Statistics for Data Science PDF

Statistics for data science covers describing data with averages and spread, probability, common distributions, sampling, the central limit theorem, confidence intervals, hypothesis tests with p-values, correlation and the basics of regression. This page explains each idea in plain language with small examples and the mistakes that cause wrong conclusions.

Key takeaways

- Describe data first: centre, spread, shape and outliers.
- Probability and distributions explain how likely different values are.
- A confidence interval gives a range of plausible values, and a p-value measures how surprising data is if there were no effect.
- Correlation does not prove causation.
- Check assumptions and sample size before trusting a test result.

What is statistics for data science?

Statistics is the science of collecting, summarising and interpreting data, and in data science it provides the methods for measuring uncertainty, comparing groups and deciding whether a pattern is likely real or due to chance.

How do I describe a dataset?

Start with the **centre**: mean (average), median (middle) and mode (most common). Then the **spread**: range, interquartile range (IQR), variance and standard deviation. Look at the **shape**: symmetric or skewed, with or without outliers. Use a histogram and a box plot. The median and IQR are robust to outliers, while the mean and standard deviation are not.

What is probability?

Probability measures how likely an event is, from 0 (impossible) to 1 (certain). Key rules: for independent events, $P(A \text{ and } B) = P(A) \times P(B)$. $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$. **Conditional probability** $P(A \mid B)$ is the chance of A given that B happened. **Bayes' theorem** updates a belief when new evidence arrives: $P(A \mid B) = P(B \mid A) \times P(A) / P(B)$.

Which distributions should I know?

Distribution	Describes	Example
Normal	Symmetric bell-shaped data	Heights, measurement error
Binomial	Number of successes in n trials	Orders that convert out of 100 visits
Poisson	Count of events in a time period	Calls per hour
Uniform	Equal chance for all values	Random number between 0 and 1
Exponential	Time between events	Time until next arrival

In a normal distribution about 68 percent of values fall within one standard deviation of the mean, 95 percent within two and 99.7 percent within three.

Why does sampling matter?

We rarely see the whole population, so we study a **sample**. A sample is useful only if it represents the population; random sampling helps. Bias in sampling, such as asking only online users, leads to wrong conclusions. A larger sample reduces random error but not bias.

What is the central limit theorem?

The distribution of sample means approaches a normal distribution as the sample size grows, even if the underlying data is not normal. This is why many tests use averages and why confidence intervals work.

What is a confidence interval?

A confidence interval gives a range of plausible values for a population quantity. A 95 percent interval means that if you repeated the sampling many times, about 95 percent of such intervals would contain the true value. It does not say the true value has a 95 percent chance of lying in this one interval. Wider intervals come from smaller samples or noisier data.

How does hypothesis testing work?

State a **null hypothesis** (no effect) and an **alternative**. Choose a significance level, often 0.05. Compute a test statistic and a **p-value**: the probability of seeing data at least this extreme if the null were true. A small p-value is evidence against the null. It is not the probability that the null is true and it says nothing about the size or importance of the effect. Report the effect size and an interval too.

Which test should I use?

Situation	Typical test
Compare two group means	t-test

Situation	Typical test
Compare more than two means	ANOVA
Two categorical variables	Chi-square test
Compare proportions	z-test for proportions
Non-normal small samples	Non-parametric tests such as Mann-Whitney

Check assumptions such as independence and approximate normality.

What are Type I and Type II errors?

A Type I error rejects a true null (a false positive) and has probability equal to the significance level. A Type II error fails to reject a false null (a false negative). **Power** is the chance of detecting a real effect, and it increases with sample size and effect size. Plan the sample size before an experiment.

What about correlation and regression?

Correlation measures the strength and direction of a linear relationship from -1 to 1. It does not show causation; a third factor or coincidence may explain it. **Linear regression** fits a line to predict a number from one or more variables, and the coefficients show the average change per unit. Check residuals and avoid extrapolating beyond the data.

What mistakes cause wrong conclusions?

Testing many things and reporting only the significant ones (multiple comparisons), stopping an experiment early when results look good, ignoring sample bias, confusing statistical with practical significance and using the mean for heavily skewed data.

How do I practise these ideas?

Take a small dataset, compute the mean, median and standard deviation by hand and with a tool, and plot a histogram. Draw random samples of different sizes and watch how the sample mean varies. Run a t-test between two groups, compute an interval and write one paragraph in plain words that a manager could read, including the limits of the result. Writing the paragraph is the real skill.

Frequently asked questions

What statistics do I need for data science?

Descriptive statistics, probability, common distributions, sampling, confidence intervals, hypothesis tests, correlation and regression basics.

What does a p-value mean?

It is the probability of seeing data at least as extreme as observed if the null hypothesis were true. It is not the probability that the hypothesis is true.

What is the central limit theorem?

The distribution of sample means approaches a normal distribution as the sample size grows, whatever the shape of the data.

Does correlation mean causation?

No. Two variables may move together because of a third factor or chance, so experiments are needed for causal claims.

Which test compares two group averages?

A t-test is the common choice, provided its assumptions about independence and approximate normality are reasonable.

Sources and further reading

- NIST/SEMATECH e-Handbook of Statistical Methods
- SciPy statistics documentation

About this PDF

These PDFs are study material for learning and revision. They are not a certificate, a degree or a promise of a job, and technology changes, so check official documentation for the latest details. Spot an error? Write to info@uncodemy.com and the trainer team will review it.

More free resources and courses: uncodemy.com | info@uncodemy.com | +91 98183 66550