Introduction to Data Science

Lecture 2. Elements of Statistics

04.10.18

Mikhail Belyaev, Maxim Panov

Skoltech
Skolkovo Institute of Science and Technology

Info about the courses

Course instructors:

- Mikhail Belyaev, m.belyaev@skoltech.ru
- Maxim Panov, m.panov@skoltech.ru

Who we are (both MB & MP)

- Researchers at CDISE
- PhD in Data Science (Candidate of Science in math)
- Have 5+ years as Data Scientists at Datadvance company:
  - Developed machine learning algorithms in the context of an industrial data analysis library intended mainly for aerospace and automotive
  - Solved a set of data analysis problem from Airbus, Astrium, Areva, Eurocopter, Force India F1 ant many others

Outline

- Introduction
- Probability
- Statistical Estimation and Inference
- Hypothesis Testing

Outline

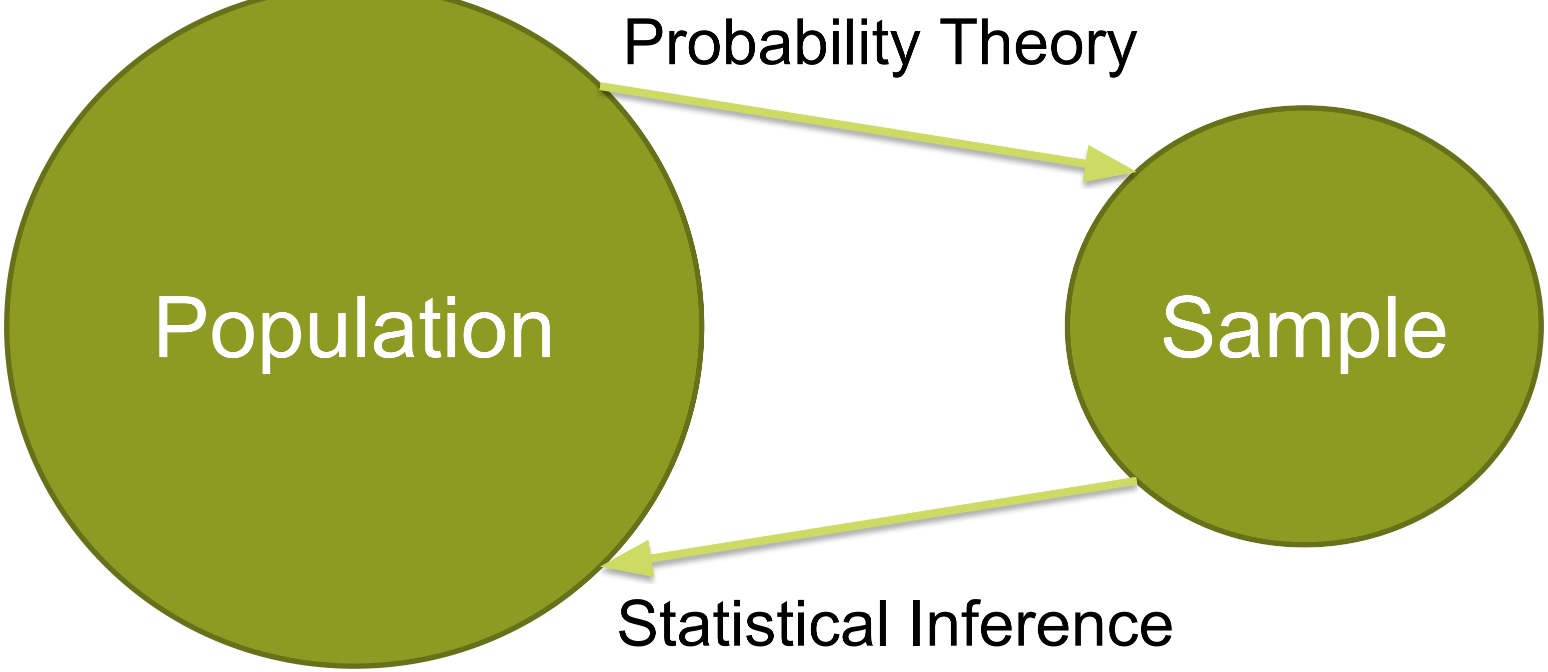
- **Introduction**
- Probability
- Statistical Estimation and Inference
- Hypothesis Testing

Introduction

- **Population:** the entire set of observations.
- **Sample:** a sub-group of the population.
- **Parameter:** the true value of a characteristic of the population
  - denoted by Greek characters: $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$.
- **Statistic:** an estimate of the parameter calculated using the sample
  - denoted by normal characters: $\bar{\boldsymbol{x}}$ and $\boldsymbol{s}^2$.

Skoltech
Skolkovo Institute of Science and Technology

Probability vs Statistics



Skoltech
Skolkovo Institute of Science and Technology

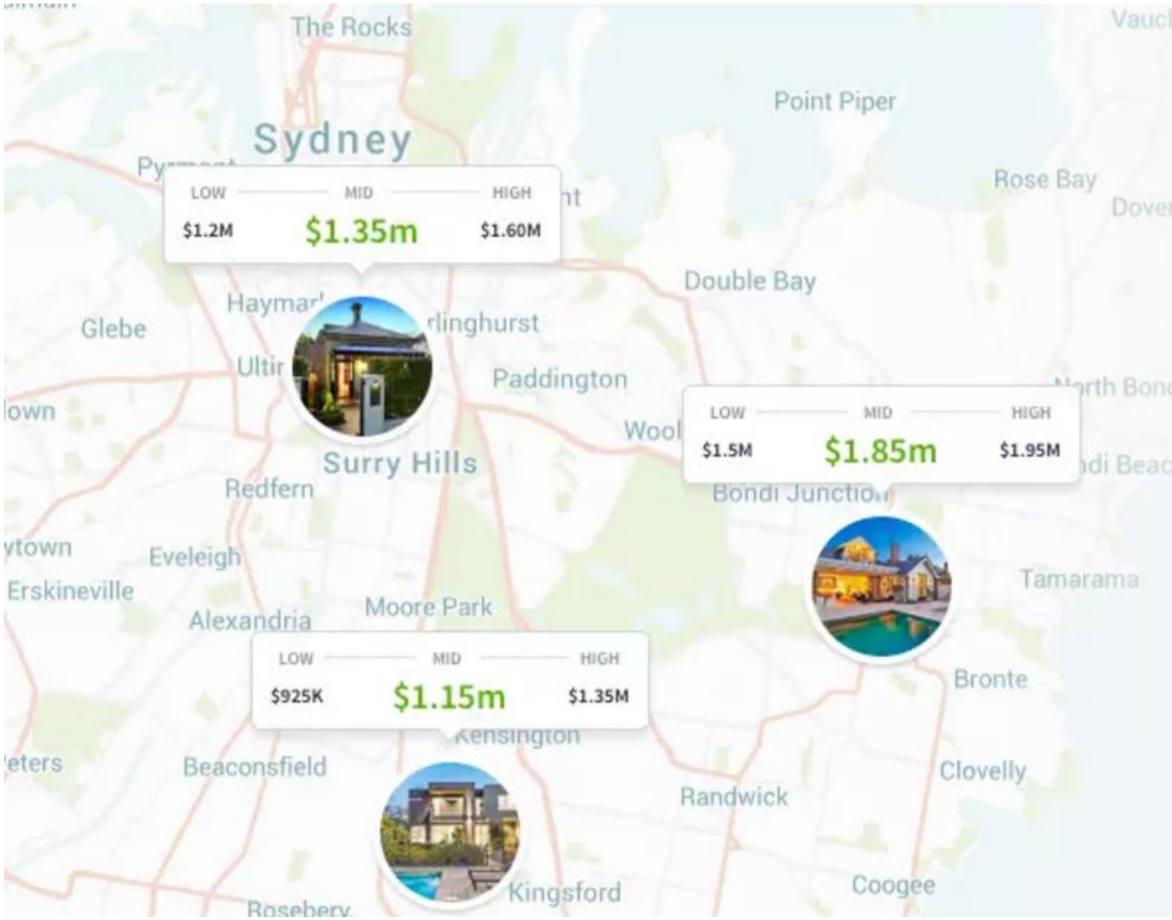
Probability

- 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.

Real estate price estimation



The question: what is *MID*?

Skoltech
Skolkovo Institute of Science and Technology

Measures of location

- **Mean** – strictly the arithmetic mean, the well known 'average'.
- **Median** – e.g. the estate price in the middle of the distribution.
- **Mode** – e.g. the estate price that occurs most often.
- These different measures can give different answers…

Mean

Estate	1	2	3	4	5	6	7	8	9	10
price	15	15	20	25	45	55	70	85	125	250

$$\mu = \frac{\sum_i x_i}{n} = \frac{705}{10} = 70.5$$

Mean estate price is therefore 70.5 million rubles.

Mean

Estate	1	2	3	4	5	6	7	8	9	10
price	15	15	20	25	45	55	70	85	125	250

$$\bar{x} = \frac{\sum_i x_i}{n} = \frac{705}{10} = 70.5$$

Mean estate price is therefore 70.5 million rubles

Median

- The price of the 'middle estate' – i.e. the one located halfway through the distribution.

Cheapest

The most expensive

This estate's price

- The median is little affected by outliers unlike the mean.

Median

- We have 10 observations in the sample, so the estate 5.5 in rank order has the median wealth. This estate is somewhere between 45M and 55M.

Estate	1	2	3	4	5	6	7	8	9	10
price	15	15	20	25	45	55	70	85	125	250

- Hence the median income is 50M per year.
- Q: what happens to the median if the richest person's income is doubled to 500M?
- Q: what happens to the mean?

Mode

- The mode is the observation with the highest frequency.
- For our data we have a single mode at 15 M.
- It is possible to have a sample or population with no mode, or more than 1 mode:
  - e.g. two modes: bimodal.

Measures of dispersion

- **Range** – the difference between smallest and largest observation. Not very informative for most purposes.
- **Variance** – based on all observations in the population or sample.

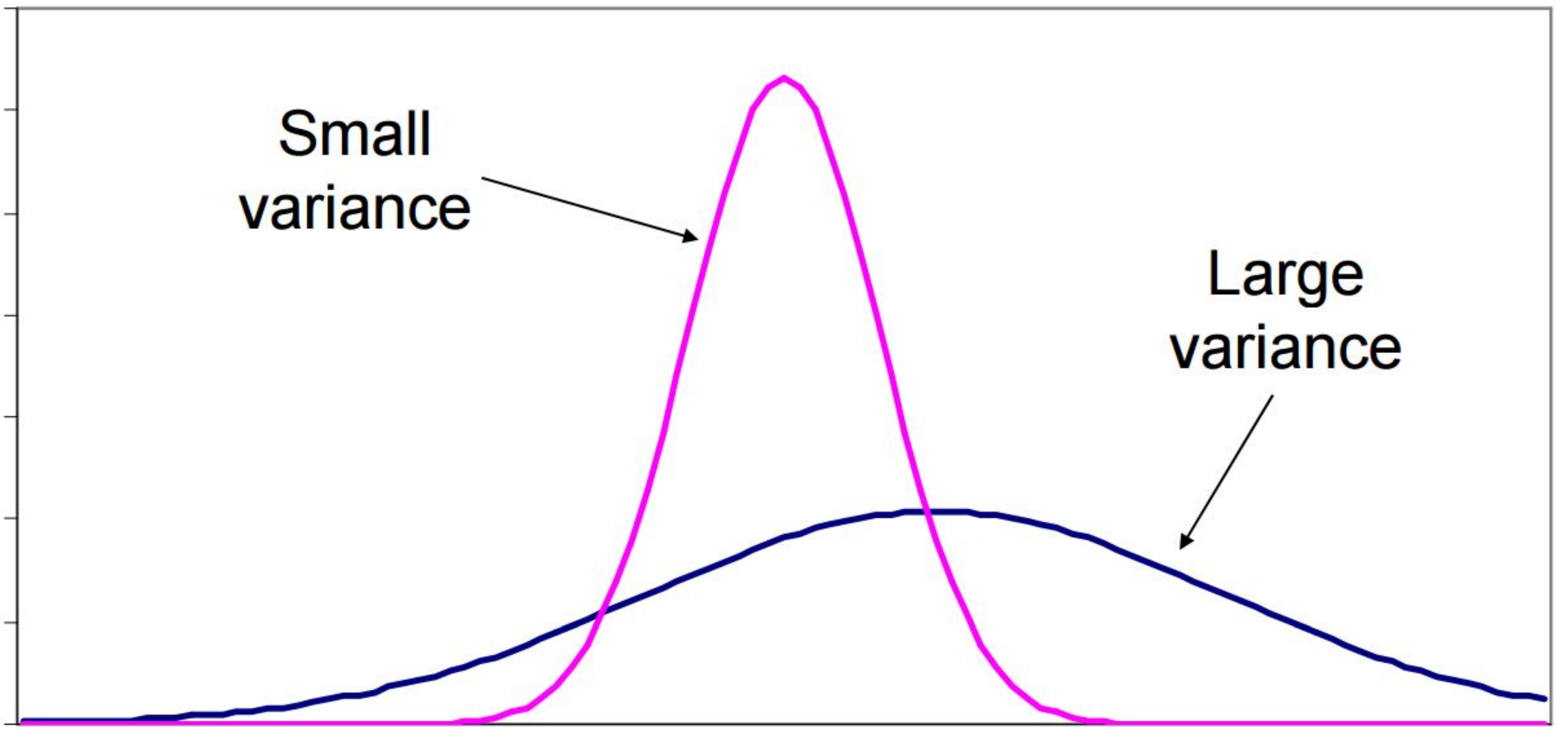
Variance

- The variance is the average of all squared deviations from the mean:

$$\sigma^2 = \frac{\sum_i (x_i - \mu)^2}{n}$$

- The larger this value, the greater the dispersion of the observations.
- **NB**: σ^2 is used for population variance; for sample variance use $s^2$and divide by $n - 1$ rather than by n.

Variance

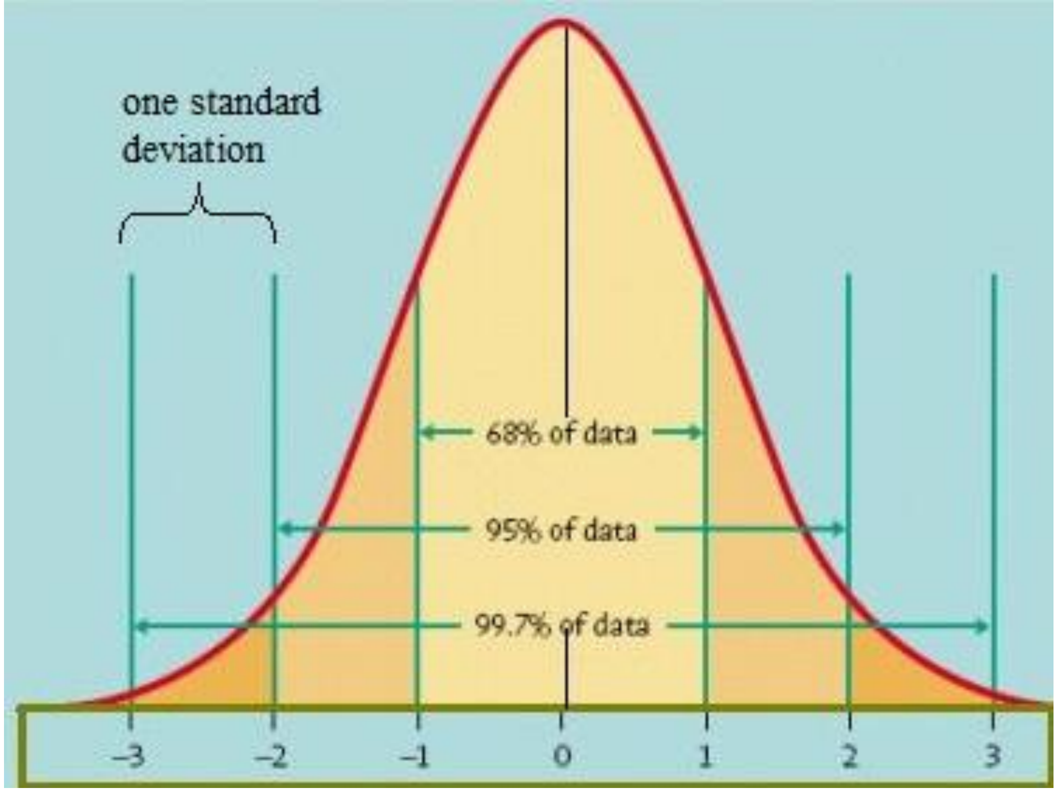


Calculating the Sample Variance

- $$s^2 = \frac{\sum_i (x_i - \bar{x})^2}{n-1}$$
- In our example: $s^2 = 5230$.
- Standard deviation: $s = \sqrt{s^2} = 72.318$.
- Standard deviation:
  - The % of obs. that lie within a given number of standard deviations above or below the mean.
  - Where a particular observation lies relative to the

Standard deviation

- 68% of observations lie within ± 1 st. devs
- 95% of observations lie within ± 2 st. devs
- 99% of observations lie within ± 3 st. devs



Skoltech
Skolkovo Institute of Science and Technology

Outline

- Introduction
- **Probability**
- Statistical Estimation and Inference
- Hypothesis Testing

Probability distributions

- With each outcome in the sample space we can associate a probability.
- **Example**: Toss a coin
  - Pr(Head) = ½
  - Pr(Tail) = ½
- This is an example of a **probability 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.

Rules for Probabilities

- $0 \leq P(A) \leq 1$
- $\sum_i P(A_i) = 1$, where i runs over all outcomes
- $P(not\ A) = 1 - P(A)$

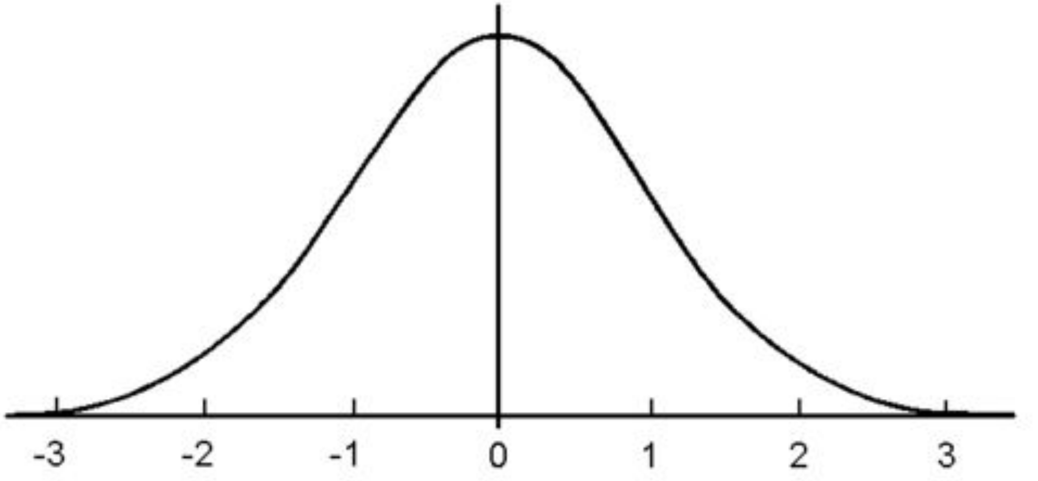
Random variables

- 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

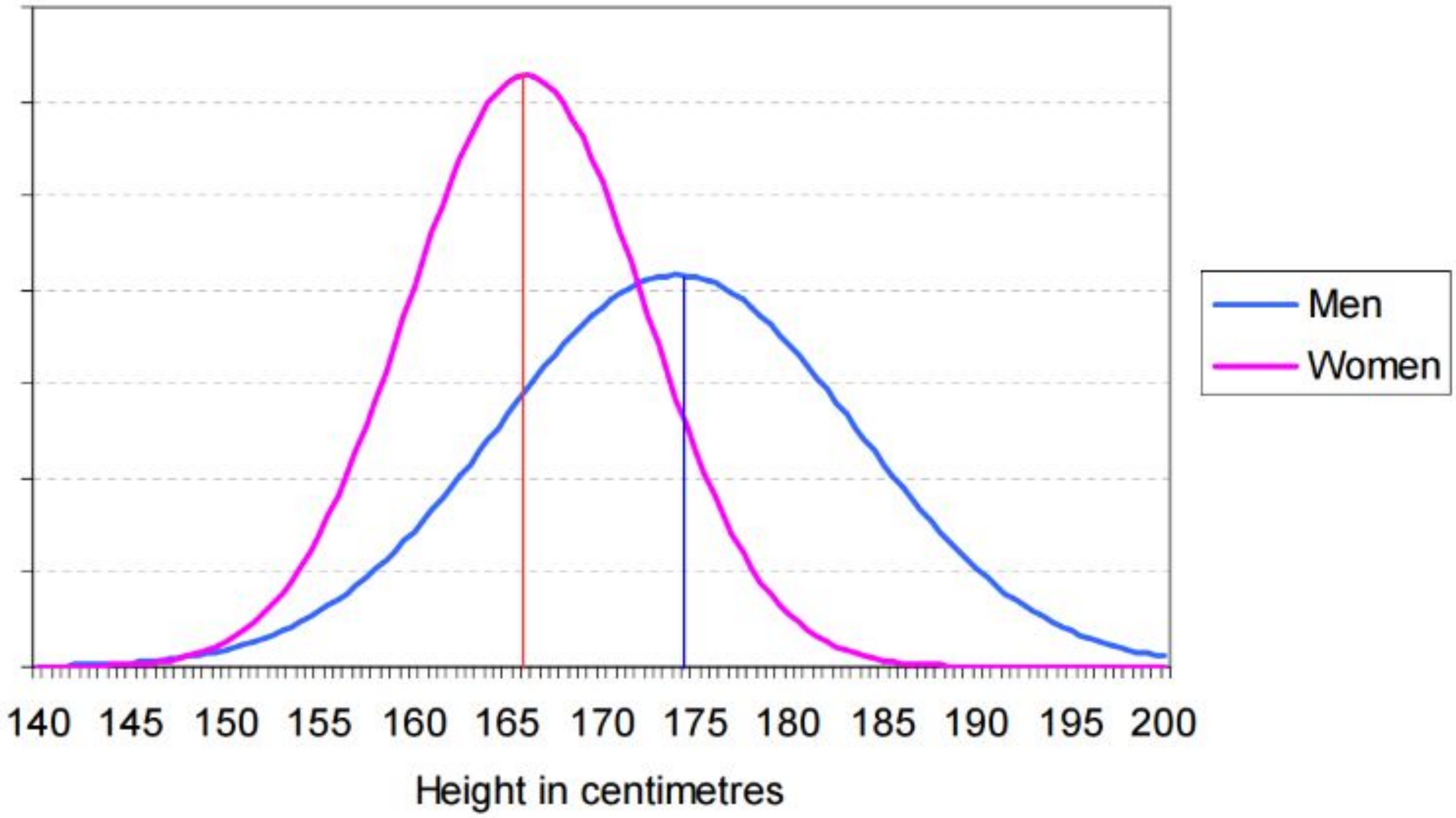
➔ The Normal distribution is

- Bell shaped
- Symmetric
- Unimodal
- and is defined for $x \in (-\infty; +\infty)$



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

Men’s and Women’s Heights



Skoltech
Skolkovo Institute of Science and Technology

Normal distribution

- The two parameters of the Normal distribution are the **mean** μ and the **variance** σ^2:

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

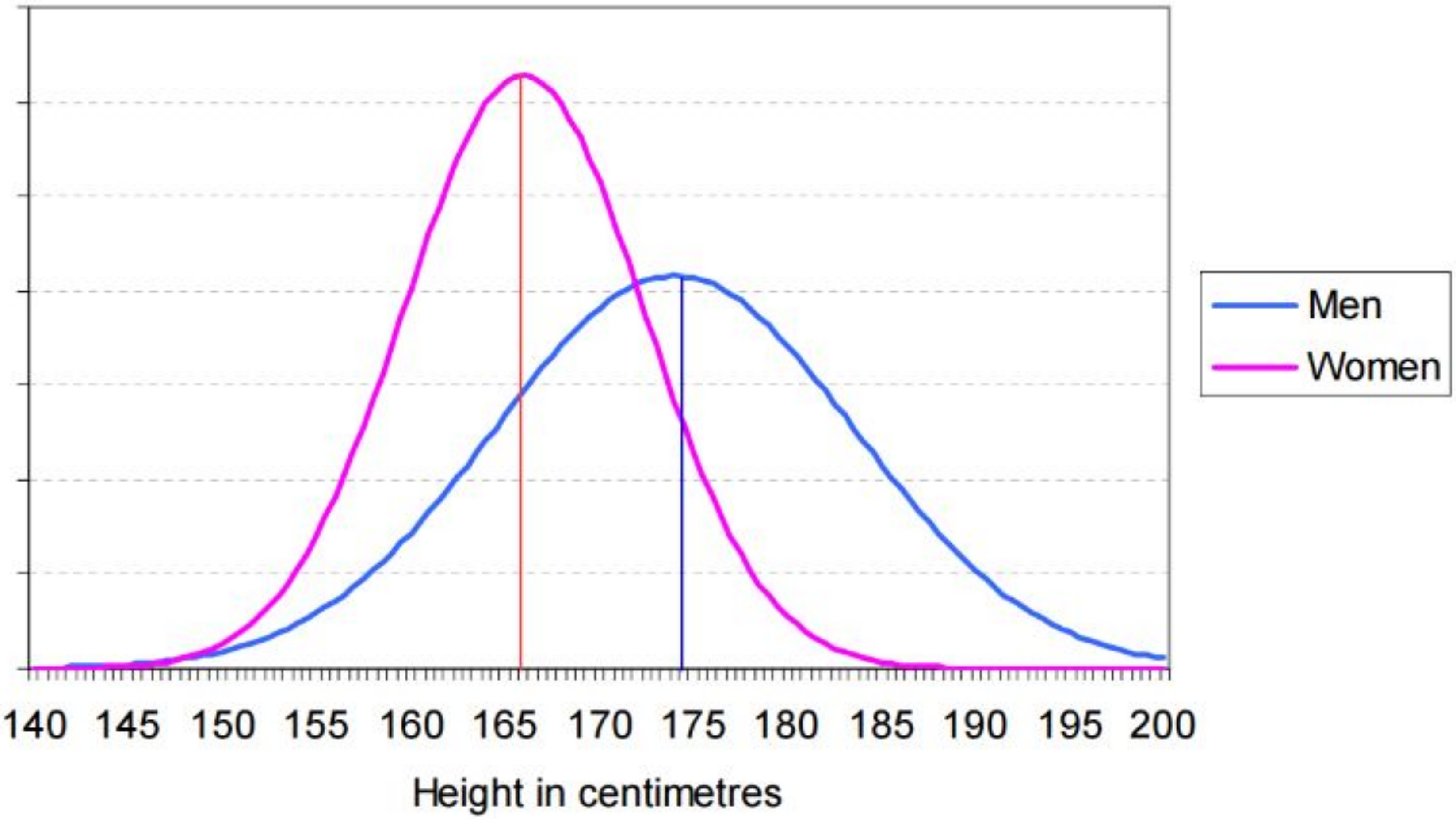
- Men's heights are Normally distributed with mean 174 cm and variance 92.16:

$$x_M \sim N(174, 92.16).$$

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

$$N(166, 40.32)$$

Men's and Women's Heights



Skoltech
Skolkovo Institute of Science and Technology

The Distribution of the Sample Mean

- If samples of size n are randomly drawn from a Normally distributed population of mean μ and variance σ^2 the sample mean is distributed as

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

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

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

- note the very small standard error: $\sqrt{40.32/50} = 0.897$

The Distributions of x and $\overline{x}$

- Note the distinction between

$$x \sim N(\mu, \sigma^2)$$

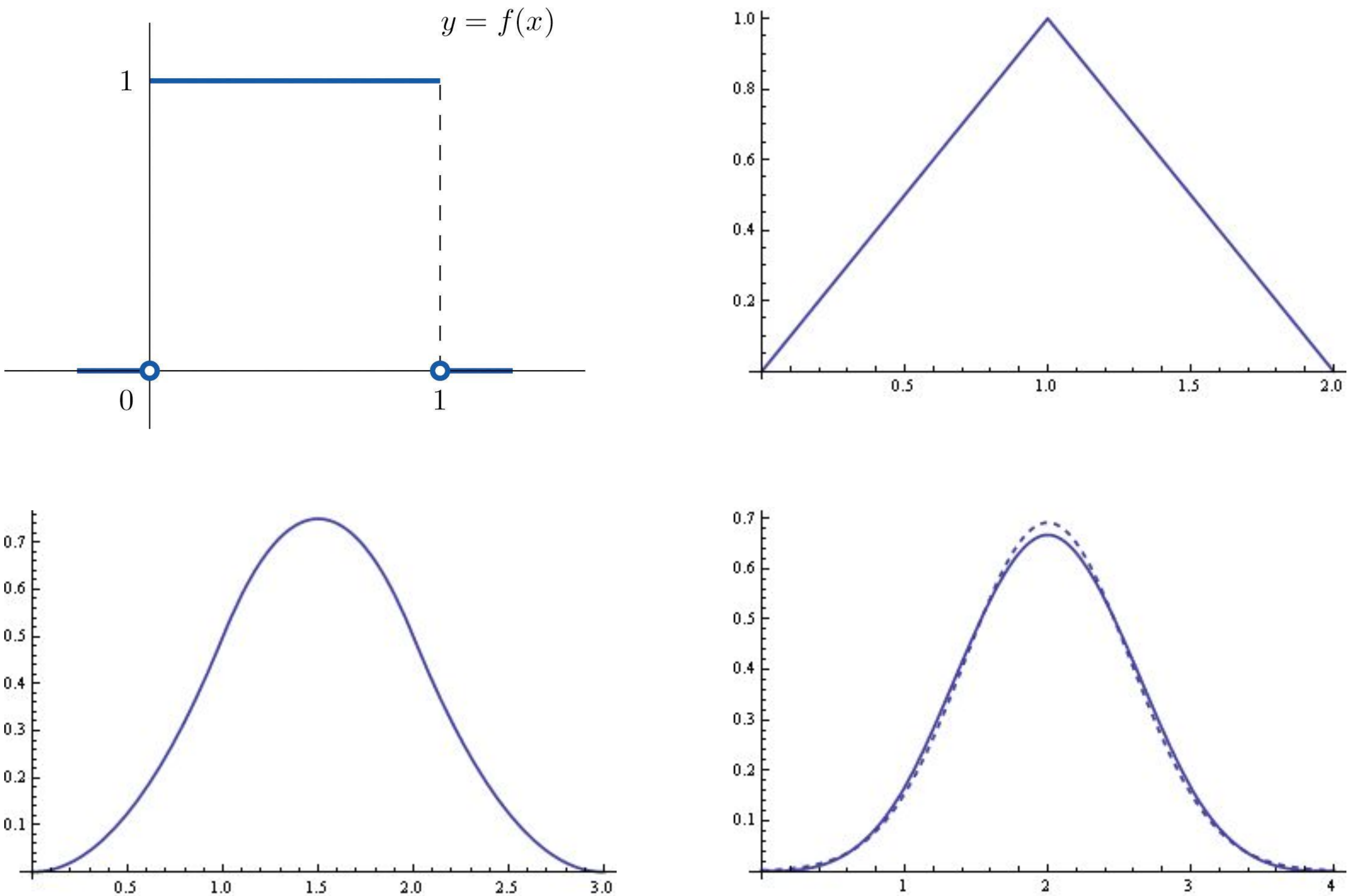
and

$$x \sim N(\mu, \sigma^2/n).$$

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

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

Sums of uniforms



The Central Limit Theorem

- If the sample size is large ($n > 25$) the population does not have to be Normally distributed, the sample mean is (approximately) Normal whatever the shape of the population distribution.
- The approximation gets better, the larger the sample size. 25 is a safe minimum to use.

Outline

- Introduction
- Probability
- **Statistical Estimation and Inference**
- Hypothesis Testing

Estimation

- Estimation is the process of using sample data to draw **inferences** about the population

Sample information $\bar{x}, s^2$ $\longrightarrow$ Population parameters μ, σ^2

Inferences

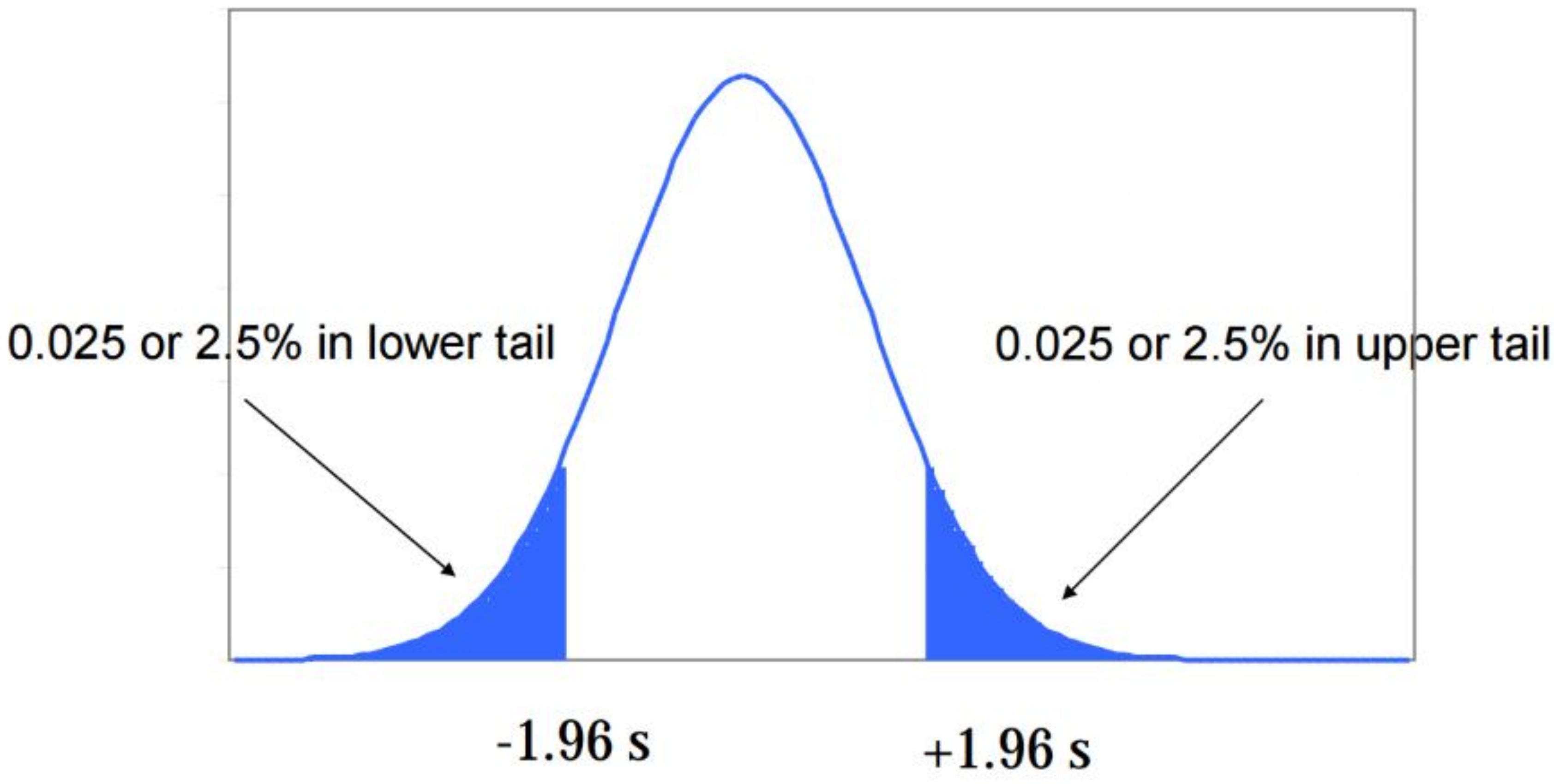
Point and Interval Estimates

- **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 ± ‘something’.
- What is the something?
- Go back to the distribution of $\bar{x}$.

Normal Distribution



The 95% confidence interval

- Recall the distribution of the sample mean

$$x \sim N(\mu, \sigma^2)$$

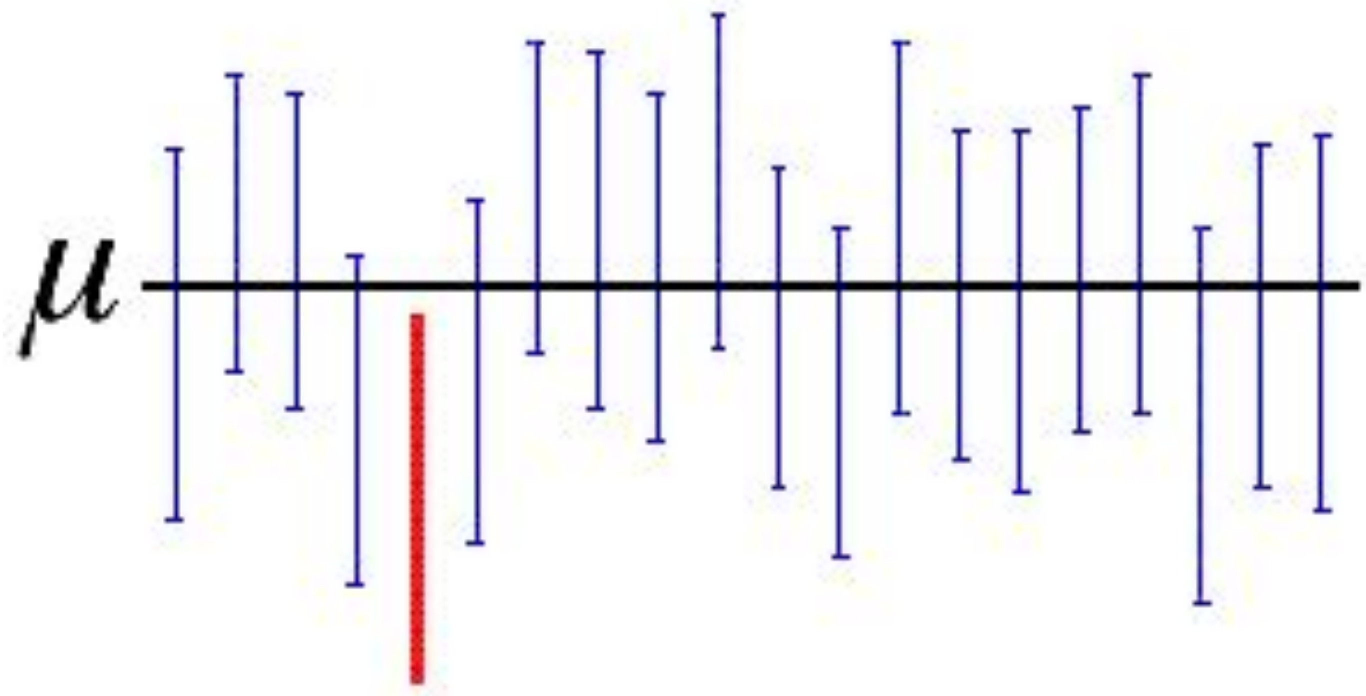
- Hence the 95% probability interval is

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

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

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

The 95% confidence interval



One sample out of 20 (5%) does not contain the true mean.

Example: Estimating Average Wealth

- Sample data:
  - $\bar{x} = 130$ (in thousands roubles),
  - $s^2 = 50000$,
  - $n = 100$.
- Estimate μ, the population mean.

Example: Estimating Average Wealth

- Point estimate: 130 (uses the sample mean)
- Interval estimate:

$$\bar{x} \pm 1.96\sqrt{s^2/n}$$

$$= 130 \pm 1.96 * \sqrt{50000/100}$$

$$= 130 \pm 43.8 = [86.2, 173.8]$$

- so we are 95% confident that the true mean lies somewhere 86,200 and 173,800 roubles.

Example: Estimating Average Wealth

- Sample data:
  - $\bar{x} = 130$ (in thousands roubles),
  - $s^2 = 50000$,
  - $n = 100$.
- Estimate μ, the population mean.

Outline

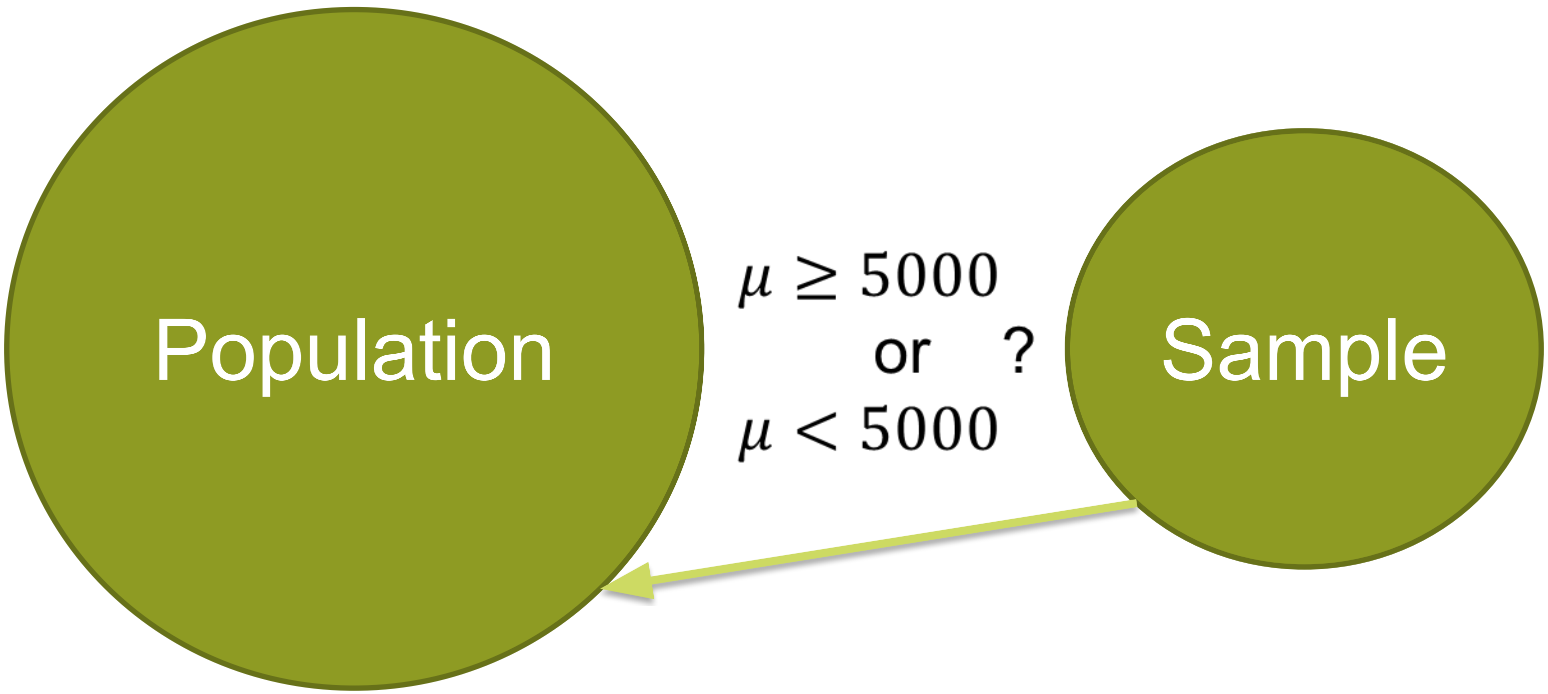
- Introduction
- Probability
- Statistical Estimation and Inference
- **Hypothesis Testing**

Hypothesis Testing

- Hypothesis testing is about making decisions.
- Is a hypothesis true or false?
- Are women paid less, on average, than men?

Probability vs Statistics

- Assume that $x \sim N(\mu, \sigma^2)$ 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 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

Type I and Type II Errors

	True situation	
Decision	H_0 true	H_0 false
Accept H_0	Correct decision	Type II error
Reject H_0	Type I error	Correct decision

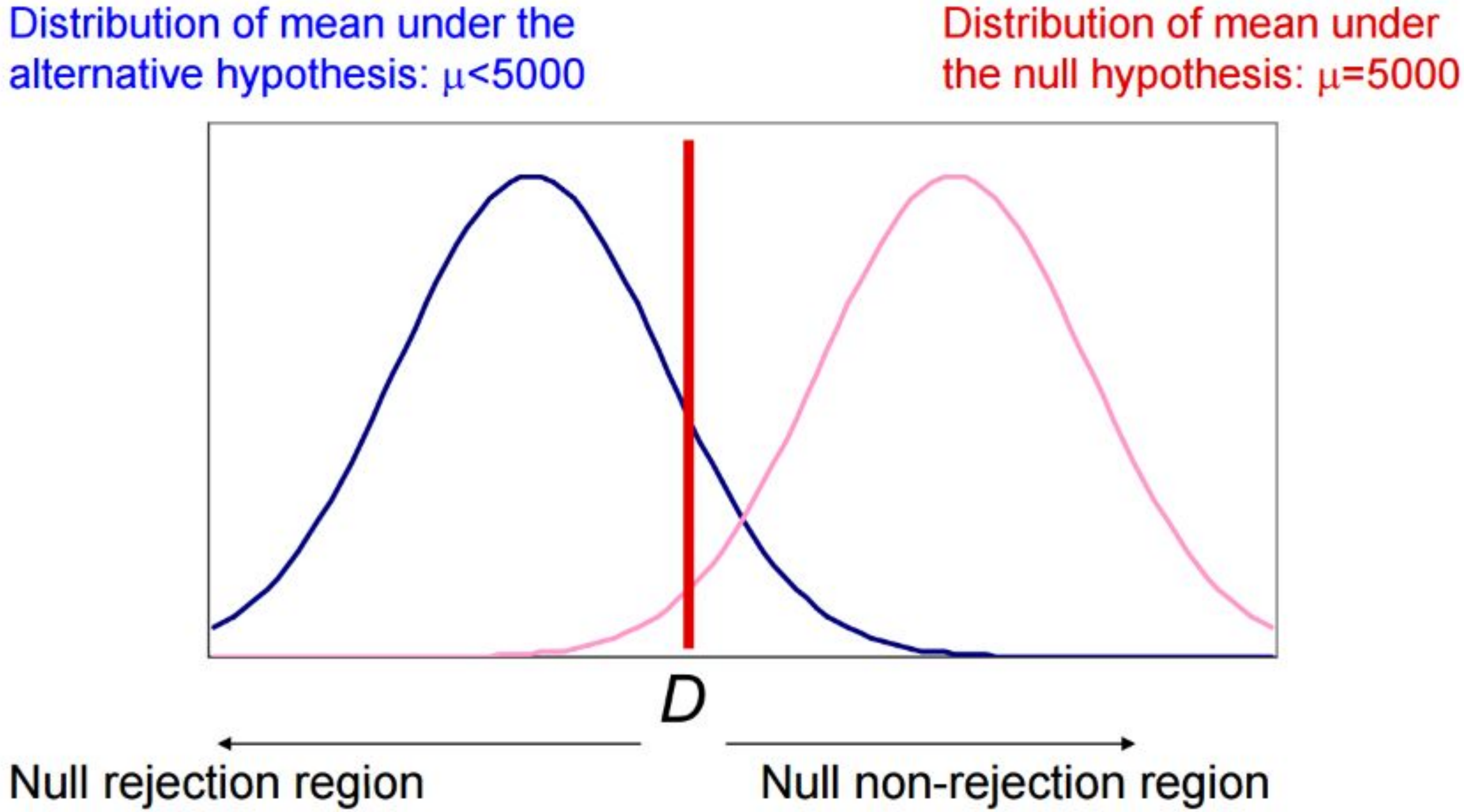
Avoiding Incorrect Decisions

- We wish to avoid both Type I and II errors.
- We can alter the decision rule to do this.
- Unfortunately, reducing the chance of making a Type I error generally means increasing the chance of a Type II error.
- Hence there is a trade off.

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?

Diagram of the Decision Rule



How to Make a Decision

- Where do we place the decision line?
- Set the Type I error probability to a particular value. By convention, this is 5%.
- There is therefore a 5% y probability that we are wrongly rejecting the null.
- This is known as the significance level (α) of the test.
- It is complementary to the confidence level ($1 - \alpha$) of estimation.

Should the Null Hypothesis be Rejected?

- Is 4,900 far enough below 5,000?
- Is it more than 1.64 standard errors below 5,000?
- Note, that this is one tailed test, so quantiles are different!
- 4,900 is 1.79 standard errors below 5,000 so falls into the rejection region (bottom 5% of the distribution).
- Hence, we can reject H at the 5% significance level or, equivalently, with 95% confidence.
- If the true mean were 5 000, here is less than a 5%

Multiple comparisons

- **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.

Why Multiple Testing Matters

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

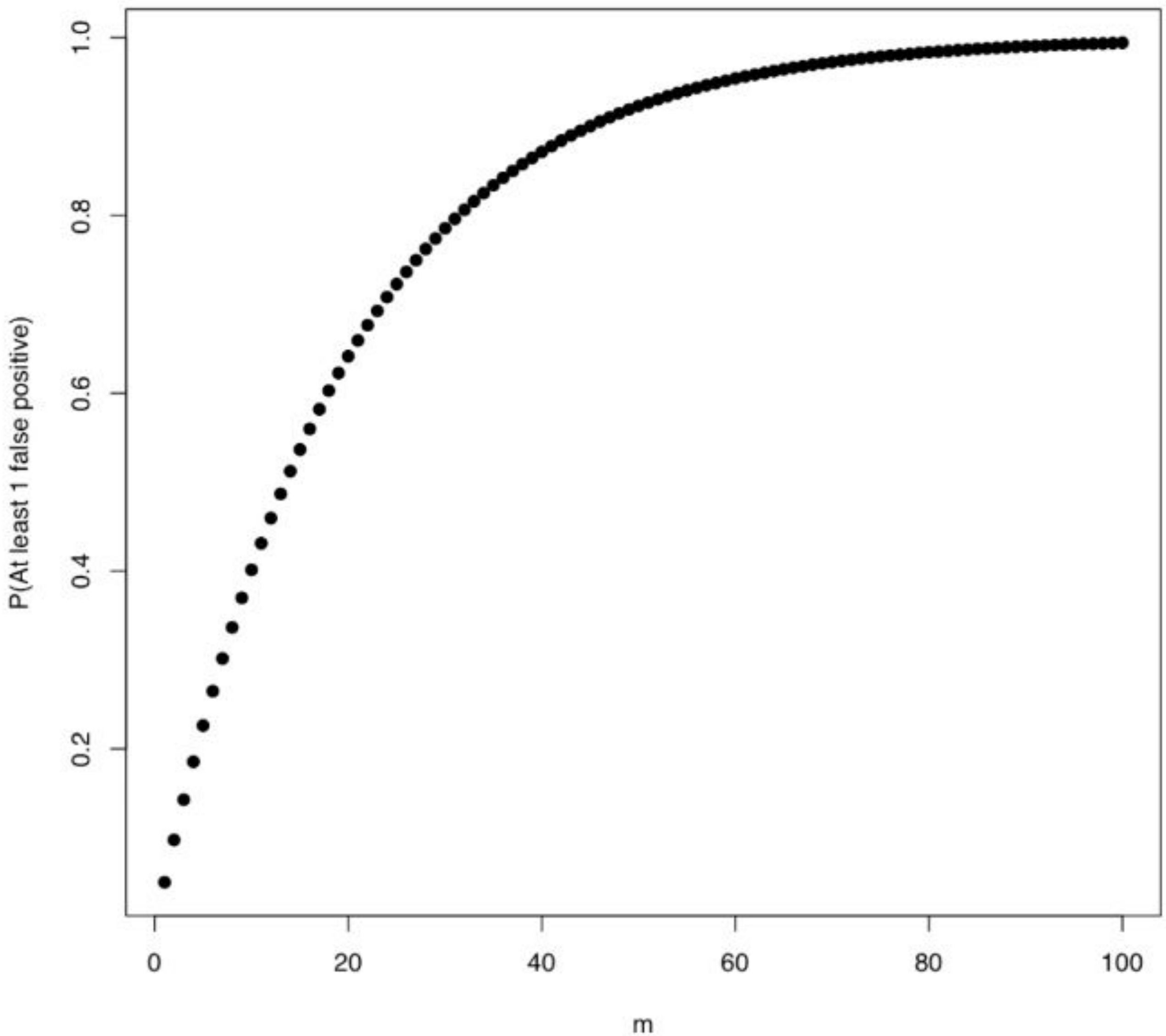
$$P(Making\ an\ error) = \alpha$$

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

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

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

Probability of At Least 1 False Positive



Skoltech
Skolkovo Institute of Science and Technology

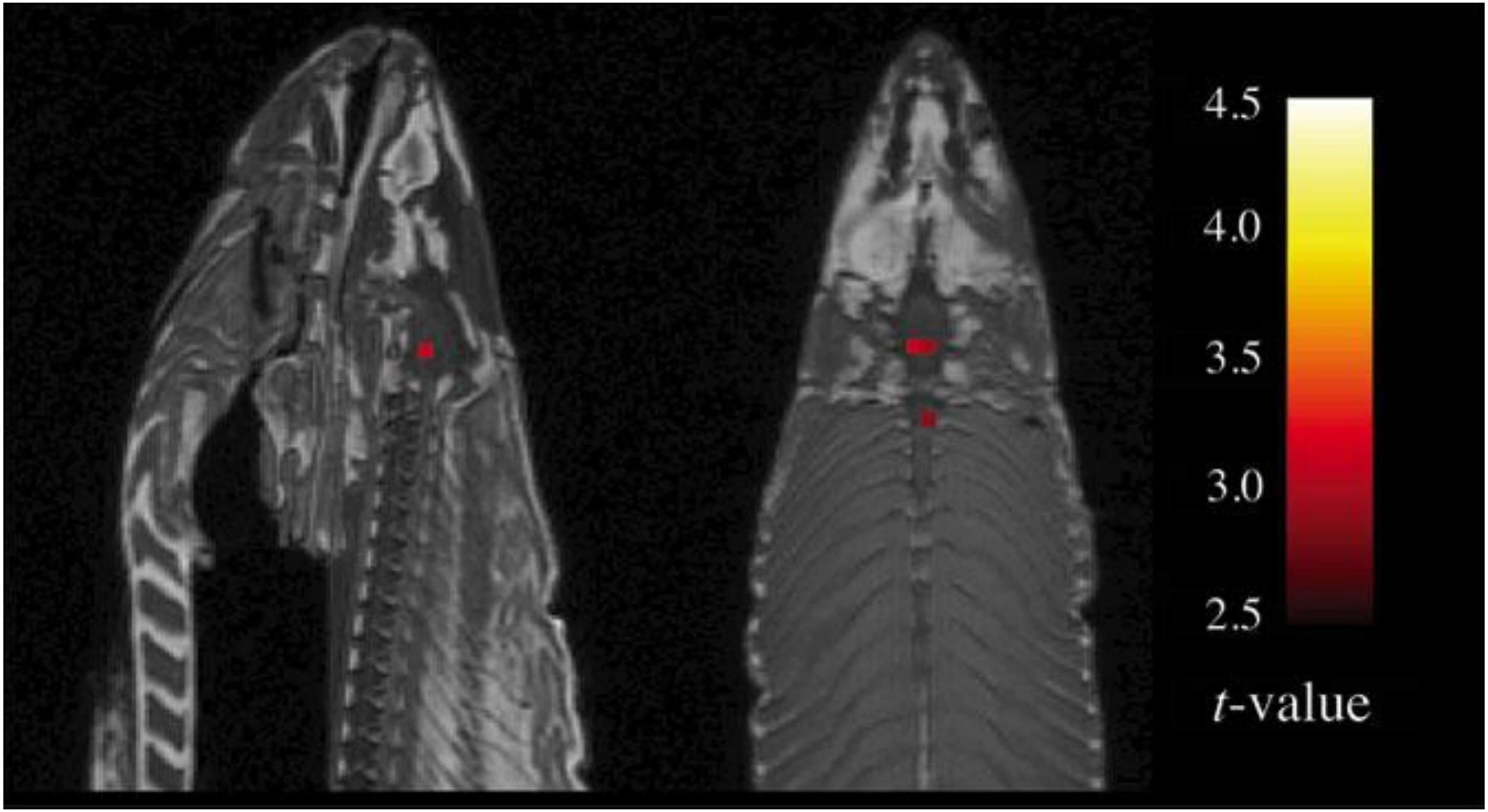
Probability of At Least 1 False Positive

- When people say “adjusting for the number of hypothesis tests performed” what they mean is controlling the Type I error rate.
- Very active area of statistics - many different methods have been described.

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



Thank you for your attention!

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

Skoltech
Skolkovo Institute of Science and Technology