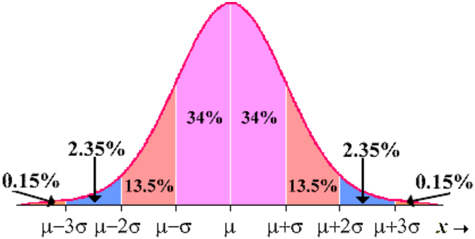
COMP229: Introduction to Data Science
Lecture 10: Hypothesis and significance

Olga Anosova, O.Anosova@liverpool.ac.uk
Autumn 2023, Computer Science department
University of Liverpool, United Kingdom

Last lecture recap

$X \sim N(\mu, \sigma^2)$ has $\phi_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.

CLT: "in the limit any average is normal".



68%-95%-99.7% rule

confidence level	90%	95%	99%
critical value z^*	$1.645 \approx 1.7$	$1.96 \approx 2$	$2.576 \approx 2.6$

Statistical inference gives an interval estimate for μ from $X \sim N(\mu, \sigma^2)$: $P(\mu \in [\bar{x} - z^* \frac{\sigma}{\sqrt{n}}, \bar{x} + z^* \frac{\sigma}{\sqrt{n}}]) = 95\%$

Quiz on confidence

Interval estimate for μ: $P(\mu \in [\bar{x} - z^* \frac{\sigma}{\sqrt{n}}, \bar{x} + z^* \frac{\sigma}{\sqrt{n}}]) = 95\%$

Match the terms to the expressions:

confidence level	$\leftrightarrow$	95%
margin of error	$\leftrightarrow$	$\pm z^* \frac{\sigma}{\sqrt{n}}$
confidence interval	$\leftrightarrow$	$[\bar{x} - z^* \frac{\sigma}{\sqrt{n}}, \bar{x} + z^* \frac{\sigma}{\sqrt{n}}]$

Statistical inference again

In the past lecture confidence intervals were used to estimate a population parameter, e.g. a mean by using a standard deviation and a data sample.

Another common type of inference (tests of significance) aims to test some *claim* about a population parameter again by using a sample.

Example 10.1. We flip a coin and get one head followed by a row of tails. How many tails should we experience before claiming that the coin is biased?

Hypothesis: null vs alternative

Definition 10.2. A **hypothesis** is an educated guess. A scientific hypothesis should be *testable*. The **null hypothesis** H_0 is a claim that some assumed fact is true and nothing new is happening.

The H_0 can be thought of as a nullifiable hypothesis. We should be able to test H_0 and either *reject* (nullify) it in favor of the **alternative** hypothesis H_a or *fail to reject* it. But: we can not say that we *accept* H_0 just because we haven't found a counterexample yet.

Often an alternative hypothesis H_a claims a new effect and is stated before formulating H_0.

Examples of H_0

1) What is H_0 in experiment with flipping a coin?

2) What is H_0 in Galileo's model of heliocentrism?

Are the following statements valid H_0 hypotheses?

3) All unicorns are pink.

4) Earth is flat.

5) Goldfish make better pets than guinea pigs.

Examples of H_a

State H_0 and H_a in the following cases:

1) A lecturer is studying the effects of number of lectures on the average of all exam marks. Average of all exam marks for the module are 60%.

2) A lecturer thinks that if module lectures will happen twice a week (instead of 3 times), the average of all exam marks will be lower. Average of all marks for the module is 60%.

Hypotheses: 1-sided vs 2-sided

Both H_0 and H_a should be stated in terms of parameters of a whole population (the mean of all exam marks), not for a small sample outcome.
Let H_0 say that the mean of exam marks is $\mu = 60$.

Then H_a may say $\mu \neq 60$ and is called *2-sided* in this case, so H_a is the union of the two hypotheses: 1) $\mu < 60$ and 2) $\mu > 60$.
H_a may not be complementary (exactly opposite) to H_0, e.g. H_a may say that $\mu > 60$ and is called *1-sided*.

6 stages of Hypothesis testing

Step 1: null hypothesis. Define the null hypothesis H_0 (e.g. no differences between groups with and without the characteristic of interest) and the alternative hypothesis H_1.

Step 2(3): statistical assumptions. State statistical assumptions about the sample (for example, assumptions about independence or distributions).

Step 3(2): choosing a test. Choose the appropriate test based on the types of variables and assumptions and define the test statistic. In practice, we usually choose the test out of the list of known tests and check if the assumptions required to perform this test hold.

The p-value of a data sample

Assume that a null hypothesis H_0 is true.

A **test statistic** (a numerical measurement of a sample) estimates how far an actual measurement diverges from an expected value in H_0.

Definition 10.3. Assuming the null hypothesis, the p-**value** is the probability to obtain a result equal to or more extreme than what was observed,

i.e. $p\text{-value} = P(\text{this or more extreme result } | H_0)$.

Alternative hypothesis H_a can be one-sided or two sided, so p-values can also be one-sided or two-sided.

Examples of p-values

Assume that students have

- identically distributed knowledge and

- do the exam independently of each other.

From CLT, we choose normal distribution as our statistic.

Let the expected average mark be 60, so $H_0 = \{\mu = 60\}$.

However, the average mark turns out to be 69.

This observation casts some doubt on the null hypothesis H_0.

To quantify our doubt, hence reject (or not to reject) H_0, we can consider

the 1-sided p-value $P(X \geqslant 69)$ or

the 2-sided p-value $P(|X - 60| \geqslant 9)$.

p values and z statistic

Let $x_1, \dots, x_n$ be numerical values drawn from normal distributions $X_1, \dots, X_n \sim N(\mu, \sigma^2)$, where a deviation σ is known, but a mean μ is unknown.

To test the null hypothesis H_0 that $\mu = \mu_0$ (a given value), we find the *test statistic* $\bar{z} = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$.

From our sample we can get $\bar{z} \neq 0$. For $\bar{z} > 0$, the 1-sided p-value to test $H_a = \{\mu > \mu_0\}$ against H_0 is $P(Z > \bar{z})$, where $Z = \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$ and the average $\bar{X} = \dfrac{1}{n}\sum\limits_{i=1}^{n} X_i$. For $\bar{z} < 0$, the p-value is $P(Z < \bar{z}) = P(Z > |\bar{z}|)$.

The 2-sided p-value against H_0 is $P(|Z| \geqslant |\bar{z}|)$.

Reject or not to reject?

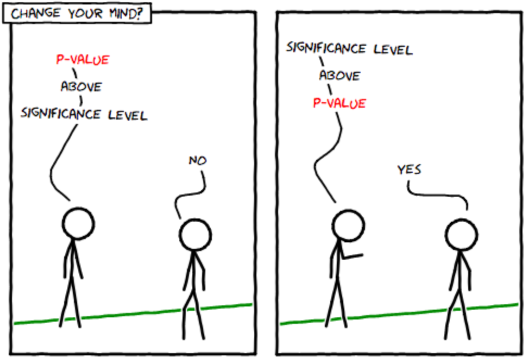
Assuming that the standard deviation of exam marks is $\sigma = 15$ and we get $n = 25$ sample marks with the sample mean $\bar{x} = 69$, shall we reject (or not) the hypothesis that the mean mark $\mu = 60$?

1) Compute the z statistic as follows:
$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{69 - 60}{15/\sqrt{25}} = \frac{9}{3} = 3.$$
2) Compute the 2-sided p-value: $P(|Z| \geqslant 3) = 0.3\%$ by the 68-95-99.7 rule.

So, what is the conclusion? Can we reject H_0?

Significance level α

Definition 10.4. If the p-value is (non-strictly) smaller than a specified **significance level** α, the data are called **statistically significant** at level α.



This image and more explanations can be found here.

If the data is statistically significant, we can **reject** the null hypothesis H_0.

Change in significance

In our example, the 2-sided p-value: $P(|Z| \geqslant 3) = 0.3\%$ by the 68-95-99.7 rule. The p-value of 0.3% for the null hypothesis $H_0 = \{\mu = 60\}$ means that the actual average of 69 is statistically significant at level 1%, or 0.5%, but not statistically significant at 0.1%.

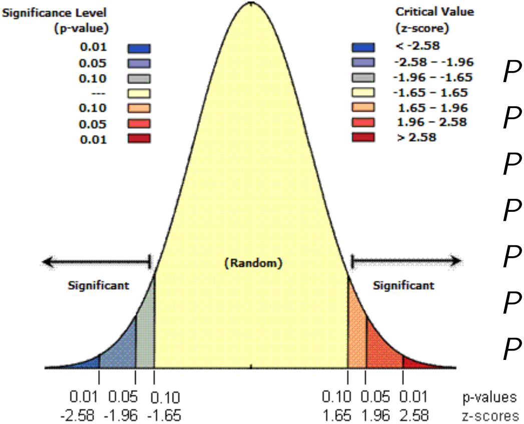
Moreover, the 1-sided p-value $P(Z \geqslant 3) = \frac{0.3\%}{2} = 0.15\%$. When is it statistically significant?

z statistic	1.645	1.96	2.326	2.576
2-sided p-value	10%	5%	2%	1%
1-sided p-value	5%	2.5%	1%	0.5%

What if $\bar{x} = 66$?

Revision: p-levels vs z-scores

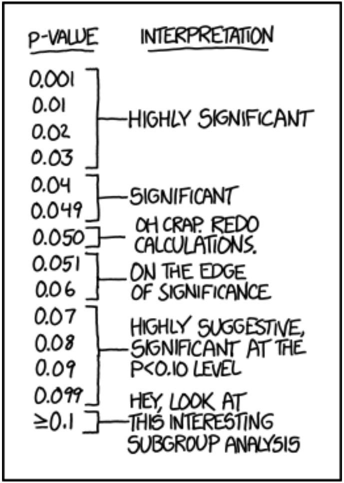
For $\bar{x} = 66$, the z-statistic is $z = \frac{66-60}{3} = 2$ and 2-sided p-value is $P(|Z| \geqslant 2) \approx 5\%$, so the 2-sided statistic is significant at level 5%, but not at smaller levels.



$$
\begin{aligned}
P(|Z| &> 1.65) = 10\%, \\
P(Z &> 1.65) = 5\%, \\
P(|Z| &> 1.96) = 5\%, \\
P(Z &> 1.96) = 2.5\%, \\
P(|Z| &> 2.32) = 2\%, \\
P(Z &> 2.32) = 1\%, \\
P(|Z| &> 2.58) = 1\%.
\end{aligned}
$$

Choosing a significance level

Conclusion: **A significance level α should be set before the study**, not after, to avoid any unjustified conclusions.



This image is courtesy of xkcd.

Traditionally α is 5% or 1%.

Otherwise there might be a strong urge to always have significant results.

Type I and type II errors

Definition 10.5. **Type I error (false positive)** is changing your mind when you shouldn't: it is the mistaken rejection of H_0 that is actually true.
Type II error (false negative) is NOT changing your mind when you should: the failure to reject a null hypothesis that is actually false.

H_0 will be rejected as soon as the significance level threshold is reached (even when H_0 is actually true), significance level is the probability of Type I Error (or false positive) of mistakenly rejecting the true statement. Lower significance level reduces this error, but increases the risk of failing to reject false H_0 (Type II Error or false negative).

6 stages of Hypothesis testing

Step 1: null hypothesis.

Step 2(3): choosing a test.

Step 3(2): checking the assumptions.

Step 4: setting a significance level, below which the H_0 will be rejected.

Step 5: calculating test statistic and p-value, the probability of getting a test statistic at least as extreme as the one we observed if H_0 was true $\left(P(\text{evidence}|H_0)\right)$.

Step 6: make a decision about the null hypothesis. Is the p-value less than the predefined significance level?

If yes, reject the H0.

If no, there is *insufficient evidence* to reject H_0; *never accept* the H_0.

Sample size

Another frequenlty overlooked value that should be pre-set before the analysis is the sample size n.
You're welcome to research this by changing the values of n in our previous examples.

Time to revise and ask questions

- The *p-value* is the probability (assuming the null hypothesis) to obtain a result equal to or more extreme than what was observed.
- If the p-value is (non-strictly) smaller than a specified significance level α, the data are called *statistically significant at level* α.

Problem 10.6. Let $n = 16$, $\sigma = 20$, $\bar{x} = 47.5$. Should we reject (or not reject) the null hypothesis H_0 that $\mu = 60$ at the significance level 1%?