UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Escola Superior d'Enginyeries Industrial,
Aeroespacial i Audiovisual de Terrassa

Statistical Inference and Hypothesis Testing

Jose M Sallan
jose.maria.sallan@upc.edu

Quantitative Research Methods

May 8, 2018

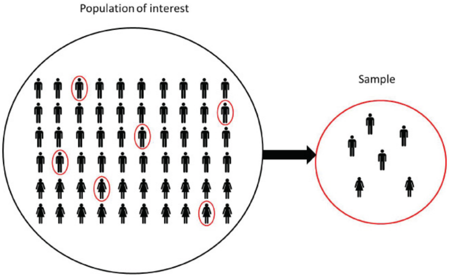
Outline

Statistical inference

Statistics is a branch of mathematics dealing with the collection, analysis, interpretation, presentation, and organization of data.

Descriptive statistics is solely concerned with properties of the observed data.

The aim of **inferential statistics** is to deduce properties about the probability distribution of a **population** analyzing data from a **sample**.



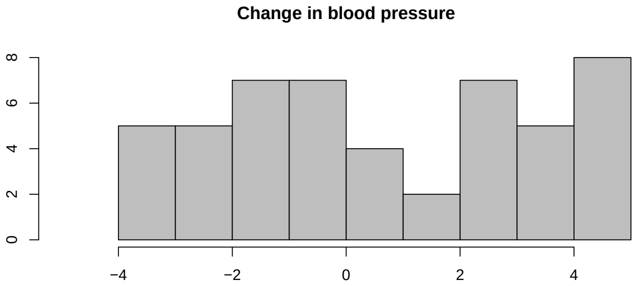
Source: `https://ies.ed.gov/blogs/nces/2016/04/05/default`

Statistical hypothesis testing

Statistical hypothesis testing is an attempt to **detect an effect** on a **population** from data taken from a **sample**.

Detecting an effect (mean)

Change in blood pressure after taking experimental drug A in $n = 50$ participants:



Has the drug any effect? $\Rightarrow$ Is the **population mean** of change of blood pressure **different from zero**?

Null and alternative hypothesis

The first step to hypothesis testing is to define null and alternative hypotesis:

- **Null hypothesis:** there is no effect in the population.
- **Alternative hypothesis:** there is effect in the population.

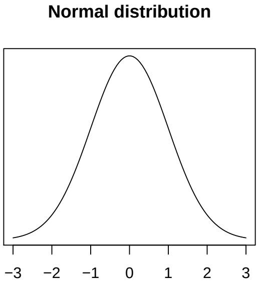
In this case, absence of effect means that the population mean μ of differences is equal to zero:

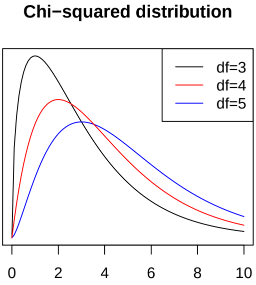
- $H_0 : \mu = 0$
- $H_1 : \mu \neq 0$

Then we need to know what is the **probability distribution of sample variable** if H_0 is true.

Probability distributions

Normal $N(\mu, \sigma)$ and the **chi-square** χ^2.





Any normal distribution can be transformed into the **standard normal distribution** $N(0, 1)$ by:

$$\bar{z} = \frac{x - \mu}{\sigma}$$

The central limit theorem

The **central limit theorem** for the sample mean establishes that the sample mean $\bar{x}$ computed with n elements of a random variable x of mean μ and variance σ follows a normal distribution with mean μ and variance $\sigma/\sqrt{n}$.

The variable:

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

follows a $N\left(0, 1\right)$ distribution.

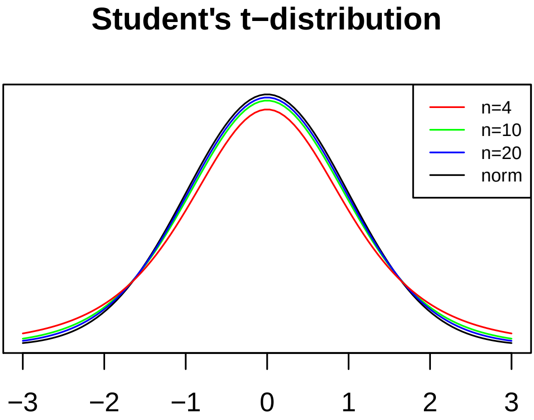
The Student's t distribution

Usually we don't know the population standard deviation σ, but the sample standard deviation s. The variable:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

Follows a Student's t-distribution of $n - 1$ degrees of freedom.

The Student's t distribution



For large enough values of n, Student's t is similar to a normal distribution.

Null and alternative hypohesis

We have to test the null hypotesis:

- $H_0 : \mu = 0$
- $H_1 : \mu \neq 0$

Now we know that, **if** H_0 **is true**, the variable:

$$t = \frac{\bar{x}}{s/\sqrt{n}}$$

follows a Student's t-distribution of $n - 1 = 49$ degrees of freedom.

Null and alternative hypohesis

For the sample of drug A, we know that:

$$\bar{x} = 0.5872 \qquad\qquad s = 2.762$$

If H_0 is true, then $\mu = 0$ so the standardized value is:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{0.5872 - 0}{2.762/\sqrt{50}} = 1.503$$

Is this a surprising value on a t_{49} distribution?

Introducing p-values

To rate the extent to an observed value of the probability distribution is surprising, we use p-values.

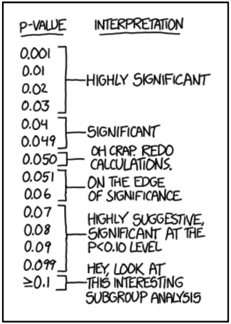
A **p-value** is the probability of getting the observed or a more extreme value, assuming that the null hypothesis is true.

The p-value for $t_{49} = 1.503$ is $p = 0.1392$.

We can observe a value of 1.503 or higher coming from a t_{49} distribution with a probability of 13.92%. **Is this p-value high or low?**

Thresholds of p-values

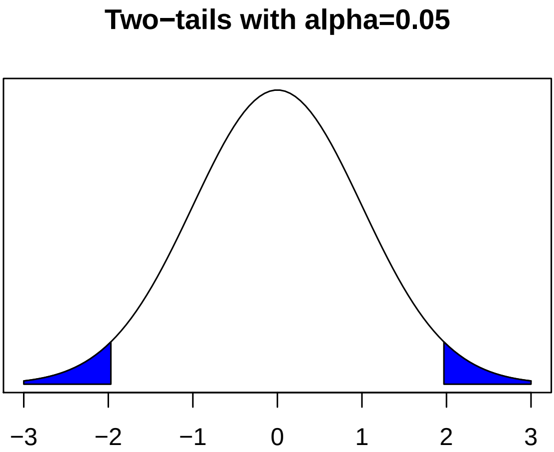
We can consider that we have found a significant effect when $p < 0.05$. This is an arbitrary value, arising from common practice.



Source: `https://xkcd.com/1478/`.

Thresholds of p-values

In a two-tailed test, we assume that values can be positive or negative, so the probability is split on both sides. This are the tails of a two-tailed test for $p = 0.05$ (values ± 1.96):



Testing the null hypothesis

In the example of drug A we have that the p-value obtained is larger than 0.05, so we **cannot reject the null hypothesis**.
We can do it faster with **R** using the `t.test` function (values are stored in vector a):

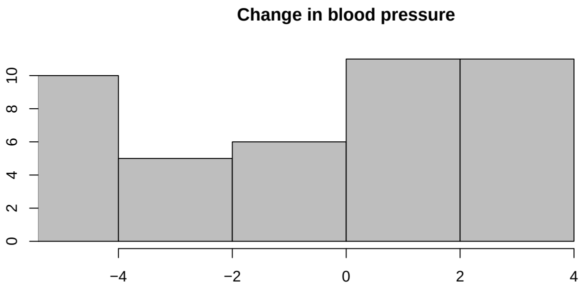
```
> t.test(a)

        One Sample t-test

data:  a
t = 1.5032, df = 49, p-value = 0.1392
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.1978011  1.3721480
sample estimates:
mean of x
0.5871735
```

Testing the null hypothesis

We have made another test with a drug B, obtaining the following values:



Testing the null hypothesis

For this sample, we obtain that $p < 0.01$, therefore we can reject the null hypothesis:

```
> t.test(b)

        One Sample t-test

data:  b
t = -2.8835, df = 49, p-value = 0.005827
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -2.449835 -0.437545
sample estimates:
mean of x
 -1.44369
```

Means comparison

Two possible contexts for mean comparison:

- **Paired data:** comparison of two observations taken from the same sample (e.g., scores before and after taking a course).
- **Non-paired data:** comparison of means of two independent samples.

See details of implementation in **R** in:
`https://www.statmethods.net/stats/ttest.html`.

Correlation

The test $H_0 : \rho = 0$ for the population Pearson correlation let us know if a relationship exists between two variables.

If H_0 is true and r is the sample correlation taken from n observations, the value:

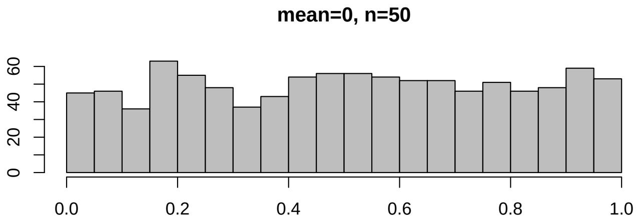
$$\frac{r}{\sqrt{1 - r^2}}\sqrt{n - 2}$$

follows a t_{n-2} distribution.

We can perform correlational analysis using functions of `psych` and `corrplot` packages.

Distribution of p-values when H_0 is true

Let's obtain 1,000 mean samples of $n = 50$ from a variable with $\mu = 0$, and let's compute the p-value of $H_0 : \mu = 0$ for each:

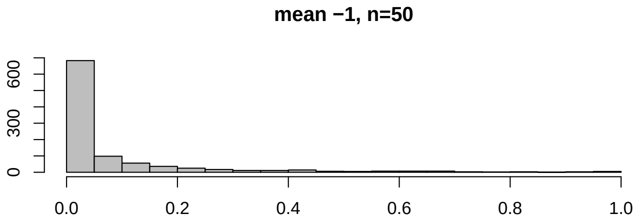


We observe that:

- p-values are uniformly distributed, if the number of samples is large enough.
- There is a probability $p = 0.05$ of rejecting H_0 when it is true. This is a **Type I error**.

Distribution of p-values when H_0 is false

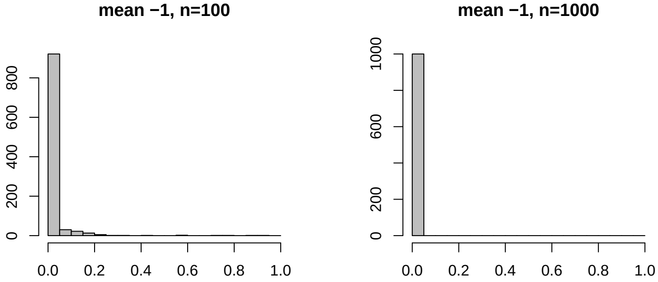
Let's obtain 1,000 mean samples of $n = 50$ from a variable with $\mu = -1$, and let's compute the p-value of $H_0 : \mu = 0$ for each:



Now most of the p-values are smaller than 0.05, but the 34.2 % are larger. In this cases we are accepting H_0 when it is false. This is a **Type II error**.

Distribution of p-values when H_0 is false

We can **reduce the Type II error** rate **increasing the sample size** (rate is 7.9% for $n = 100$ and 0% for $n = 1000$):



Type I and type II errors

	H_0 true	H_0 false
Accept H_0: non-significant finding.	Correct inference $(1-\alpha)$: true negative.	Type II error β: false negative.
Reject H_0: significant finding.	Type I error α: false positive.	Correct inference $(1-\beta)$: true positive.

- α is the Type I error rate. Can be controlled setting a p-value.
- β is the Type II error rate, and $1-\beta$ the statistical power of the test.

Type I and type II errors

Let's consider that:

- There is a 50% probability that an effect really exists.
- We have fixed a Type I eror rate of $\alpha = 0.05$ and $1 - \beta = 0.8$.

	H_0 true **50%**	H_0 false **50%**
Accept H_0: non-significant finding.	Correct inference: true negative. $0.5 * 0.95 = \mathbf{0.475}$	Type II error: false negative. $0.5 * 0.2 = \mathbf{0.1}$
Reject H_0: significant finding.	Type I error: false positive. $0.5 * 0.05 = \mathbf{0.025}$	Correct inference: true positive. $0.5 * 0.8 = \mathbf{0.4}$

... the most likely result is true negative.

Learn more at: `http://rpsychologist.com/d3/NHST/`

Multiple tests

Performing an analysis with multiple tests may lead to **error inflation**. If we set a Type I error rate of α, the probability of having at least one Type I error when performing k tests is:

$$1 - (1 - \alpha)^k$$

Some values of errors:

k	error
1	0.05
2	0.0975
5	0.226
10	0.401

Multiple tests

Type I error inflation can be corrected being more exigent with α when multiple test are performed.

Bonferroni correction: set Type I error rate to α/k. It is considered too conservative, and can yield to higher Type II errors. Other methods can be found in:
`https://en.wikipedia.org/wiki/Family-wise_error_rate`.

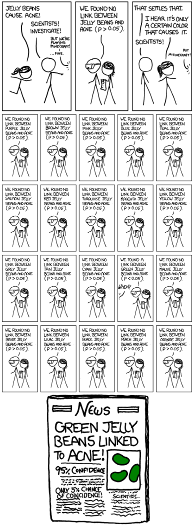
Replication crisis

Hypothesis testing is a powerful tool to make scientific discoveries through statistical inference.

A careless use of hypothesis testing may lead to misleading results, as unnoticed Type I or Type II error rates:

- **p-hacking**: making multiple (unreported) statistical tests until a "good" p-value appears (e.g., optional stopping).
- **HARKing**: hypothesizing after the results are known.
- **Publication bias**: journal editors discourage publication of non-significant results and replication studies.

https://en.wikipedia.org/wiki/Replication_crisis



Source: `https://xkcd.com/882/`.

Replication crisis

Remedies for replication crisis:

- Larger samples, lower Type I error rates.
- Encourage replication studies.
- Tackling publication bias with pre-registration of studies.
- Sharing raw data in online repositories.