# Practical 10: t-test

ADS2

Semester 1, 2023/24

## Learning objectives

After completing this practical you will be able to:

- Explain how a one-sample t-test works and can be calculated using R functions as well as manually.
- Discuss how sampling can affect the result of one-sample t-test
- Discuss how sample size can affect the result of one-sample t-test…
- … and generalise these discussion to any statistical test!

## Introduction

In the lectures you have learned about one-sample t-test and how it can help us answering the question of whether the true mean of a **population** is equal to a certain value. In real life, however, you never have access to the full population data, and you have to work on a **sample** of that population. How does that affect the reliability of your conclusions?

## Today's case study

It's the year 2300… humans have built the first stable colony in space, on planet Orion Alpha 10. After about 2 years, the colony is home to just more than one hundred thousand people; Dr Spiff has been tasked to check that their physiology has not been altered by the long time in space.

After a fairly uneventful space trip, he lands on the planet and gets to work. The first question he wants to answer is whether the core body temperature of the colony's inhabitants is maintained at 37°C, just like on Earth.

Unfortunately, he doesn't have the capacity of sampling everyone and, even more unfortunately, Dr Spiff has not been paying attention during his statistics lectures, so he does not quite remember how to appropriately sample a population!

So… let's give Dr Spiff a hand!

## The data

On Learn you will find a file called *OrionTemp.csv*. This contains the core body temperature of all the inhabitants of the colony. In real life you will never have access to the data from the whole population, but since we cannot fly you to space this is where we will generate our samples for today!

Start by reading the file in R. You can examine the data, for instance using (I am assuming you read it into a variable called `temp` but you can use any name you want)

```r
head(temp)
summary(temp)
```

Use the *sample* function to get a sample of 10 individuals from the whole population.

```r
mysample <- sample(temp$Temperature, size = 10)

mysample
```

```
##  [1] 36.81 37.10 37.14 37.17 37.15 37.07 36.97 36.89 36.89 37.00
```

The `sample` function has a parameter called `replace` which, if set to `TRUE` allows sampling with replacement; that means that we could take the same observation multiple times. **Do you think that would be a good thing to do in this case? Why?**

We can now test whether the mean of our sample is different from 37°, by using a one sample t-test.

We will use the `t.test` R function. To perform a one sample t-test we need to specify the parameter `mu`, which represents the mean we are testing against. We can also specify the parameter `alternative` to test whether the test is one- or two-sided. In this case, since we are not testing for a change in a specific direction we use a two-sided test. If we wanted to test whether the temperature was higher (or lower) than 37°C then we would use a one-sided test (`alternative = "greater"` or `alternative = "lower"`)

```r
t.test(mysample, mu = 37, alternative = "two.sided")
```

- What is the mean of your sample?
- What null hypothesis is the t-test testing?
- What is the p-value? What does that mean?
- Is the temperature in your sample statistically different from 37°C?
- Try to calculate the t-value using the formula you learned in class
- Generate and plot the null distribution for your test. The t-test uses a t-distribution as a null, that you can recreate in R using the `dt` function, to which you need to pass the number of degrees of freedom (parameter `df`). Do you remember how to calculate the degrees of freedom? *Hint*: the null is defined between $-\infty$ and $+\infty$, but you can simply calculate it between, for example -10 and 10. What is the area under your null distribution be? Why?
- The null distribution shows the probability of getting a certain t-value (or bigger), if the mean of the population is not different from the reference value. Find your t-value in the distribution and calculate the area under the curve above that t-value. That is your p-value!

Now, repeat the sampling and the t-test for 5 times. **Do the 5 tests agree with each other? Why?**

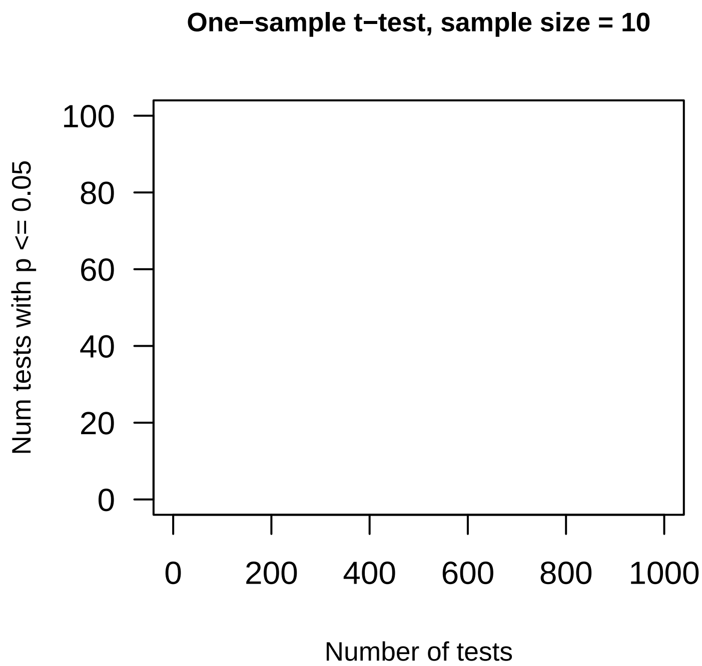
This needs some more thorough inspection. Let's do a simulation!

We want to establish how many times we get a statistically significant result depending on the number of times we repeat the test.

**Can you guess what the result will be?**

Now, try and create a simulation that will fill the following graph.

**Is the result what you expected? Can you think of a better graph to plot?**




Here are some hints:

- You will have to use two nested `for` loops, one to calculate multiple t-tests, and the other to calculate the number of p-values <= 0.05 at each different number of tests.
- If you save the result of `t.test` into a variable, you can then access the p-value in this way

```
res <- t.test(...)
res$p.value # This is the p value!
```

- You can store the p-values for each point into a variable that you initialise (outside the `for` loop) to `NULL`. You can then use `c` to add new values to it. For example

```
p.values <- NULL
for (...)
  {
  <calculate the test here>
  p.values <- c(p.values, test$p.value)
  }
```

Finally, try answering the question: has the core body temperature of the colony's inhabitants deviated from 37°C? Do this by running a t-test on the whole population. Now try runnning 100 tests with different sample sizes (e.g. from 5 to 100) and find the percentage of erroneous results from your tests.

**Does this depend on the sample size? Is there any advantage in using a bigger sample size?**