Mean comparison tests

Probabilités et statistique pour la biologie (STAT1)

Jacques van Helden

2019-10-10

Contents

Contents

We present here one of the most popular applications of statistics: the mean comparison test.

This test is used in a wide variety of contexts, and we will apply it here to two data types:

1. **Artificial data** generated by drawing samples in two populations following normal distributions, and whose means will be either identical or different, depending on the case. The goal will be to understand how to run the test and how to unterpret the results, in conditions where we control all the parameters (we know whether or not we created populations with equal or different means).

2. **Transcriptome data** obtained with microarrays. We will test whether a give gene is expressed differentially between two groups of biological samples (e.g. patients suffering from two different types of cancers).

   Note: the microarray technology has been replaced by NGS, and RNA-seq has now been widely adopted for transcriptome studies. However, differential analysis with RNA-seq data relies on more advanced concepts, which will be introduvced in other courses.

The hypotheses

General principle:

- We observe a difference between sample means, and we would like to know whether this difference results from sampling effects or from an actual difference between the population means.
- We oppone the ***null hypothesis*** (H_0) according to which there is no difference between the two populations, and the ***alternative hypothesis*** (H_1), which states that there is a difference.
- We evaluate the **P-value**, i.e. the probability that a random sampling under the null hypothesis would return a difference at least as high as what we observe.
- If this P-value is very weak (below a given threshold α), we reject the null hypothesis (RH_0), else we (temporarily) accept it (AH_0).

Two-tailed test

In the ***two-tailed test***, we are a priori interested by a difference in either direction ($\mu_1 > \mu_2$ or $\mu_2 > \mu_2$).

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

One-tailed test

In the ***one–tailed test***, we are specifically interested by detecting differences with a given sign. The null hypothesis thus covers both the equality and the differences with the opposite sign.

Positive difference (*right-tailed test*):

$$H_0 : \mu_1 \leq \mu_2$$
$$H_1 : \mu_1 > \mu_2$$

Negative difference (*left-tailed test*):

$$H_0 : \mu_1 \geq \mu_2$$
$$H_1 : \mu_1 < \mu_2$$

Assumptions

- There are several possible methods to test the mean equality.
- The choice of a method depends on the nature of the data.
- Before running the test, it is crucial to answer some questions in order to choose the appropriate method.
- This amounts to **check the assumptions**, i.e. a series of conditions of applicability for the selected method.

Normality assumption

Do the two populations to be compared follow normal distributions?

- If so, we can run **parametric tests** (which rely on a normality assumption).
- Else, we must use non-parametric tests.

  Why? Because the tables used to evaluate the probability of risk are derived from on mathematical models relying on a normality assumption.
- In case of non-normality, **do we dispose of large-sized samples?** If so, we can use parametric tests despite the non-normality.

Why? Because, by virtue of the **Central Limit Theorem**, the sample means tend towards a normal distribution even though the original populations are not normal.

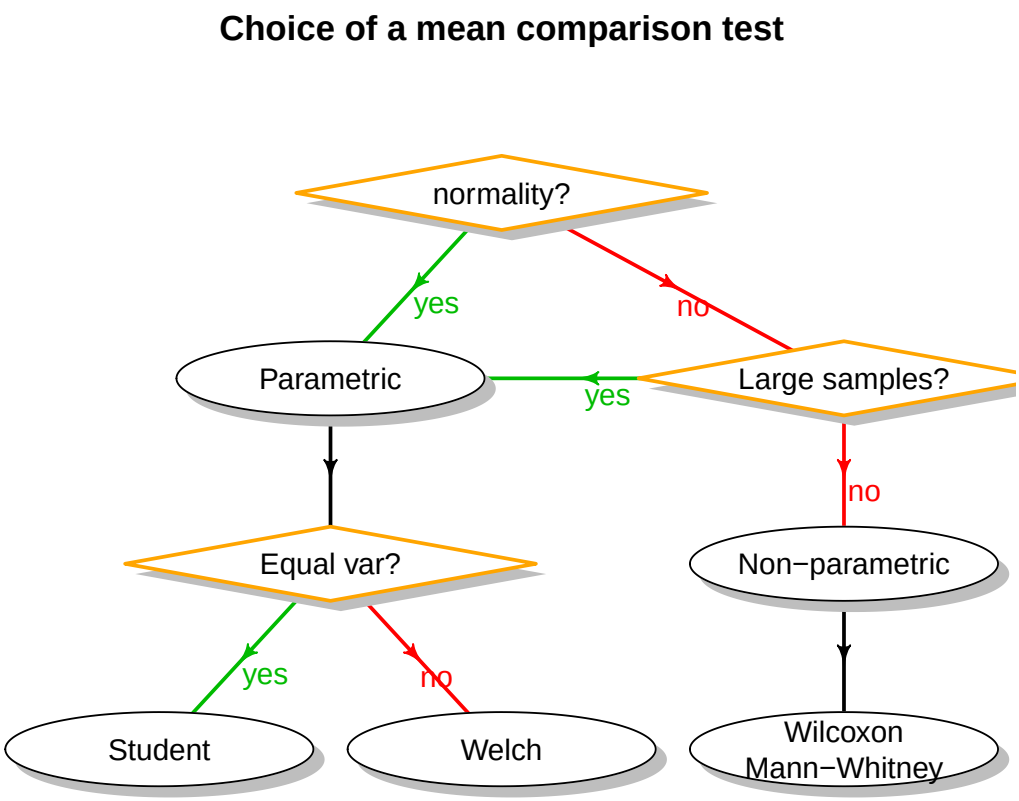
Homoscedasticity assumption (equality of population variances)

For parametric tests, the second question is: **do the two population have the same variance?**

- Yes → we can use Student test
- No → we must use Welch test

  Why? Student probability distribution was computed based on a homoscedasticity hypothesis. Welch test is a variation on Student test, which corrects the probabilities in case of **heteroscedasticity (unequal variances)** by modifying the number of degrees of freedom as a function of the difference between variances.

Flowchart for the choice of a mean comparison test



Student test

Assumptions: **normalité** (or large samples), **homoscedasticity**.

Student's statistics:

$$t_S = \frac{\hat{\delta}}{\hat{\sigma}_\delta} = \frac{\bar{x}_2 - \bar{x}_1}{\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Population and sample parameters

For a finite population, the ***population parameters*** are computed as follows.

Parameter	Formula
Population mean	$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$
Population variance	$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$
Population standard deviation	$\sigma = \sqrt{\sigma^2}$

Where x_i is the i^{th} measurement, and N the population size.

However, in practice we are generally not in state to measure x_i for all individuals of the population. We thus we have to ***estimate*** the population parameters (μ, σ^2) from a ***sample***.

Sample parameters are computed with the same formulas, restricted to a subset of the population.

Parameter	Formula
Sample mean	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Sample variance	$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$
Sample standard deviation	$s = \sqrt{s^2}$

where n is the sample size (number of sampled individuals), and $\bar{x}$ the sample mean.

Sampling issues: why *n-1* and not simply *n*?

The sample mean is an ***unbiased estimator*** of the population mean. Each time we select a sample we should expect some random fluctuations, but if we perform an infinite number of sampling trials, and compute the mean of each of these samples, the mean of all these sample means tends towards the actual mean of the population.

On the contrary, the sample variance is a ***biased estimator*** of the population variance. On the average, the sample variance under-estimates the population variance, but this can be fixed by multiplying it by a simple correcting factor: $n/(n-1)$.

Parameter	Formula
Sample-based estimate of the population mean	$\hat{\mu} = \bar{x}$
Sample-based estimate of the population variance	$\hat{\sigma}^2 = \frac{n}{n-1}s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
Sample-based estimate of the population standard deviation	$\hat{\sigma} = \sqrt{\hat{s}^2}$

Greek symbols (μ, σ: population parameters; Roman symbols ($\bar{x}$, s): sample parameters; the "hat" symbol $\hat{}$

reads "***estimation of***".

R `var()` and `sd()` functions

Beware: the **R** functions `var()` and `sd()` directly compute an estimate of the population variance ($\hat{\sigma}^2$) and standard deviation ($\hat{\sigma}$), respectively, instead of computing the variance ($\bar{x}$) and standard deviation (s) of the input data.

```
x <- c(1, 5)

n <- length(x) # Gives 2
sample.mean <- mean(x) # Gives 3
sample.var <- sum((x - sample.mean)^2) / n # Gives 4
pop.var.est <- sum((x - sample.mean)^2) / (n - 1) # Gives 8
r.var <- var(x) # Gives 8

kable(data.frame(sample.var = sample.var, pop.var.est = pop.var.est, r.var = r.var))
```

sample.var	pop.var.est	r.var
4	8	8

P-value

This statistics can be used to compute an ***P-value*** (P), which measures the probability to obtain, ***under the null hypothesis***, a t_S statistics at least as extreme as the one observed. Extreme refers here to the tail(s) of the distribution depending on the orientation of the test.

In the case of hypothesis testing, the P-value can be interpreted as an evaluation of the risk of **first kind error**, which corresponds to the risk of rejecting the null hypothesis whereas it is true. ## Degrees of freedom for Student tst

The shape of the Student distribution depends on a parameter named ***degrees of freedom*** (ν), which represents the number of independent variables used in the equation.

In a two-sample t-test (as in our case), the degrees of freedom are computed as the total number of elements in the respective samples (n_1, n_2) minus the number of means estimated from these elements (we estimated the means of group 1 and group 2, respectively). Thus:

$$\nu_S = n_1 + n_2 - 2$$

In classical texbooks of statistics, the p-value can be found in Student's t table.

With R, the p-value of a t statistics can be computed wiht the function $pt()$.

Student distribution

Decision

- Classically, one choses *a priori* a threshold on p-value, denoted α (e.g. $\alpha = 0.05$).
- If $P < \alpha$, the **null hypothesis is rejected** (RH_0) and the test is declared **positive**.
- If $P \geq \alpha$, the **null hypothesis is accepted** (AH_0) and the test is declared **negative**.

Orientation of the test	Decision criterion
Right-tailed	RH_0 if $t_S > t_{1-\alpha}^{n_1+n_2-2}$
Left-tailed	RH_0 if $t_S < t_{alpha}^{n_1+n_2-2} = -t_{1-\alpha}^{n_1+n_2-2}$

Orientation of the test	Decision criterion
Two-tailed	$RH_0 \quad \text{if} \quad \lvert t_S \rvert > t_{1-\frac{\alpha}{2}}^{n_1+n_2-2}$

The two-tailed test splits the risk symmetrically on the left and right tails ($\frac{\alpha}{2}$ on each side).

Welch *t* statistics

Welch's t-test defines the t statistic by the following formula.

$$t_W = \frac{\bar{x}_2 - \bar{x}_1}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Where:

- the indices 1 and 2 denote the respective populations
- $\bar{x_i}$ is the mean of sample i,
- s_i^2 the sample variance,
- n_i the sample size.

Welch degrees of freedom

The Welch test corrects the impact of heteroscedacity by adapting the degrees of freedom (ν) of the Student distribution according to the respective sample variances.

$$\nu_W = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2 \cdot (n_1 - 1)} + \frac{s_2^4}{n_2^2 \cdot (n_2 - 1)}}$$

- You generally don't have to compute this by yourselve, it is done automatically in R.
- Under homoscedasticity, Student and Welch degrees of freedom are equal.
- Under homoscedasticity, the ***power of the test*** (capacity to reject H_0 under H_1) is higher for Student than Welch.

Summary table: notations

Greek symbols (μ, σ) denote population-wide statistics, and roman symbols ($\bar{x}$, s) sample-based statistics.

Symbol	Description
μ_1, μ_2	Population means
$\delta = \mu_2 - \mu_1$	Difference beyween population means
σ_1, σ_2	Population standard deviations
n_1, n_2	Sample sizes
$\bar{x}_1, \bar{x}_2$	Sample means
$d = \bar{x}_2 - \bar{x}_2$	Effect size, i.e. difference between sample means
s_1^2, s_2^2	Sample variances
s_1, s_2	Sample standard deviations

Summary table: formulas

The "hat" ($\hat{}$) symbol is used to denote sample-based estimates of population parameters.

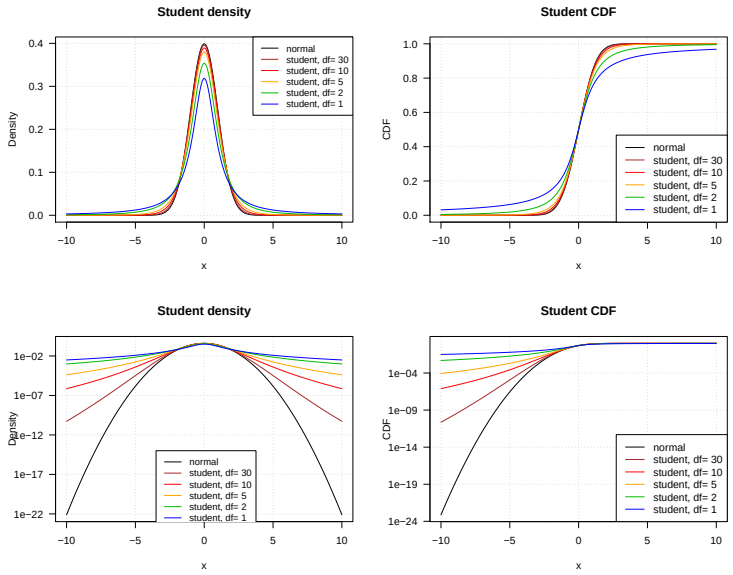


Figure 1: Student density (left) and CDF (right) on linear (top) or logarithmic (bottom) scale. The log scale highlights the rather low convergence in the tails.

Symbol	Description
$d = \hat{\delta} = \hat{\mu}_2 - \hat{\mu}_1 = \bar{x}_2 - \bar{x}_1$	d = Effect size (difference between sample means), estimator of the difference between population means δ.
$\hat{\sigma}_p = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1+n_2-2}}$	Estimate of the pooled standard deviation under variance equality assumption
$\hat{\sigma}_\delta = \hat{\sigma}_p \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$	Standard error about the difference between means of two groups whose variances are assumed equal (Student).
$t_S = \frac{\hat{\delta}}{\hat{\sigma}_\delta} = \frac{\bar{x}_2 - \bar{x}_1}{\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$	Student t statistics
$\nu_S = n_1 + n_2 - 2$	degrees of freedom for Student test
$t_W = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$	Welch t statistics
$\nu_W = \frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2 \cdot (n_1-1)} + \frac{s_2^4}{n_2^2 \cdot (n_2-1)}}$	degrees of freedom for Welch test

Exercise 1

A researcher analysed the level of expression of a gene of intrerest in 50 patients ($n_p = 50$)

Un chercheur a analysé, à l'aide de biopuces, le niveau d'expression de l'ensemble des gènes à partir d'échantillons sanguins prélevés chez 50 patients ($n_p = 50$) et chez 50 sujets sains ("contrôles", $n_c = 50$). He obtains the following results.

- for the patients, a mean expression level of $m_p = 21$
- for the controls, a mean expression level of $m_t = 10$
- the sample standard deviations are identical for the two groups: $s_p = s_c = s = 15$

In order to test whether the observed difference between the means is significant, the researcher decides to run a Student test.

a. Is the choice of this test appropriate? Justify. Which alternatives would have been conceivable?

- The sample standard deviations are identical for the two groups: $s_r = s_s = s = 2.8$.

a. Which method would you recommend to test mean equality? Justify.

b. Which are the assumptions of this test?

c. Assuming that these conditions are fulfilled, formulate the null hypothesis and compute the t statistics.

d. Evaluate the corresponding P-value basd on the Student's t table.

e. Based on these results, which decision would you take? Justify your answer