# Introduction.
# Main problems and methods in Applied Statistics

Maxim Panov

Skoltech

November, 2021

Skoltech

# Outline

# Outline

# Learning Outcomes

### Knowledge

- How the ideas from mathematical statistics can be applied in modern methods of data analysis and processing.

### Skill

Be able to

- formulate in mathematical terms a real-world problem,
- built a corresponding probabilistic model,
- select an appropriate statistical method.

### Experience

- Obtain a sufficient experience during practical exercises and project activities to become a qualified user of statistical methods.

# Course Structure

Lec. 1 Introduction. Main problems and methods in Applied Statistics

Lec. 2 Elements of probability theory. Statistical functionals and distances.

Lec. 3 Parametric estimation

Lec. 4 Confidence intervals and bootstrap

Lec. 5 Hypothesis testing 1

Lec. 6 Hypothesis testing 2

Lec. 7 Regression and design of experiments

Lec. 8 Nonparametric estimation

Lec. 9 Bayesian estimation

Lec.10 MCMC and sampling

Bonus AB testing

# Assignments, Exam, Project

- Assignments (4 * 12.5 % = 50%)

HW 1: Parametric estimation and confidence intervals // Deadline 10.11.2021
HW 2: Hypothesis testing // Deadline 24.11.2021
HW 3: Regression and nonparametric estimation // Deadline 04.12.2021
HW 4: Bayesian estimation // Deadline 11.12.2021

- Midterm exam (20%) // Date: 18.11.2021

- Final project (30%) // Implementation and investigation of recent methods on intersection of statistics and machine learning.

# Course Logistics

- Lectures (Tuesday and Thursday at 4 pm)
  - Online via Zoom.
- Practical seminars (Tuesday and Thursday at 5:30-6:30 pm)
  - Online via Zoom.
- Contact us (TA office hours: Monday, Wednesday and Friday, 10am - 6pm)
  - Piazza for discussions: http://piazza.com/skoltech.ru/fall2021/ma030416;
  - Canvas for main announcements;
  - Telegram channel for rapid information: https://t.me/sk_applied_statistics.

# Course Instructor and Teaching Assistants

**Course instructor:**

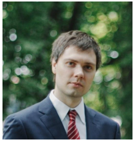
- Maxim Panov, m.panov@skoltech.ru
- Assistant Professor at Skoltech
- PhD in Statistics
- Head of Statistical Machine Learning group

Maxim Panov

**Teaching Assistants:**

- Researchers at Skoltech;
- Perform research in intersection of Machine Learning and Statistics.

Daria Kotova



Alexander Fishkov

# Outline

# Population and Sample

**Population:** the entire set of observations.

**Sample:** a sub-group of the population.

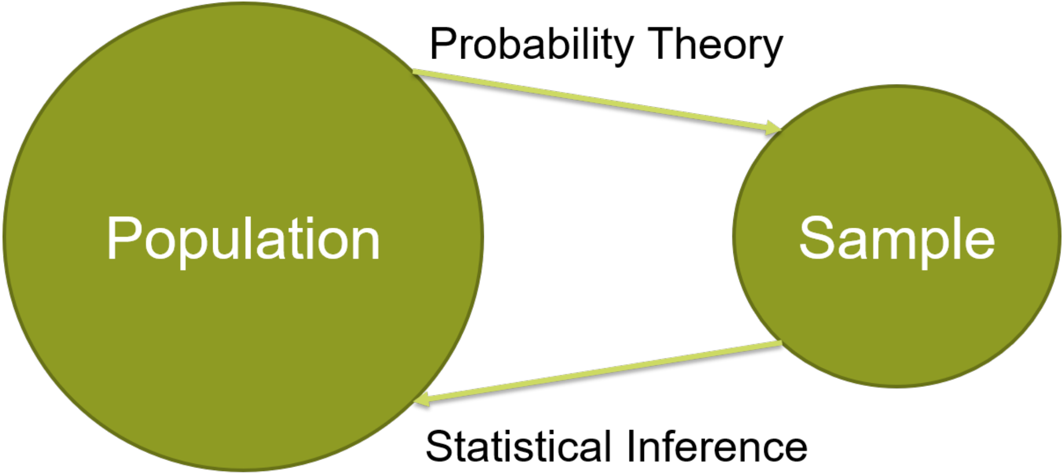
**Parameter:** the true value of a characteristic of the population

- usually denoted by Greek characters: $\mu$ and $\sigma^2$.

**Statistic:** an estimate of the parameter calculated using the sample

- denoted by normal characters: $\bar{x}$ and $s^2$.

# Probability vs Statistics



# Probability

- Probability underlies statistical inference – the drawing of conclusions from a sample of data.
- If samples are drawn at random, their characteristics (such as the sample mean) depend upon chance.
- Hence to understand how to interpret sample evidence, we need to understand chance, or probability.

# Probability Distributions

- With each outcome in the sample space we can associate a probability.
- Example: Toss a coin
  - $\Pr(\text{Head}) = \frac{1}{2}$,
  - $\Pr(\text{Tail}) = \frac{1}{2}$.
- This is an example of a probability distribution (the particular case of Bernoulli distribution).

# Definition of Probability

The probability of an event A may be defined in different ways:

- **The frequentist view:** the proportion of trials in which the event occurs, calculated as the number of trials approaches infinity.

- **The subjective view:** someone's degree of belief about the likelihood of an event occurring.

**Axioms of Probability:**

- $0 \leq P(A) \leq 1$;
- $\sum_{i=1}^{n} P(A_i) = 1$, where $i$ runs over all outcomes;
- $P(not\ A) = 1 - P(A)$.

# Statistical model

Let the data sample consists of independent identically distributed random variables: $X_1, \dots, X_n \sim F$.

Our goal is to infer $F$ or some feature of $F$ given a sample.

**Example (Parametric Estimation)**

Suppose that $X_1, \dots, X_n$ are independent and distributed according to Normal distribution with unknown mean $\mu$ and known variance $\sigma^2 = 1$.

The goal is to estimate the parameter $\mu$ from the data.

**Question:** What can be used as an estimate of $\mu$?

# Random Variables

One of the possible estimates:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

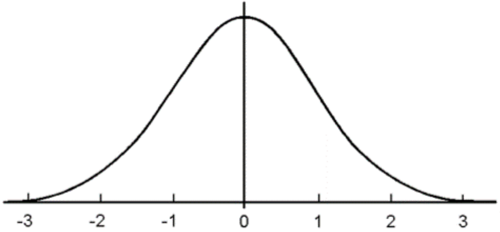
which is usually called *sample mean*.

- Most statistics (e.g. the sample mean) are random variables.
- Many random variables have well-known probability distributions associated with them.
- To understand random variables, we need to know about probability distributions

# Normal Distribution

The Normal distribution is

- Bell-shaped;
- symmetric;
- unimodal;
- defined for $x \in (-\infty; +\infty)$.

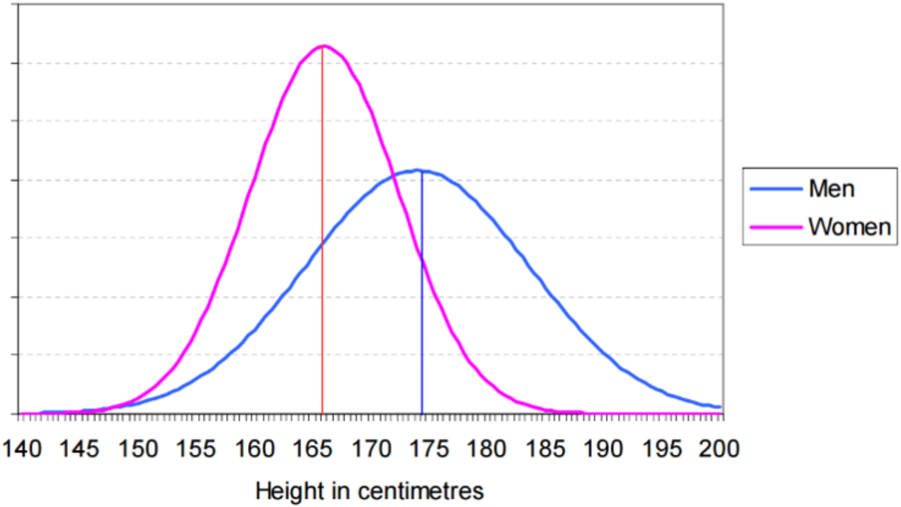



The density of the normal distribution:

$$f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right).$$

Normal distribution is the case when many small independent factors influence a variable.

# Men's and Women's Heights



# Normal Distribution

- Two parameters of the Normal distribution are the mean $\mu$ and the variance $\sigma^2$:

$$x \sim \mathcal{N}(\mu, \sigma^2).$$

- Men's heights are Normally distributed with mean 174 cm and variance 92.16:

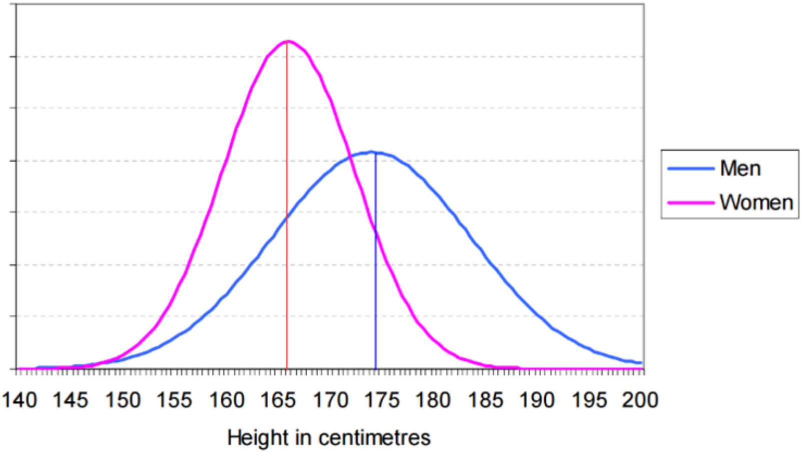
$$x \sim \mathcal{N}(174, 92.16).$$

- Women's heights are Normally distributed with mean 166 cm and variance 40.32:

$$x \sim \mathcal{N}(166, 40.32).$$

# Men's and Women's Heights

- Men: $x \sim \mathcal{N}(174, 92.16)$.
- Women: $x \sim \mathcal{N}(166, 40.32)$.



# Outline

# Statistical Inference

We have just discussed how one can do estimation based on the sample from probability distribution.

**Question:** can we say something about the properties of the obtained estimates?


Statistical inference aims to draw the conclusions about the properties of the estimates.

# Distribution of the Sample Mean

- If samples of size $n$ are randomly drawn from a Normally distributed population of $\mu$ and variance $\sigma^2$ the sample mean is distributed as

$$\bar{x} \sim \mathcal{N}(\mu, \sigma^2/n).$$

- E.g. if samples of 50 women are chosen, the sample mean is distributed

$$\bar{x} \sim \mathcal{N}(166, 40.32/50).$$

- Note the very small standard error: $\sqrt{40.32/50} \simeq 0.897$.

# Distribution of the Sample Mean

- Note the distinction between
$$x \sim \mathcal{N}(\mu, \sigma^2).$$
and
$$\bar{x} \sim \mathcal{N}(\mu, \sigma^2/n).$$

- The former refers to the distribution of a typical member of the population and the latter to the distribution of the sample mean.

- **Question:** what if individual $x$ is not Normally distributed?

# Distribution of the Sample Mean

- Note the distinction between
$$x \sim \mathcal{N}(\mu, \sigma^2).$$
and
$$\bar{x} \sim \mathcal{N}(\mu, \sigma^2/n).$$

- The former refers to the distribution of a typical member of the population and the latter to the distribution of the sample mean.

- **Question:** what if individual $x$ is not Normally distributed?

- **Answer:** in the case of the sample mean *Central Limit Theorem* can help.

# Maximum Likelihood Estimation

Let $X_1, \ldots, X_n$ are i.i.d. with PDF $f(x;\theta)$.

**Definition**

The **likelihood function** is defined by

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f(X_i;\theta).$$

The **log-likelihood function** is defined by

$$\ell_n(\theta) = \log \mathcal{L}_n(\theta).$$

We treat likelihood as a function of the parameter $\mathcal{L}_n \colon \Theta \to [0,\infty)$.

**Definition**

The **maximum likelihood estimator (MLE)** is defined by

$$\widehat{\theta}_n = \arg\max_{\theta} \mathcal{L}_n(\theta) = \arg\max_{\theta} \ell_n(\theta).$$

# Properties of the MLE

1. The MLE **is consistent**: $\widehat{\theta}_n \xrightarrow{\mathbb{P}} \theta_*$;
2. The MLE **is asymptotically Normal**: $(\widehat{\theta} - \theta_*)/\widehat{\sigma} \rightsquigarrow \mathcal{N}(0, 1)$;
3. The MLE is asymptotically optimal or efficient: roughly, this means that among all well-behaved estimators, the MLE has the smallest variance, at least for large samples.

**Remark**

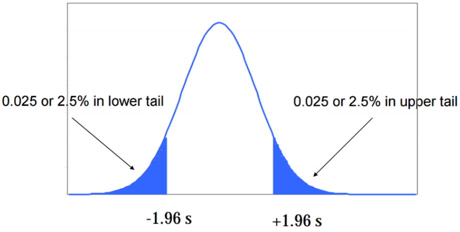
- *Aforementioned properties of the MLE hold if function $f(x; \theta)$ is sufficiently regular.*
- *In sufficiently complicated problems, these properties will no longer hold and the MLE will no longer be a good estimator.*

# Beyond point estimates

- Point estimate – a single value
  - E.g. the temperature tomorrow will be 23°.
- Interval estimate – a range of values, expressing the degree of uncertainty
  - E.g. the temperature tomorrow will be between 21° and 25°.

**Estimating a mean:**

- Point estimate: use the sample mean.
- Interval estimate: sample mean $\pm$ "something".
- What is something?
- Go back to the distribution of $\bar{x}$.



# The 95% confidence interval

- Recall the distribution of the sample mean

$$\bar{x} \sim \mathcal{N}(\mu, \sigma^2/n).$$

  Hence the 95% probability interval is

$$P\left(\mu - 1.96\sqrt{\sigma^2/n} \leq \bar{x} \leq \mu + 1.96\sqrt{\sigma^2/n}\right) = 0.95.$$

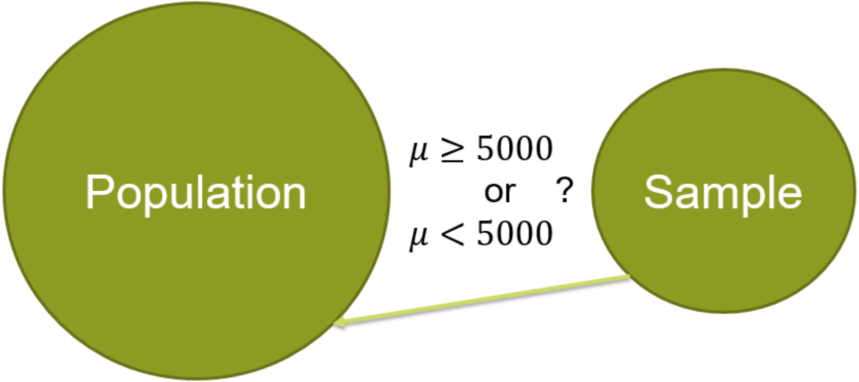
- Rearranging this gives the 95% confidence interval for our estimate of the true population mean:

$$\left[\bar{x} - 1.96\sqrt{\sigma^2/n} \leq \mu \leq \bar{x} + 1.96\sqrt{\sigma^2/n}\right].$$

# Outline

# Hypothesis Testing

Assume that $x \sim \mathcal{N}(\mu, \sigma^2)$ and $\sigma = 500$.



# Principles of Hypothesis Testing

- The null hypothesis is initially presumed to be true.
- Evidence is gathered, to see if it is consistent with the hypothesis, and tested using a decision rule.
- If the evidence is consistent with the hypothesis, the null hypothesis continues to be considered "true".
- If not, the null is rejected in favour of the alternative hypothesis.

# Two Possible Types of Error

- Decision making is never perfect and mistakes can be made.
- Type I error. Rejecting the null when it is true:
  - shows a patient to have a disease when in fact the patient does not have the disease;
  - a fire alarm going on indicating a fire when in fact there is no fire.
- Type II error. Accepting the null when it is false:
  - a blood test failing to detect the disease it was designed to detect, in a patient who really has the disease;
  - a fire breaking out and the fire alarm does not ring.

# Outcomes of Hypothesis Testing

| | | True hypothesis | |
|---|---|---|---|
| | | $H_0$ | $H_1$ |
| Hypothesis test | $H_0$ | OK | Type II error |
| Result | $H_1$ | Type I error | OK |

- We wish to avoid both Type I and II errors by altering the decision rule.
- Unfortunately, reducing the chance of making a Type I error generally means increasing the chance of a Type II error $\rightarrow$ There is a trade off.

# Example: Normal Distribution

## Example

- Let $X_1, X_2, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, where $\sigma$ is known.
- Consider 2 hypotheses $H_0: \mu \geq 5000 \quad H_1: \mu < 5000$.
- Let us introduce the test which rejects $H_0$ if $T = \bar{x} < c$.
- Thus,

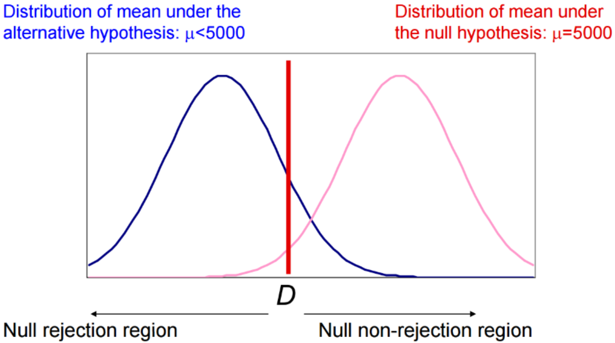
$$P_\mu(\bar{x} < c) = P_\mu\left(\frac{\sqrt{n}(\bar{x} - \mu)}{\sigma} < \frac{\sqrt{n}(c - \mu)}{\sigma}\right)$$

$$= P\left(Z < \frac{\sqrt{n}(c - \mu)}{\sigma}\right) = \Phi\left(\frac{\sqrt{n}(c - \mu)}{\sigma}\right),$$

where $Z$ has standard Normal distribution.

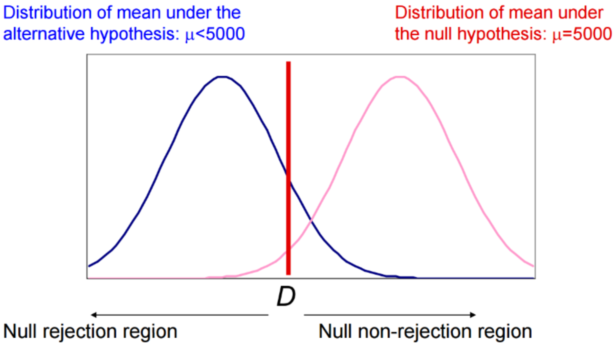
# Example: How Long do Batteries Last?

- A well known battery manufacturer claims its product lasts at least 5000 hours, on average.
- A sample of 80 batteries is tested. The average time before failure is 4900 hours, with standard deviation 500 hours.
- Should the manufacturer's claim be accepted or rejected?



# Should the Null Hypothesis be Rejected?

- Is 4900 far enough below 5000?
- 4900 is 1.79 standard errors below 5000 so falls into the rejection region (bottom 5% of the distribution).
- Hence, we can reject $H_0$ with 95% confidence.
- If the true mean were 5 000, here is less than a 5% chance of obtaining sample evidence such as $\bar{x} \leq 4900$ from a sample of $n = 80$.



# Multiple Testing

- Genomics = Lots of Data = Lots of Hypothesis Tests
- A typical microarray experiment might result in performing 10000 separate hypothesis tests. If we use a standard significance level of 0.05, we'd expect 500 genes to be deemed "significant" by chance.

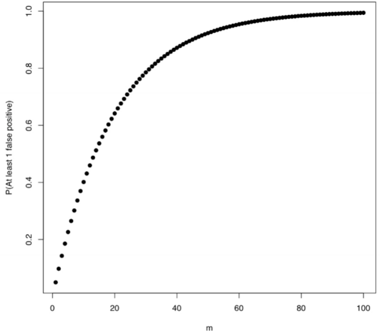
In general, if we perform $m$ hypothesis tests, what is the probability of at least 1 false positive?

$$P(Making\ an\ error) = \alpha,$$

$$P(Not\ making\ an\ error) = 1 - \alpha,$$

$$P(Not\ making\ an\ error\ in\ m\ tests) = (1 - \alpha)^m,$$

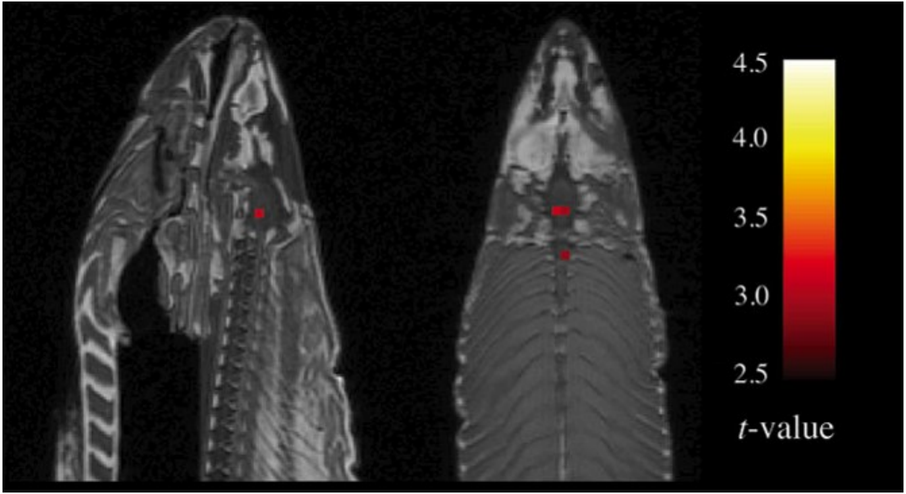
$$P(Making\ at\ least\ 1\ error\ in\ m\ tests) = 1-(1-\alpha)^m.$$



# The Dead Salmon Study

- Neuroscientist purchased a whole Atlantic salmon.
- He took it to a lab put it into an fMRI machine used to study the brain.
- So, as the fish sat in the scanner, they showed it "a series of photographs depicting human individuals in social situations".
- Salmon "was asked to determine what emotion the individual in the photo must have been experiencing".
- The salmon "was not alive at the time of scanning".

# The Dead Salmon Study



# Conclusions and Outlook

- This is not and advanced course.
  - If you had a decent undergrad course on Statistics, then you probably will know a significant share of the content.
- However, we have some topics that go beyond standard courses:
  - Bootstrap;
  - Robust statistics;
  - Design of experiments;
  - Bayesian estimation and MCMC.
  - ...
- We don't plan to go deep into theory, but we will provide detailed justifications behind (almost) all the methods used.
- We plan to focus on the hands-on experience of statistical modelling in Python.

# Literature

Some useful books on statistics and statistical learning:

- Wasserman "All of the Statistics"
- James, Witten, Hastie, Tibshirani "Introduction to Statistical Learning"
- Hastie, Tibshirani, Friedman "Elements of Statistical Learning"
- Ben-David, Shalev-Shwartz "Understanding Machine Learning"

# Thank you for your attention![1]

[1]And many thanks for wonderful lectures by Paula Surridge (School of Sociology, Politics and International Studies University of Bristol), which partially inspired these slides.