INTRODUCTION TO DATA SCIENCE

Introduction to statistics and statistical hypothesis testing

Maciej Świtała, PhD

Autumn 2025

UNIVERSITY OF WARSAW | FACULTY OF ECONOMIC SCIENCES

1. Introduction to statistics
   - Basic concepts
   - Data sources, data collection methods
   - Data cleaning and validation
   - Population vs. sample, sampling methods
   - Random variables and their distributions
   - The most common statistical distributions

2. Statistical hypothesis testing
   - Type I and type II errors
   - The hypothesis testing procedure
   - Example: one-sample t-test
   - Example: two-sample t-test
   - Example: two-sample t-test (left- and right-tailed)
   - What is the p-value?

Introduction to statistics

What is statistics?

Statistics, in its intuitive sense, is a **set of methods** that allow us to draw conclusions from data - that is, to transform **raw data** into **knowledge** and ultimately use that knowledge to make **decisions**.

Stages of a typical analytical project

1. Identify the research or business problem.
2. Formulate the project objective.
3. Collect data.
4. Clean and validate data.
5. Perform exploratory data analysis.
6. Model building and validation.
7. Inference.

Research vs. business problem

Research problem:

- an issue that motivates conducting a study,
- should concern phenomena of significant importance - whose better understanding supports socio-economic development,
- facilitates understanding of social issues and supports decision-making and conflict resolution.

Business problem:

- a situation, challenge, or difficulty that limits the efficiency, profitability, or growth of an organization and requires action to solve it,
- has a practical dimension - it is not about discovering new knowledge but about improving organizational performance, financial results, service quality, customer satisfaction, operational efficiency, or competitiveness.

Project objective; research questions vs. hypotheses

- The **project objective** specifies how the analyst intends to address the research or business problem - it identifies the purpose of the analysis and what should result from it.
- Research questions and hypotheses narrow the scope of inquiry:
  - a **research question** is a clear, focused, specific, and usually open-ended question that the researcher aims to answer to gain insight or understanding of a phenomenon.
  - a **research hypothesis** is a detailed, testable statement predicting the outcome of a study, based on existing theory or prior research.
- Typically, research questions are formulated in qualitative studies (based on non-numeric data such as surveys, interviews, focus groups, observations), while research hypotheses are used in quantitative studies (based on numeric data).

Data sources

Primary data (collected directly):

- surveys and questionnaires,
- individual interviews,
- focus groups,
- direct or participatory observation,
- laboratory or field experiments,
- psychometric tests and measurements.

Secondary data (existing sources):

- corporate databases (CRM, ERP, sales logs),
- public/government data (statistics, demographics),
- scientific and industry publications,
- social media and web data,
- data from partners or suppliers,
- historical or archival data.

Data collection methods

Quantitative methods:

- standardized surveys,
- controlled experiments,
- analysis of transactions and system logs,
- KPI monitoring,
- technical tests and measurements,
- GPS data analysis.

Qualitative methods:

- in-depth interviews,
- participant observation,
- case studies,
- content analysis (texts, media),
- ethnographic and field research.

Channels and technologies for data collection

Offline/traditional:

- paper surveys,
- face-to-face interviews,
- field observations.

Online/digital:

- online surveys, mobile forms,
- system logs, app data,
- social listening, media crawling,
- external APIs.

Automatic sensors:

- IoT, RFID, GPS,
- image recognition cameras,
- machine and production monitoring.

Types of data structures

1. **Cross-sectional data** - many entities observed once, e.g.
   - household income survey in a given year,
   - customer satisfaction survey after a purchase.
2. **Time series** - one entity observed multiple times over intervals, e.g.
   - monthly revenues of a company from 2015–2025,
   - daily stock price of a listed company.
3. **Panel data** - many entities observed repeatedly over time, e.g.
   - annual sales data for 50 firms from 2010–2020,
   - exam results of students across schools for five years.

Data cleaning and validation

Data cleaning - the process of detecting and correcting (or removing) inaccurate records from a dataset, including:

- detection and handling of missing data,
- identifying and treating outliers,
- removing duplicates,
- correcting data entry errors,
- ensuring consistent data types and units.

Data validation - verifying that the data meets quality and accuracy standards:

- logical consistency (e.g., age cannot be negative),
- range and domain checks,
- referential integrity between tables,
- comparing with external or reference data.

Basic concepts

A **population** is the entire set of individuals, objects, or events possessing some common observable feature, about which we want to draw conclusions.

A **sample** is a subset of the population selected for observation or measurement, usually to make inferences about the entire population.

Sampling methods

Random sampling:

- simple random sampling - each unit has an equal chance of selection,
- systematic sampling - selecting every k-th unit from an ordered list,
- stratified sampling - dividing the population into subgroups (strata) and sampling within each,
- cluster sampling - selecting entire groups or clusters.

Non-random sampling:

- convenience sampling,
- judgmental or expert-based sampling,
- snowball sampling (for hidden populations).

Random variables

A **random variable** assigns a real number to each outcome of a random experiment; it can take discrete or continuous values.

Types of random variables:

- **discrete random variable:** countable number of outcomes, example: number of defective items in a batch.
- **continuous random variable:** takes values from an interval of real numbers, example: body temperature, height, income.

Probability distributions

Probability distribution describes how probabilities are assigned to possible values of a random variable.

We describe distiburions with:

1. **Probability Mass Function (PMF)** (for discrete distributions):

$$\sum_i P(X = x_i) = 1, \quad P(X = x_i) \geq 0$$

2. **Probability Density Function (PDF)** (for continuous distributions):

$$\int_{-\infty}^{\infty} f(x)\, dx = 1, \quad f(x) \geq 0$$

Cumulative Distribution Function (CDF)

$$F(x) = P(X \leq x)$$

The most important properties:

- applicable for both discrete and continuous distributions,
- for discrete variables it is a step function,
- for continuous variables: smooth, increasing function,
- $0 \leq F(x) \leq 1$,
- $\lim_{x \to -\infty} F(x) = 0$, $\lim_{x \to \infty} F(x) = 1$.

Parameters of a random variable distribution

- Distributions have parameters that determine their final mathematical form.
- Several basic types of parameters are distinguished, which define the distribution's shape, location, and scale:
  1. **location parameters** – determine the center of the distribution, e.g. the mean, or the shift of the normal distribution, i.e. μ in $N(\mu, \sigma^2)$.
  2. **scale parameters** – describe the spread of values around the center, e.g. the variance of the normal distribution, i.e. σ^2 in $N(\mu, \sigma^2)$.
  3. **shape parameters** – define the general shape of the distribution, its asymmetry, tail heaviness, or peakedness, e.g. p in the Bernoulli or Binomial distribution, or the number of degrees of freedom in the t-Student, Chi-square, and F-Snedecor distributions.
  4. **mixed/additional parameters**, e.g. λ in the Poisson or Exponential distribution, describing the rate or intensity of events.

The most common statistical distributions

Discrete distributions:

- Bernoulli,
- Binomial,
- Poisson.

Continuous distributions:

- Uniform,
- Normal (Gaussian),
- Exponential,
- Chi-square,
- Student's t,
- F-Snedecor.

Notation of the most common statistical distributions

Discrete distributions:

- Bernoulli: $X \sim \mathrm{Bern}(p)$,
- Binomial: $X \sim \mathrm{Bin}(n, p)$,
- Poisson: $X \sim \mathrm{Poisson}(\lambda)$.

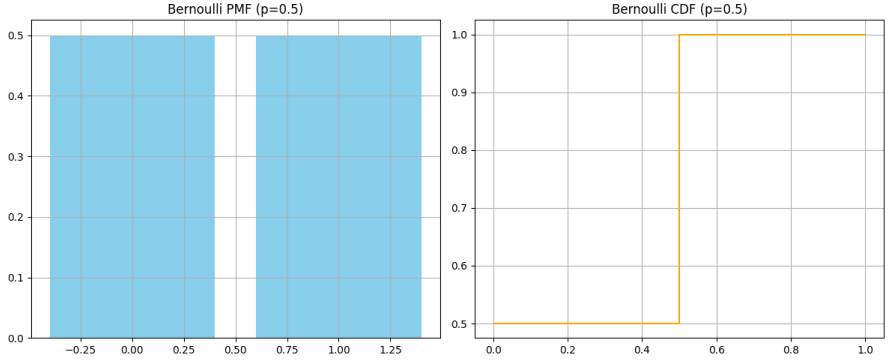
Continuous distributions:

- Uniform: $X \sim U(a, b)$,
- Normal: $X \sim N(\mu, \sigma^2)$,
- "Standard" normal: $Z \sim N(0, 1)$,
- Exponential: $X \sim Exp(\lambda)$,
- Chi-square: $X \sim \chi^2_k$,
- Student's t: $X \sim t_\nu$,
- F-Snedecor: $X \sim F_{d_1, d_2}$.

Bernoulli distribution

PMF:

$$P(X = x) = p^x(1-p)^{1-x}, \quad x \in \{0, 1\},\ 0 \leq p \leq 1$$

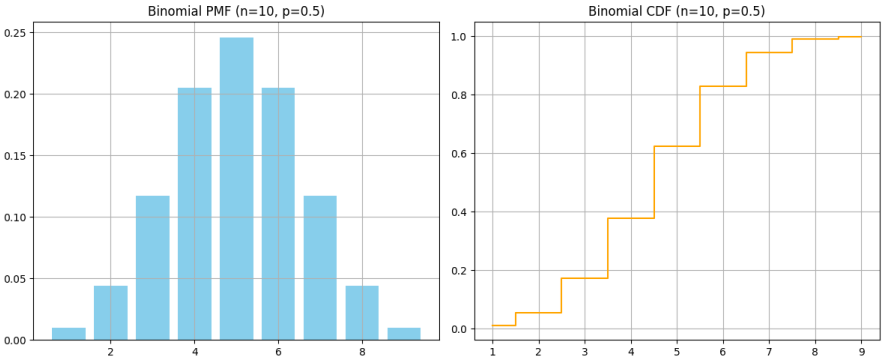


Example: a coin toss, success = heads, failure = tails.
Intuition: models an event with only two possible outcomes — success or failure.

Binomial distribution

PMF:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n$$



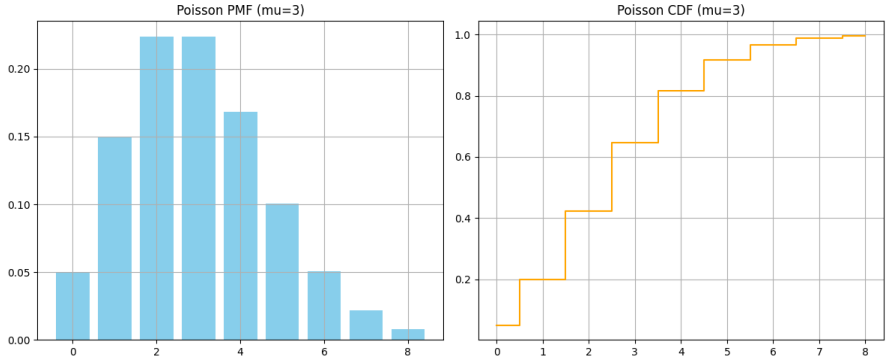
Example: number of heads in 10 coin tosses.

Intuition: the sum of n independent Bernoulli trials — how many successes occur in n trials.

Poisson distribution

PMF:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \ldots$$

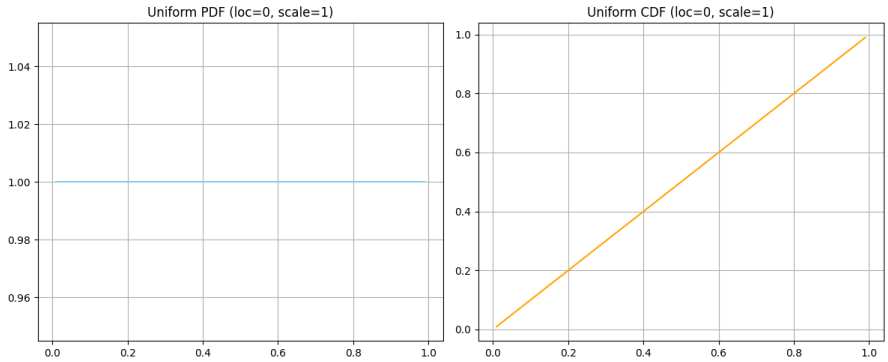


Example: number of phone calls received in one hour.
Intuition: models rare events occurring over a fixed time or space interval.

Uniform distribution

PDF:

$$f_X(x) = \frac{1}{b-a}, \quad x \in [a, b]$$

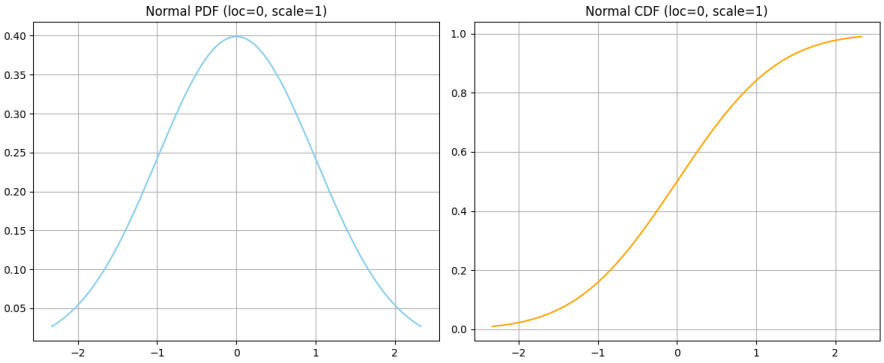


Example: drawing a random number from the interval [0,1].
Intuition: all values within the interval are equally likely.

Normal (Gaussian) distribution

PDF:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big), \quad x \in \mathbb{R}$$

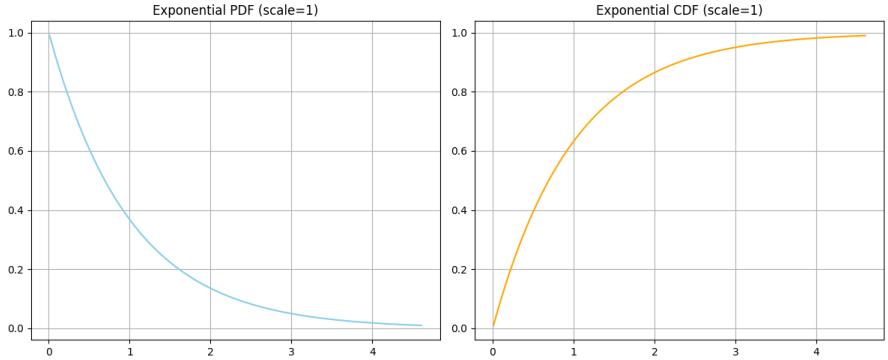


Example: human height in a population.
Intuition: symmetric "bell-shaped" curve, with most values concentrated around the mean.

Exponential distribution

PDF:

$$f_X(x) = \lambda e^{-\lambda x}, \quad x \geq 0$$



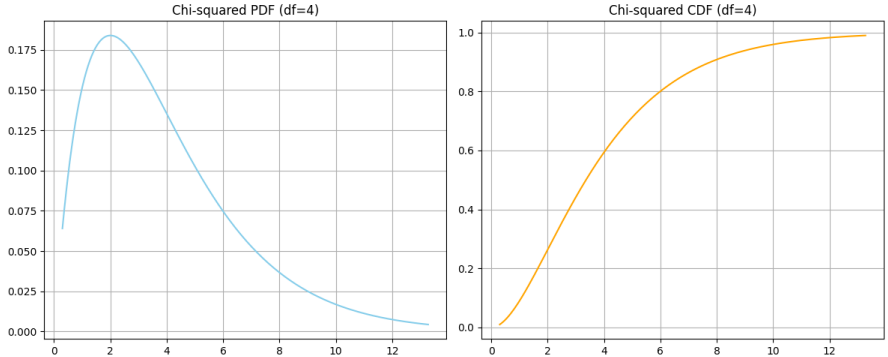
Example: waiting time for a bus to arrive.
Intuition: models the time between events; small values are more probable.

Chi-square distribution

PDF:

$$f_X(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}, \quad x \geq 0$$

$$\text{where } \Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt, \quad z > 0$$



Example: goodness-of-fit or variance tests in statistics.

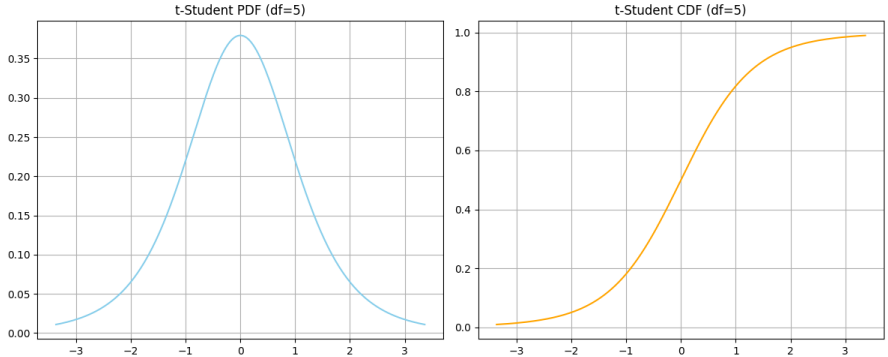
Intuition: the sum of squares of k independent standard normal variables.

Student's t distribution

PDF:

$$f_X(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \quad x \in \mathbb{R}$$

$$\text{where } \Gamma(z) = \int_0^\infty t^{z-1} e^{-t}\, dt, \quad z > 0$$



Example: testing means with a small sample size.

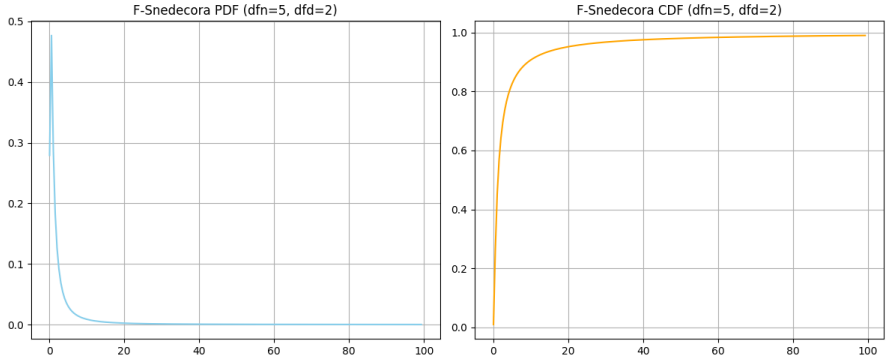
Intuition: similar to the normal distribution, but with heavier tails for small

F-Snedecor distribution

PDF:

$$f_X(x) = \frac{\sqrt{\frac{(d_1 x)^{d_1} d_2^{d_2}}{(d_1 x + d_2)^{d_1 + d_2}}}}{x\, B(d_1/2, d_2/2)}, \quad x \geq 0$$

$$\text{where } B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1} dt, \quad a, b > 0$$



Example: analysis of variance (ANOVA).

Intuition: compares the variances of two independent samples.

Relationships between distributions

Derivations of key distributions:

- **Chi-square:**

$$X_1, \ldots, X_k \sim N(0,1) \quad \Rightarrow \quad \chi^2_k = \sum_{i=1}^{k} X_i^2$$

- **Student's t:**

$$Z \sim N(0,1), \quad Y \sim \chi^2_\nu \text{ independent} \quad \Rightarrow \quad T = \frac{Z}{\sqrt{Y/\nu}} \sim t_\nu$$

- **F-Snedecor:**

$$X \sim \chi^2_{d_1}, \quad Y \sim \chi^2_{d_2} \text{ independent} \quad \Rightarrow \quad F = \frac{X/d_1}{Y/d_2} \sim F_{d_1,d_2}$$

Intuition:

- Chi-square = sum of squared standard normal variables,
- Student's t = standard normal divided by the square root of a scaled Chi-square variable,
- F-Snedecor = ratio of two scaled Chi-square variables.

Statistical hypothesis testing

What is statistical hypothesis testing?

- Hypothesis testing is a statistical method used to make inferences about a **population** based on **sample** data.
- It allows us to decide whether there is **enough evidence to reject a claim** (hypothesis) about a population parameter.
- There is a formal framework to do the above; we consider two competing hypotheses in it:

H_0 : null hypothesis (status quo)

H_1 : alternative hypothesis (research hypothesis)

Type I and type II errors

	fail to reject H_0	**reject H_0**
H_0 is true	correct decision	type I Error (α)
H_0 is false	type II Error (β)	correct decision

The most important ideas:

- **type I error** - rejecting H_0 when it is true,
- **type II error** - failing to reject H_0 when it is false,
- **significance level (α):** probability of type I error, it is usually assumed (i.e., we assume what type I error we can accept when rejecting true H_0), common choices are: 0.05, 0.01, or 0.10,
- **power of the test ($1 - \beta$):** probability of correctly rejecting H_0 when H_1 is true.

The hypothesis testing procedure

1. Formulate H_0 and H_1.
2. Choose a significance level α.
3. Select an appropriate test statistic.
4. Determine the rejection region or critical value.
5. Compute the test statistic from the sample.
6. Make a decision: reject or fail to reject H_0.
7. Interpret the result in context.

Example: one-sample t-test

Considered scenario:

- you have one group of numbers (e.g., test scores, heights, etc.),
- you want to check if the mean of your sample is different from a specific number.

Assumptions:

- continuous data: the variable is measured on a continuous scale,
- independence: each observation is independent from the others,
- normality: the population (or the sample) should be approximately normally distributed.

Numeric example:

- sample data: $[102, 98, 101, 105, 97, 99, 100, 103, 95, 104]$,
- population mean: 100,
- question: is the sample mean different from 100?

Statistical framework:

- null hypothesis (H_0): $\mu = 100$ (no difference),
- alternative hypothesis (H_1): $\mu \neq 100$ (there is a difference).

Example: one-sample t-test - cont'd

Test statistic:

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$$

Inference:

$$|t| > t_{\alpha/2,n-1} \Rightarrow \text{ reject } H_0 \text{ in favour of } H_1$$

Intuition:

- the test compares the **difference between your sample mean and 100** relative to the **spread of your data**, i.e., the test statistic shows how large the difference between the sample mean and 100 is (in standard-error units),
- if the difference is big compared to the spread $\rightarrow$ not likely by chance, if it's small or your data are noisy $\rightarrow$ could be just random.

Example: one-sample t-test - cont'd

Numeric example:

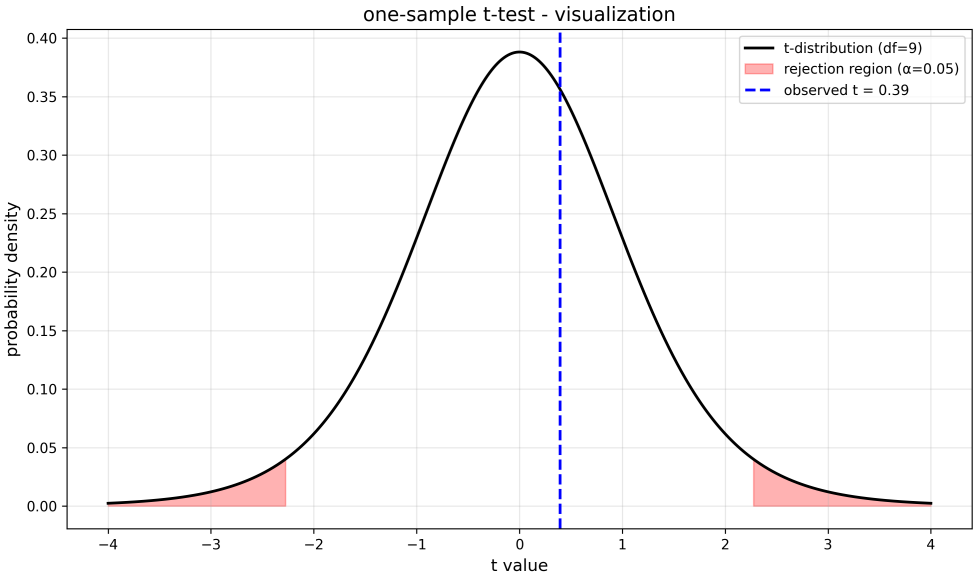
- sample data: $[102, 98, 101, 105, 97, 99, 100, 103, 95, 104]$,
- population mean: 100,
- $n = 10$, $\bar{X} = 100.4$, $s = 3.03$, $\mu_0 = 100$,
- test statistic computation:

$$t = \frac{100.4 - 100}{3.03/\sqrt{10}} \approx 0.42$$

- citical value, assumming $\alpha = 0.05$:

$$t_{0.025,9} = 2.262 \Rightarrow |t| < t_{0.025,9} \Rightarrow \text{ fail to reject } H_0$$

Example: one-sample t-test - cont'd



Example: two-sample t-test

Considered scenario:

- you have two independent groups (e.g., men vs women, treatment vs control),
- you want to check if their means are significantly different.

Assumptions:

- continuous data,
- independence: observations in each group are independent,
- normality: both groups come from approximately normal populations,
- equal variances (for the standard version; if not, use Welch's test).

Numeric example:

- group A (control): $[102, 98, 101, 105, 97]$,
- group B (treatment): $[110, 108, 112, 107, 111]$,
- question: is there a significant difference between group means?

Example: two-sample t-test - cont'd

Statistical framework:

- null hypothesis (H_0): $\mu_1 = \mu_2$ (no difference),
- alternative hypothesis (H_1): $\mu_1 \neq \mu_2$ (difference exists).

Test statistic (equal variances):

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \quad s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

$$t \sim t_{n_1+n_2-2}$$

Inference:

$$|t| > t_{\alpha/2, n_1+n_2-2} \Rightarrow \text{ reject } H_0$$

Intuition:

- compares the difference between two sample means relative to the spread within groups,
- large $|t| \rightarrow$ unlikely that observed difference is due to chance,
- small $|t| \rightarrow$ the means are similar given the variability.

Example: two-sample t-test - cont'd

Numeric computation:

- $\bar{X}_1 = 100.6$, $s_1 = 3.2$, $n_1 = 5$,
- $\bar{X}_2 = 109.6$, $s_2 = 1.9$, $n_2 = 5$,
- pooled variance:

$$s_p = \sqrt{\frac{(4)(3.2^2) + (4)(1.9^2)}{8}} = 2.68$$

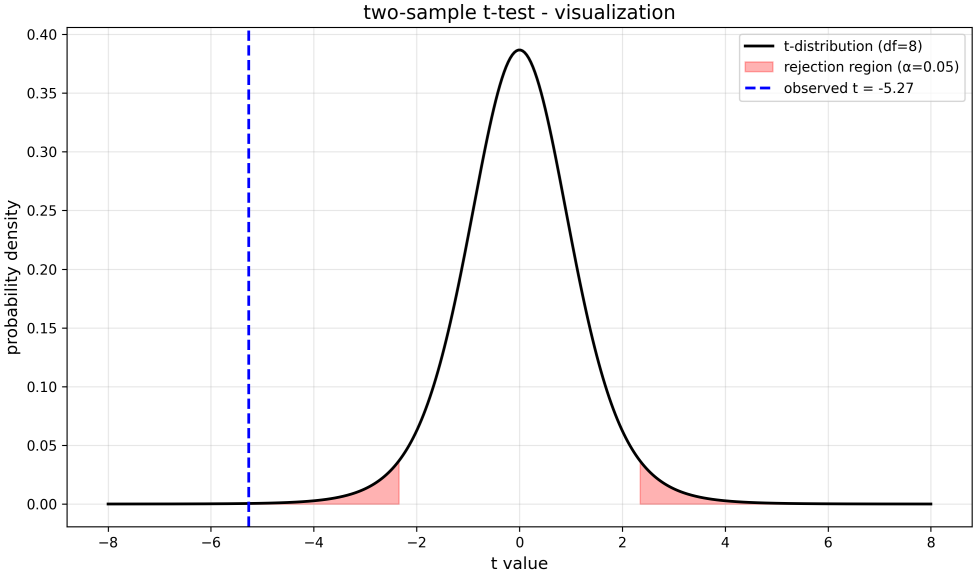
- test statistic:

$$t = \frac{100.6 - 109.6}{2.68\sqrt{\frac{1}{5} + \frac{1}{5}}} = -6.0$$

- critical value for $\alpha = 0.05$, $df = 8$: $t_{0.025,8} = 2.306$

$$|t| > t_{0.025,8} \Rightarrow \text{ reject } H_0$$

Example: two-sample t-test - cont'd



Example: two-sample t-test (left-tailed)

Statistical framework:

- null hypothesis (H_0): $\mu_1 = \mu_2$,
- alternative hypothesis (H_1): $\mu_1 < \mu_2$.

Test statistic:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$t \sim t_{n_1+n_2-2}$$

Decision rule:

$$t < -t_{\alpha, n_1+n_2-2} \Rightarrow \text{ reject } H_0$$

Example:

$$t = -6.0, \quad t_{0.05,8} = 1.860$$

$$t < -t_{0.05,8} \Rightarrow -6.0 < -1.860 \Rightarrow \text{reject } H_0$$

Conclusion: mean of group 1 is significantly smaller than mean of group 2.

Example: two-sample t-test (right-tailed)

Statistical framework:

- null hypothesis (H_0): $\mu_1 = \mu_2$,
- alternative hypothesis (H_1): $\mu_1 > \mu_2$.

Test statistic:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$t \sim t_{n_1+n_2-2}$$

Decision rule:

$$t > t_{\alpha, n_1+n_2-2} \Rightarrow \text{ reject } H_0$$

Example:

$$t = -6.0, \quad t_{0.05,8} = 1.860$$

$$t > t_{0.05,8} \Rightarrow -6.0 > 1.860 \Rightarrow \text{do not reject } H_0$$

Conclusion: no evidence that mean of group 1 is significantly greater than mean of group 2.

What is the p-value?

Definition

The **p-value** is the probability of obtaining a result at least as extreme as the one observed, assuming that the null hypothesis (H_0) is true.

- Small p-value ($< \alpha$) $\Rightarrow$ evidence **against** H_0.
- Large p-value ($> \alpha$) $\Rightarrow$ data are **consistent** with H_0.

Note: α is a preassumed significance level, commonly $\alpha = 5\%$, sometimes $\alpha = 10\%$, or $\alpha = 1\%$, or $\alpha = 0.1\%$.

Interpreting p-values obtained from statistical tests 1/3

Example (revisited) - one-sample t-test:

$$\text{sample} = [102, 98, 101, 105, 97, 99, 100, 103, 95, 104]$$

$$n = 10, \qquad \bar{X} = 100.4, \qquad s = 3.03, \qquad \mu_0 = 100$$

$$t = \frac{100.4 - 100}{3.03/\sqrt{10}} \approx 0.42$$

degrees of freedom: $\nu = n - 1 = 9$

Two-sided p-value:

$$p = 2(1 - F_{t,9}(|t|)) \approx 0.68$$

Conclusion:

$$p \approx 0.68 > 0.05 \Rightarrow \textbf{fail to reject } H_0$$

i.e., no evidence that the mean differs from 100

Interpreting p-values obtained from statistical tests 2/3

Example (revisited) - two-sample t-test:

$$\text{group A} = [102, 98, 101, 105, 97], \quad \text{group B} = [110, 108, 112, 107, 111]$$

$$n_1 = n_2 = 5, \quad \bar{X}_1 = 100.6, \quad \bar{X}_2 = 109.6, \quad s_1 = 3.2, \quad s_2 = 1.9$$

$$s_p = \sqrt{\frac{(4)(3.2^2) + (4)(1.9^2)}{8}} \approx 2.68$$

$$t = \frac{100.6 - 109.6}{2.68\sqrt{1/5 + 1/5}} \approx -6.0$$

$$\text{degrees of freedom: } \nu = n_1 + n_2 - 2 = 8$$

Two-sided p-value:

$$p = 2(1 - F_{t,8}(|t|)) \approx 0.0002$$

Conclusion:

$$p \approx 0.0002 < 0.05 \Rightarrow \textbf{reject } H_0$$

i.e., strong evidence that the means of group A and B differ.

Interpreting p-values obtained from statistical tests 3/3

Example - two sample t-test, left-tailed:

$$H_0 : \mu_1 = \mu_2, \quad H_1 : \mu_1 < \mu_2$$

$$t \approx -6.0, \quad \nu = 8$$

Left-tailed p-value:

$$p = F_{t,8}(t) \approx 0.0001$$

Conclusion:

$$p \approx 0.0001 < 0.05 \Rightarrow \textbf{reject } H_0$$

i.e., mean of group A is significantly smaller than mean of group B.

Thank you for your attention!

Maciej Świtała, PhD
ms.switala@uw.edu.pl