` —> `compare` —> `comparison.r`

Paired T-Test

A **paired t-test** or **dependent t-test** can be used to determine if two sets of data that belong to the same sample or group have the same **mean** or average.

The basic syntax to perform a paired t-test is:

```
t.test(x, y, paired = TRUE)
```

Where:
* `x` and `y` represent numeric vectors
* `paired` represents a logical value specifying that we want to compute a paired t-test (`TRUE`)

For example, if we have two numeric vectors below which represent mice weight before and after a treatment:

```
# Weight of the mice before treatment
before <- c(200.1, 190.9, 192.7, 213, 241.4, 196.9, 172.2,
        185.5, 205.2, 193.7)
# Weight of the mice after treatment
after <- c(392.9, 393.2, 345.1, 393, 434, 427.9, 422, 383.9,
        392.3, 352.2)
```

We can use:

```
print(t.test(before, after, paired = TRUE))
```

Which returns:

```
    Paired t-test


data:  before and after
t = -20.883, df = 9, p-value = 6.2e-09
alternative hypothesis: true difference in means is not equal to
        0
95 percent confidence interval:
 -215.5581 -173.4219
sample estimates:
mean of the differences
                -194.49
```

Null Hypothesis, Alternative Hypothesis & P-Value

There are three important **key terms** that you should know and they are: **null hypothesis**, **alternative hypothesis**, and **p-value**. Any time a statistic test is conducted, the goal is always to determine if the test shows a **significant** difference between two sets of data. The null hypothesis is rejected in favor of the alternative hypothesis if the result shows a **significant** difference between the data. On the other hand, we fail to reject the null hypothesis if the result shows **no significant** difference in the data. To determine if there is significance or not, we look at the **p-value**.

- A p-value of **0.05 or below** means there is a **significant difference** between the data, in which case we will **reject the null hypothesis** in favor of the alternative hypothesis.
- A p-value of **greater than 0.05** means there is **no significant difference** between the data, in which case we will **fail to reject the null hypothesis**.
- When it comes to these hypotheses, we always refer to either **rejecting** the null hypothesis (in other words, the alternative hypothesis is supported) or **failing to reject** the null hypothesis (null hypothesis is supported). The **null hypothesis** is always the hypothesis being tested.

Given our data from before:

```
	Paired t-test

data:  before and after
t = -20.883, df = 9, p-value = 6.2e-09
alternative hypothesis: true difference in means is not equal to
        0
95 percent confidence interval:
 -215.5581 -173.4219
sample estimates:
mean of the differences
                -194.49
```

The p-value is calculated to be `6.2e-09` which is **significantly less than 0.05**. Therefore, we will reject the null hypothesis in favor of the alternative hypothesis that there **is** a significant difference between the weights of the mice sample **before** and **after** the treatment they were given.

Independent T-Test

Independent T-Test

Independent t-test works almost the same way. However, the tags that are needed are slightly different from those of a paired t-test. **Note** that the samples (or vectors) for an independent t-test are not related to each other. The basic syntax to perform an independent t-test is:

```
t.test(x, y, paired = FALSE, var.equal = FALSE)
```

Where:
* `x` and `y` represent numeric vectors
* `paired` represents a logical value specifying that we want to compute an independent t-test (`FALSE`)
* `var.equal` represents a logical variable indicating whether to treat the two variances as being equal (**variance** is how spread out the data set is)
* Specify `TRUE` if the data within the data sets all occurs within approximately the same range. `TRUE` will compute the `Two Sample t-test` in RStudio.
* Specify `FALSE` if the data within the data sets are all sporadically spread out. `FALSE` will compute the `Welch Two Sample t-test` which tries to help **normalize** the data sets. In statistics, it is important to have data that is normalized.

For example, below are two vectors of data involving weights of men and weights of women. These groups do not share any data between each other making them **independent**:

```
women_weight <- c(38.9, 61.2, 73.3, 21.8, 63.4, 64.6, 48.4,
        48.8, 48.5)
men_weight <- c(67.8, 60, 63.4, 76, 89.4, 73.3, 67.3, 61.3,
        62.4)
```

Since the weights between both groups appear to fall within the same range, we can use:

```
print(t.test(women_weight, men_weight, paired = FALSE, var.equal
        = TRUE))
```

Which returns:

```
	Two Sample t-test

data:  women_weight and men_weight
t = -2.7842, df = 16, p-value = 0.01327
alternative hypothesis: true difference in means is not equal to
        0
95 percent confidence interval:
 -29.748019  -4.029759
sample estimates:
mean of x mean of y
 52.10000  68.98889
```

If you have doubts regarding the variances of the data sets, you can set `var.equal = TRUE` to `FALSE` or vice versa to double check. If the results are similar, then it doesn't matter whether `Two Sample t-test` or `Welch Two Sample t-test` is conducted.

```
print(t.test(women_weight, men_weight, paired = FALSE, var.equal
        = FALSE))
```

Which returns:

```
	Welch Two Sample t-test

data:  women_weight and men_weight
t = -2.7842, df = 13.114, p-value = 0.01538
alternative hypothesis: true difference in means is not equal to
        0
95 percent confidence interval:
 -29.981920  -3.795858
sample estimates:
mean of x mean of y
 52.10000  68.98889
```

Looking at both p-values from the `Two Sample t-test` (`0.01327`) and the `Welch Two Sample t-test` (`0.01538`), we can say with 95% confidence that we will reject the null hypothesis in favor of the alternative hypothesis that there is a **significant** difference between the weights of women and men because the p-values are both less than or equal to `0.05`.

ANOVA

ANOVA

The **one-way analysis of variance (ANOVA)** is used to compare means in a situation where there are more than two groups. The null hypothesis in this case is that there is **no significant** difference in data between the groups. The alternative hypothesis is that **at least one group** has a **significant** difference in data compared to the others. The basic syntax to perform an ANOVA test is:

```
aov(formula = x ~ y, data = df)
```

Where:
* `x` represents a numeric vector
* `y` represents a character vector (groups)
* `df` represents the data frame in which the vectors are derived from (if any is provided)

For example, below are two vectors of data. One called `size` and one called `pop`. The data from the `size` vector represents the size of the population represented in `pop`. Particularly, there are 3 groups or categories, `"A"`, `"B"`, and `"C"`.

```
size <- c(3,4,5,6,4,5,6,7,7,8,9,10)
pop <- c("A","A","A","A","B","B","B","B","C","C","C","C")
```

We can use:

```
print(aov(size ~ pop))
```

Which returns:

```
Call:
   aov(formula = size ~ pop)

Terms:
                     pop Residuals
Sum of Squares  34.66667  15.00000
Deg. of Freedom        2         9

Residual standard error: 1.290994
Estimated effects may be unbalanced
```

Additionally, you can call `summary()` on the `aov()` function to get even more details regarding the data.

```
print(summary(aov(size ~ pop)))
```

Which returns:

```
            Df Sum Sq Mean Sq F value  Pr(>F)   
pop          2  34.67  17.333    10.4 0.00457 **
Residuals    9  15.00   1.667                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

`print(summary(aov(size ~ pop)))` is preferred because it allows us to see the `Pr(>F)` value which acts the same as the p-value. Since `Pr(>F)` is `0.00457` which is much less than `0.05`, we reject the null hypothesis in favor of the alternative hypothesis that **at least one group** has data that is significantly different from the others.

Another Example:

If we modified the data to something like this:

```
size <- c(3,4,5,5,4,4,3,5,3,4,3,3)
pop <- c("A","A","A","A","B","B","B","B","C","C","C","C")
```

What do you think calling `print(summary(aov(size ~ pop)))` will reveal?

Result:

```
              Df Sum Sq Mean Sq F value Pr(>F)
pop          2  2.167  1.0833   1.773  0.224
Residuals    9  5.500  0.6111
```

The new `Pr(>F)` is now `0.224` which is much higher than `0.05`. This means that we now fail to reject the null hypothesis. This also means that there is **no significant difference** in data between the three groups. Why? Because if you take a look at the modified data, all of the elements between the groups ranged from `3` to `5` which hints that there isn't really a difference between them.

TukeyHSD

In our original data sets:

```
size <- c(3,4,5,6,4,5,6,7,7,8,9,10)
pop <- c("A","A","A","A","B","B","B","B","C","C","C","C")
```

we determined at least **one** of the groups is significantly different from the others. However, the ANOVA test did not allow us to determine **which** group or groups were different. Luckily, the `TukeyHSD()` function allows us to do just that. Applying the `TukeyHSD()` like so:

```
size <- c(3,4,5,6,4,5,6,7,7,8,9,10)
pop <- c("A","A","A","A","B","B","B","B","C","C","C","C")

print(TukeyHSD(aov(size ~ pop)))
```

which results in:

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = size ~ pop)

$pop
    diff        lwr      upr     p adj
B-A    1 -1.5487408 3.548741 0.5402482
C-A    4  1.4512592 6.548741 0.0045122
C-B    3  0.4512592 5.548741 0.0231730
```

we can determine based on the `p adj` or "p-adjusted value" that the **biggest difference** exists between groups `C` and `A` (`C-A`) because their computed `p adj` value of `0.0045122` is significantly less than `0.05`. Groups `C` and `B` (`C-B`) also show a significant difference in data because their `p adj` value is `0.0231730` is less than `0.05`.

On the other hand, there is **no significant** difference between groups `B` and `A` (`B-A`) given their `p adj` value of `0.5402482` is much greater than `0.05`.