Group of
Horribly
Optimistic
STatisticians

Statistics

Intro to Data Science

Maksymilian Norkiewicz & Jędrzej Ogrodowski

Agenda

1. Why do we need statistics?
2. Descriptive statistics
3. Population vs Sample
4. Theory of estimation
5. Hypothesis testing
6. Basic statistical testing methods

Why do we need statistics for Data Science?

Understanding basics of statistical inference, distributions, hypothesis testing is necessary for consciously developing models and describing the phenomena being studied.

This knowledge will come handy when we need to debug our code - bugs in Data Science are hard to see!

Example

We want to build a model predicting production outcome in **3 different chambers**.

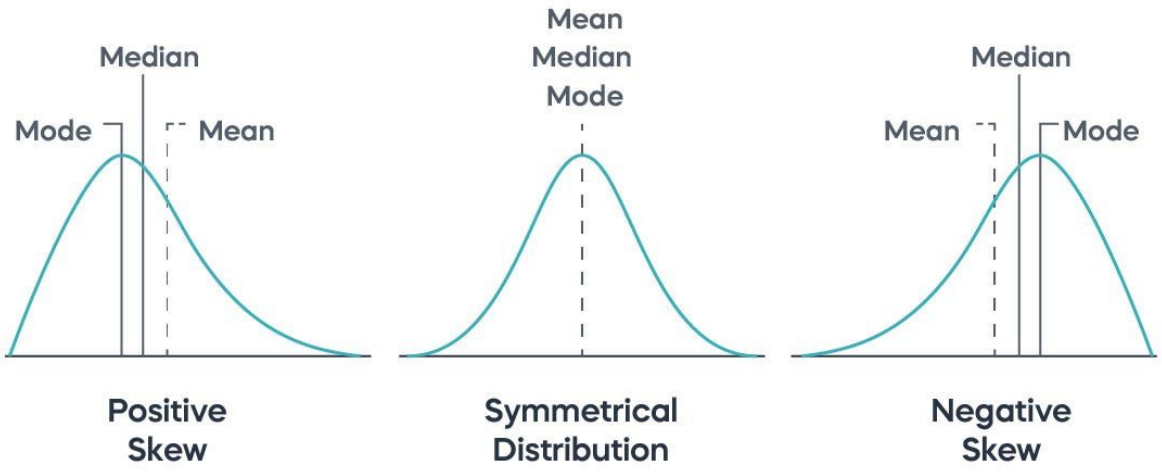
Questions we could ask:

- Are all chambers producing from the *same distribution*?
- Would differences affect our predictions?
- Is our sample **representative** of real production?
- How do we **formulate hypotheses** to test this?

Descriptive statistics - how to describe our data?

- With descriptive statistics we can summarize and understand data. We can think of this as of three categories:
  - Where data **centres** (Measure of Central Tendency)
  - How **spread out** the data is (Measure of Variability)
  - How the data is **distributed** (Measure of distribution)
- Basic statistics:
  - Measure of Central Tendency: mean, median, mode, quantile
  - Measure of Variability: standard deviation, iqr, variance, range
  - Measure of Frequency (Distribution): frequency / count tables, histograms

Example on central tendency



https://medium.com/@nitesh.py/understanding-measures-of-central-tendency-mean-median-and-mode-cabb73175b29

Quantiles

- Quantiles divide dataset into equal size, ordered parts.
- Quartile is special case of quantile that divides data in 4 segments with cut-points:
  Q0 - minimum value
  Q1 - 25% <= Q1 and 75% >= Q1
  Q2 - 50% <= Q2 and 50% >= Q2 (median value)
  Q3 - 75 <= Q3 and 25% >= Q3
  Q4 - maximum value
- Percentile is special case of quantile that divides data in 100 equal segments.

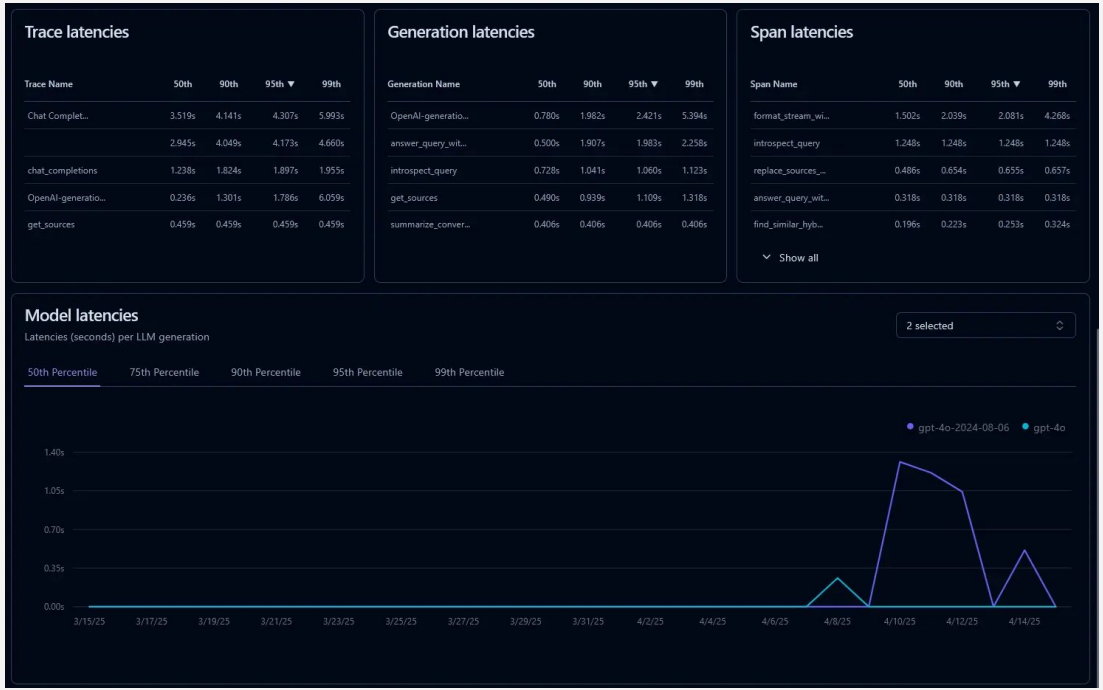
How quantiles could be used?

- When measuring whether an agent system responds fast enough, we should focus more on distribution based latency, not just average values.

  We could consider 50th percentile (or Q2) as typical response time

  To filter out occasional latency spikes, we can use 95th percentile or 99th percentile (standard in LangFuse, OpenAI dashboards, Azure AI Foundry)

How quantiles could be used?



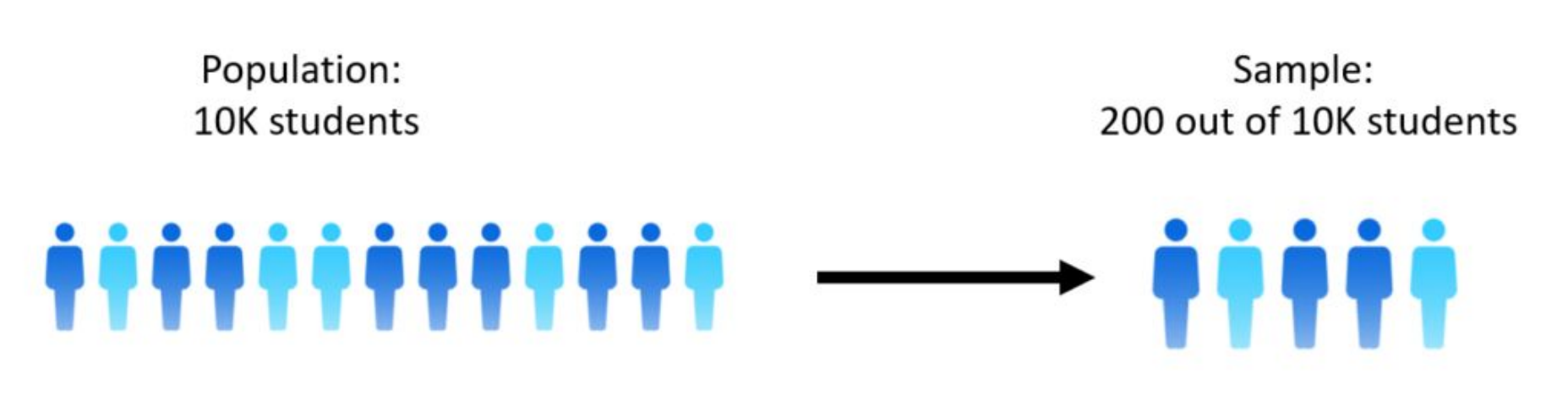
Population vs Sample

- We want to check house pricing dependent on location, neighbourhood etc. Does it mean that we have to include **every** house to make our work reliable?

  Answer is **no**.

  When performing research we work with some examples, that we believe, are representative for the whole case.

Population vs Sample



Population vs Sample

- Population:
  - collection of **all** elements that are subject of study (object population)
  - set of **all** possible observable values of characteristics describing the studied phenomenon (event population)
- Sample:
  - A **subset** of population available for scientist and forming the basis of the study.

Overall, preparing good dataset is challenging but crucial for good performance later.

Sampling factors

- Availability
- Application
- Bias
- Cost
- Representation

What's make a good sample?

- Covers up all studied parameters of population
- Unbiased - no skew to some specific subgroup
- Appropriately large - to capture variability
- Balanced - ideally each group should be similarly represented

- Example: when testing hearing aid production phases we should make tests for distortions etc. How to prepare representative sample when we know that some devices might have different characteristics, flaws? Is our sample suitable for testing procedure?

Mistakes

- One common trap related to sampling is lack of monitoring parameters of sampled data. E.g. we want to build model for predicting salaries based on some parameters but when sampling - majority of chosen samples are people from Warsaw. We introduced bias, Warsaw is not representative of all Poland.

Mistakes


Gdzie w Polsce zarabia się najwięcej?
Średnia płaca brutto miastach i województwach, w złotych
10200,97
Gdańsk
POMORSKIE
8452,93
WARMIŃSKO-MAZURSKIE
7094,23
Olsztyn
8742,48
ZACHODNIOPOMORSKIE
7542,02
Szczecin
8376,10
7548,67
Bydgoszcz
7661,30
Toruń
PODLASKIE
6820,03
Białystok
6635,83
KUJAWSKO-POMORSKIE
7091,29
MAZOWIECKIE
9568,24
Gorzów Wielkopolski
6641,63
8590,69
Poznań
LUBUSKIE
7253,95
Zielona Góra
7483,49
WIELKOPOLSKIE
7628,21
Warszawa
10044,91
7681,50
Łódź
LUBELSKIE
6861,25
ŁÓDZKIE
7492,73
DOLNOŚLĄSKIE
8471,57
Wrocław
9063,25
Lublin
7517,27
6788,39
Kielce
OPOLSKIE
7830,29
ŚLĄSKIE
8344,29
ŚWIĘTOKRZYSKIE
6856,01
Opole
7979,73
Katowice
8945,58
8287,05
Rzeszów
10652,31
Kraków
PODKARPACKIE
7026,00
MAŁOPOLSKIE
8717,73
300
GOSPODARKA
Źródło: GUS, dane za kwiecień 2024

Selected sampling methods

- Convenience sampling
- Snowball sampling
- Random sampling
- **Stratified sampling**
- **Importance sampling**

Stratified sampling

A common use of sampling is when we split data into training and test sets. Random sampling is simple, but it does not guarantee that the selected samples reflect the true proportions in the dataset. To avoid imbalanced or unrepresentative splits, we use a stratified approach.

- We divide the dataset into groups called strata
  (e.g., by class labels, categories, or other key features)
- Then we perform sampling within each stratum, preserving its original proportion

Stratified sampling example

- Commonly used function to split dataset use such approach (if we define parameter)

```
val_df, test_df = train_test_split(
    temp_df,
    test_size=0.33,
    stratify=temp_df['labels'],
    random_state=42
)
```

Importance sampling

- Allows us to sample from one distribution when we have access only to another distribution.
- Imagine you have to sample x from distribution P(x) but P(x) is really expensive. So we choose cheaper distribution Q(X) and then we sample from it and weigh sample by P(X) / Q(X). Q(x) can be any distribution as long as Q(x) > 0 whenever P(x) != 0.

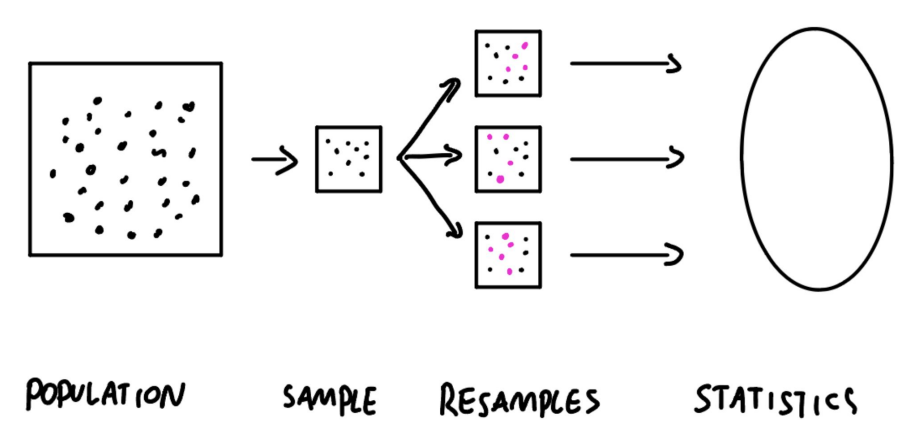
$$E_{P(x)} = \sum_x P(x)x = \sum_x xQ(x) \cdot \frac{P(x)}{Q(x)} = E_{Q(x)} \frac{P(x)}{Q(x)} x$$

Resampling

- We use resampling to create new samples based on existing ones. Two most popular methods are **Bootstrapping** and **Cross-Validation.**
- Bootstrapping is used to quantify uncertainty associated with given estimate (e.g. estimation of mean)
- More on **Cross-Validation** on later meetings.

Bootstrapping

- Bootstrapping estimate the uncertainty of a statistic when the underlying population distribution is unknown. Core idea is to generate many new samples by sampling with replacement from original dataset.



Bootstrapping

- Step 1: From your original dataset of size n, randomly select n observations with replacement. This means some data points will be chosen multiple times, and others not at all.
- Step 2: Compute the statistic of interest for this new bootstrap sample.
- Step 3: Repeat steps 1 and 2 a large number of times (e.g. 1,000 or 10,000). You now have a distribution of your statistic (the bootstrap distribution).
- Step 4: Use this bootstrap distribution to calculate standard errors, confidence intervals, or bias.

This idea is used in concept called Bagging.

Estimation

- We want to estimate population parameters based on available observations. So in other words, try to mirror sample parameters on whole population.
- For example, it is desired to estimate the proportion of a population of voters who will vote for a particular candidate. That proportion is the parameter; the estimate is based on a small random sample of voters. Alternatively, it is desired to estimate the probability of a voter voting for a particular candidate, based on some demographic features, such as age.

Terminology

- Parameter vs Statistic: parameter is value **describing** population, statistic is value **calculated** from sample
- Estimator: **rule** used to calculate an estimate from sample
- Estimate: **actual value** calculated by estimator

Hypothesis testing

- Hypothesis test provides statistical framework for answering yes or no questions about the data.
- Steps of test:
  1. Define null and alternative hypotheses
  2. We construct a test statistics
  3. We compute p-value
  4. Decision whether to reject the null hypothesis

Hypothesis testing

Null hypothesis (H0) - assumes that there is no difference or relationship between the variables

Alternative hypothesis (H1) - assumes that there is a real difference or relationship between the variables.

Significance level (α) - this is the threshold set by the researcher before analysis, usually at 0.05 (5%). It means that if the probability of the result happening by chance is less than 5%, we reject the null hypothesis and consider the result statistically significant.

What is p-value?

When we compute a test statistic, we want to know:
How strong is the evidence against the H0?

The p-value helps us answer this.
It is defined as:

- The probability of observing a test statistic as extreme or more extreme than the one we got, assuming H0 is true.

Small p-value means that the observed result would be unlikely under H0 - giving us evidence against the null hypothesis.

Type I and Type II errors

- Type I error: we reject H0 when H0 is true
- Type II error: we do not reject H0 when Ho is false
- Power of hypothesis: probability of correctly rejecting H0

	Null Hypothesis is TRUE	Null Hypothesis is FALSE
Reject null hypothesis	Type I Error (False positive)	Correct Outcome! (True positive)
Fail to reject null hypothesis	Correct Outcome! (True negative)	Type II Error (False negative)

Type I and Type II errors

- Type I error:
  A test says a patient **has a disease** but the patient is actually **healthy**.

- Type II error:
  A test says a patient is **healthy** but they **actually have the disease**.

Basic statistical tests

- When we perform analysis, we aim to **justify our intuition with evidence**.
- Data often contains **uncertainty**, so we rely on **mathematical methods** to understand whether observed changes are meaningful or just noise.
- After updating a website layout, average sales may go up — but **is the difference real or random?**
- Does it **actually matter** for the business?

Basic statistical tests

A typical analytical task is assessing whether two groups differ in a meaningful way.
Key comparisons include:

- Differences in means
- Differences in distributions
- Differences in proportions

These methods help determine whether the observed impact is **statistically significant** or just random.

Basic statistical tests

- Start: I want to compare groups!
  - Are you comparing MEANS or AVERAGES?
    - Comparing TWO groups?
      - Data is Normal & N is large? -> Z-test
      - Data is Normal? -> T-test (Independent or Paired)
      - Data is NOT Normal? -> Mann-Whitney U / Wilcoxon Rank-Sum (Independent) or Wilcoxon Signed-Rank (Paired)
    - Comparing THREE OR MORE groups?
      - Data is Normal? -> ANOVA
      - Data is NOT Normal? -> Kruskal-Wallis Test
  - Are you comparing PROPORTIONS or CATEGORIES?
    - Two categorical variables? -> Chi-Square Test
    - Two proportions? -> Z-test for Proportions
  - Are you measuring the RELATIONSHIP between two variables?
    - Both variables are Continuous? -> Correlation Test
    - One ordinal, one continuous? -> Spearman's Correlation

Statistical significance vs business value

Even if we detect a statistically significant difference, we still need to ask:
Is it actually worth the effort?

- **Is it real?**
  Use p-values and confidence intervals to confirm the effect isn’t due to chance.
- **How big is it?**
  Look at the actual difference and effect size to understand practical impact (thats where domain expertise kicks in).
- **Should we care?**
  Evaluate the business context.

Example

- Question: Is there a difference in the expected blood pressure of laboratory mice in control group and in treatment group? (20 samples are used)
- STEP 1:
- STEP 2:
- STEP 3:
- STEP 4:

Example

- Question: Is there a difference in the expected blood pressure of laboratory mice in control group and in treatment group? (20 samples are used)
- STEP 1: H0 = there is no difference [...] H1 = ~H0
- STEP 2:
- STEP 3:
- STEP 4:

Example

- Question: Is there a difference in the expected blood pressure of laboratory mice in control group and in treatment group? (20 samples are used)
- STEP 1: H0 = there is no difference [...] H1 = ~H0
- STEP 2: T-test
- STEP 3:
- STEP 4:

Example

- Question: Is there a difference in the expected blood pressure of laboratory mice in control group and in treatment group? (20 samples are used)
- STEP 1: H0 = there is no difference [...] H1 = ~H0
- STEP 2: T-test
- STEP 3: pvalue = 0.02
- STEP 4:

Example

- Question: Is there a difference in the expected blood pressure of laboratory mice in control group and in treatment group? (20 samples are used)
- STEP 1: H0 = there is no difference [...] H1 = ~H0
- STEP 2: T-test
- STEP 3: pvalue = 0.02
- STEP 4: pvalue = 0.02 < 0.05 = alpha - we reject H0

References

- Introduction to Statistical Learning, G. James
- Designing Machine Learning Systems, C. Huen