University of BRISTOL

Spring Semester 2020

Lecture 4: Hypothesis testing

EMAT30007 Applied Statistics

Nikolai Bode & Filippo Simini

Department of Engineering Mathematics

Outline of the lecture

In this lecture you will learn:

- How to formulate and test a hypothesis.
- How to compute p-values and interpret the results of hypothesis tests.
- How to use the z-test to test if a population has a given mean.
- How to use the t-test to test if a population has a given mean.
- How to evaluate and interpret the errors of a test.

Hypothesis testing

In most quantitative disciplines, including science, engineering, medicine, knowledge is extracted from data through a process called hypothesis testing, which is a fundamental part of statistical inference. It works as follows:

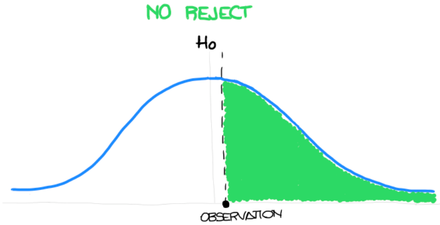
- Researchers formulate a hypothesis, for example “this coin is fair”.
- Relevant data is collected, for example the coin is tossed 10 times.
- Data is analysed and the hypothesis is either rejected or not. For example, given that the observed number of heads is not significantly different from the number of heads expected for a fair coin, we conclude that we cannot reject the hypothesis that the coin is fair.

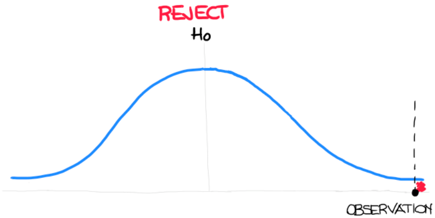
In this lecture we’ll learn how to test a hypothesis using data.

Significance test

We use significance tests to determine whether a given empirical estimate, such as a sample mean, is a likely outcome of a sampling distribution specified by a given hypothesis.

That is, an estimated quantity is compatible with a hypothesis if any deviation of our estimate from the value we expect (if the hypothesis were true) could reasonably be ascribed to the randomness introduced in selecting the sample.

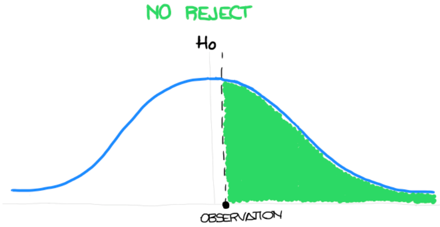


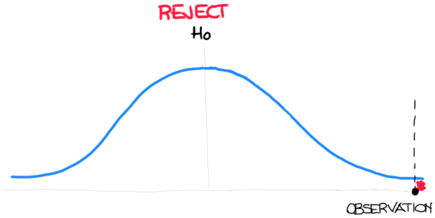


Rationale

The reasoning of tests goes like this:

1. Choose an estimator for the quantity that is the subject of your hypothesis.
2. Construct the sampling distribution for the estimator under the assumption that the hypothesis is true.
3. Compute the probability to draw from the sampling distribution a more extreme value (larger or smaller) than the one observed: if this probability is very small, it means that it would be very rare to observe the estimated quantity by chance, and so there is evidence to reject our hypothesis.





Formulate a null hypothesis

The first step is to formulate a testable hypothesis, called null hypothesis, H_0.

A well-defined null hypothesis should specify:

- A quantity of interest in a population and its value if H_0 is true.
- An estimator for the quantity of interest, called test statistic.

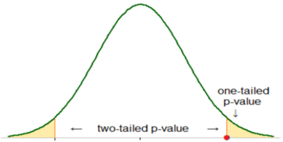
Examples of null hypotheses

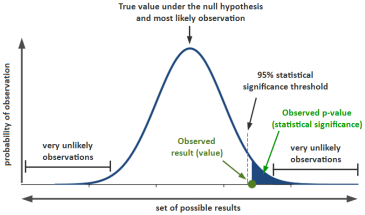
- "The mean height of students at the University is 173 cm"
  - Quantity and value: mean, $\mu_0 = 173$.
  - Test statistic: mean of heights, $\hat{\mu} = \bar{X}$.
- "This drug does not cure that disease"
  - Quantity and value: mean number of non-treated patients who recover, p_0.
  - Test statistic: mean number of treated patients who recover, $\hat{p} = \bar{X}$.
- "Global temperature and CO_2 concentration are not correlated"
  - Quantity and value: Pearson correlation, $c_0 = 0$ (no correlation).
  - Test statistic: sample correlation coefficient, $\hat{c} = Corr(T, CO_2)$.

p-value

After specifying the null hypothesis, we compute the p-value, p, which is the probability to observe a more extreme value of the test statistic, under the assumption that the null hypothesis is true.

- A two-tailed p-value is the probability to observe a more extreme value on **either** side of the sampling distribution: $p = 2\min\{Pr(X \leq x), Pr(X \geq x)\} = 2\min\{F_X(x), 1 - F_X(x)\}$, where x is the value of the test statistics and F_X is the sampling distribution's CDF under a true H_0.
- A right-tailed p-value (or left-tailed p-value) is the probability to observe a more extreme value only on the **right** (or **left**) tail of the sampling distribution:
  $p = Pr(X \geq x) = 1 - F_X(x)$ (right p-val),
  $p = Pr(X \leq x) = F_X(x)$ (left p-val).




True value under the null hypothesis and most likely observation
probability of observation
95% statistical significance threshold
Observed p-value (statistical significance)
very unlikely observations
Observed result (value)
very unlikely observations
set of possible results

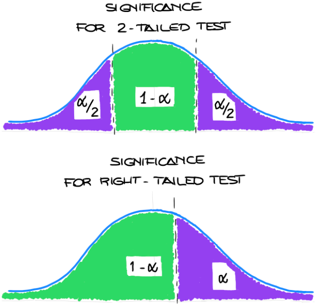
A low p-value (e.g. $p = 0.01$) means that there is a small chance (1%) to observe a value of the test statistic at least as extreme, when H_0 is true.

Reject an hypothesis

An observation is statistically significant (i.e. it is very unlikely to have occurred given the null hypothesis) if its p-value, p, is smaller than a specified significance level (or threshold) α.

Comparing the p-value p and the significance threshold α we decide whether an observation is statistically significant and we have enough evidence to reject the null hypothesis:

- If $\boldsymbol{p \leq \alpha}$, the observation is statistically significant and **the null hypothesis is rejected**.
- If $\boldsymbol{p > \alpha}$, there isn't enough evidence to disproof H_0 and **we don't reject the null hypothesis**.



Reject H_0 if observation falls in purple area.

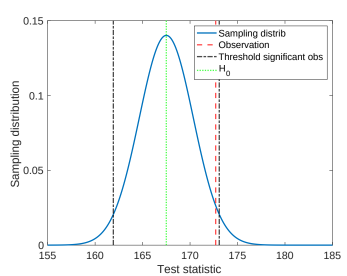
z-test for a population mean (with known variance)

Example: *z-test for mean student heights (two-tailed)*

We measure the height of $n = 10$ students chosen at random and we observe the sample $x = [176, 165, 189, 180, 172, 169, 162, 161, 183, 170]$ cm.

Assuming that a student's height is a Normal RV with standard deviation $\sigma = 9$ cm, We want to know if the mean height of the students in the University is **significantly different** from 167.5 cm, at a significance level of 5%.

- H_0: μ is not different from $\mu_0 = 167.5$
- Test statistic: $\hat{\mu} = \bar{X}$. Observation: $\bar{x} = 172.7$
- Sampling distribution: $\bar{X} \sim N(\mu_0, \sigma^2/n)$
- Two-tailed P-value: $p = 0.0987$ using MATLAB's `2*(1-normcdf(172.7, 167.5, 9/10`$^{0.5}$`))`
- Significance threshold: $\alpha = 0.05$



Result: $p > \alpha$, so we don't reject H_0 because the observed value is not among the 5% farthest values from μ_0 in both directions (the highest 2.5% and lowest 2.5%).

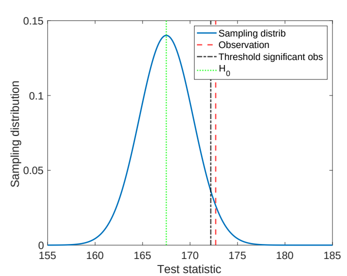
z-test if a population mean exceeds a threshold

Example: *z-test for mean student heights (right-tailed)*

We measure the height of $n = 10$ students chosen at random and we observe the sample $x = [176, 165, 189, 180, 172, 169, 162, 161, 183, 170]$ cm.

Assuming that a student's height is a Normal RV with standard deviation $\sigma = 9$ cm, We want to know if the mean height of the students in the University is **significantly higher** than 167.5 cm, at a significance level of 5% (in both directions).

- H_0: μ is not higher than $\mu_0 = 167.5$
- Test statistic: $\hat{\mu} = \bar{X}$. Observation: $\bar{x} = 172.7$
- Sampling distribution: $\bar{X} \sim N(\mu_0, \sigma^2/n)$
- Right-tailed P-value: $p = 0.0493$ using MATLAB's `1-normcdf(172.7, 167.5, 9/10`$^{0.5}$`)`
- Significance threshold: $\alpha = 0.05$



Result: $p < \alpha$, so we reject H_0 because the observed value is among the 5% highest values (farthest from μ_0 in the right direction).

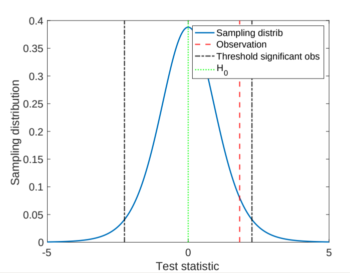
t-test for population mean (with unknown variance)

Example: *t-test for mean student heights (two-tailed)*

We measure the height of $n = 10$ University students chosen at random and we observe the sample $x = [176, 165, 189, 180, 172, 169, 162, 161, 183, 170]$ cm.

Assuming that a student's height is a Normal RV with unknown standard deviation, we want to know if the mean height of the students in the University is **significantly different** from 167.5 cm, at a significance level of 5% (in both directions).

- H_0: μ is not different from $\mu_0 = 167.5$
- Test statistic: $T = \frac{\bar{X}-\mu_0}{(S/\sqrt{n})}$ where $\bar{X}$ is the sample mean and S the sample standard deviation. $\bar{x} = 172.7$ and $s = 9.24$, observation $t = 1.61$
- Sampling distribution: t-distribution with $n-1$ DoF
- Two-tailed P-value: $p = 0.1422$ using MATLAB's `2*(1-tcdf((172.7-167.5)/(9.24/10^0.5),9))`
- Significance threshold: $\alpha = 0.05$



Result: $p > \alpha$, we don't reject H_0.

Welch's t-test for equal means of two populations

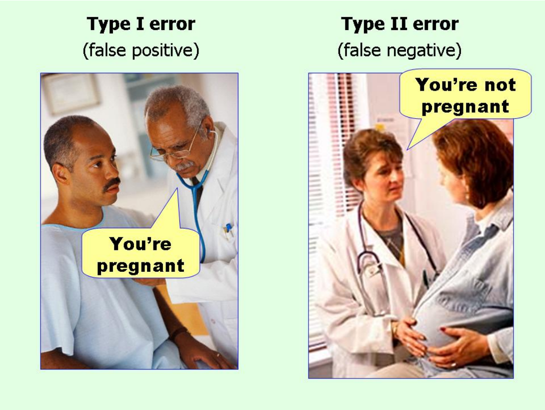
Example: *Do students at UoB and UWE have the same mean height?*

We measure the height of $n = 10$ University of Bristol students and we observe a sample mean and standard deviation of $\bar{x} = 172.7$cm and $s_x = 9.24$cm respectively. The mean and standard deviation of a sample of $m = 15$ students at the University of the West of England are $\bar{y} = 169.8$cm and $s_y = 8.95$cm.

Assuming that heights of university students follow Normal distributions that may have different means and variances in different universities, we want to know if the mean height of the students in UoB and UWE are **significantly different**, at a significance level of 1%.

- H_0: means at UoB and UWE are not different: $\mu_X - \mu_Y = 0$.
- Test statistic: $T = \frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sqrt{\frac{(n-1)S_X^2+(m-1)S_Y^2}{n+m-2}\left(\frac{1}{n}+\frac{1}{m}\right)}}$, observation $t = 0.715$
- Sampling distribution: t-distribution with $r = n + m - 2$ DoF
- Two-tailed P-value: $p = 0.48 > \alpha = 0.01$, we don't reject H_0. Data are compatible with the hypothesis that students in the two universities have the same mean height.

Type I and II errors



effectsizefaq.com

Null hypothesis:

"You are not pregnant"

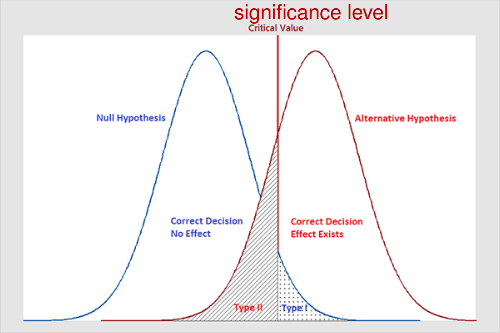
Errors:

- False positive (type I): reject the null hypothesis when it is true (false discovery).
- False negative (type II): don't reject the null hypothesis when it is false (failed discovery).

Type I and II errors

		reality	
		H_0 true	H_0 false
test result	H_0 not rejected	**TN**: True Negative	**FN**: False Negative
	H_0 rejected	**FP**: False Positive	**TP**: True Positive

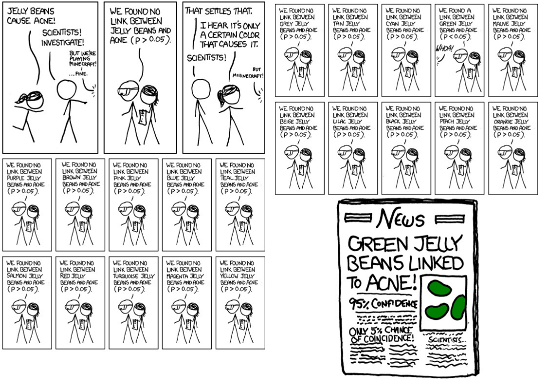
Confusion matrix



https://statisticsbyjim.com/hypothesis-testing/types-errors-hypothesis-testing/

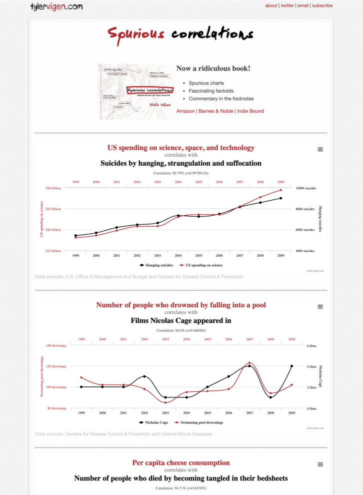
- The significance level α is the probability to reject the null hypothesis if it is true, hence it is the probability of type I errors (false positive): $\alpha = \frac{FP}{TN+FP}$.
- The probability of type II errors (false negative) is $\beta = \frac{FN}{FN+TP}$.
- Decreasing the significance level (α) to reduce the probability of type I errors will increase the probability of type II errors (and viceversa).

False positive (type I errors)


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

https://imgs.xkcd.com/comics/significant.png

With a significance level of $\alpha = 0.05$ we expect a false discovery every 20 experiments.


tylervigen.com

about | twitter | email | subscribe

Spurious correlations

Now a ridiculous book!

- Spurious charts
- Fascinating factoids
- Commentary in the footnotes

Amazon | Barnes & Noble | Indie Bound

US spending on science, space, and technology
correlates with
Suicides by hanging, strangulation and suffocation

Correlation: 99.79% (r=0.99789126)

Data sources: U.S. Office of Management and Budget and Centers for Disease Control & Prevention

Number of people who drowned by falling into a pool
correlates with
Films Nicolas Cage appeared in

Correlation: 66.6% (r=0.666004)

Data sources: Centers for Disease Control & Prevention and Internet Movie Database

Per capita cheese consumption
correlates with
Number of people who died by becoming tangled in their bedsheets

Correlation: 94.71% (r=0.947091)

https://www.tylervigen.com/spurious-correlations

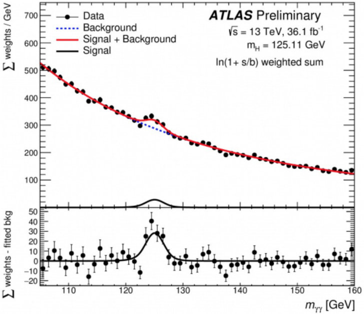
Spurious correlations

False positive (type I errors)

To reduce the chance of announcing false discoveries, particle physicists set their significance thresholds at *very* low levels:

- $\alpha = 1 - F_{N(0,1)}(3) = 0.0013$
  or once every 741 times:
  “evidence” level

- $\alpha = 1 - F_{N(0,1)}(5) = 0.0000003$
  or once every 3,486,914 times:
  “discovery” level



https://medium.com/@chris.m.pease/the-higgs-boson-and-5-sigma-eec238b43f93

Evidence of the discovery of the Higgs boson by the ATLAS experiment

https://www.science20.com/quantum_diaries_survivor/fundamental_glossary_higgs_broadcast-85365

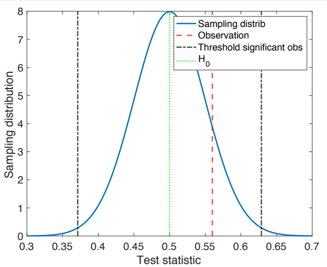
z-test for a population proportion

Example: *z-test if a coin is fair (two-tailed)*

A coin is tossed 100 times and we observe 56 heads.

Is the coin fair at a significance level of 1%?

- H_0: p (head probability) is not different from $p_0 = 0.5$
- Test statistic: $\hat{p} = \bar{X}$. Observation: $\bar{x} = 0.56$
- Sampling distribution: $\bar{X} \sim N(p_0, p_0(1-p_0)/n)$
- Two-tailed P-value: $p = 0.23$ using MATLAB's `2*(1-normcdf(0.56,0.5,(0.25/100)`$^{0.5}$`))`
- Significance threshold: $\alpha = 0.01$



Result: $p > \alpha$, so we don't reject H_0 because the observed value is not among the 1% farthest values from p_0 in both directions (the highest 0.5% and lowest 0.5%).

How many tosses are needed before 0.56 becomes significant at 1%?

$$0.56 = p_0 + z_{1-\alpha/2}\sqrt{p_0(1-p_0)/n} \quad \Rightarrow \quad n = \left(\frac{\sqrt{p_0(1-p_0)}}{(0.56-p_0)} z_{1-\alpha/2}\right)^2 \simeq 461$$

False negative (type II errors)

Computing the probability of type II errors is very difficult in general, because it requires to know the probability of all outcomes (the sampling distribution) when the null hypothesis is false.
We can estimate type II errors when we can fully specify a well-defined alternative hypothesis, H_1, that must be true if the null hypothesis is false.

In some cases it is preferrable to reduce the probability of false negatives, for example in a test for a deadly disease: We would rather be wrong when telling a healthy person they're ill (false positive) rather than be wrong when telling an ill person they're healthy (false negative).

To reduce the probability of false negative errors we can:

- Increase the significance level (α).
- Increase the sample size (this would reduce the variance of the sampling distribution).

Probability of type II errors

A deadly disease affects 5% of the population. A test is developed that claims to be 90% accurate, that is, it predicts the correct outcome in 9 out of 10 tests, and the probability of a false positive, α, is 0.1. The test is tested on 1000 random individuals.

1. Fill out all the elements of the confusion matrix.
2. What is the probability of type II errors?
3. What is the probability to have the disease if the test is negative?
4. What is the probability of not having the disease if the test is positive?

$$\begin{cases} \frac{TP+FN}{TP+TN+FP+FN} & = 0.05 \\ \frac{TP+TN}{TP+TN+FP+FN} & = 0.9 \\ \frac{FP}{TN+FP} & = \alpha \\ TP+TN+FP+FN & = 1000 \end{cases}$$

		reality		
H_0: “no disease”		H_0 true (no disease)	H_0 false (disease)	
test result	H_0 not rejected (no disease)	TN = 855	FN = 5	860
	H_0 rejected (disease)	FP = 95	TP = 45	140
		950	50	1000

2. The probability of type II errors, i.e. to fail to detect the disease, is $\frac{FN}{FN+TP} = 10\%$.
3. The probability to have the disease if the test is negative is $\frac{FN}{TN+FN} = 0.6\%$.
4. The probability to not have the disease if the test is positive is $\frac{TP}{TP+FP} = 32\%$.