DATA SCIENCE RETREAT®
SINCE 2014

Introduction to Statistics

July 7 2023

Dr. Harald Stein
LinkedIn: https://www.linkedin.com/in/harald-stein-phd-1648b51a
ResearchGate: https://www.researchgate.net/profile/Harald-Stein

Agenda

1. Technical Prerequisites
2. Descriptive statistics
3. Combinatorics
4. Distributions
5. Conditional probability
6. Sampling
7. Hypothesis testing
8. Model estimation

09:30 – 10:00	**Technical readiness, Introduction**
10:05 – 11:30	**Descriptive statistics, Combinatorics**
11:35 – 13:00	**Distributions, Conditional probability**
13:00 – 13:30	Break
13:30 – 15:30	**Sampling, Hypothesis testing**
15:35 – 17:00	**Modell estimation**
17:05 – 17:30	**Buffer: Exercises & Discussion**

Let‘s start

1. Technical Prerequisites

means installation of Python environment with Jupyter

Install the environment:

- GitHub Link: https://github.com/MathforDataScience/DSR_Statistics1/
- List of packages: requirements.txt
- How to install: readme.txt
- Start Jupyter Notebook

2. Descriptive statistics

involves the organization, summarization, and visualization of data. It provides simple summaries about the sample and the measures.

- Types of Data
- Measures: central tendency, dispersion
- Measures: Skewness and kurtosis
- Quantiles
- Graphical representation of data
- Quiz

Types of Data

Definition and examples of Qualitative and Quantitative data

Type of Data	Subtype of data	Example
Quantitative / Numerical: • can be measured • or counted	**Discrete:** • can only take certain values	• number of students in a class
	Continuous: • can take any value within certain range	• Age, Height, Weight, Temperature, Income
	Interval	• Temperature • IQ scores • Years
	Ratio	• Probability
Qualitative / Categorical: • non-numerical • often relates to subjective qualities, characteristics, or descriptions	**Nominal:** • No order, hierarchy or sequence	• Colors • Types of cuisine • Genders • Blood types
	Ordinal: • can be arranged in a particular order	• Survey responses • Educational level • Military rank

... other types of data: binary, time-series, geodata, textual, multimedia, etc.

Measures of Central Tendency

calculate typical central points that describe or represent dataset

Arithmeric Mean
- Sum of all data points divided by number of data points
- E.g.: (1+2+3)/3 = 2

Median
- middle value in an ordered dataset
- E.g.: For {1, 2, 3}, the median is 2
- **Interesting because of outlier resistance!**

... Other measures of centrality:
- Geometric mean: e.g. population growth, investment return
- Harmonic mean: e.g. F-score (Machine Learning, binary classification, evaluation of results)

Measures of Dispersion

provide information about how data values are distributed around central tendency

Measure of Dispersion	Definition	Example
Range	The difference between the highest and lowest values in a dataset	For {1, 2, 3, 4}, the range is 4 - 1 = 3
Sum of Squares	$SS = \sum (x - \mu)^2$	For {1, 2, 3}, SS: (1+0+1) = 2
Variance (population)	$\sigma^2 = \frac{\sum (x-\mu)^2}{n}$	For {1, 2, 3}, pop. variance: (1+0+1)/3 = 0.67
Variance (sample)	$s^2 = \frac{\sum (x-\mu)^2}{n-1}$	For {1, 2, 3}, sample variance: (1+0+1)/2 = 1
Standard Deviation	The square root of the variance	The square root of 0.67 is approximately 0.82

Skewness and Kurtosis

Describe asymmetry of distributions diviation from normal distribution

Skew

- Measure of asymmetry of the probability distribution
- of a real-valued random variable about its mean
- Positive (negative): Long tail on the right (left)

Kurtosis

- defines how heavily the tails of a distribution differ from the tails of a normal distribution
- Outliers may facilitate positive kurtosis

Quantiles

statistical measures that divide a dataset into intervals of equal probability

Quantiles …

- are used to understand data distribution and identify outliers
- partition data into several subsets containing equal number of observations.

Several types of quantiles:

- Quartiles: divide the data into four equal parts.
- Deciles: These divide the data into ten equal parts.
- Percentiles: These divide the data into 100 equal parts.
- Median: separates data into two halves, with 50% of the data falling below and 50% above the median.

Graphical Representation of Data

Input:
- Categorical
- Nominal
- unordered

Output:
- Numerical
- Continuous
- Absolute

Input:
- Numerical
- continuous

Output:
- Numerical
- continuous

Input:
- Categorical
- Nominal
- ordered

Output:
- Numerical
- Continuous
- Absolute

Input:
- Categorical
- Nominal
- Unordered or ordered

Output:
- Numerical
- Continuous
- relative

Exercises

Exercise 1:
Question: Given the array numbers = np.array([1, 2, 3, 4, 5, 6, 7, 8, 9, 10]), compute the mean of the numbers.

Exercise 2:
Question: Given the array numbers from Exercise 1, compute the median of the numbers.

Exercise 3:
Question: Compute the variance of the numbers.
a) Use population variance b) Use sample variance
ages = [21, 23, 25, 27, 29, 31, 33]
b) Let's say we take a sample of 5 ages from the original group: sample_ages = [21, 23, 27, 31, 33]

Exercise 4:
Question: Given the array numbers from Exercise 1, compute the standard deviation of the numbers.

Exercise 5:
Question: Given the array numbers from Exercise 1, compute the first quartile (25th percentile), the second quartile (50th percentile, or median), and the third quartile (75th percentile) of the numbers.

Exercise 6:
show with python plots for bar chart, histogram, line diagram, pie diagram in a 2x2 subfigure matrix

3. Combinatorics

is concerned with the study of countable, discrete structures and includes the arrangement, combination, and permutation of sets.

- Permutation
- k-permutation
- Combination
- Pascal‘s triangle
- Quiz

Permutation

is an arrangement of objects in a specific order.

Table 5-1 Possible Rearrangements for Four People Sitting in a Straight Line

Rearrangement #	*Listing*	*Rearrangement #*	*Listing*
1	Tim, Syd, Elena, Mark	13	Elena, Tim, Syd, Mark
2	Tim, Syd, Mark, Elena	14	Elena, Tim, Mark, Syd
3	Tim, Elena, Syd, Mark	15	Elena, Syd, Tim, Mark
4	Tim, Elena, Mark, Syd	16	Elena, Syd, Mark, Tim
5	Tim, Mark, Syd, Elena	17	Elena, Mark, Tim, Syd
6	Tim, Mark, Elena, Syd	18	Elena, Mark, Syd, Tim
7	Syd, Tim, Elena, Mark	19	Mark, Tim, Syd, Elena
8	Syd, Tim, Mark, Elena	20	Mark, Tim, Elena, Syd
9	Syd, Elena, Tim, Mark	21	Mark, Syd, Elena, Tim
10	Syd, Elena, Mark, Tim	22	Mark, Syd, Tim, Elena
11	Syd, Mark, Tim, Elena	23	Mark, Elena, Tim, Syd
12	Syd, Mark, Elena, Tim	24	Mark, Elena, Syd, Tim

The number of permutations

- without repetition
- of a set of n objects is given by:

$$\boldsymbol{P(n) = n!}$$

For $n = 4$ there are

$$\boldsymbol{P(4) = 4! = 4 * 3 * 2 * 1 = 24}$$

possibilities.

Permutation

is an arrangement of objects in a specific order.

Figure 5-3: Possible rearrangements of four people sitting around a table discussing probability.

Example

- 4 people sit at a table
- Restriction: The position of one of them is fixed.
- How many possible arrangements are there?

$$P(3) = 3! = 3 * 2 * 1 = 6$$

K-permutation

is an ordered arrangement of k elements selected from that set. In German it is called “Variation”

The number of k-permutations

- without repetition
- of a set of n objects and k selections is given by

$$\boldsymbol{P(n,k) = \frac{n!}{(n-k)!}}$$

For $n = 4$ and $k = 2$ there are

$$\boldsymbol{P(4,2) = \frac{24}{2} = 12}$$

possibilities.

File:Kpermutation.png - Wikimedia Commons

Combination

is a selection of items from a larger set where the order of selection does not matter.

The number of combinations

- without repetition
- of a set of n objects and k selections is given by

$$\boldsymbol{P(n,k) = \frac{n!}{(n-k)!k!} = \binom{n}{k}}$$

In the example of German lottery there are

- 49 balls
- 6 selected balls

i.e. there are

$\boldsymbol{P(49,6) = \binom{49}{6} = 13.983.816}$ possibilities.

Pascal‘s triangle

is a triangular array of numbers in which each number is the sum of the two directly above it. It is used for binomial expansions, combinatorics, probability calculations, etc.

Binomial theorem

- Coefficients from binomial expansion:

$$(a+b)^0 = 1$$
$$(a+b)^1 = 1\mathrm{a} + 1\mathrm{b}$$
$$(a+b)^2 = 1a^2 + 2\mathrm{ab} + 1b^2$$
$$(a+b)^3 = 1a^2 + 3a^2\mathrm{b} + 3\mathrm{a}b^2 + 1b^2$$

- Sum of rows → 2^n

Coming in next chapter:

- $\binom{n}{k}$: crucial for explaining distributions like binomial distribution

Quiz

4. Distributions

describe the probabilities of occurrences of different possible outcomes in an experiment

- Overview of important distributions
- Way to… binomial distribution
- Poisson distribution
- Normal distribution, standard normal distribution
- Students‘ t-distribution
- χ^2- distribution
- F-Distribution
- Quiz

Overview of important distributions

Most common distributions are based on the binomial distribution.

Uniform distribution

is a probability distribution that describes the number of successes in a fixed number of independent binary experiments, each with the same probability of success.

Expected value:

- Arithmetic mean

Use Cases:

- Generating Random Numbers:
  e.g.: fair dice, fair coin tosses
- Quality Control and Manufacturing:
  e.g.: check the uniformity of products

Be careful:

- Sum of 2 dice is not uniformly distributed!
  (it is multinomial distribution, out of scope)

Bernoulli distribution

is a discrete probability distribution for a random variable which can take a binary outcome: one (success) or zero (failure); example: toss of coin (that is not necessarily fair)

Formula:

$$P(x;p) = p^x * (1-p)^{1-x} \; for \; x\epsilon[0;1]$$

p: probability of success. E.g. if the coin is fair, the probability of getting a head (considered a success) would be 0.5

x: outcome of Bernoulli trial; it can take only two values: 1(success) or 0(failure)

$P(x;p)$ **:** probability of outcome x, given probability p of success.

Binomial distribution

Example for probability calculation:
Basketball player that is known for 60% prob of making free shots makes 7 of 10

Step 1:

- P(Make the shot) = 60 %
- P(Miss the shot) = 40 %
  = 1 – 60 %

Binomial distribution

Example for probability calculation:
Basketball player that is known for 60% prob of making free shots makes 7 of 10; Step 2

Step 2:

- P(miss, miss) = 60 %
- P(miss, make) = 48 %
- P(make, make) = 36 %

Assumption:
Independence of shots

Binomial distribution

Example for probability calculation:
Basketball player that is known for 60% prob of making free shots makes 7 of 10; Step 3

Step 3:

- P(miss, miss, miss) = 8 %
- P(miss, miss, make) = 28.8 %
- P(miss, make, make) = 43.2 %
- P(make, make, make) = 21.6 %

Binomial distribution

Example for probability calculation:
Basketball player that is known for 60% prob of making free shots makes 7 of 10; Step 4

$$\boldsymbol{P(7\ of\ 10) = 120 \times 0.6^7 \times 0.4^3 \approx 21.5\%}$$

Binomial distribution

Example for probability calculation - Generalization:
Basketball player that is known for 60% prob of making free shots makes k of n

$$P(k \text{ of } n) = \binom{n}{k} 0.6^k 0.4^{n-k}$$

This example

Binomial distribution

is repeated Bernoulli distribution, i.e. probability distribution that describes number of successes in fixed number of independent binary experiments, each with same probability of success.

Formula:

$$P(k\ of\ n) = \binom{n}{k} p^k (1-p)^{n-k}$$

Parameters:

- n: number of trials
- p: probability of success on each trial
- k (input factor on x-axis): number of actual positive outcomes

Use cases:

- Quality Control: In a manufacturing company the probability of producing a defective item is 0.05, and it manufactures 200 items per day
- Spam email classifier

Poisson distribution

probability distribution of a given number of independent events occurring in a fixed interval of time or space, given fixed average rate of occurrence.

Events occur with a known constant mean rate and independently of the time since the last event.

Relation to binomial distribution:

1. The number of trials (n) is very large, i.e. continuous.
2. The probability of success (p) is small.
3. The average number of successes (np) is moderate.

Parameter:

λ = np (the mean of the distribution)

Use cases:

- Number of calls received in a given hour, sales per hour, etc.
- emails arriving in inbox per hour, given average arrival rate

AI use cases:

- Sport bets: goals/baskets per match
  (Prediction of football, basketball results)

Normal distribution

is a continuous probability distribution characterized by a symmetric bell-shaped curve, defined by its mean (average) and standard deviation, also known as the Gaussian distribution

Figure 9-1: Three normal distributions.

Parameters:

- Mean: where the peak of the bell occurs
- Standard deviation: determines the width of the bell

Use cases:

- Human Heights: most individuals have heights around the average, very tall or small people are much less frequent
- Test Scores: most students will score around average, fewer students have very high/low scores.

AI use case:

- Supervised learning: Prediction errors (differences between actual and predicted values) are often normally distributed, e.g. a house price predictor based area, number of rooms, etc.

Standard normal distribution

is a special case of the normal distribution with a mean of zero and standard deviation of one.

Figure 9-4: Transforming numbers on the normal distribution to numbers on the Z distribution.

z-scoring" or standardization

- subtracting the mean score
- and dividing by the standard deviation

Properties:

- mean (μ): 0
- standard deviation (σ): 1
- symmetric, bell-shaped

Why is it useful?

- This makes it easier to compare scores from different tests or datasets.

Students‘ t-distribution

is used when sample size is small or when population standard deviation is unknown.
It is similar to normal distribution, but with heavier tails, e.g. more flat.

Figure 9-1: *t*-distributions for different sample sizes

If $n < 30$ (number of samples), t-distribution gives more accurate confidence intervals and hypothesis tests compared to normal distribution.

Use case:

- new drug is tested on small (say, 15 patients). Then, population standard deviation is unknown, and sample that is used for estimation is small.

AI use case:

- Any models with very small sample, e.g. for healthcare predictions.

χ^2 distribution

is a probability distribution of the sum of squares of independent standard normal random variables.

It's the distribution of a sum of the squares of k independent standard normal random variables.

Parameter:

k: independent standard normal random variables, square each and sum them up

Use cases:

- χ^2-tests of independence: They use χ^2 -distribution.
  Is preference for having cat or dog is independent from city?
- Several other hypothesis tests (see below)

Use case in AI:

- Validation of model performance

F-distribution

is ratio of 2 independent χ^2-distributions divided by their respective degrees of freedom

Degrees of freedom (df):

- dfn (df of numerator):
  number of groups being compared minus 1
  <u>between group comparison</u>
- dfd (df of denominator):
  total number of observations minus the number of groups
  <u>within group comparison</u>

Use Cases

- is used in Analysis of Variance (ANOVA)
- regression analysis
- comparing variances of ≥ 3 samples

Quiz

5. Conditional probability

is the probability of an event occurring given that another event has already occurred

- Formula of Bayes
- Table with conditional probabilities
- Quiz

Bayes‘ theorem

describes how to update the probabilities of hypotheses when given evidence. It’s is used to calculate the conditional probability, i.e., the probability of an event based on prior knowledge.

From conditional probabilities to Bayesian theorem:

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

$$P(B|A) = \frac{P(B \text{ and } A)}{P(A)}$$

$$P(A \text{ and } B) = P(B \text{ and } A)$$

Thus: $P(A|B)P(B) = P(B|A)P(A)$

Or:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Table with conditional probabilities

Use Case: How many visitors make a purchase depending on whether they visit website A or B

	Purchase	No Purchase	Total (MP)
Website A	80 (A1) JP: 4 % **CP: 8 %**	920 (A2) JP: 46 % CP: 92 %	$Total_A$ **1000** $P(A)$ =50%
Website B	90 (B1) JP: 4.5 % CP: 9 %	910 (B2) JP: 45.5 % CP: 91 %	$Total_B$ **1000** $P(B)$ =50%
Total (MP)	Purchase: 170 (T1) P(Purchase) = 8.5%	Non-purchase: 1830 (T2) P(Non-Purchase) = → 91.5 %	$Total = Total_A + Total_B$ 2000 $P(Total) = 100\%$

$$\boldsymbol{P(Purchase\ given\ Website\ A) = \frac{P(Website\ A\ given\ Purchase) * P(Purchase)}{P(Website\ A)}}$$

$$\boldsymbol{0.08 = \frac{80/_{170} * 0.085}{0.5}}$$

Joint probabilities (JP):

- P(A and Purchase) = A1; T = 80 / 2000 = 0.04
- P(B and Purchase) = B1; T = 90 / 2000 = 0.045

Conditional probabilities (CP):

- P(Purchase | A) = A1; TA = 80 / **1000** = 0.08
  (probability of a purchase given the user saw website A)
- P(Purchase | B) = B1; TB = 90 / **1000** = 0.09
  (probability of a purchase given the user saw website B)

Marginal probabilities (MP):

- P(A) = TA; T = 1000 / 2000 = 0.5
  (probability a visitor sees website A)
- P(B) = TB; T = 1000 / 2000 = 0.5
  (probability a visitor sees website B)
- P(Purchase) = T1; T = 170 / 2000 = 0.085
  (probability a visitor makes a purchase)

Quiz

6. Sampling

is the process of selecting a subset of individuals from a statistical population to estimate characteristics of the whole population

- Central Limit Theorem (CLT)
- Confidence intervals
- Resampling
- Bootstrapping
- Quiz

Central limit theorem

states that for large number of iid random variables, the distribution of their sum (or average) tends toward a normal distribution, irrespective of shape of original population distribution

Definition:

- For population with mean μ and standard deviation σ
- take sufficiently large random samples from the population
- with replacement (independent)
- Independent and identically distributed (i.i.d) random variables
- Sample size is large enough (typically n > 30)

→ distribution of sample means is appr. normally distributed, regardless of population distribution

Why is it important?

- This principle enables us to make statistical inferences about populations based on sample data
- foundation for hypothesis testing, confidence intervals, etc.

Confidence intervals

are a range of values estimated to contain the value of unknown population parameter with certain degree of confidence

Properties:

- Central Limit Theorem allows to use normal distribution for confidence interval estimation when the sample size is large, even if the original population is not normally distributed.
- Confidence intervals are widely used tool in hypothesis testing, regression analysis, etc.

Usual distributions:

- Standard normal distribution
- t-distribution

Confidence intervals

… need careful interpretation. Assumption: $\alpha = 5\%$

Assumption:
population parameter μ (average of population) is fixed

Wrong interpretation:
- With a chance of 95% the confidence interval contains μ

Correct interpretation:
- If sampling process is repeated many times or infinitely
- confidence intervals are generated each time
- 95% of these confidence intervals contain μ

Resampling

Is set of methods of statistical inference that involve drawing repeated samples from original data samples

What is it good for?	• Versatile method to make statistical inferences about a dataset • It helps to validate models by using random subsets (bootstrap) or sequences (cross-validation) dataset • It extends possibilities for hypothesis testing and estimation
Example: Bootstrap	• Many resamples (with replacement) of the observed dataset are generated • Arbitrary distribution is transformed into normal distribution • convenient for estimation and hypothesis testing.
Example: Cross-validation	• a form of resampling used to prevent overfitting in machine learning models • dataset is divided into 'k' subsets; used for training the model • respective residual data is used as test set

Bootstrapping

is a resampling method that involves drawing repeated samples from the original data set, with replacement. Each sample drawn is the same size as the original dataset.

Procedure:

- Draw sample from the original dataset with replacement
- Often of same size like original dataset
- Compute desired statistic from this resample
- Repeat the process multiple times (often thousands or tens of thousands of times), each time storing the statistic.
- Use distribution of these bootstrapped statistics to make inferences about the population.

Why does it make sense?

- It provides method to estimate variability or confidence intervals of statistic when underlying distribution is unknown or complex.
- It's a simple and powerful method requiring very few assumptions

AI Use case:

- Random forests (like XGBoost):
  Classification or numeric prediction
  based on many resampled pruned decision trees

Quiz

7. Hypothesis testing

describe the probabilities of occurrences of different possible outcomes in an experiment

- Hypothesis Testing
- One-sided vs. Two-sided tests
- p-Value
- Errors: False positives and false negatives
- A/B Testing
- Examples: t-, χ^2-, ANOVA- tests
- Quiz

Hypothesis testing

is a statistical method used to assess the evidence from sample data and make conclusions about a population by comparing competing hypotheses.

It involves formulating two competing hypotheses

- the null hypothesis (H_0): initial claim, status quo, which usually means: No effect/difference
- and the alternative hypothesis (H_1); we want evidence for the alternative hypothesis
- We assess evidence from the data, in order to determine whether to reject or fail to reject the null hypothesis.

Conditional probability

Hypothesis testing

involves specifying significance level, analyzing sample data using statistical test rejecting or not rejecting null hypothesis

Workflow:

1. **State Hypotheses:**
   Null Hypothesis (H_0)
   and Alternative Hypothesis (H_1)

2. **Specify critical value α:**
   usually 0.05 for 2-sided tests

3. **Calculate test statistic t with:**

- t-test
- χ^2-test
- z-test
- ANOVA-test, and many more

4. **Interpret results comparing t and α :**

- Reject null hypothesis (support H_1) or not (support H_0)?

One-sided vs. Two-sided tests

One-sided hypothesis test checks if parameter is greater/less than a certain value.
Two-sided test checks if parameter is not equal to a certain value, regardless of direction.

Two-sided (or Two-tailed) Tests

- … assess if a parameter is either greater or less than the null hypothesis value.
- **Example:**
  Company is testing if a website redesign affects,
  i.e. either increases or decreases user time on site.
  - H_0: Website redesign has no effect on user time on site.
  - H_1: Website redesign affects user time on site.

One-sided (or One-tailed) Tests

- … check if parameter is <u>either</u> greater or less than null hypothesis value, but <u>not both</u>.
- **Example:**
  Pharmaceutical company is testing a new drug and wants to know if it reduces disease recovery time (not interested if it increases).
- H_0: Drug does not reduce recovery time.
- H_1: Drug reduces recovery time.

p-value

is a tool for deciding whether to reject or not reject the null hypothesis

Example for rejecting H_0, the 2 groups are different

Interpreting the P-value

- Small p-value (typically ≤ $\alpha = 0.05$)
  suggests strong evidence against the null hypothesis,
  "we reject the null", "we support H_1"
- Large p-value $\alpha = 0.05$)
  suggests weak evidence against the null hypothesis,
  "we fail to reject the null", "we support H_0"
- If we reject the null, there is typically significant difference between the 2 groups

Properties

- p-values can be affected by the size of the sample.
- If you reject H_0 and perform the same experiment millions of times, risk H_0 actually being true ("false positives") is p
- Figure: in 0.72 of 100 times you should have selected H_0.

Errors: False positives and false negatives

matter, as they represent incorrect conclusions that can impact decisions based on test results

- Selecting H_1 is the "positive" statement
- Selecting H_0 is the "negative" statement

Type I error means "false positive":
- When H_0 is <u>true</u> and therefore "selecting H_1 is incorrect"
- precise formulation: H_0 is incorrectly rejected

Type II error means false negative:
- When the H_0 is <u>false</u>, but "we select it"
- precise formulation: we fail to reject it

Connection to significance level α:
- The maximal probability of making Type I error is defined by α, such as 0.05 or 0.01.

Connection to p-value:
- p-value is the probability of false positives

What is A/B-Testing?

also known as split testing, is method of comparing two versions of webpage, email, or other user experience to determine which one performs better.

https://medium.com/towards-data-science/data-science-you-need-to-know-a-b-testing-f2f12aff619a

- **Identify goal:**
  Your goal might be to increase website conversions, improve email click-through rates, or boost the number of sign-ups.
- **Create Variants:**
  Develop two versions of element you want to test - the control (A) and the variant (B). The versions should be identical except for one change.
- **Split audience:**
  Randomly divide audience into two equal groups.
  One sees version A, the other sees B.
- **Conduct test:**
  Release both versions at the same time and collect data on how each version performs in relation to your goal.
- **Analyze the Results:**
  Use statistical analysis to determine which version performed better.
- **Implement Changes:**
  If B performs significantly better, consider replacing version A with B.

A/B Testing: Choosing the Right Test Type

Tests types are nuanced, slight differences in hypothesis determine selection

Test Type	Scipy method	Use Case, compare:	Examples
1. Two sample t	ttest_ind	• means of • continuous data between • two groups of independent samples	• Company tests new website B and compares it with old one A • Visitors are divided into 2 groups that only see one of them • Visit durations are measured **→ Does the subgroup stay longer on new website B?**
2. Paired t	ttest_rel	• means of • continuous data for the • same group at different times	• Like 1 • But: The same group is tested • Visit durations are measured, first with website A later with B **→ Does the same group stay longer on new website B?**
3. χ^2 for independence	chi2_contingency	• categorical data between • two or more groups	• A company has employees in 4 experience levels • On each level a part of them received promotion (yes/no) **→ Does experience level have impact on promotion?**
4. One factor ANOVA	f_oneway	• means of • continuous data between • ≥ three groups of independent samples	• Education program in company, • 3 learning styles: self-paced online courses, instructor led, mixed • Employees are divided in 3 subgroups **→ Does the style have impact on learning success?**

Two sample t-tests

are statistical tests used to compare the means of continuous data between two groups and determine if they are significantly different from each other.

Variant A:

```
group_A_scores = np.random.normal(75, 10, 30)  # mean=75, std_dev=10, n=30
group_B_scores = np.random.normal(80, 10, 30)  # mean=80, std_dev=10, n=30
```

Variant B:

```
group_A_scores = np.random.normal(75, 10, 30)  # mean=75, std_dev=10, n=30
group_B_scores = np.random.normal(90, 10, 30)  # mean=80, std_dev=10, n=30
```

The formula for the t-statistic in a two-sample t-test is:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

With:

$\bar{X}_1$ and $\bar{X}_2$ are the sample means of the two groups

s_1^2 and s_2^2 are the sample variances of the two groups.

n_1 and n_2 are the sample sizes of the two groups

df = n1 + n2 – 2

Two sample t-tests

are statistical tests used to compare the means of continuous data between two groups and determine if they are significantly different from each other.

H0: The same group does not stay significantly longer on new website B

→ **e-commerce website can yield higher average purchase value if it selects price B (if alpha = 5%)**

H1: The same group stays longer on new website B

→ **e-commerce website can yield higher average purchase value if it selects price B (if alpha = 5%)**

Paired t-tests

are used when you want to compare the means of continuous data for the same group at two different times or under two different conditions

The formula for the t-statistic in a paired t-test is:

$$\boldsymbol{t = \frac{\bar{D}}{s_D / \sqrt{n}}}$$

With:

- $\bar{D}$ is the mean of the differences between the paired observations.
- s_D is the standard deviation of the differences between the paired observations.
- n is the number of pairs.

→ **New layout makes the same group of users spend more time on the site**

Paired t-tests

are used when you want to compare the means of continuous data for the same group at two different times or under two different conditions

t_stat = 1.23186

t_stat = 5.231

H0: The same group does not stay significantly longer on new website B

H1: The same group stays longer on new website B

χ^2 tests

statistical tests used to determine if there's a significant association between two categorical variables in a sample.

Observed	2007	2008	2009	2010	2011	Total	
Freshman	560	495	553	547	512	2667	
Sophomore	369	385	358	361	393	1866	
Junior	209	226	248	268	285	1236	
Senior	267	277	304	328	340	1516	
Unclassified	64	70	93	77	126	430	
Total	**1469**	**1453**	**1556**	**1581**	**1656**	**7715**	**7715**

Observed	2007	2008	2009	2010	2011	Total
Freshman	507,818924	502,287881	537,893973	546,536228	572,462994	2667
Sophomore	355,301879	351,432016	376,344264	382,390927	400,530914	1866
Junior	235,344653	232,781335	249,282696	253,287881	265,303435	1236
Senior	288,658976	285,514971	305,754504	310,667012	325,404537	1516
Unclassified	81,8755671	80,9837978	86,7245625	88,117952	92,2981205	430
Total	**1469**	**1453**	**1556**	**1581**	**1656**	**7715**

alpha	0,05
df	16
Chi^2 Critical value	26,2962276

Ch^2-Computation				
Observed	**Expected**	**O - E**	**(O - E)^2**	**(O - E)^2/E**
560	507,82	52,18	2.722,86	5,36188107
369	355,30	13,70	187,64	0,52811009
209	235,34	-26,34	694,04	2,94903983
267	288,66	-21,66	469,11	1,62513998
64	81,88	-17,88	319,54	3,9027015
495	502,29	-7,29	53,11	0,10574256
385	351,43	33,57	1.126,81	3,2063373
226	232,78	-6,78	45,99	0,19755237
277	285,51	-8,51	72,50	0,2539437
70	80,98	-10,98	120,64	1,48972779
553	537,89	15,11	228,19	0,42423241
358	376,34	-18,34	336,51	0,89416013
248	249,28	-1,28	1,65	0,00660017
304	305,75	-1,75	3,08	0,01006783
93	86,72	6,28	39,38	0,45409414
547	546,54	0,46	0,22	0,00039354
361	382,39	-21,39	457,57	1,19660723
268	253,29	14,71	216,45	0,85454721
328	310,67	17,33	300,43	0,96705621
77	88,12	-11,12	123,61	1,40276589
512	572,46	-60,46	3.655,77	6,38604364
393	400,53	-7,53	56,71	0,14159872
285	265,30	19,70	387,95	1,46230552
340	325,40	14,60	213,03	0,65465452
126	92,30	33,70	1.135,82	12,3059567
			Chi^2-val	46,7812601

The test statistic for a chi-square test is calculated as:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Where

O_i are the observed frequencies

E_i are the expected frequencies

χ^2 tests

statistical tests used to determine if there's a significant association between two categorical variables in a sample.

Chi^2_stat = 0.697102

H0: There is no significant difference …

Chi^2_stat = 16.780310

H1: There is significant difference …

… between the signup rates depending on the versions A, B and C of landing page

Oneway – analysis of variance test

statistical tests used to compare the means of three or more groups and determine if they are significantly different from each other.

Group 1	Group 2	Group 3	
82	71	64	
93	62	73	
61	85	87	
74	94	91	
69	78	56	
70	66	78	
53	71	87	
Mean	**Mean**	**Mean**	**Overall mean**
71,71	75,29	76,57	74,52

N_Total	21
Num_Groups	3
N_per_Group	7
N_per_Group - 1	6
alpha	5%

	df	SS	MS	F_stat
Between (SSC)	2	88,66667	44,33333333	**0,283726128**
Within (SSE)	18	2812,571	156,2539683	
Total (SST)	20	2901,238		

- The larger the differences between the group samples the larger F_stat
- Exceeding F_crit, we “reject the null” (H1)

F_crit(5%, 2, 18)	3,554557146
F_stat	0,283726128

- **F_stat < F_crit**

→ H0 is not rejected (“H0”)

Oneway – analysis of variance test

statistical tests used to compare the means of three or more groups and determine if they are significantly different from each other.

F_stat = 0,283726128

H0: There is no significant difference …

F_stat = 4,5

H1: There is significant difference …

… between the signup rates depending on the versions A, B and C of landing page

Quiz

8. Model estimation

is the process of determining the parameters of statistical model that best fit observed data

- From Variance to Covariance to Correlation
- Parameter estimation in linear models
- Overfitting vs. underfitting
- Estimation of linear model
- Linear regression + model evaluation
- Logistic regression + model evaluation
- Quiz

From Variance to Covariance to Correlation

Covariance is a generalization of variance to multiple dimensions or variables.
Correlation is a normalized form of covariance to have values between -1 and +1.

https://medium.com/@raghavan99o/scatter-matrix-covariance-and-correlation-explained-14921741ca56

Variance (o^2):
- Measures how far a set of numbers spread out from mean.
- In the figure: The diagonal

Covariance:
- Measures how much two random variables vary together.
- In the figure: The triangles above and below the diagonal

Correlation:
- Standardized measure of the relationship between two variables
- ranging from -1 (perfect negative correlation) to +1 (perfect positive correlation), e.g. it is the normalized covariance.

Correlation

is the standardized measure of the relationship between two variables

https://en.wikipedia.org/wiki/Correlation

Positive Correlation:

- When two variables increase or decrease together.
- For example, the more time spent studying (variable 1), the higher the grades (variable 2). Correlation value ranges from 0 to +1.

Negative Correlation:

- When one variable increases as the other decreases.
- For example, the more time spent watching TV (variable 1), the lower the grades (variable 2). Correlation value ranges from 0 to -1.

Parameter estimation in linear models

involves finding the coefficients that best fit the observed data, i.e. represent the relationship between two correlated variables (X and Y)

https://vitalflux.com/generalized-linear-models-explained-with-examples/

Simple linear regression model:

$Y = \beta_0 + \beta_1 * X + \varepsilon$

β_0 (intercept) and β_1 (slope) are the parameters to be estimated.

- β_0 : Value of Y when X = 0.
- β_1 : Change in Y for each one-unit change in X.

Use Cases:

- Generalization of sample data to overall population data
- Forecasting like prediction of house prices based on independent features

Ordinary Least Squares (OLS) estimation

represents a relationship between two correlated variables (X and Y) and is used in linear regression analysis

https://gregorygundersen.com/blog/2020/01/04/ols/

Ordinary Least Squares (OLS) estimation:

- Method used in linear regression to estimate the model's parameters
- by minimizing the sum of the squares of the differences between the observed and predicted values of the dependent variable.
- Separation systematic relations from white noise

IID & BLUE:

If the residuals are identically and independently distributed (iid) the OLS estimator is used, because it provides best linear unbiased (BLUE) estimation.

What if IID-ness of residuals is violated?

- Heteroscedasticity: Variance is not identical
- Autocorrelation: Residuals depend on precedessors (typical example: time series)
- Transformations can be used in order to make OLS estimation BLUE.

Overfitting

is a modeling error where a model is trained too well on the training data

Problems associated with Overfitting:

Poor Generalizability:
Overfit models perform well on training data but poorly on new, unseen data, which limits their predictive accuracy.

Complexity:
Overfitting often results in overly complex models that have too many parameters.

Unreliable Predictions:
Overfit models are sensitive to minor fluctuations in data, leading to unreliable, inconsistent predictions over time.

Overfitting vs. underfitting

In opposite to overfitting, underfitting is when the model fails to capture important trends in the training data, resulting in poor performance on both the training and unseen data.

Underfitting:
Occurs when a model cannot adequately capture the underlying structure of the data. It's oversimplified, has poor predictive performance on both training and unseen data due to high bias.

What is worse?
Both overfitting and underfitting lead to poor predictions and indicate that the model may not be the right choice or needs fine-tuning.

R-squared: Goodness of fit

is a statistical measure in regression models that represents the proportion of the variance in the dependent variable that is predictable from the independent variables

$$R^2 = 1 - \frac{SSR}{SST}$$

Properties:

- Ranges from 0 to 1.
- Value of 1 → independent variable(s) perfectly predict dependent variable.
- Value of 0 → no predictive power.

- Higher R-Squared values → better fit, however very high R-Squared indicates overfitting.

- Only interpretable in the context of the data and domain. Low R-Squared could be acceptable in some fields where prediction is difficult.

Adjusted R-squared:

(…)

Statsmodels: Summary for OLS

provides a comprehensive summary of the results from fitted OLS model

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.669
Model:                            OLS   Adj. R-squared:                  0.667
Method:                 Least Squares   F-statistic:                     299.2
Date:                Mon, 01 Mar 2021   Prob (F-statistic):           2.33e-37
Time:                        16:19:34   Log-Likelihood:                -88.686
No. Observations:                 150   AIC:                             181.4
Df Residuals:                     148   BIC:                             187.4
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -3.2002      0.257    -12.458      0.000      -3.708      -2.693
x1             0.7529      0.044     17.296      0.000       0.667       0.839
==============================================================================
Omnibus:                        3.538   Durbin-Watson:                   1.279
Prob(Omnibus):                  0.171   Jarque-Bera (JB):                3.589
Skew:                           0.357   Prob(JB):                        0.166
Kurtosis:                       2.744   Cond. No.                         43.4
==============================================================================
```

Model Info

Goodness of fit

Coefficients Table

Autocorrelation, e.g. basis of time series

multicollinearity

skewness and kurtosis

Statsmodels: Summary – Coefficients table

provides detailed information about each parameter in the model

	coef	std err	t	P>|t|	[0.025	0.975]
const	-3.2002	0.257	-12.458	0.000	-3.708	-2.693
x1	0.7529	0.044	17.296	0.000	0.667	0.839

coef:
- Estimates of the parameters

1.std err:
- Standard error of the estimate of the coefficient
- not standard deviation!
- Smaller values indicate a more precise estimate.

t:
- t-statistic value
- Ratio: coefficient / std. err
- measure of how statistically significant the coefficient is.

P>|t|:
- p-value associated with the t-statistic.
- Smaller p-value (<0.05, traditionally) suggests that null hypothesis can be rejected, i.e. independent variable **is a significant predictor** of dep. variable.

[0.025 0.975]
- 95% confidence intervals for the coefficient.
- If zero is not in this interval, that suggests the variable is a significant predictor at the 0.05 level.

Logistic regression

is a statistical model to predict a binary outcome (0 or 1, true or false) based on one or more predictor variables, using the logistic function to model the probability of the binary outcome.

Fig 12: Typical logistic regression curve

https://medium.com/@cmukesh8688/logistic-regression-sigmoid-function-and-threshold-b37b82a4cd79

Logistic function (= sigmoid function):
central to logistic regression:
$P(z) = \frac{1}{1+e^{-z}}$ with $z = \beta_0 + \beta_1 x$

- It converts any input into a value between 0 and 1, which can be interpreted as a probability.
- For given input 'x', the output represents probability that 'x' belongs to class 1 (usually coded as positive outcome).
- Simple rounding can be used for predicting 0 or 1

It is based on …
Logged odds:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

→ Linear model is projected on S-curve

Statistics perspective: Summary for Logistic regression

provides a comprehensive summary of the results from fitted logistic regression model

```
                           Logit Regression Results
==============================================================================
Dep. Variable:                  Class   No. Observations:                  768
Model:                          Logit   Df Residuals:                      760
Method:                           MLE   Df Model:                            7
Date:                Fri, 22 Oct 2021   Pseudo R-squ.:                 0.05922
Time:                        01:36:12   Log-Likelihood:                -467.33
converged:                       True   LL-Null:                       -496.74
Covariance Type:            nonrobust   LLR p-value:                 2.583e-10
==========================================================================================
                                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------
Number of times pregnant       0.1284      0.029      4.484      0.000       0.072       0.185
Plasma glucose concentration   0.0129      0.003      4.757      0.000       0.008       0.018
Diastolic blood pressure      -0.0303      0.005     -6.481      0.000      -0.039      -0.021
Triceps skinfold thickness     0.0002      0.006      0.032      0.974      -0.012       0.012
2-Hour serum insulin           0.0007      0.001      0.942      0.346      -0.001       0.002
Body mass index               -0.0048      0.011     -0.449      0.653      -0.026       0.016
Diabetes pedigree function     0.3203      0.240      1.335      0.182      -0.150       0.790
Age                           -0.0156      0.008     -1.852      0.064      -0.032       0.001
==========================================================================================
```

ML perspective: Train test split

is a method where the dataset is divided into subsets: the training set, used to train the model, and the testing set, used to evaluate the model's performance on unseen data

https://medium.com/@rathinavel.mph/how-the-train-and-test-samples-are-split-d7e46a8e2361

Training models vs. analytical solution

- Logistic regression does not have analytical solution in opposite to linear regression.
- Therefore squared error minimizing parameters are approximated, e.g. with Newton's approximation method
- However: Danger of overfitting

Training Set:

- Portion of data used to create the model, i.e. to learn/approximate the parameters
- Usually 70-80% of the entire dataset

Testing Set:

- Subset of data that the model has not seen during learning phase
- Used to evaluate performance of the model on unseen data and check for overfitting. Typically 20-30% of entire dataset

Logistic regression

is a statistical model to predict a binary outcome (0 or 1, true or false) based on one or more predictor variables, using the logistic function to model the probability of the binary outcome.

Logistic function (= sigmoid function):
central to logistic regression:
$P(z) = \frac{1}{1+e^{-z}}$ with $z = \beta_0 + \beta_1 x$

- It converts any input into a value between 0 and 1, which can be interpreted as a probability.
- For given input 'x', the output represents probability that 'x' belongs to class 1 (usually coded as positive outcome).
- Simple rounding can be used for predicting 0 or 1

It is based on …
Logged odds:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

→ Linear model is projected on S-curve

Fig 12: Typical logistic regression curve

https://medium.com/@cmukesh8688/logistic-regression-sigmoid-function-and-threshold-b37b82a4cd79

Confusion matrix

is a tabular representation to visualize the performance of a predictive model on a set of test data for which the true values are known.

1. **True Positives (TP):** These are cases in which we predicted yes (they have the condition), and they do have the condition.
2. **True Negatives (TN):** We predicted no, and they don't have the condition.
3. **False Positives (FP):** We predicted yes, but they don't actually have the condition. (Also known as a "Type I error.")
4. **False Negatives (FN):** We predicted no, but they actually do have the condition. (Also known as a "Type II error.")

1.Accuracy: It is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. Accuracy = (TP + TN) / (TP + FP + FN + TN) Why is it misleading?

1.Precision: What proportion of positive identifications was actually correct? Precision = TP / (TP + FP) What is it good for?

Accuracy rate and error rate

example

		Prediction	
		NEG	POS
Actual	NEG	43	12
	POS	4	41

Type II Error
(False Negatives)

Type I Error
(False Positives)

Accuracy Rate and Error Rate:

$$AR = \frac{Correct}{Total} = \frac{TN + TP}{Total} = \frac{84}{100} = 84\%$$

$$ER = \frac{Incorrect}{Total} = \frac{FP + FN}{Total} = \frac{16}{100} = 16\%$$

Cumulative Accuracy Profile (CAP) curve

illustrates the cumulative number of positive outcomes along different thresholds of classification

•Y-axis: Cumulative percentage of positive samples.

X-axis: Cumulative percentage of total samples.

Key Components

- Random Model line: Represents a random guess model.
- Perfect Model line: Achieves 100% identification of true positives with no false positives.
- Logistic Regression Model line: Represents the actual model's cumulative accuracy profile.

50% Rule

- Consider the point where x=50%, trace it up to the Logistic Regression model line. From there, trace horizontal line to y-axis.
- y < 60%: model might not be beneficial.
- y is between 60% and 70%: model is good.
- y > 70%: model possibly too good, indication of overfitting

Quiz

Outlook – Machine learning

is

- A-B testing with Reinforcement Learning
- Supervised Learning …
- Non parametric tests
- Logistic regression: ROC-AUC-Curve, F-Score, etc.

Links

i.e. sources for self-learning

Variance:
https://www.youtube.com/watch?v=JgMFhKi6f6Y

Quantiles
https://docs.eggplantsoftware.com/epp/9.4.0/analyzer/analyzer-understanding-charts-events.htm

Miracles of Pascal‘s triangle:
https://www.youtube.com/watch?v=J0I1NuxUcpQ

Distributions		
	Uniform probability distributions	https://www.youtube.com/watch?v=aCW8wm6nrRw
	Binomial distribution	https://www.youtube.com/watch?v=6YzrVUVO9M0
	Chi^2	https://www.youtube.com/watch?v=JCOeytq7F0E

Links

i.e. sources for self-learning

Conditional probability		
	Bayes' Theorem of Probability With Tree Diagrams & Venn Diagrams	https://www.youtube.com/watch?v=OByl4RJxnKA
	What is the Bayes' Theorem?	https://medium.com/mlearning-ai/what-is-the-bayes-theorem-545a2ef0b91c
	Data Science. Bayes theorem	https://luminousmen.com/post/data-science-bayes-theorem
	Bayesian Machine Learning in Python: A/B Testing	https://www.udemy.com/course/bayesian-machine-learning-in-python-ab-testing

Links

i.e. sources for self-learning

Sampling		
	Central limit theorem	https://www.geeksforgeeks.org/python-central-limit-theorem
	Confidence intervals	https://data.library.virginia.edu/the-intuition-behind-confidence-intervals/
	Resampling	https://www.linkedin.com/pulse/resampling-methods-pranshu-jaryal
	Bootstrapping	https://www.youtube.com/watch?v=9IFcRTyhd5Y
		https://dgarcia-eu.github.io/SocialDataScience/2_SocialDynamics/025_Bootstrapping/Bootstrapping.html
		https://janhove.github.io/teaching/2016/12/20/bootstrapping

Hypothesis testing		
	Bayesian Machine Learning in Python: A/B Testing	https://www.udemy.com/course/bayesian-machine-learning-in-python-ab-testing

Links

i.e. sources for self-learning

Model estimation		
	Scatter matrix , Covariance and Correlation Explained	https://medium.com/@raghavan99o/scatter-matrix-covariance-and-correlation-explained-14921741ca56
		https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148#:~:text=Logistic%20regression%20transforms%20its%20output,to%20return%20a%20probability%20value.
	Linear Regression Diagnostics	https://www.youtube.com/watch?v=HLAg1yBfNk8&list=PLlbbWgBRF8EePgK40-i7aGU2_ky1yujgL&index=11
	Machine Learning Classifier evaluation using ROC and CAP Curves	https://towardsdatascience.com/machine-learning-classifier-evaluation-using-roc-and-cap-curves-7db60fe6b716
	Classification Model Performance Evaluation using AUC-ROC and CAP Curves	https://medium.com/geekculture/classification-model-performance-evaluation-using-auc-roc-and-cap-curves-66a1b3fc0480
	Deep Learning Prerequisites: Logistic Regression in Python	https://www.udemy.com/course/data-science-logistic-regression-in-python/

Subtitle demo text

About Us

Lorem Ipsum is simply dummy text has been of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy text ever since the 1500s, when an unknown printer took a galley of type and scrambled it to make a type Letraset sheets containing Lorem Ipsum passages, and more PageMaker including versions of Lorem Ipsum.

Lorem Ipsum is simply dummy text has been of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy text ever since

Subtitle demo text

Section Breaks

AUC [redacted] Under The Curve)

[redacted] ability to distinguish between classes. An AUC of 1 indicates perfect classification, [redacted] than random guessing.

https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-classification-in-python/

ROC (Receiver Operating Characteristic) curve: performance measurement tool for classification problems at various thresholds settings.

[redacted] the trade-off between the True Positive Rate (TPR) [redacted] Positive Rate (FPR).

[redacted] known as sensitivity) on the y-axis [redacted] the x-axis.

[redacted] the classifier will rank a[redacted]r than a negative o[redacted]

An AUC value of 1 in[redacted]ation, while AUC value of 0.5 suggests the [redacted]etter than random guessing.