Computational Biology Group
D BSSE
ETH zürich

Introduction to Bayesian Statistics with R

1: Notes - t-test recap

Jack Kuipers

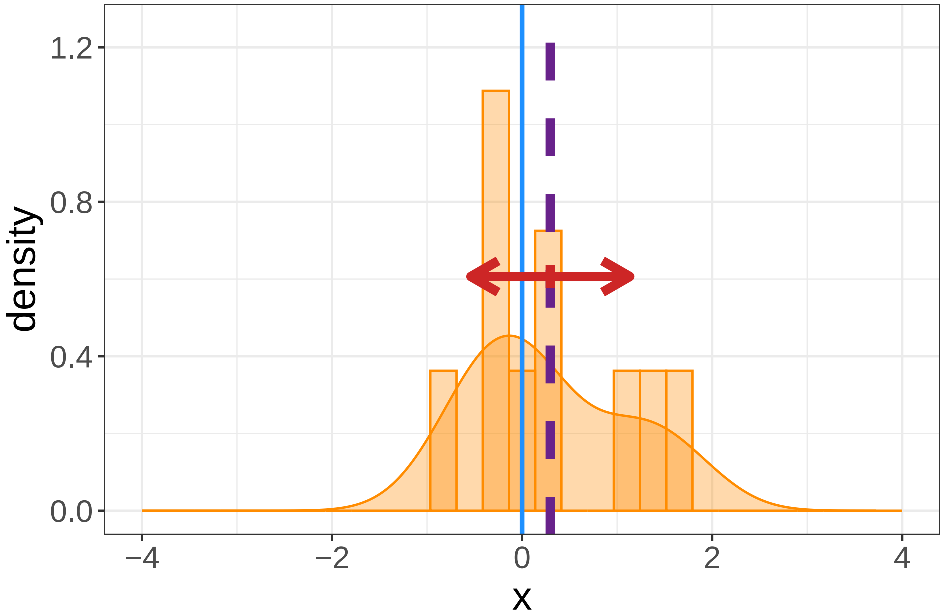
17 December 2024

Before we get into Bayesian statistics, we start with these notes with a quick recap of the basic workhorse of classical statistical inference: the *t*-test.

Is the mean 0?

Given a sample of data, we can start asking whether the mean of the underlying distribution is 0.

Of course, we don't normally know the true distribution, we only have access to a random sample from it. Consider the following sample of 10 data points, onto which we draw the sample mean (purple dashed line) and the sample standard deviation (red arrows):



From the plot it is very hard or impossible to tell. Maybe the true mean is 0, and we randomly have more positive values or maybe the mean is positive, or even negative. What we do know is that, because of the CLT, the sample mean will become Gaussian for larger sample sizes and its uncertainty, the standard error, will shrink.

If that sentence is clear to you, you can skip through the next two sections, where we give a bit more background.

Distribution of the sample mean

In particular, although the sample mean typically changes with each new sample, since the sample mean, which we're going to call m from now on, is defined

$$m = \frac{1}{N}\sum_{i=1}^{N} x_i$$

as the average of the sample of $x_i, i = 1, \ldots, N$, we can work out its mean and variance.

The mean of m is the same as for x

$$E[m] = E\left[\sum_{i=1}^{N} \frac{x_i}{N}\right] = \sum_{i=1}^{N} E\left[\frac{x}{N}\right] = \sum_{i=1}^{N} \frac{E[x]}{N} = E[x]$$

so that the estimate of the mean of the sample mean is itself (Can you get your head around that sentence?)

$$\widehat{\mu_m} = m = \overline{x}$$

while its variance

$$V[m] = V\left[\sum_{i=1}^{N} \frac{x_i}{N}\right] = \sum_{i=1}^{N} V\left[\frac{x}{N}\right] = \sum_{i=1}^{N} \frac{V[x]}{N^2} = \frac{V[x]}{N}$$

is the variance of x divided by N. This means that the estimate of the variance of the sample mean is also the sample variance divided by N

$$\widehat{\sigma_m^2} = \frac{\widehat{\sigma^2}}{N} = \frac{s^2}{N}$$

To help reduce confusion (or at least to try to help reduce confusion), the standard deviation of the sample mean is called the **standard error** and is related to the standard deviation of the sample itself via

$$s_m = \frac{s}{\sqrt{N}}$$

One key thing to notice is that as the sample size N increases, the standard error and hence the uncertainty in the sample mean decreases. This fact of course lies behind statistical testing.

The central limit theorem

Along with the standard error, which is the uncertainty in the sample mean, decreasing with increasing sample size, the distribution of the sample mean itself becomes a **normal** or **Gaussian** distribution.

A normally distributed rv X has a continuous state space

$$X \in \mathbb{R}$$

with a probability of occurring in an infinitesimal dx of

$$P(x < X \leq x + \mathrm{d}x) = f(x)\mathrm{d}x$$

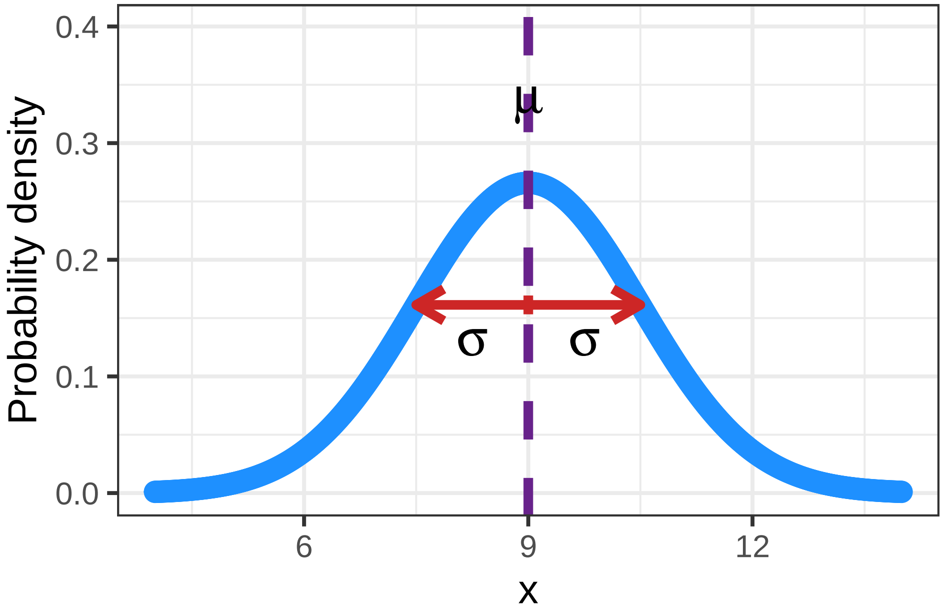
where $f(x)$ is the **probability density**

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The two parameters of the normal distribution are the

- mean μ
- variance σ^2

which control the location and spread of the curve respectively



A shorthand for X being normally distributed with mean μ and variance σ^2 is writing

$$X \sim \mathcal{N}(\mu, \sigma^2)\,, \qquad P(x < X \leq x + \mathrm{d}x) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}} \mathrm{d}x$$

so that our sample mean will approach

$$m \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{N}\right)$$

for larger sample sizes.

A simpler question

Now we return to our original question of whether the mean of the underlying distribution is 0. However, because answering this question from sampled data is quite difficult, we can try instead to answer a much simpler question: If the true population mean were 0, what would be the probability of observing a sample mean as far away from 0 (or further) than our one.

This may take some time to digest, so let's break it down in the following sections.

t-statistic

When we compute our sample mean we get a number, 0.297 in the case above, but this is meaningless without knowing something of the scale of variability we might expect. Luckily we have an estimate for this from our standard error, or 0.264 in this case. What we care about is how far our sample mean is from 0, in units of the standard error (cf the rescaling of the Gaussian from lecture 1). For this we define the **t-statistic**:

$$t = \frac{m}{s_m} = \sqrt{N}\frac{m}{s}$$

where s_m is the sample standard error and m is the sample mean, while s is the sample standard deviation.

Next we recall that the standard error is a random variable since it is computed from the sample. When we divide the sample mean (which is Gaussian for larger sample sizes) by the sample standard deviation, we actually get a Student-t distribution when the true mean is 0 (and the underlying distribution is Gaussian).

Student-t distribution

A **Student-t** distributed T has a continuous state space:

$$T \in \mathbb{R}$$

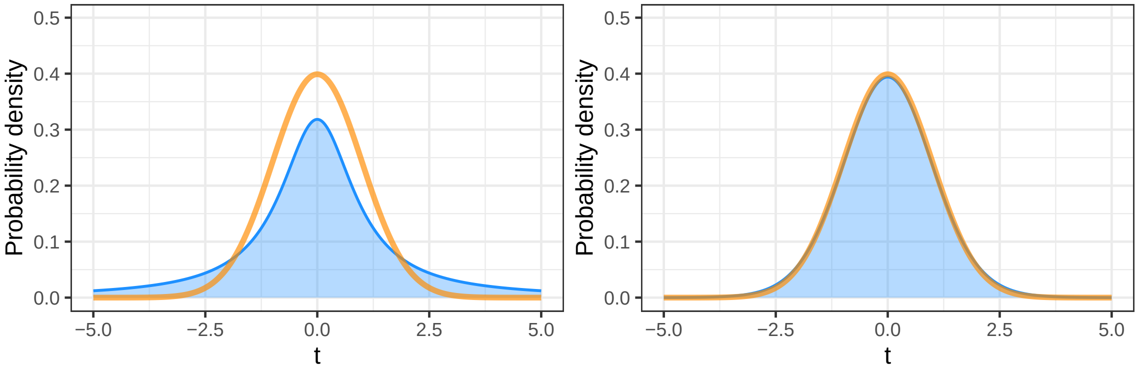
with probability of occurring in an infinitesimal dt of

$$P(t < T \leq t + \mathrm{d}t) = f(t)\mathrm{d}t$$

with probability density

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

The Student-t has one parameter $\nu > 0$ which is called the **degrees of freedom**. We show what it looks like for $\nu = 1$ (left) and $\nu = 20$ (right) below, where we have overlaid a Gaussian on top (in orange).



In the limit $\nu \to \infty$ it becomes exactly a standard Gaussian (remember the rescaling), but for lower degrees of freedom, which correspond to smaller sample sizes we have heavier tails and extreme values more often than under a Gaussian. This is precisely because we divide by the sample standard error, which is a noisy estimate from data and more noisy for smaller sample sizes.

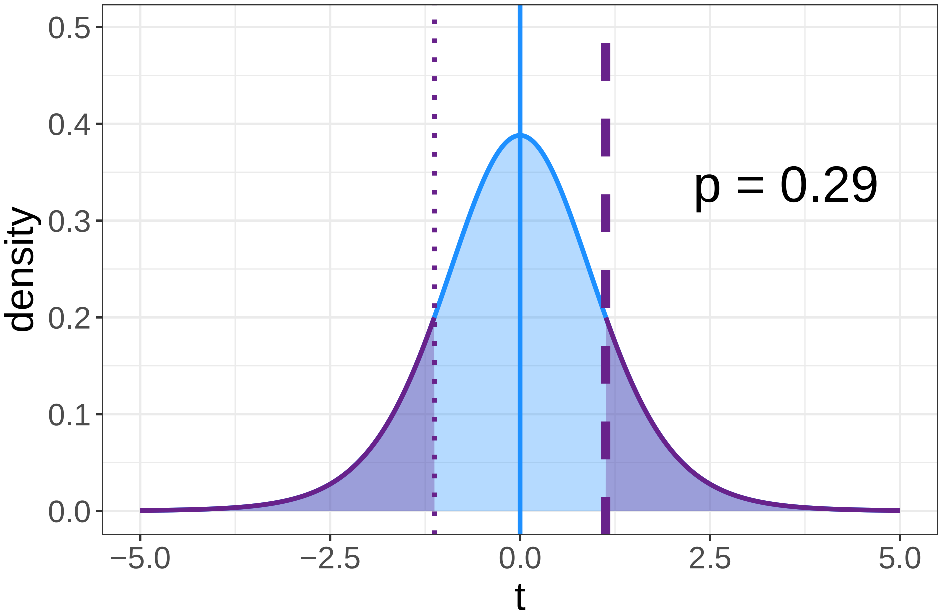
p-values

Now we can compare our t-statistic (sample mean divided by sample standard error) which was 1.13 for the sample at the top of this document, to what we would expect when the true mean is 0. Assuming the true mean is 0, is also called the **null hypothesis**.

For a sample size of N, we would expect a t-distribution with $(N-1)$ degrees of freedom, or

$$t_0 \sim \mathcal{T}_{N-1}$$

under the null hypothesis when the true mean is 0. The **p-value** is the probability of observing our t-statistic, or more extreme, under the null distribution. For the sample above we obtain the value 0.29, and we can visualise this as the area under the curve below. In the plot the dashed purple line is the t-value, the dotted line is its negative and the purple area covers more extreme values in both directions giving the p-value.



Why do we look at more extreme values, rather than just our t-statistic? The probability of any particular value for a continuous distribution is 0. To get a probability we need to take a range of values, so we look at values further from 0.

Why do we also look at the reflected area for negative t above? We would consider both positive and negative means as being different from 0, so we consider both possibilities although we can only observe one.

t-test

The steps above are known as a t-test, and what we have done are the following:

Defined the null hypothesis H_0

- that there is no effect, or the true mean $\mu_0 = 0$

Computed the t-statistic

$$t = \frac{m - \mu_0}{s_m} = \sqrt{N}\frac{m - \mu_0}{s}$$

Compared our observed t to the Student-t distribution which occurs under the null to compute the p-value

- the probability of observing a more extreme t-statistic assuming the null hypothesis

$$p = P(|t_0| > |t|)\,, \qquad t_0 \sim \mathcal{T}_{N-1}$$

t-tests in R

To run such a test in R we use the `t.test(...)` function into which we simply need to give a vector of samples

```
# Run a t-test with null hypothesis of mean 0
t.test(normal_samples) # the samples above were stored as normal_samples
```

```
##
##  One Sample t-test
##
## data:  normal_samples
## t = 1.1253, df = 9, p-value = 0.2896
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.3003473  0.8949409
## sample estimates:
## mean of x
## 0.2972968
```

R prints out a summary of the test run, and the t-statistic along with the p-value. These are stored in a list format, so we can access the individual components with the $

```
# extract the t-statistic
t.test(normal_samples)$statistic
```

```
##        t
## 1.125305
```

```
# extract the p-value
t.test(normal_samples)$p.value
```

```
## [1] 0.2895729
```

Computing p-values with `t.test(...)` is relatively easy, it is however much more complicated to interpret them properly. In case you'd like a recap, have a look at the next two sections, otherwise skip ahead.

p-value distributions

First we should have a good understanding of how p-values are distributed when the null is true, which in our case is that the true mean is indeed 0.

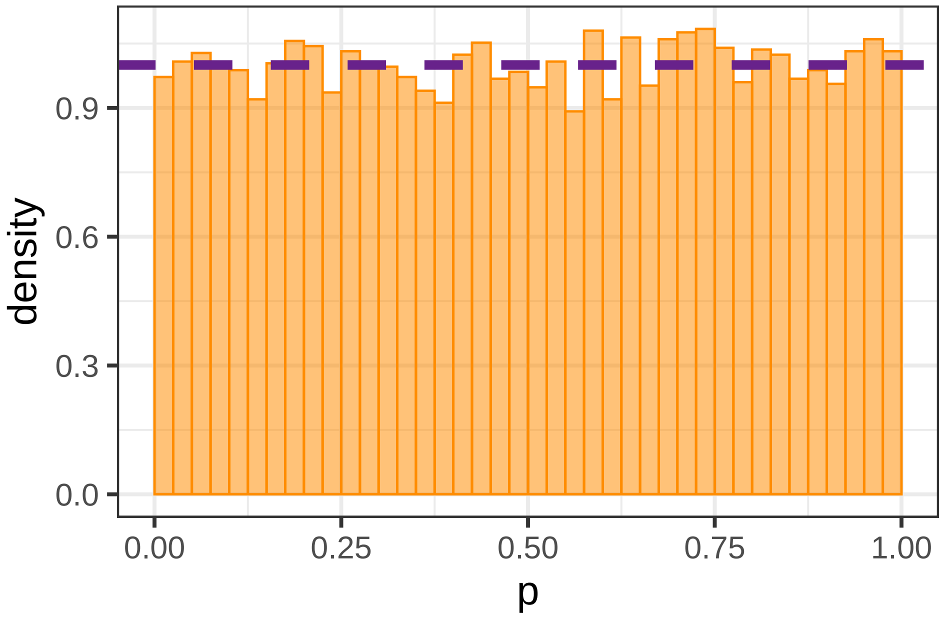
What distribution do p-values follow under the null? It is often surprising, but the distribution is uniform, which means completely flat between 0 and 1 (remember that p-values are probabilities).

If you need convincing, we can easily run a simple experiment. Let's sample 10 observations from a standard normal with mean 0, and compute the p-value

```
t.test(rnorm(10))$p.value
```

```
## [1] 0.7584735
```

Now, I'm going to do this 10,000 times and plot the results in a histogram:



If this still seems weird, these links give a more mathematical reasoning:

[https://www.r-bloggers.com/2016/04/a-simple-proof-that-the-p-value-distribution-is-uniform-when-the-null-hypothesis-is-true-2/]

[https://joyeuserrance.wordpress.com/2011/04/22/proof-that-p-values-under-the-null-are-uniformly-distributed/]

For a more intuitive reasoning, let's imagine we sample a very large number, say a million times from the null and get a million random values of t_0. We then take the absolute values $|t_0|$. There will be more values near 0, and fewer further away in line with the probability density function. Now we sort the $|t_0|$ in the order in which they occur from 0 upwards and just focus on their ranks.

Next we take our observed sample and get a value $|t|$. Our p-value is defined as the fraction of $|t_0|$ which are bigger than $|t|$, which is the fraction that have a higher rank in ordering.

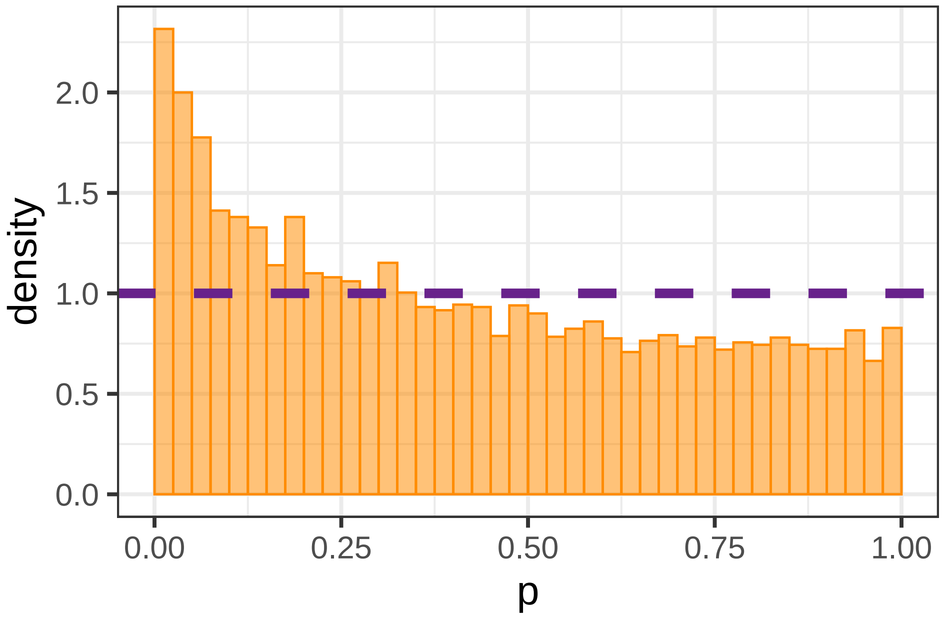
Now our assumption is that our observed sample is from the null distribution, and that it comes from the same distribution as all the t_0. What is the distribution of the rank of $|t|$? Well t is from the same distribution as all the t_0 samples so they are all interchangable and it is equally likely to have any rank. With a uniform distribution of the rank of $|t|$, the distribution of p-values under the null is also uniform.

Our original sample at the start of the document actually came from a distribution with a true mean of -0.25. I can create more repetitions of the sampling process

```
t.test(rnorm(10, -0.25))$p.value
```

```
## [1] 0.1592496
```

and again plot the distribution of p-values when the mean is not 0, but -0.25 with a sample size of 10:



We observe an enrichment of lower p-values near 0, which becomes more pronounced when the true mean (relative to the true standard deviation) is further from our null value of 0 and when the sample size increases (so the standard error shrinks). This is related to the **power** of the test, the fraction of times we would reject the null when a given alternative is true. The difference of the true mean from 0, divided by the true standard deviation is also called the **effect size**.

p-value summary

As the effect size or sample size increase:

- p-values tend to become lower

Low p-values may then hint at non-zero effects, but each p-value

- is a random variable
- is observed once (per experiment)
- does not tell us the true effect size

These bullet points and the fact that a p-value is a random variable should not be underestimated. Although you can see some difference in the overall distributions above when the mean was -0.25 compared to the null when it was 0, remember that when you run your experiment or analyse your data, you typically get one p-value which is a single random sample from some distribution. Please be very careful not to hope that a single number can tell us the truth of the underlying distribution or effect size.

The **p-value** is the probability of observing a more extreme statistic **assuming** the null hypothesis

- nothing more, nothing less
- conditioned on the null

In general, it is best to think of a p-value as a measure of *surprise*

- how surprising would the observed data be, if there were no effect?

Independent groups - two sample Welch's t-test

With these warnings ringing in your ears, or echoing in the background if you skipped ahead, let's get back to testing.

Normally we don't want to compare a single sample to a mean of 0, but to compare two groups to see if there is a difference between them. For example, between the control and treatment arm for testing a new drug, we want to see if there is a difference on average.

The null hypothesis is therefore that there is no difference.

When comparing the difference between two groups, we should keep in mind that if $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ the distribution of their difference, $(X_1 - X_2) \sim \mathcal{N}(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$ has a variance which is the sum of the variances of the individual groups.

In particular, for each group we compute the sample mean and variance

- Group 1 with N_1 subjects, mean m_1, variance s_1^2
- Group 2 with N_2 subjects, mean m_2, variance s_2^2

then the rescaled t-statistic is

$$t = \frac{m_1 - m_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

Under the null, this is approximately Student-t distributed

$$t_0 \sim T_\nu \,, \qquad \nu = \frac{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)^2}{\frac{\left(\frac{s_1^2}{N_1}\right)^2}{N_1 - 1} + \frac{\left(\frac{s_2^2}{N_2}\right)^2}{N_2 - 1}}$$

with a hideous formula for the number of degrees of freedom. When the group have equal size and variance, $\nu = N_1 + N_2 - 2$, and you can think of the -2 as arising because each of the two estimated standard deviations are random variables. In general, unequal variances or sample sizes will reduce ν and add uncertainty (heavier tails) to the null.

We don't need to worry so much about the formulas since `R` will handle them internally for us, and to run such a t-test we simply give two sample vectors to `t.test(...)`

For example if we create simulated data with the -0.25 difference we had before:

```r
# Generate some Gaussian samples with mean -0.25 and control with mean 0
test_samples <- rnorm(50, mean = -0.25, sd = 1)
control_samples <- rnorm(50, mean = 0, sd = 1)
```

we just pass both vectors to `t.test(...)`

```r
# Run a two sample t-test with null hypothesis of equal means
t.test(test_samples, control_samples)
```

```
##
##  Welch Two Sample t-test
##
## data:  test_samples and control_samples
## t = -2.2674, df = 91.187, p-value = 0.02573
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.86514760 -0.05716553
## sample estimates:
##  mean of x  mean of y
## -0.2332504  0.2279062
```

Data example

More generally we might have a data frame, like we have here as an example biometric data of Swiss army conscripts from Kaspar Staub (*University of Zurich*), which is a subset of historical data from Zurich.

```
# Read in the swiss army data from file
swiss_army_df <- read.csv("./data/Zurich_data.csv")
head(swiss_army_df)
```

```
##   Sex Height Weight
## 1   M  165.5     59
## 2   M  171.0     57
## 3   M  164.0     64
## 4   M  183.0     72
## 5   M  177.0     54
## 6   M  171.0     53
```

In the data we have three variables, *Sex*, *Height* and *Weight*.

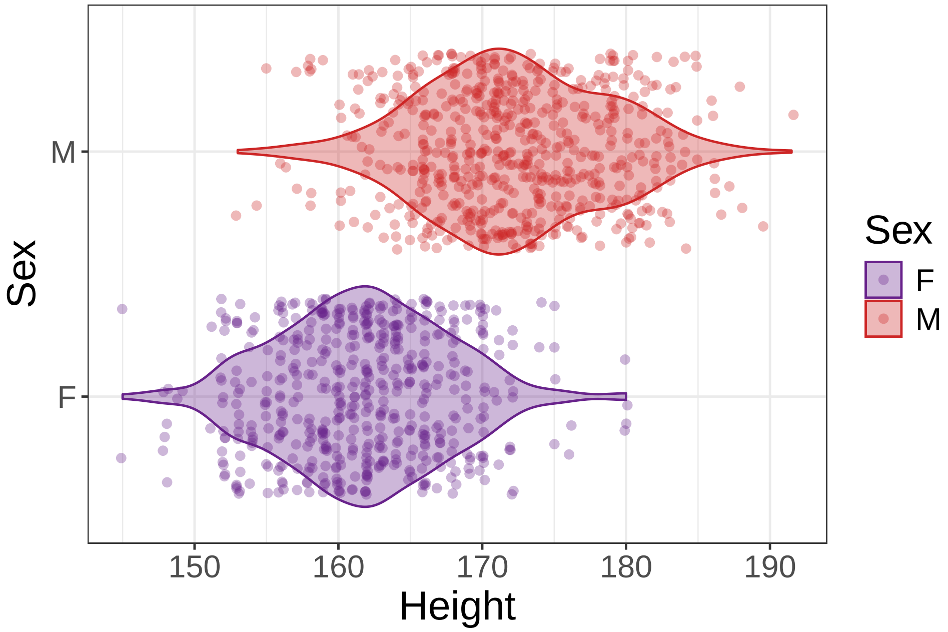
Sex and gender are complex attributes, and current standard for statistical surveys tend to define *Sex* as the sex recorded at birth or in infancy with the three categories, *female*, *male* and *other*. In surveys, *Gender* refers to identity, expression or experience and as a minimum may include category options of *female*, *male*, *non-binary*, *other* and the options to self-describe and not to respond. In historical data, as here, *Sex* often only has two levels recorded. This simplification is not intended to overlook or invalidate the diverse realities of sex and gender but is a practical limitation of the data itself.

With that in mind, we can test if there is a difference in average height between the two sexes in the dataset with a t-test:

```
# use the regression formulation
t.test(Height ~ Sex, swiss_army_df)
```

```
##
##  Welch Two Sample t-test
##
## data:  Height by Sex
## t = -29.278, df = 1152.2, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group F and group M is not equal to 0
## 95 percent confidence interval:
##  -11.151169  -9.750496
## sample estimates:
## mean in group F mean in group M
##        161.7041        172.1550
```

The difference is very significant with a very low p-value and you can clearly see the difference in means from a violin plot of the data



Although the distributions overlap, we are testing the difference in means so we compare the difference in units of the standard error which is very small with the large sample size.

Summary

We can now summarise our whistlestop recap of *t*-tests:

***p*-values** are a measure of surprise

- how unusual would the data be **under** the null?

For normal data the rescaled sample mean

- follows a **Student *t*-distribution**
- becomes more normal with larger sample sizes (CLT)

***t*-tests** can be employed to examine average differences

- one sample, paired samples and two samples

With more than two groups, **ANOVA** is a generalisation of *t*-tests.

t-tests assume no outliers and normality

- can check with **EDA** and **Q-Q plots**

→ Exercises 1