

Lecture 4: Hypothesis testing EMAT30007 Applied Statistics

Nikolai Bode

Department of Engineering Mathematics

Spring Semester

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.

Spring Semester

Lecture 4, Topic 1 – Hypothesis testing

@ntent covered

- Significance tests
- Null Hypotheses
- p-values

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.
- Lata 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.

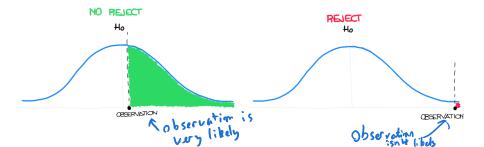
In this lecture we'll learn how to test a hypothesis using data.



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.

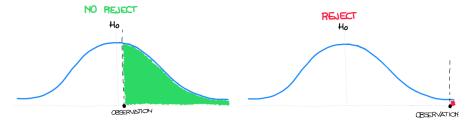




Rationale

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:
 - \blacktriangle 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"
 - \blacktriangleright Quantity and value: mean number of non-treated patients who recover, p_0 .
 - ▶ Test statistic: mean number of treated patients who recover, $\hat{p} = \bar{X}$.
- - ightharpoonup Quantity and value: Pearson correlation, $c_0 = 0$ (no correlation).
 - ▶ Test statistic: sample correlation coefficient, $\hat{c} = Corr(T, CO_2)$.



p-value

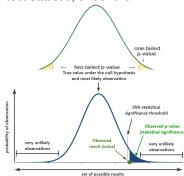
After specifying the null hypothesis, we compute the p-value, p, which is the probability to observe a more extreme value of the test statistic, under the assumption that the null hypothesis is true.

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:

$$\begin{aligned} p &= Pr(X \geq x) = 1 - F_X(x) \text{ (right p-val)}, \\ p &= Pr(X \leq x) = F_X(x) \text{ (left p-val)}. \end{aligned}$$



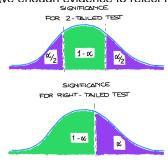
A low p-value (e.g. p=0.01) means that there is a small chance (1%) to observe a value of the test statistic at least as extreme, when H_0 is true.

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 $p \le \alpha$, the observation is statistically significant and the null hypothesis is rejected.
- If $p > \alpha$, 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.

Lecture 4, Topic 2 – Examples for hypothesis tests

Ontent covered

- ► z—tests
- ► t-tests
- ▶ Welch's t−test



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.

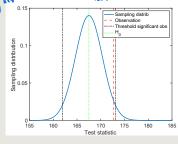
Assuming that a student's height is a Normal RV with standard deviation $\sigma = 9$ cm. We want to know if the mean height of the students in the University is significantly different

 $H_0\colon \mu \text{ is not different from } \mu_0=167.5 \text{ May have the first state of the first st$ Key Test statistic: $\hat{\mu} = \bar{X}$. Observation: $\bar{x} = 172.7$

Ke Sampling distribution: $\bar{X} \sim N(\mu_0, \sigma^2/n)$

 \not Two-tailed P-value: p = 0.0677 using MATLAB's $2*(1-normcdf(172.7, 167.5, 9/10^{0.5}))$

K 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

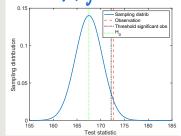
Example: *z-test for mean student heights (right-tailed)*

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).

- $\not\vdash H_0$: μ is not higher than $\mu_0=167.5$
- **W** Test statistic: $\hat{\mu} = \bar{X}$. Observation: $\bar{x} = 172.7$
- **K** 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)*

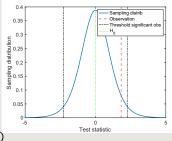
We measure the height of n=10 University students chosen at random and we observe the sample x=[176,165,189,180,172,169,162,161,183,170] cm.

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).

- \not H_0 : μ is not different from $\mu_0=167.5$
- $\text{ Test statistic: } T = \frac{\bar{X} \mu_0}{(S/\sqrt{n})} \text{ where } \bar{X} \text{ is the sample mean and } S \text{ the sample standard deviation.}$

$$\bar{x}=172.7$$
 and $s=9.24$, observation $t=1.61$

- $\ensuremath{\mathbb{K}}$ Sampling distribution: t-distribution with n-1 DoF
- $\begin{tabular}{ll} & \textbf{Two-tailed P-value: } p=0.1422 \ \text{using MATLAB's} \\ & 2*(1-tcdf((172.7-167.5)/(9.24/10^{0.5}),9)) \\ \end{tabular}$
- & 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 $\bar{x}=172.7 \mathrm{cm}$ and $s_x=9.24 \mathrm{cm}$ respectively. The mean and standard deviation of a sample of m=15 students at the University of the West of England are $\bar{y}=169.8 \mathrm{cm}$ and $s_y=8.95 \mathrm{cm}$.

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%.

- $\not\vdash H_0$: means at UoB and UWE are not different: $\mu_X \mu_Y = 0$.
- $\text{ Test statistic: } T = \frac{(\bar{X} \bar{Y}) (\mu_X \mu_Y)}{\sqrt{\frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}(\frac{1}{n} + \frac{1}{m})}}, \text{ observation } t = 0.715$
- $\ensuremath{\mathsf{k}}$ Sampling distribution: t-distribution with r=n+m-2 DoF
- We 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.

Lecture 4, Topic 3 – Type I and type II errors

Ontent covered

- Type I and type II errors
- Confusion matrix
- Reducing the probability of tyoe I and type II errrors



Type I and II errors

Type I error (false positive)



Type II error (false negative)



effectsizefaq.com

Null hypothesis:

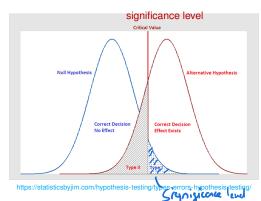
- "You are not pregnant" Errors:
- False positive (type I): reject the null hypothesis when it is true (false discovery).
- False negative (type II): don't reject the null hypothesis when it is false (failed discovery).



Type I and II errors

	reality			
		H₀ true	H_0 false	
test	H₀ not rejected	TN: True Negative	FN: False Negative	
result	<i>H</i> ₀ rejected	FP: False Positive	TP : True Positive	

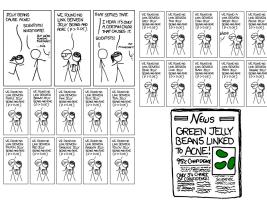
Confusion matrix



- 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): $\alpha = \frac{FP}{TN+FP}$.
- \checkmark The probability of type II errors (false negative) is $\beta = \frac{FN}{FN+TP}$.
- Lecreasing 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)



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

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



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

Spurious correlations



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.

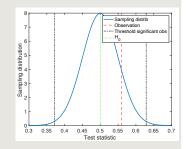
Is the coin fair at a significance level of 1%?

 H_0 : p (head probability) is not different from $p_0=0.5$

K Test statistic: $\hat{p} = \bar{X}$. Observation: $\bar{x} = 0.56$

 \swarrow Sampling distribution: $\bar{X} \sim N(p_0, p_0(1-p_0)/n)$

 $\begin{tabular}{ll} \begin{tabular}{ll} \be$



& 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:

- \normalfont{k} 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?

	· · · · · · · · · · · · · · · · · · ·	-	_		-		
	TP+FN	0.05	H_0 : "no disease"		reality		
	$\overline{TP+TN+FP+FN}$	= 0.05			H_0 true	H_0 false	
	$\frac{TP+TN}{TP+TN+FP+FN}$	= 0.9			(no disease)	(disease)	
{	FP			H_0 not rejected	TN = 855	FN = 5	860
	$\overline{TN+FP}$	$= \alpha$	test result	(no disease)			
	TP+TN+FP+FN	= 1000		H_0 rejected	FP = 95	TP = 45	140
				(disease)			
					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 $\frac{FP}{TP+FP}=68\%$.