UNIVERSITY OF TWENTE.

Chapter 4: Hypothesis Testing

Part 1: Introduction

Statistical Techniques for CS/BIT 202001033

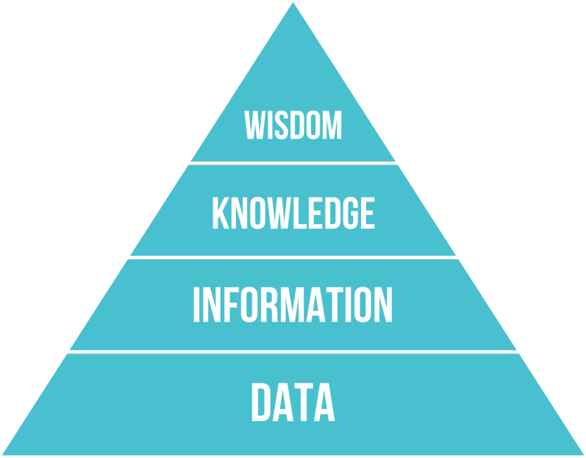
November 26, 2021

From data to knowledge

So far, we have covered:

Descriptive statistics	$\longrightarrow$	**analyze data**
Point estimation	$\longrightarrow$	**transform data to information**
Confidence intervals	$\longrightarrow$	**draw conclusions and quantify uncertainty**

From data to knowledge



Overview

Question

How do we gain knowledge from the world?

Through

- careful and systematic observation
- rigorous skepticism
- performing measurements and experiments
- formulating, testing and modifying hypotheses

These are the principles of the **scientific method**.

Steps in the scientific method

- Formulation of a question
- Formulation of a hypothesis
- Make a prediction (a consequence of the hypothesis)
- Test the prediction
- Analyze the results
- Make conclusions

In a similar way, we will define a **test procedure** for statistical tests.

We gain knowledge from the world through the scientific method.

We gain knowledge from data through statistics.

But we need to apply statistics **properly**!

Otherwise our results are invalid and we are just fooling ourselves!

Two situations

- Statistical tests to answer **research questions**
- Statistical tests to check/challenge **model assumptions**

Today's discussion only applies to research questions!

Overview

▶ **Question:** How to treat patients infected with Covid-19?

The drug hydroxychloroquine (HCQ) has been used to treat several diseases.

It is known that in some cases it may have antiviral effects against some viruses.

- **Hypothesis:** Hydroxychloroquine is effective in the treatment of Covid-19.

If a treatment is effective, patients will tend to recover faster.

We could look at the *number of days* that patients remain hospitalized, and compare two groups: those treated with HCQ and those treated with the standard care.

Warning

The distribution of number of days is highly skewed!

Alternative

We fix a time window of 14 days from hospital admission. On day 14, we test whether or not the patient is still infected using a SARS-CoV-2 PCR test.

- **Prediction:** Patients receiving HCQ plus standard care will have a higher rate of negative PCR tests 14 days after admission compared to patients receiving standard care only.

How do we test the prediction?

How much is enough evidence?

- **Test:** We can design the following experiment:

- We know that typically 50% of patients treated with the standard care test negative after 14 days.
- We randomly choose 100 new patients to treat with HCQ plus standard care.
- Patients closely monitored for a period of 14 days.

Results

The rate of negative PCR test by 14 days are as follows:

Standard care + HCQ: **54%**.

Standard care only: **50%**.

Question

Does this prove that hydroxychloroquine is a more effective treatment?

We need to account for the randomness (uncertainty) involved!

- Errors in the experimental design
- Other variables not included in the model
- **Can we generalize from a sample to the population?**

This is where statistics come in!

- **Status quo:** *"The existing state of affairs"*

Before the experiment, the medical community **does not** regard HCQ as an effective treatment against Covid-19.

Suppose *we* believe that hydroxychloroquine is an effective treatment.

Do *we* have to prove that HCQ is effective?

Or does *the medical community* need to prove that HCQ is ineffective?

- **Burden of proof:** Whoever challenges the status quo has the burden of proof.

Each observation $X_1, \ldots, X_{100}$ corresponds to a patient treated with HCQ plus standard care, and represents a Bernoulli trial:

$X_i = 1$ (success) if the patient has a negative Covid-19 test on day 14.

$X_i = 0$ (failure) if the patient has a positive Covid-19 test on day 14.

We assume the observations are independent and random

$$X = \sum_{i=1}^{100} X_i \sim \text{Binom}(100, p).$$

Probability model

The data we have recorded follows a binomial distribution,

$$X \sim \text{Binom}(100, p),$$

with **unknown** parameter p.

Null and alternative hypotheses

We now state two competing hypotheses:

Null hypothesis (H_0)

HCQ + standard care is not more effective than standard care alone. Thus,

$$p = 0.5.$$

This is the status quo.

Alternative hypothesis (H_1)

HCQ + standard care is more effective than standard care alone. Thus,

$$p > 0.5.$$

This is the the researcher's hypothesis, challenging the status quo.

Standing assumption

The random variable X is binomial: $X \sim \mathrm{Binom}(100, p)$.

Proof by contradiction

Suppose we want to prove a mathematical proposition P.

- Assume the opposite is true (assume P is false).
- If we arrive to a contradiction, our assumption must have been absurd.
- Conclusion: We have **proved** that P is true.

Strategy

- Assume H_0 is true. Then $X \sim \mathrm{Binom}(100, 0.5)$.

Let Z be the Z-score of the random variable X under H_0 (using $p = 0.5$).

- If Z is surprisingly large, then
  - either, our assumption $p = 0.5$ is absurd,
  - or, we are observing an event with a very low probability of occurring.
- Conclusion: The data **suggests** that p is larger than 0.5.

Decision rule

Under H_0, $X \sim \text{Binom}(100, 0.5)$. Thus, $n = 100, p = 0.5$ and

$$\text{E}(X) = np = 50, \quad \text{Var}(X) = np(1-p) = 25$$

Because $n = 100$ is quite large, we invoke the Central Limit Theorem:

$$Z = \frac{X - 50}{\sqrt{25}} \approx N(0,1) \quad \text{(under } H_0\text{)}.$$

We will use this random variable as the **test statistic**.

Decision rule and significance level

Let c be such that $\mathbf{P}(Z \geq c) = 0.05$ (**significance level 5%**).

If z, the observed value of Z, satisfies

$$z \geq c,$$

then we **reject** the hypothesis that $p = 0.5$.

We would then conclude that the data suggests $p > 0.5$.

In our experiment, 54 of the 100 HCQ-treated patients recovered within 14 days. This means that $x = 54$.

Analysis

The observed statistic is:

$$z = \frac{54 - 50}{\sqrt{25}} = \frac{4}{5} = 0.8$$

From a normal table we get:

$$c = 1.645 \quad \text{(meaning } \mathbf{P}(Z \geq c) = 0.05\text{)}.$$

We note that $0.8 < 1.645 \longrightarrow$ The value $z = 0.8$ is not surprising!

Alternative analysis

Under H_0 we have $Z \approx N(0, 1)$.
Thus, if H_0 is true, we have:

$$\mathbf{P}(Z \geq 0.8) = 0.21.$$

The number 0.21 is called the **p-value** $\longrightarrow$ Large p-value is not surprising!

Conclusion

There is not enough evidence to reject the null hypothesis. We say:

- We **fail to reject** the null hypothesis, or
- we **retain** the null hypothesis.

In words

From the data collected, there is no statistically significant evidence (at the 5% level) that HCQ improves the probability of recovery within 14 days, compared to the standard care.

The higher observed recovery rate (54% vs 50%) could be reasonably explained by random chance.

Warning

Note that we don't claim that "we have proved H_0".

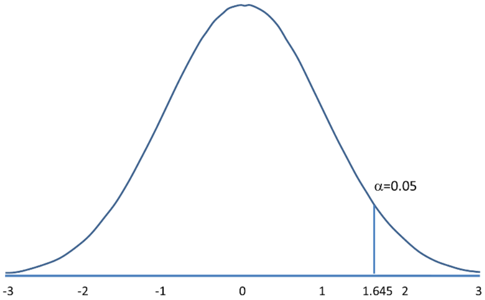
Testing procedure in 8 steps

1. Give a probability model and statistical assumptions.
2. State the null hypothesis and alternative hypothesis using parameters in the model.
3. Give the proper test statistic.
4. State the distribution of the test statistic under H_0.
5. Compute the observed value of the test statistic.
6. State the rejection rule.
7. State the statistical conclusion: reject H_0, or fail to reject H_0.
8. Draw conclusions and explain them in words.

Overview

We reject the null hypothesis whenever we observe a value of the test statistic with a low probability of occurring.

In our example we took $\alpha = 5\%$ as the **level of significance**.



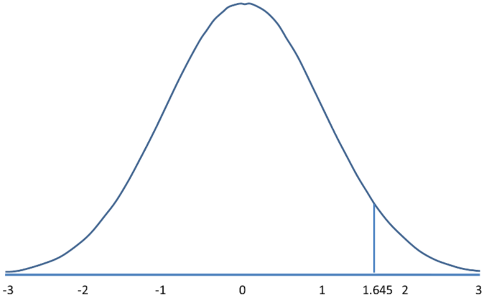
Thus the **critical value** is $c = 1.645$. We also write $z_\alpha = 1.645$.

The p-value is the probability of obtaining a value of the test statistic equal or more extreme than the one observed assuming H_0 is true.

$$\mathbf{P}(Z \geq z \mid H_0 \text{ true}) = \mathbf{P}(N(0,1) \geq z) = 1 - \Phi(z)$$

Equivalent rules:

- Reject H_0 if $z \geq z_\alpha$,
- Reject H_0 if $\mathrm{p} \leq \alpha$.

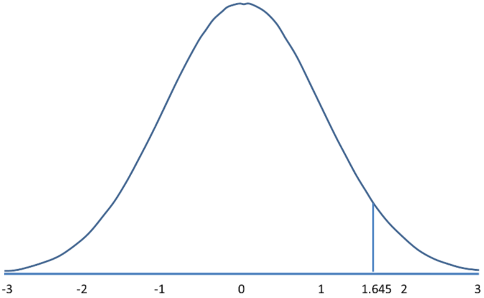


Assume we want to test:

$$H_0: \ p = 0.5 \quad \text{vs} \quad H_1: \ p \neq 0.5$$

That is,

$$H_0: \ p = 0.5 \quad \text{vs} \quad H_1: \ (p > 0.5 \text{ or } p < 0.5)$$



Instead of z_α, we use the critical values $\pm z_{\alpha/2} = \pm 1.960$.

Overview

As with any decision-making process, in a statistical test it is possible to make a wrong decision.

There are two types of errors:

Definition

A **Type I error** is committed if we reject the null hypothesis when it is in fact true.

Definition

A **Type II error** is committed if we retain the null hypothesis when it is in fact false.

- The probability of committing a Type I error is denoted by α.
- The probability of committing a Type II error is denoted by β.

Warning

The reader denotes probability of Type II error as $1 - \beta$ (instead of β).

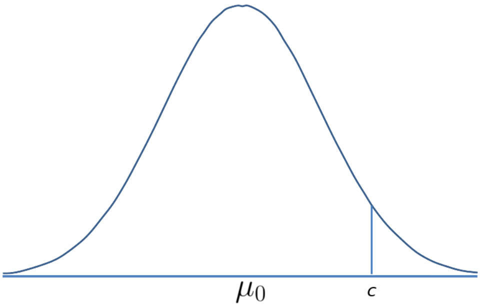
Error types

Decision	In reality	
	$\mathbf{H_0}$ **is true**	$\mathbf{H_1}$ **is true**
Reject H_0	Type I error	Correct decision
Accept H_0	Correct decision	Type II error

Type I error

A Type I error is only committed when H_0 is true.

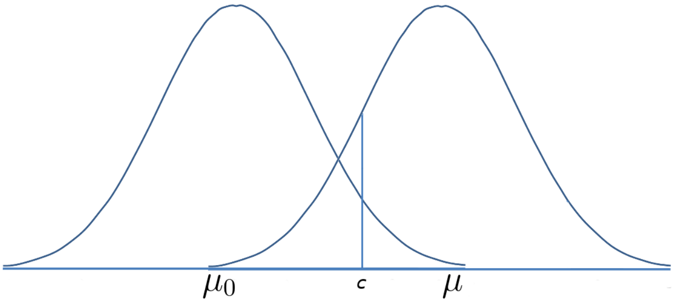
So assume H_0 is true.



We control α, the probability of committing a Type I error by specifying the significance level.

A Type II error is only committed when H_0 is false.

So assume H_0 is false, and let $\mu \neq \mu_0$ be the true mean.



We can't control β as easily, since we don't know the value of μ.

- The further apart μ is from μ_0, the better the test can correctly reject H_0.
- A bigger sample size reduces variance, increasing the power of the test.

Power of a test

The power of a test is $1 - \beta$. This is a function $power(\mu)$ that depends on μ.