Introduction to hypothesis testing

Joaquín Rapela

Gatsby Computational Neuroscience Unit
University College London

January 14, 2024

Contents

Contents

Background

- Last Spring 2023 I helped in the discussion sessions of this course.
- Suggested to Klara Olofsdotter (SWC PhD program coordinator) and Sonja Hofer (SWC PhD program faculty coordinator) to ask SWC PhD students to take this course. They liked the idea.
- I volunteered to lead discussions and do grading with Gatsby Unit PhD students and postdoctoral scholars.

A few motivations to run this course

1. Gain more teaching experience.
2. Provide SWC PhD students with essential neural data-analysis tools.
3. Contribute to better interactions between the SWC and the Gatsby Unit.

Course structure

Week 01	Jan 11	The t-test and randomisation tests	Joaquin Rapela	tutorial
Week 02	Jan 18	Power spectra	Joaquin Rapela Yousef Mohammadi Joe Zimminski	tutorial
Week 03	Jan 25	Spectrograms and coherence	Joaquin Rapela Yousef Mohammadi Joe Zimminski	tutorial
Week 04	Feb 01	Circular statistics	Joaquin Rapela	tutorial
Week 05	Feb 08	Singular value decomposition	Will Dorrell	tutorial
Week 06	Feb 15 Feb 16	Linear regression	Lior Fox	lecture tutorial
Week 07	Feb 22 Feb 23	Linear dynamical systems	Aniruddh Galgali Joaquin Rapela	lecture tutorial
Week 08	Feb 29	no class (CoSyNe)		
Week 09	Mar 07 Mar 08	Artificial neural networks	Erin Grant	lecture tutorial
Week 10	Mar 14 Mar 15	Experimental control with Bonsai	Goncalo Lopes Joaquin Rapela	lecture tutorial
Week 11	Mar 21 Mar 22	Reinforcement learning		lecture tutorial
Week 12 Week 15	Mar 28 Apr 25	Project development		
Week 16	May 02	Project presentations		

Teaching assistants: Kira Dusterwald, Sihao (Daniel) Liu

Graded worksheets

Every Thursday we will assign you a worksheet that is due on the second Monday after the assignment.

Contents

Working examples

Example 1

We know that the average running speed of control mice is 1 cm/sec. The sample average running speed of a cohort (n=100) of transgenic mice is $\bar{x} = 2.7$ cm/sec and the sample standard deviation is $s = 10$ cm/sec. Is the average running speed of mice in the transgenic cohort larger than that of control mice? Test with a confidence level $\alpha = 0.01$.

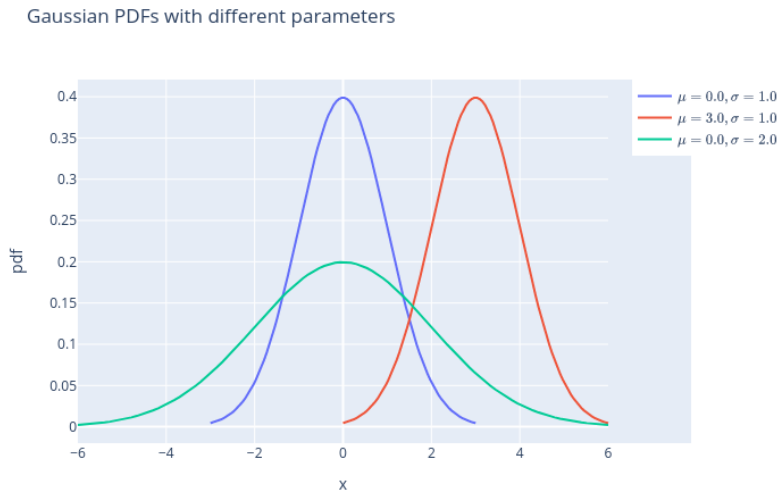
Working examples

Example 2

We want to study the effect of a new drug on visual electrophysiology in humans. We know that the mean peak evoked response potential (ERP) over V1 during the first 200 ms after stimuli presentation is 2 mV. The sample mean peak ERP for a group of 50 medicated subjects is $\bar{x} = 1.3\,\text{mV}$ and the sample standard deviation is $s = 2.6$ mV. Does taking the new drug change the mean evoked ERP over V1? Provide your test p-value.

Statistical remarks

1. Random data (e.g., observed data) is characterised using probability distributions. For example a:
   - Normal distribution with parameters mean μ and variance σ^2, $\mathcal{N}(\mu, \sigma^2)$,



   - Exponential distribution with rate parameter λ, $\mathcal{E}(\lambda)$,
   - Poisson distribution with expected rate parameter λ, $\mathcal{P}(\lambda)$,
   - Binomial distribution with number of observation parameter n and with a success probability parameter p, $\mathcal{B}(n, p)$.

Statistical remarks

2. One branch of statistics, **estimation theory**, provides tools to estimate parameters of distributions from observations.
3. Another branch of statistics, **hypothesis testing**, provides tools to make statistically-informed decisions about values of parameters of distributions.

Statistical remarks

4. To estimate parameters, or to make decisions about them, we use observations, $x_1, \ldots, x_N$, that are **independent and identically distributed**.

Statistical remarks

5. To estimate parameters, or to make decisions about them, we use observations, $x_1, \ldots, x_N$, that are **independent and identically distributed**.

Example 1

An observation is the average speed of a transgenic mouse during an experimental session. We assume that the average speeds of all mice are samples from a common probability density function (identically distributed) and that average speeds are independent across mice (independent).

Statistical remarks

6. To estimate parameters, or to make decisions about them, we use observations, $x_1, \ldots, x_N$, that are **independent and identically distributed**.

Example 1

An observation is the average speed of a transgenic mouse during an experimental session. We assume that the average speeds of all mice are samples from a common probability density function (identically distributed) and that average speeds are independent across mice (independent).

Example 2

An observation is the peak ERP of a medicated cohort subject. We assume that these ERPs are samples from the same probability density function (identically distributed) and that these ERPs are independent across subjects (independent).

Statistical remarks

7. A goal of statistics is to **infer properties of the population** (e.g., the effect of the genetic manipulation on the running speed of mice) from **properties of the sample** (e.g., the effect of the manipulation on the running speed of the 100 sampled mice).

Statistical remarks

Theorem (Central Limit Theorem)

Let $X_1, \ldots, X_n$ be independent and identically distributed random variables with mean μ and finite variance σ^2. Let n be large. Then the sample mean

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{N} X_i \tag{1}$$

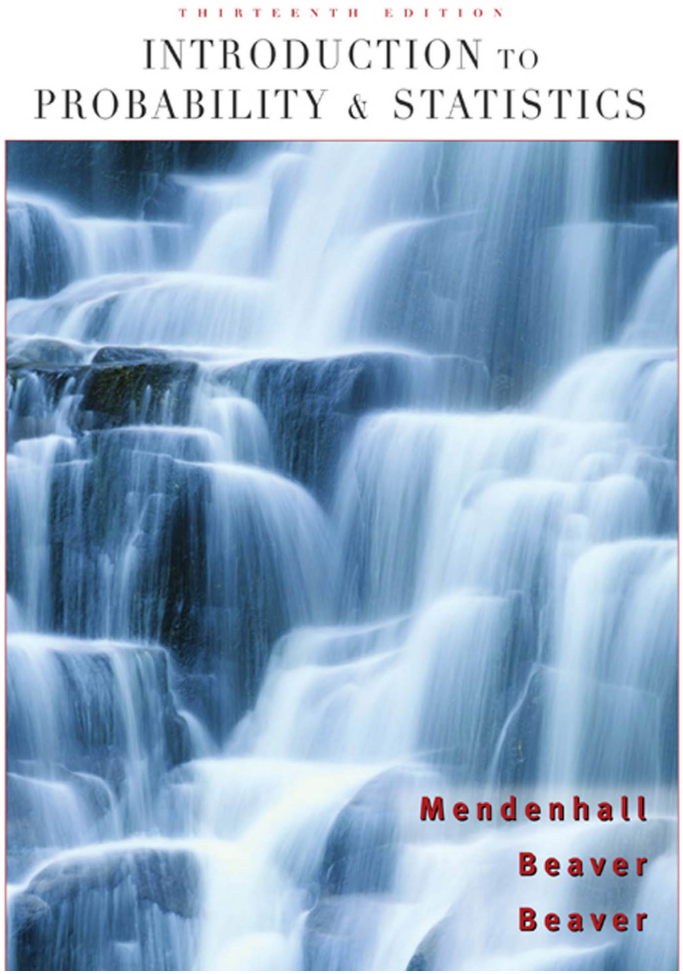
is distributed as $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$.

Note: if σ^2 is unknown, we can estimate σ^2 with the sample variance $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$. Then, for large n, $\bar{X}$ is approximately distributed as $\mathcal{N}(\mu, \frac{s^2}{n})$.

Contents

Main source

Chapter 9 "Large-sample test of hypothesis" and chapter 10 "Inference from small samples" from



Null and alternative hypothesis

- In hypothesis testing we work with a **null hypothesis**, $\mathcal{H}_0$, and an **alternative hypothesis**, $\mathcal{H}_a$, collect a sample of data $x_1, \ldots, x_N$, and test if this data provides sufficient statistical evidence in favour of the alternative hypothesis. If this happens we reject the null hypothesis.
- However, if the collected data does not provide sufficient statistical evidence in favour of the alternative hypothesis, we do not accept the null hypothesis, but we say that we failed to reject it. **Hypothesis tests do not prove null hypothesis, they only provide statistical evidence to reject it, or fail to reject it.**

Null and alternative hypothesis

- In hypothesis testing we work with a **null hypothesis**, $\mathcal{H}_0$, and an **alternative hypothesis**, $\mathcal{H}_a$, collect a sample of data $x_1, \ldots, x_N$, and test if this data provides sufficient statistical evidence in favour of the alternative hypothesis. If this happens we reject the null hypothesis.
- However, if the collected data does not provide sufficient statistical evidence in favour of the alternative hypothesis, we do not accept the null hypothesis, but we say that we failed to reject it. **Hypothesis tests do not prove null hypothesis, they only provide statistical evidence to reject it, or fail to reject it.**

The hypothesis that we aim to prove should be the alternative one.

Null and alternative hypothesis

Example 1

Is the average running speed of the transgenic cohort larger than that of the control cohort (i.e., 2 cm/sec)?

$\mathcal{H}_0$: the average running speed of the transgenic cohort is 2 cm/sec.

$\mathcal{H}_a$: the average running speed of the transgenic cohort is lager than 2 cm/sec.

Null and alternative hypothesis

Example 2

Is the mean peak visual ERP in the first 200 ms post stimuli different in medicated than in control subjects (i.e., 2 mV)?

$\mathcal{H}_0$: the mean peak visual ERP in medicated subjects is 2 mV.

$\mathcal{H}_a$: the mean peak visual ERP in medicated subjects is different from 2 mV.

One- and two-tailed tests of hypothesis

One-tailed test of hypothesis directionality is suggested by the alternative hypothesis.

Example 1

It is a one-tailed hypothesis test because the alternative hypothesis requires that the mean speed of the transgenic mice be larger (directionality) than that of the control mice.

One- and two-tailed tests of hypothesis

One-tailed test of hypothesis directionality is suggested by the alternative hypothesis.

Example 1

It is a one-tailed hypothesis test because the alternative hypothesis requires that the mean speed of the transgenic mice be larger (directionality) than that of the control mice.

Two-tailed test of hypothesis directionality is not suggested by the alternative hypothesis.

Example 2

It is a two-tailed hypothesis test because the alternative hypothesis requires that the visual ERP of the medicated subjects be different (no directionality) than that of the control subjects.

Test statistic and its sampling distribution

- To perform a hypothesis test we propose a **test statistic**, a function of the sample data, like the sample mean in Eq. 1.
- Because the sample data is random, the test statistic is also random. To perform hypothesis tests we need to know the distribution of the test statistic, which is called the **sampling distribution**.

Test statistic and its sampling distribution for the working examples

Because both examples are tests for the population mean, μ, because the sample mean, $\bar{x}$, is a good estimator of the population mean, and because both examples use a large number of samples, we will use the standardised sample mean as our test statistic:

$$Z = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

From the central limit theorem, we know the sampling distribution of this test statistic under the null hypothesis:

$$Z \sim \mathcal{N}(0, 1)$$

Rejection region

The **reject region** is a region of low probability under the null hypothesis, which is consistent with alternative hypothesis. The **non-reject region**, is a region of large probability under the null hypothesis, that is inconsistent with the alternative hypothesis.

Rejection region

Example 1

We want to reject the null hypothesis that the average running speed of the transgenic mice equals 2 cm/sec if their sample mean speed is much larger than 2 cm/sec. That is, we want to reject the null hypothesis if z is large and positive. How large is large?

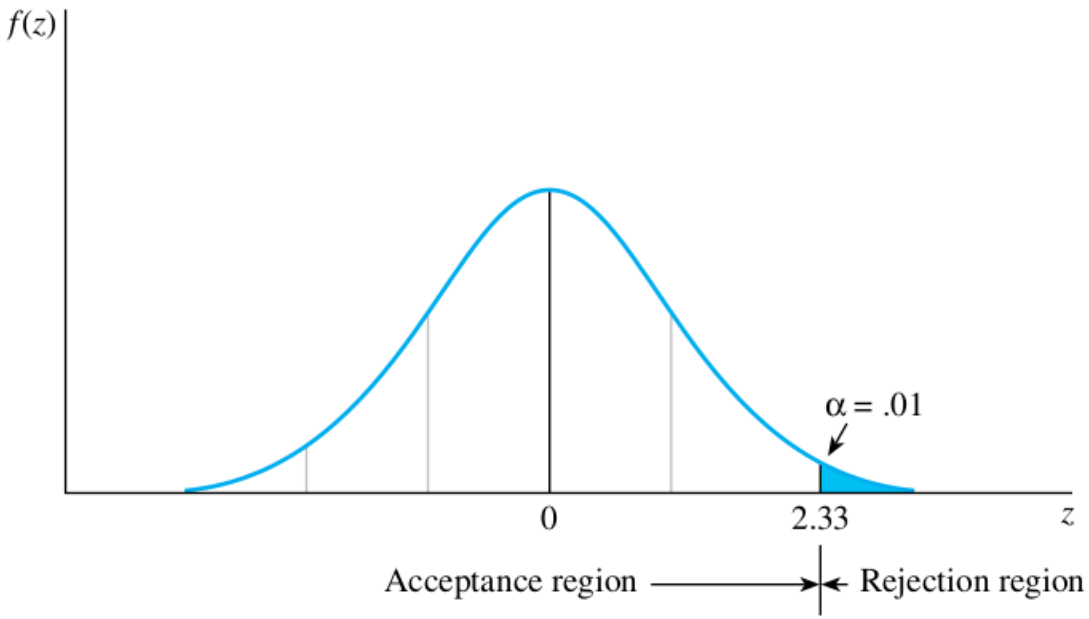
To answer this question we define the **Type I error** of a test, as the error of rejecting the null hypothesis when it is valid. We also define the **significance level of the hypothesis test**, α, as the probability of Type I error admitted by the test.

When designing a test we first decide on its significance level α. We then reject the null hypothesis if z is larger than the value z_α that leaves α probability to its right, We call this value the **critical value** of the test (see figure on next slide).

Rejection region

Example 1

FIGURE 9.3
The rejection region for a right-tailed test with $\alpha = .01$



Mendenhall et al., 2009

Rejection region

Example 2

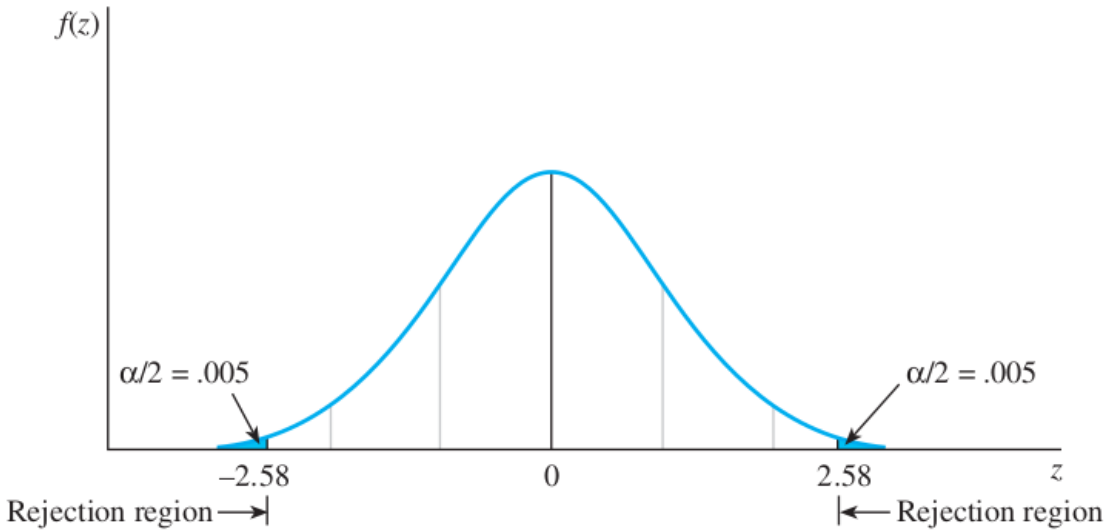
We want to reject the null hypothesis that the drug has not effect on the average visually evoked ERP if the sample mean visually evoked ERP of medicated subjects is much smaller or much larger than the mean visually evoked ERP of control subjects (2 mV). How small is small and how large is large?

We will reject the null hypothesis if the standardised mean, z, is larger than the value $z_{\alpha/2}$ that leaves $\alpha/2$ probability to its right or smaller than the value $-z_{\alpha/2}$ that leaves $\alpha/2$ probability to its left (see figure on next slide).

Rejection region

Example 1

FIGURE 9.4
The rejection region for a two-tailed test with $\alpha = .01$



Mendenhall et al., 2009

Large-sample hypothesis test for the mean

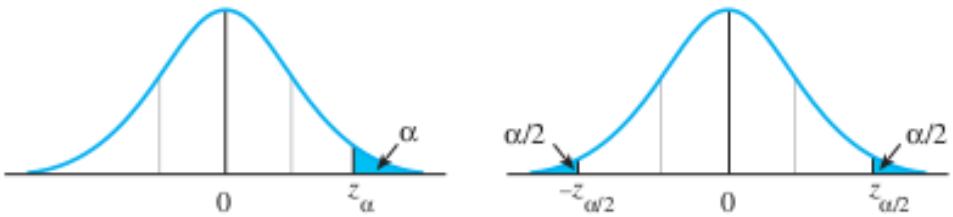
LARGE-SAMPLE STATISTICAL TEST FOR μ

1. Null hypothesis: $H_0 : \mu = \mu_0$
2. Alternative hypothesis:

One-Tailed Test	Two-Tailed Test
$H_a : \mu > \mu_0$ (or, $H_a : \mu < \mu_0$)	$H_a : \mu \neq \mu_0$

3. Test statistic: $z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$ estimated as $z = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$
4. Rejection region: Reject H_0 when

One-Tailed Test	Two-Tailed Test
$z > z_\alpha$ (or $z < -z_\alpha$ when the alternative hypothesis is $H_a : \mu < \mu_0$)	$z > z_{\alpha/2}$ or $z < -z_{\alpha/2}$



Mendenhall et al., 2009

Complete hypothesis test for example 1

Example 1

We know that the average running speed of control mice is 1 cm/sec. The sample average running speed of a cohort (n=100) of transgenic mice is $\bar{x} = 2.7$ cm/sec and the sample standard deviation is $s = 10$ cm/sec. Is the average running speed of mice in the transgenic cohort larger than that of control mice? Test with a confidence level $\alpha = 0.01$.
Relevant quantities: $\mu_0 = 1$, $n = 100$, $\bar{x} = 2.7$, $s = 10$, $\alpha = 0.01$.

1. identify the null hypothesis H_0: the average running speed of the transgenic cohort is 2 cm/sec.
2. identify the alternative hypothesis H_a: the average running speed of the transgenic cohort is larger than 2 cm/sec.
3. select a test statistic: standardised sample mean Z.

Complete hypothesis test for example 1

Example 1

4. set the rejection region: right-tailed hypothesis test, with critical value $z_{0.01} = 2.3263$.
5. calculate the observed value of the test statistic.

$$z_{\text{obs}} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{2.7 - 1}{10/\sqrt{100}} = 1.7$$

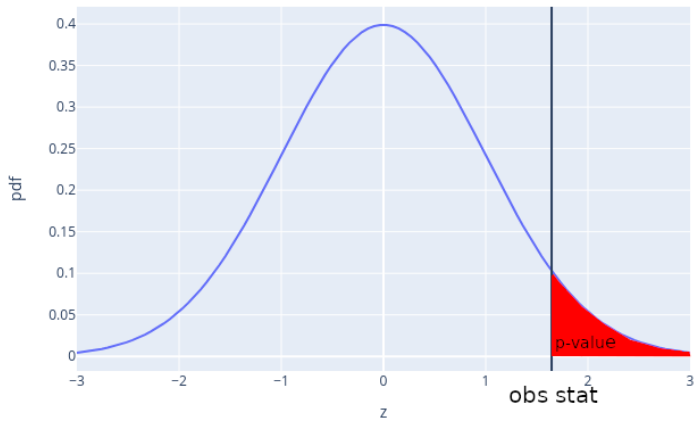
6. draw your conclusion: $z_{\text{obs}} = 1.7 < 2.3263 = z_{0.01}$. Thus, there is not enough statistical evidence to reject the null hypothesis with a confidence level $\alpha = 0.01$.

Would you reject the null hypothesis with confidence level $\alpha = 0.05$?

p-value

Definition (p-value)

The p-value (or observed significance level of a statistical test) is the probability that the test statistic is as extreme or more extreme than the observed test statistic value, when the null hypothesis is true.



(a) left-sided test

(b) two-sided test

Figure 1: p-values are the area of the red-coloured regions

p-value

Notes:

- small p-values indicate that the probability of obtaining a test statistic as extreme or more extreme as the observed test statistic value is small, showing that our data supports the alternative hypothesis.
- large p-values show that our data does not support the alternative hypothesis.
- if the p-value is smaller than a pre-assigned confidence level α, then the null hypothesis can be rejected and you can say that the result is statistically significant at confidence level α.

Complete hypothesis test for example 2

Example 2

Worksheet problem 1.

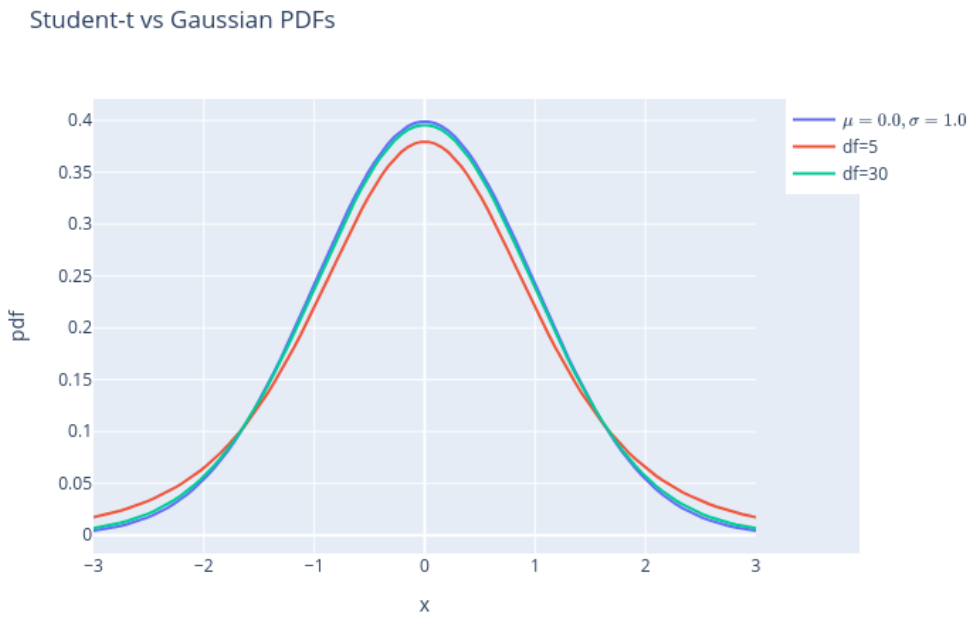
Taxonomy of hypothesis tests for the the mean

- The previous tests worked well because the sample statistic for the sample mean was approximately normal. This happens when
  - samples, of any size, are Gaussian, with known variance.
  - the number of samples is sufficiently large (e.g., $n > 30$) and samples are independent and identically distributed from any distribution with mean μ and finite variance σ^2, known or estimated by the sample variance s^2 (central limit theorem).
- What if samples are Gaussian, with unknown variance and we can only collect a small number n of samples?
  Then the sample distribution of the mean follows a student-t distribution with n-1 degrees of freedom.
- What if samples are non-Gaussian and we can only collect a small number n of samples?
  Use a resampling method (e.g., bootstrap of permutation test).

Student-t distribution

The Student-t distribution:

- has a similar shape as the standard Normal distribution,
- has heavier tails than the standard Normal distribution,
- approaches the standard normal distribution as the number of degrees of freedom increase.



Small-sample hypothesis test for the mean (t-test)

SMALL-SAMPLE HYPOTHESIS TEST FOR μ

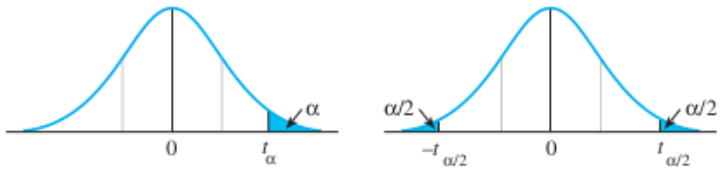
1. Null hypothesis: $H_0 : \mu = \mu_0$
2. Alternative hypothesis:

One-Tailed Test	Two-Tailed Test
$H_a : \mu > \mu_0$ (or, $H_a : \mu < \mu_0$)	$H_a : \mu \neq \mu_0$

3. Test statistic: $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$
4. Rejection region: Reject H_0 when

One-Tailed Test	Two-Tailed Test
$t > t_\alpha$ (or $t < -t_\alpha$ when the alternative hypothesis is $H_a : \mu < \mu_0$)	$t > t_{\alpha/2}$ or $t < -t_{\alpha/2}$

or when p-value $< \alpha$



The critical values of t, t_α, and $t_{\alpha/2}$ are based on $(n - 1)$ degrees of freedom. These tabulated values can be found using Table 4 of Appendix I or the **Student's *t* Probabilities** applet.

Assumption: The sample is randomly selected from a normally distributed population.

Mendenhall et al., 2009

Two types of errors

Definition (Type I error)

The error of rejecting the null hypothesis when it is true is called type I error. The probability of type I error is denoted by the symbol α (Figure 2).

Definition (Type II error)

The error of not rejecting the null hypothesis when the alternative one is true is called type II error. The probability of type II error is denoted by the symbol β (Figure 2).

Definition (Power of a statistical test)

The power of a statistical test is the probability of rejecting the null hypothesis when the alternative one is true and it is equal to $1 - \beta$.

Calculation of β

Example 1

Calculate β for the alternative hypothesis $\mathcal{H}_a : \mu_a = 5.0$ (using, as before, $\mathcal{H}_0 : \mu_0 = 1.0$, $\alpha = 0.01$, $\bar{x} = 2.7$, $s = 10$, and $n = 100$). See Figure 2.
We will first express the rejection region in terms of $\bar{X}$. Before we calculated the rejection region in terms of Z as $Z > z_\alpha$ which is

$$z_\alpha < Z = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

Then the rejection region in therms of $\bar{X}$ is

$$\bar{X} > \mu_0 + z_\alpha \, s/\sqrt{n}$$

Now

$$\beta = P(\text{not reject } \mathcal{H}_0 | \mathcal{H}_a \text{ is true}) = P(\bar{X} < \mu_0 + z_\alpha \, s/\sqrt{n} | \mu = \mu_a) = P(\frac{\bar{X} - \mu_a}{s/\sqrt{n}} < \frac{\mu_0 - \mu_a}{s/\sqrt{n}} + z_\alpha | \mu = \mu_a)$$

$$= P(Z < \frac{\mu_0 - \mu_a}{s/\sqrt{n}} + z_\alpha) = P(Z < \frac{1.0 - 2.0}{10/\sqrt{100}} + 2.326) = P(Z < -1.6736) = 0.047$$

The worksheet contains an optional exercise illustrating how to find the minimum sample size n to achieve a given probability of type I error α and probability of type II error β for a given effect size.

Calculation of β

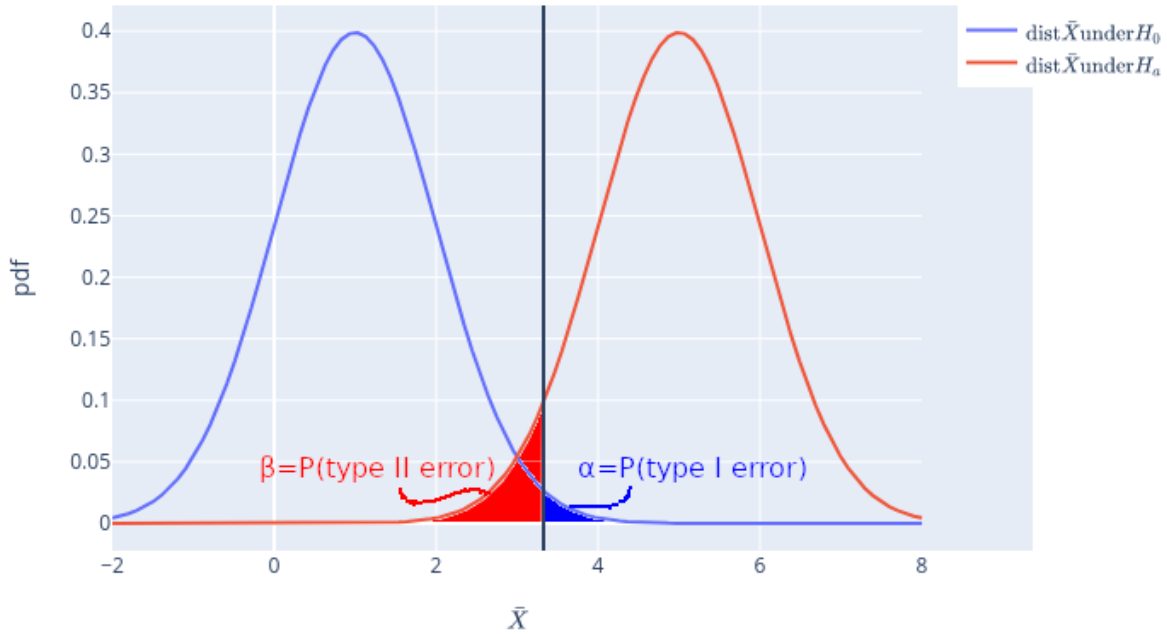


Figure 2: Probabilities of type I and type II errors (α and β, respectively) for the example in the previous page.

Lecture notes and worksheet

They can both be found in the class repository at
https://github.com/joacorapela/neuroinformatics24
The lecture notes appear here and the worksheets appear here.

Summary