Lecture 1: Introduction

Paul Röttger

Applied Analytical Statistics

20th of January 2026

Departmental Lecturer at the Oxford Internet Institute
and Associate Member at Nuffield College

Background in **Natural Language Processing**
and **Computational Social Science**

Research focus on **societal impacts of AI**,
AI safety, and **LLMs for social science**

paulrottger.com | paul.rottger@oii.ox.ac.uk | @paul_rottger

Do LLMs give different medical advice depending on user demographics?

Can social media posts predict offline violence?

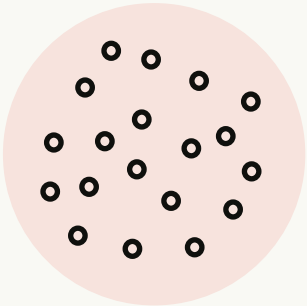
As researchers, we want to learn **truths about our world**.
We ask **general** questions that concern millions of people.

Which workers benefit most from using AI?

Do AI chat assistants give better answers in some languages than others?

500 chatbot responses to identical symptom descriptions with varied patient profiles

50k geolocated tweets from a conflict region matched to incident reports

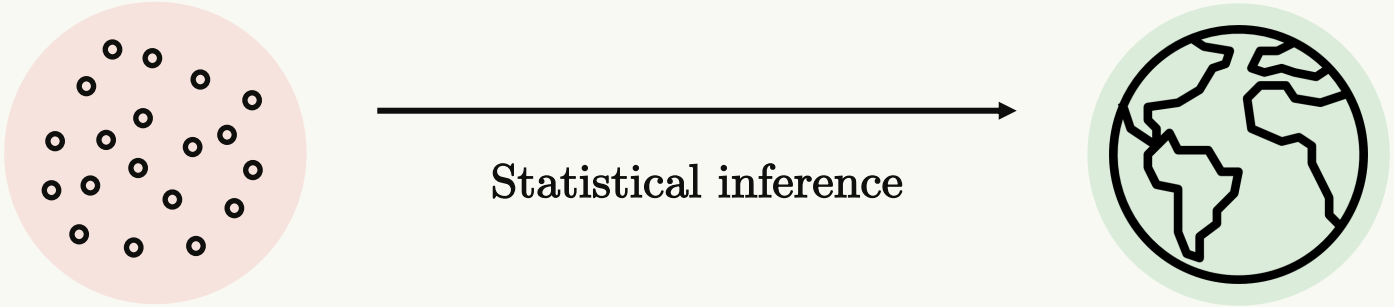


However, we **never observe everything**.
We always work with a **particular sample of data.**

Activity logs from 200 employees randomly assigned to use AI tools or not

LLM answers to 1k benchmark questions translated into 20 languages

Statistical methods are the **bridge** from the specific to the general!



With statistics, we can:

Draw **conclusions** that reach beyond our immediate observations.
Quantify our **uncertainty** about those conclusions.
Distinguish genuine **patterns** from **noise**.

Learning outcomes | What this course will teach you

Statistical thinking as a framework for learning from data.

- How to describe and summarise data effectively.
- How to express confidence and uncertainty.
- How to draw valid conclusions from data.

Regression as a tool for analysing the relationship between things we observe in the world.

- How to design the right regression model for your research.
- How to evaluate model fit and assumptions.
- How to interpret regression results and distinguish cause from correlation.

Academic writing for quantitative / computational social science research.

- How to present and discuss results of statistical analysis.
- How to structure a research paper.

Assumed knowledge | Things you should know/revise

We are going to focus on applications rather than mathematical foundations of statistics. Therefore, all you will need is some GCSE / high school level maths:

Basic algebra, including sums like $\sum_{i=1}^{n} x_i$

Linear functions like $f_1(x) = a + bx$ and their graphical intuition.

Exponential functions like $f_2(\mathrm{x}) = a^x$

Logarithmic functions like $f_3(\mathrm{x}) = \ln(\mathrm{x})$ as the inverse of exponentials.

Simple differentiation like $f_1'(\mathrm{x}) = b$, $f_2'(\mathrm{x}) = \ln(a)\, a^x$, $f_3'(\mathrm{x}) = 1/x$

Course structure | Eight weeks of condensed fun :)

Week 1:	**Introduction**
Week 2:	Statistical Inference & Uncertainty
Week 3:	Hypothesis Testing
Week 4:	Univariate Linear Regression
Week 5:	Multivariate Linear Regression
Week 6:	Logistic Regression
Week 7:	Causality
Week 8:	Recap & Paper Writing

Tutorials | Hands-on practice sessions

Mikhail Korneev
(Politics DPhil)

Every week of term **right after the lecture**.
(i.e. Tuesdays, 3:30pm-5pm)

Practical exercises in R covering lecture concepts.
Solutions released the day after.

Attendance highly recommended!

Mikhail will also manage the Async Q&A Google Doc. Send questions **that you cannot answer yourself** to mikhail.korneev@reuben.ox.ac.uk and we will post answers for everyone.

Summative | Writing a research paper

Goal: Apply the statistical methods learned in this course to answer a quantitative social science research question.

This is an **ideal practice run for your MSc thesis!** (although summative and thesis should not be about the exact same question)

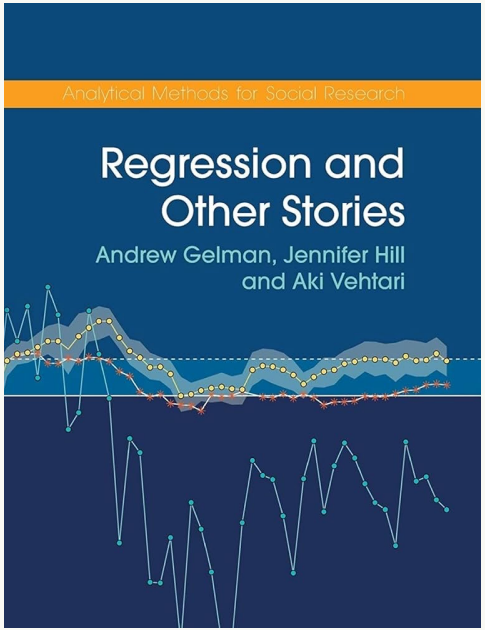
You will write a 4,000-word research paper, structured much like your thesis, just with more focus on methods and analysis, **due Thursday Week 0 TT** (TBC).

A first proposal (half-page abstract) is **due on Friday of Week 4 HT**. You will receive feedback from us and from your fellow students. More details on Canvas.

Reading list | Access via ORLO / Canvas

https://oxford.alma.exlibrisgroup.com/leganto/public/44OXF_INST/lists/50906138150007026?auth=SAML

Main textbooks | Regression and Other Stories (RegOS)



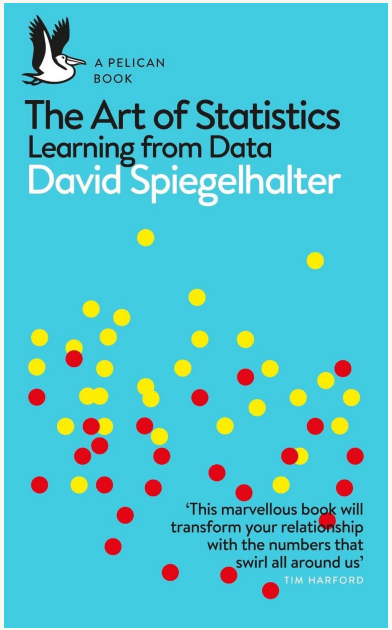
Our main resource for learning about regression.

Focus is applied rather than mathematical.

Several chapters are essential reading.

Access the latest version on the RegOS website.

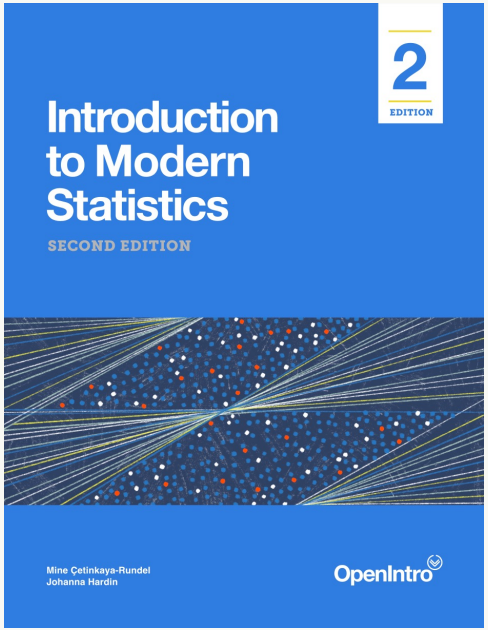
Main textbooks | The Art of Statistics (ArtOS)



More accessible intro to core concepts.

Recommended reading, especially if this is your first statistics course ever / in a long time.

Main textbooks | Introduction to Modern Statistics (IMS2)

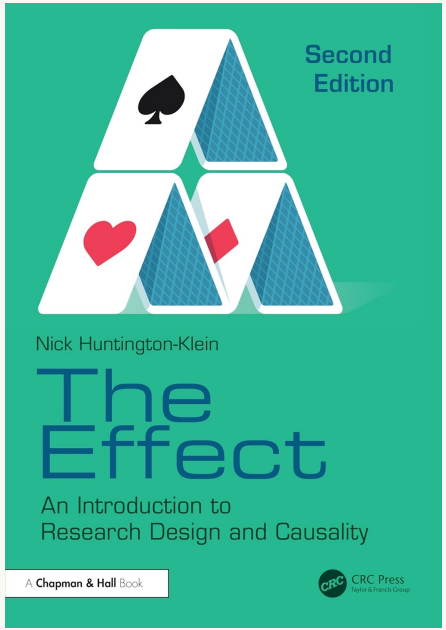


Very practical book with many exercises.

Essential only for Week 3 (Hypothesis Testing).

Recommended if you want more hands-on practice.

Main textbooks | The Effect (TEff)



Essential only for Week 7 (Causality).

Highly recommended as further reading for anyone interested in causal inference.

Using AI for this course | What a time to be alive...

Use AI to practice, learn, improve:

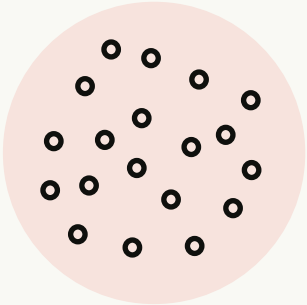
- ”Explain [concept X] with an example from [my research area].”
- “Create some exercises to practice [topic Y]. Then give me the solutions.”
- “Give me critical feedback on [this text] that I wrote."

Do not use AI to outsource your thinking. As a researcher, do not “just do things”.

“[...] To live well with machines is to insist that they serve our efforts at growth rather than replace them, that they enlarge the field for judgment instead of shrinking it. The task of becoming the person you want to be — the kind who can judge, discern, and act — cannot be outsourced. It has to be practiced by each of us, with all the false starts and frustrations that practice entails.” – Harry Law: The Worst Time to Have a Problem

Rest of today | Describing data and recapping probability

Most of this course will be about **statistical inference**, i.e. drawing conclusions about a general population based on a particular sample of data.



Today, instead, we focus on **describing the data that we are working with**, and **how it came into existence**.

We do this because the sample of data we are working with is the **foundation** for all analyses that follow.

We will then finish the lecture with a **brief probability recap** to set us up for next week.

Data collection | How did our sample come to exist?

The data we work with is typically a **sample of a larger population**.
This sample is the result of **sampling decisions** + **constraints** (data access, time, money).

How our sample was generated determines **what RQs we can credibly ask and answer:**

RQ: Which workers benefit most from using AI?

Data: Activity logs from 200 software engineers at OpenAI.

Better RQ: Do benefits from AI for software engineers depend on their level of experience?

No amount of fancy statistics can overcome fundamental limitations in our sample!

Data collection | Can we expect our results to generalise?

Ecological validity describes the extent to which we can expect results from our sample to generalise to the wider population, based on how the sample came into existence.

All samples are **limited and biased** in some ways. Research design requires researcher judgment about which kinds of generalisation are plausible or not.

Data: Activity logs from 200 software engineers at OpenAI.

Better RQ: Do benefits from AI for software engineers depend on their level of experience?

✓ Plausible generalisation: Software engineers in similar tech companies
✗ Not plausible: Other jobs that require different skills and expertise

When writing a paper, we discuss ecological validity to highlight the limits of our claims.

Data collection | Are we measuring what we care about?

Construct validity describes the extent to which a test or measurement accurately assesses the abstract, theoretical concept (construct) that it is supposed to.

In our example, the construct of interest is worker productivity:

RQ: Do benefits from AI for software engineers depend on their level of experience?

The data we work with is complex and multi-faceted:

Data: Activity logs from 200 software engineers at OpenAI.

We **operationalise** productivity by recording specific measurements, e.g. lines of code. These measurements are usually **imperfect proxies** for the construct of interest.

When writing a paper, we discuss construct validity to highlight the limits of our claims.

Data collection | How reliable are our measurements?

A reliable measure is one that is **precise** and **stable**.
  Precision concerns the **consistency of repeated measurements under identical conditions**.
  Stability concerns the **consistency of measurements across time**.

Returning to our example, using "lines of code written" as our measure:

Data: Activity logs from 200 software engineers at OpenAI.

This is a **precise** measure because it is factual record. However, there is some **instability** because random factors (e.g. mood) influence lines of code written on a specific day. We can reduce the impact of this instability by measuring many people across multiple days.

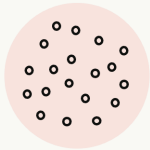
When writing a paper, we describe and discuss measurement reliability to show that we measure **real differences among people or things** rather than **random measurement error**.

Descriptive statistics | What's the point?

So far, we focused on how our data came into existence and what it can plausibly tell us.

We now move to the next stage: **describing what our specific sample looks like.**

Our goals for this stage are to:

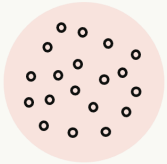


1. Summarise complex data into interpretable quantities.
2. Understand central tendencies and variation in our data.
3. Identify anomalies and potential data issues.

In the coming weeks, **knowing the shape of our data** will help us make the right modelling decisions and draw valid conclusions.

Descriptive statistics | Sample size

The first statistic we report about our sample should always be the **sample size (N).**



Activity logs from 19 software engineers at OpenAI.

For small N, summary statistics are highly sensitive to individual observations.

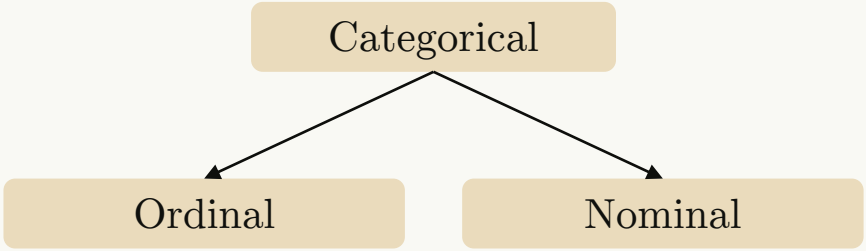
Next week, we will learn that sample size also affects certainty in downstream conclusions.

When writing a paper, state exact sample sizes and be explicit about missing data.

In general, it is a good habit to always mentally pair a statistic with the N it is based on.

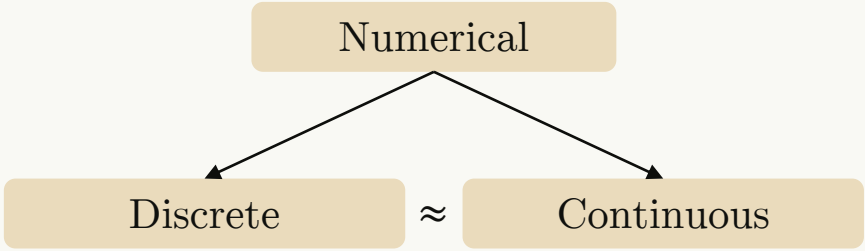
Descriptive statistics | Types of data

How we best describe our data depends on what type of data we are working with.



Ordered categories:
- Income level
- Education level
- Likert scales

Unordered categories:
- Political party
- Gender
- Personality



Counted items:
- Number of likes
- Lines of code

Measured attributes:
- Time spent on app
- % clickthrough
- Sentiment score

In the following weeks, we will learn that data types also determine modelling choices.
When writing a paper, it is therefore useful to be explicit about data types.

Categorical data | How is my variable distributed?

The **distribution** of a variable describes how often different values occur.
For categorical data, we can easily show the full distribution in a **frequency table**.

Data: Tweets in 8 languages labelled for hate speech (Tonneau et al., 2024)

ordered categories

Category	N	%	
Hateful	1,014	0.42	
Offensive	12,129	5.05	
Neither	226,857	94.52	← modal category
Total	240,000		

When writing a paper, text may be cleaner: ”We annotated 240,000 posts, of which 1,014 (0.42%) were labelled ‘hateful’, 12,129 (5.05%) ’offensive’, and 226,857 (94.52%) ‘neither’.”

Categorical data | 2-way comparison

Contingency tables help us explore the association between two categorical variables.

unordered categories

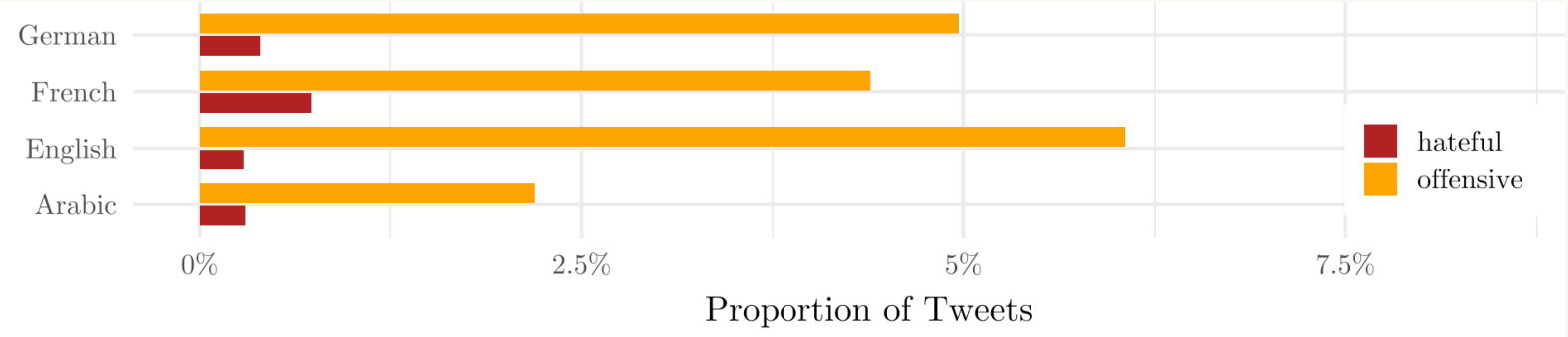
Language	% Hateful	% Offensive	% Neither
Arabic	0.30	2.20	97.51
English	0.29	6.05	93.66
French	0.73	4.39	94.87
...			
Total	0.42	5.05	94.52

Contingency tables may be reported as raw counts or normalised percentages.
When writing a paper, they are not "main body material" but can be useful in the appendix.

Categorical data | Bar charts

Data visualisation helps us **explore** and **understand** our data.
It also helps us **communicate** information about our data to our readers.

This bar chart is much easier to interpret than the contingency table on the previous slide:

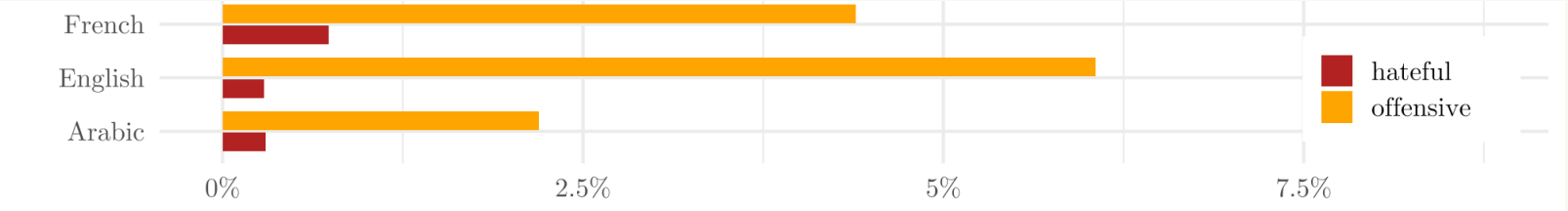


When writing a paper, remember that your readers will often look at visuals before text.
A good figure is worth a thousand words! (Great tips here)

Interlude | Claims based on descriptive statistics

RQ: How does the **prevalence** of **online hate speech** differ across languages?

We now understand the shape of our data. What claims can we make?



✗ "There is more French hate speech than Arabic hate speech on social media."

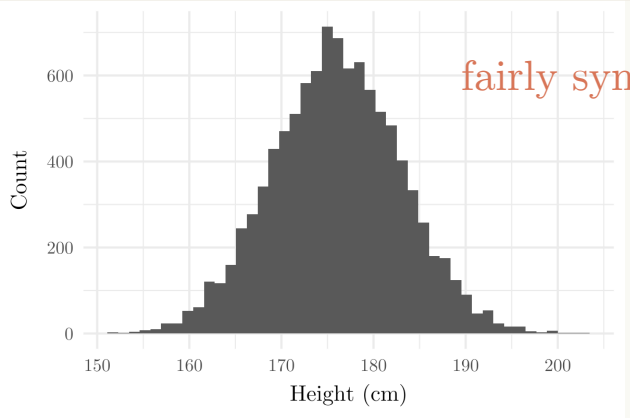
✓ "In our sample, more French content was labelled as hateful than Arabic content."

Describing our sample is useful but does not allow us to answer the population-level RQ. This is the difference between descriptive stats (today) and inferential stats (next weeks).

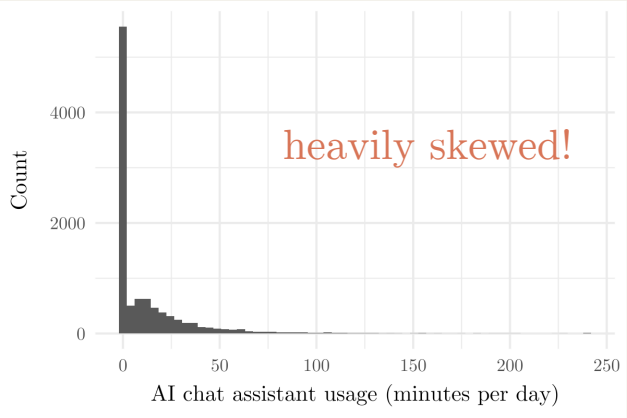
Continuous data | Visualise first!

For continuous data, **visualisation is often the best first step** because the shape of the distribution determines how we can use and interpret simpler summary statistics.

Data: Height of male adults in the UK*



Data: Daily usage of AI chat assistants*



*simulated

Most observational data is not symmetric, which complicates description.

Continuous data | Summary statistics

Visualisation is a great first step, but summary statistics play a complementary role.
They help us **condense complex information** and **enable comparison** using common metrics.

Measures of central tendency tell us where the centre of our data is.

- Median: The middle observation when all values are ordered.
- Mode: The most frequent observation.

Measures of spread tell us how much our observations vary around that centre.

- Range: The distance between the smallest and largest observed values.
- Interquartile range (IQR): The range of the middle 50% of the data.
- Standard deviation (SD): The average distance of observations from the mean.

Continuous data | Standard deviation

Standard deviation is the **least obvious** but, arguably, **most important** of these concepts.

For a sample of n observations $x_1, x_2, \ldots, x_n$ the standard deviation SD is defined as:

measured in original units of our data

$$SD = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

deviation from sample mean, squared to remove sign

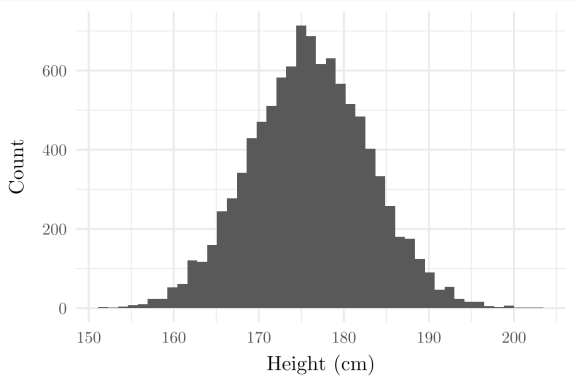
more weight to larger deviations!

We will encounter variants of this formula again throughout this course, when quantifying uncertainty, testing hypotheses, in regression models...

Continuous data | Summary statistics for symmetric data

For symmetric distributions, summary statistics are informative and tell a consistent story.

	Statistic	Height (cm)
central tendency	Mean	176.2
	Median	176.0
spread	Range	51.2
	Inter-quartile range (IQR)	9.4
	Standard deviation	7.0



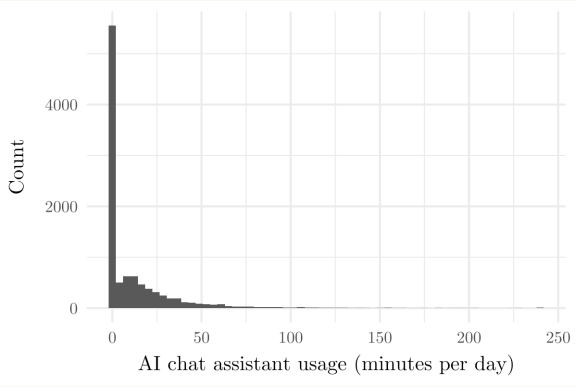
Mean + standard deviation provide a clear and compact summary of symmetric data.

Continuous data | Summary statistics for skewed data

For strongly skewed distributions, some summary statistics become misleading.

Vulnerable to outliers

Statistic	AI Usage (mins)
Mean	12.1
Median	0
Range	240
Inter-quartile range (IQR)	16.0
Standard deviation	23.0

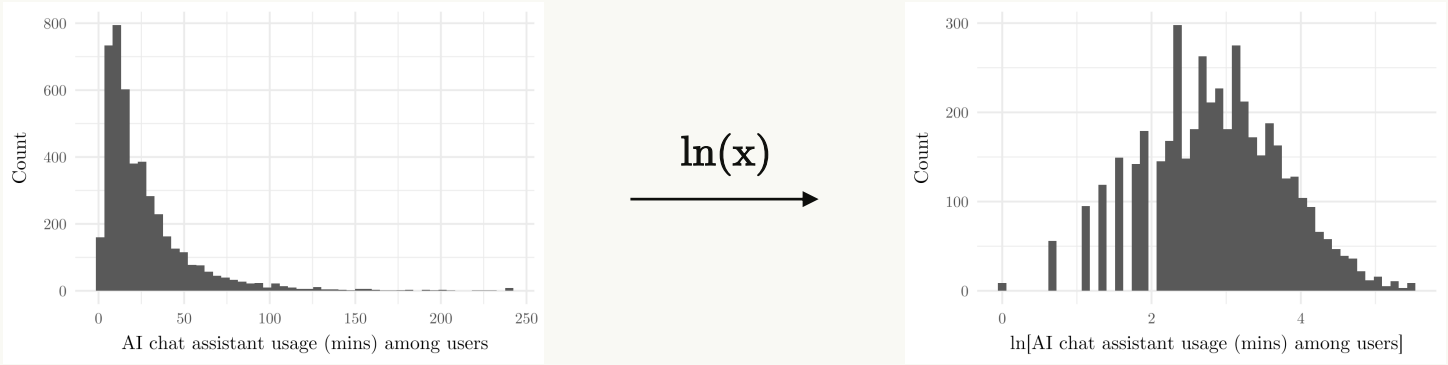


Median + IQR provide a more robust summary for very assymetric data.

Continuous data | Transformations

When data is heavily skewed, **rescaling** can make patterns easier to describe and analyse.

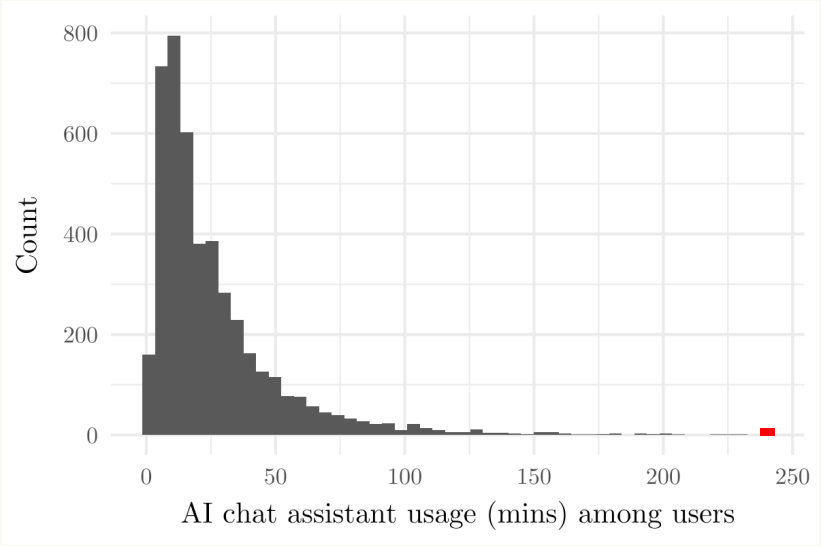
The **log transformation** is particularly useful for strictly positive, right-skewed data.
It reduces skewness and makes differences among typical observations more visible.



In coming weeks, transformations will help us better align data with modelling assumptions.

Continuous data | Outliers

Extreme values can mess with our summary statistics but deserve special attention.



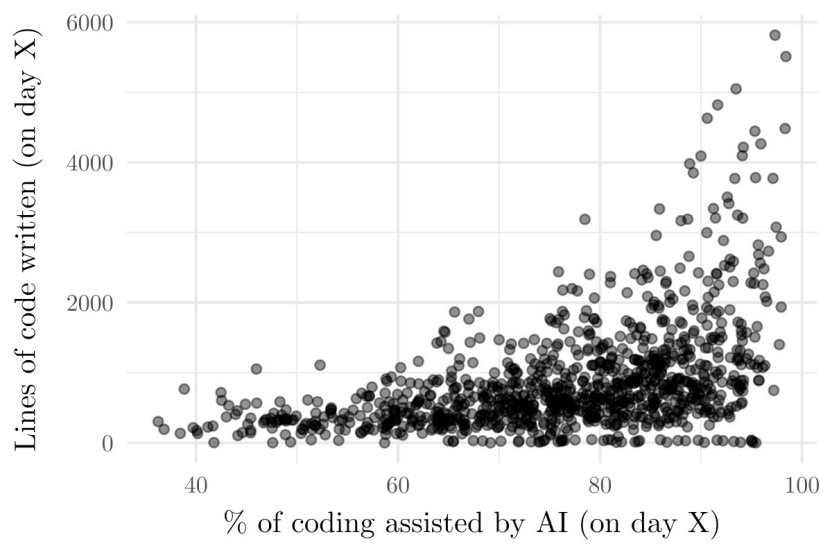
Outliers may point to **measurement errors**.

Outliers may be **rare but meaningful cases**.

Automatic removal is a missed opportunity for better understanding our data.

Continuous data | Exploring associations

Much of this course will be about investigating the relationship between different variables. For two continuous variables, a **scatterplot** is the most useful starting point.



Data: Activity logs from 600 software engineers at OpenAI*

There seems to be a **clear positive association** between how much SEs use AI and how much code they write.

🤔

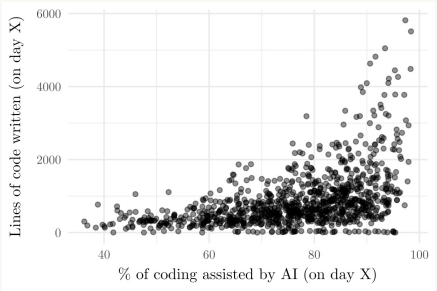
*simulated

Continuous data | Correlation

As a summary statistic for association between two continuous variables, we use **correlation**.

Pearson's correlation (**r**) measures the strength and direction of a **linear** relationship.
This is calculated based on raw values and therefore sensitive to outliers and assymetry.

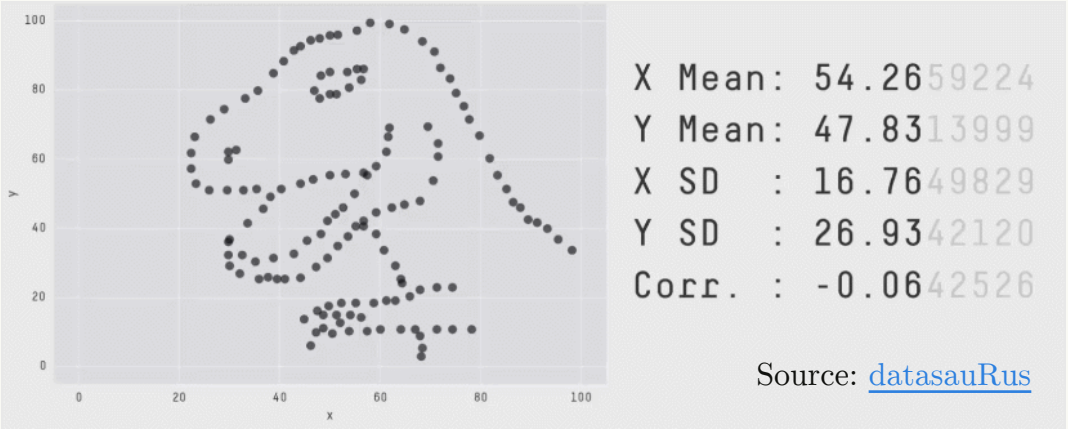
Spearman's correlation (ρ) measures the strength and direction of a **monotonic** relationship.
This is calculated based on ranks and therefore more robust to outliers and skewness.



In our example, r = 0.47 and ρ = 0.52 (on a scale from -1 to 1).
This matches our visual intuition: clear positive association.

Summary statistics | A word of warning

All summary statistics **compress information** and can therefore be misleading!



When writing a paper, a good rule of thumb is to report summary statistics for continuous data with at least one accompanying visualisation (e.g. a histogram in the appendix).

Describing data | Key principles

Explain data collection.
Describe sample selection, constraints, measurement, and reliability.

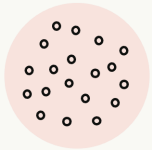
Visualise before you summarise.
Use visualisations to understand the shape of your data and spot anomalies.

Choose statistics that fit the data.
Select summary statistics that match your variable type and distribution

Treat summaries with caution.
Remember that summaries compress information, and compression comes at a cost.

Probability | Why do we care?

Probability theory is the basis for statistical inference.
It gives us a formal language to **reason about uncertainty**.



When working with samples, outcomes are never deterministic.
Random variation is unavoidable.

Probability will let us:

- Quantify how likely different outcomes are.
- Formalise assumptions about random processes.
- Distinguish systematic patterns from random fluctuation.

But before we get there, we need to cover some basics.
Everything we cover today will come in very handy next week!

Probability | What exactly is probability?

An event is something that may or may not occur (e.g. rolling an even number with a dice).

In this course, we take a **frequentist** approach to statistics, based on "**objective" probability.**

> The probability of an event is the proportion of times this event occurs in an infinite sequence of identical repetitions of a random process.

hypothetical concept: most realistic situations are not infinitely repeatable

Some people prefer a **Bayesian** approach to statistics, based on "**subjective" probability**.

> The probability of an event is how likely we believe this event is to happen after updating our prior beliefs based on new evidence.

secret sauce: additional information beyond what we observe in the data

Probability | Basic notation

Example: Rolling a fair 6-sided dice

$S = \{1, 2, 3, 4, 5, 6\}$

The **sample space** S describes the set of all possible outcomes.

An **event** A describes a particular subset of outcomes.

$A =$ *"rolling an even number"*

$\Pr(\text{"}even\text{"}) = 1/2$

The probability of an event $\Pr(A)$ is the long-run proportion of trials in which A happens.

For any event A, we know that $0 \leq \Pr(A) \leq 1$ and $\Pr(A^c) = 1 - \Pr(A)$

For two events A and B, $\Pr(A \cap B)$ is the probability that both occur,
$\Pr(A \cup \mathrm{B})$ is the probability that one occurs.

$\Pr(\text{"}even\text{"} \cap \text{"}odd\text{"}) = 0$

$\Pr(\text{"}even\text{"} \cup \text{"}odd\text{"}) = 1$

Probability | Independent events

Example: Rolling two fair 6-sided dice

Two events A and B are **independent** if knowing that one occurred tells us **nothing** about the probability of the other.

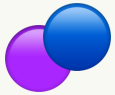
$$\Pr(\text{"}d2\ even\text{"}|\text{"}d1\ odd\text{"}) = \Pr(\text{"}d2\ even\text{"}) = 1/2$$

Formally, independence holds if and only if $\Pr(A|B) = \Pr(A)$

The probability of a sequence of independent events is $\Pr(A \cap B) = \Pr(A)\Pr(B)$

$$\Pr(\text{"}d2\ even\text{"} \cap \text{"}d1\ odd\text{"}) = \Pr(\text{"}d2\ even\text{"})\Pr(\text{"}d1\ odd\text{"}) = \frac{1}{2} * \frac{1}{2} = \frac{1}{4}$$

Probability | Dependent events



Example: Drawing two lottery balls (1-10) from an urn without replacement

Two events A and B are **dependent** if one occurring changes the probability of the other. Formally, $\Pr(A|B) \neq \Pr(A)$

$$\Pr(\text{"even"}|\text{"3"}) \neq \Pr(\text{"even"})$$

When A and B are dependent, then $\Pr(A \cap B) = \Pr(A)\Pr(B|A)$

$$\Pr(\text{"even"} \cap \text{"3"}) = \Pr(\text{"even"})\Pr(\text{"3"}|\text{"even"}) = \frac{5}{10} * \frac{1}{9}$$

From $\Pr(A \cap B) = \Pr(B \cap A)$, we can derive **Bayes' rule:** $\Pr(A|B) = \frac{\Pr(A)\Pr(B|A)}{\Pr(B)}$

$$\Pr(\text{"even"}|\text{"3"}) = \frac{\Pr(\text{"even"})\Pr(\text{"3"}|\text{"even"})}{\Pr(\text{"3"})} = \frac{(5/10)(1/9)}{1/10}$$

Probability | Random variables and distributions

So far, we have talked about probabilities of **events**.
In coming weeks, we will want to work with **numerical outcomes** of random processes.

Example: Tossing a fair coin $S = \{H, T\}$

$$X = \begin{cases} 1 \text{ if } H \\ 0 \text{ if } T \end{cases}$$

A **random variable** X assigns a numerical value to each outcome in the sample space S.

A **probability distribution** describes how probability is distributed over the values of a random variable.

For discrete random variables, this is the set of probabilities $\Pr(X = x)$ for all $x \in S$
We formalise this as the **probability mass function** (PMF) $f_X(x) = \Pr(X = x)$

$$f_X(x) = \begin{cases} 1/2 \text{ if } x = 1 \\ 1/2 \text{ if } x = 0 \end{cases}$$

Probability | Bernoulli distribution

Some probability distributions (usually those that are useful) have special names ✨

The **Bernoulli distribution** is the discrete probability distribution of a random variable X which takes the value 1 with probability p and the value 0 with probability $1 - p$.

$$f_X(k, p) = \begin{cases} p \text{ if } x = 1 \\ 1 - p \text{ if } x = 0 \end{cases}$$

This is just a generalised version of our single **coin**!

success vs. failure, correct vs. incorrect, etc.

Why is this useful? Many **real-world outcomes** can be modelled as binary variables. We will come back to this throughout the course.

... but what if we want to model a series of outcomes instead of a single event?

Probability | Binomial distribution

The **binomial distribution** is the discrete probability distribution of a random variable X that describes the number of successes in a sequence of n independent Bernoulli trials, where each trial succeeds with probability p and fails with probability $1 - p$.

$$f_X(k, n, p) = \Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \text{ where } \binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$

many coins!

It is also often useful to express $\Pr(X \leq k)$ using the **cumulative distribution function** (CDF).

$$F_X(k, n, p) = \Pr(X \leq k) = \sum_{i=0}^{k} \Pr(X = i) = \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i}$$

n correct answers -> % correct, etc.

Why do we care? Binomials will help us reason about counts and proportions.

Probability | Continuous random variables

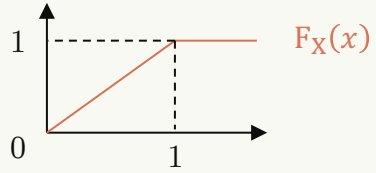
All previous examples were discrete random variables (coins, dice, lottery balls).
However, in this course, we will also encounter a lot of **continuous random variables**.

Example: Selecting a real number uniformly at random from the interval [0,1]

Here, we can only assign probabilities over intervals, not exact values.
We do this using a **probability density function** (PDF) instead of the discrete PMF.

For the **uniform distribution** in our example:

$$f_X(x) = \begin{cases} 1, & 0 \le x \le 1 \\ 0, & \text{otherwise} \end{cases} \quad \text{and} \quad F_X(x) = \begin{cases} 0, & x < 0 \\ x, & 0 \le x \le 1 \\ 1, & x > 1 \end{cases}$$



And now we have all ingredients to cover the most important probability distribution 

Probability | Normal distribution

The **normal distribution** is a special type of continuous probability distribution for a random variable X with mean μ and variance σ^2.

Why is it important? The normal distribution arises naturally as a model for many continuous measurements and as an approximation to more complex random processes.

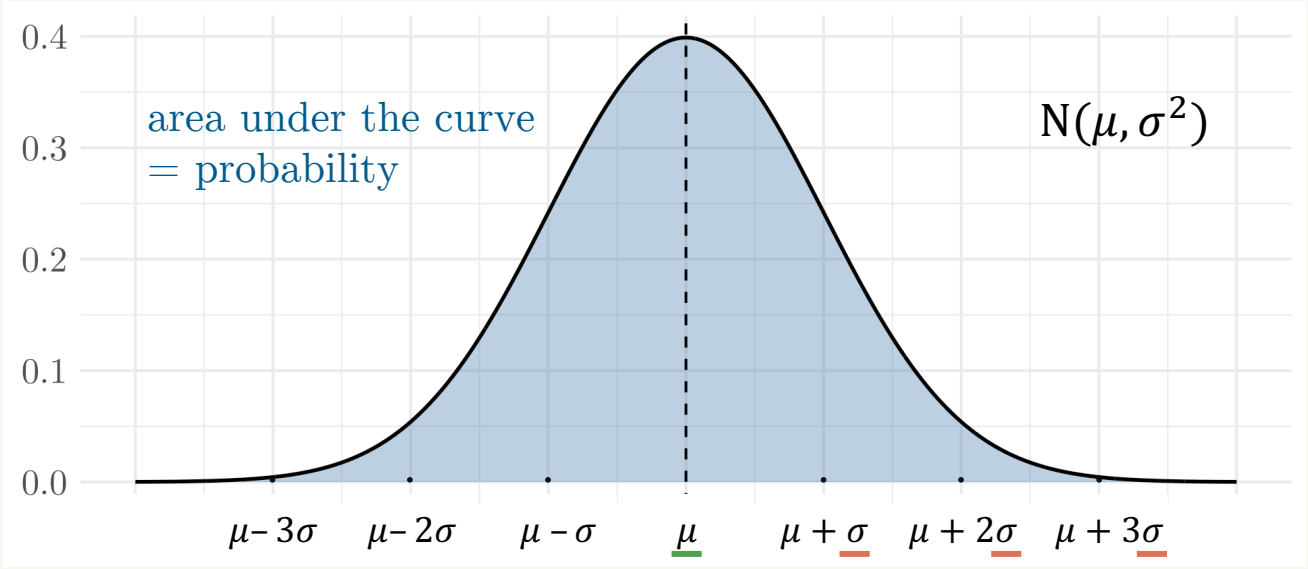
A normal distribution $\mathrm{N}(\mu, \sigma^2)$ is described by the PDF $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(x-\mu)^2}{2\sigma^2}}$

The normal CDF $\mathrm{F}_X(x)$ is generally more useful but has no closed form expression.

$\mathrm{F}_X(x) = \Pr(\mathrm{X} \leq \mathrm{x}) = \Phi\left(\frac{\mathrm{x}-\mu}{\sigma}\right)$ where $\frac{\mathrm{x}-\mu}{\sigma}$ is known as the z-score.

measures how many standard deviations x lies from the mean.

Probability | Anatomy of the normal distribution



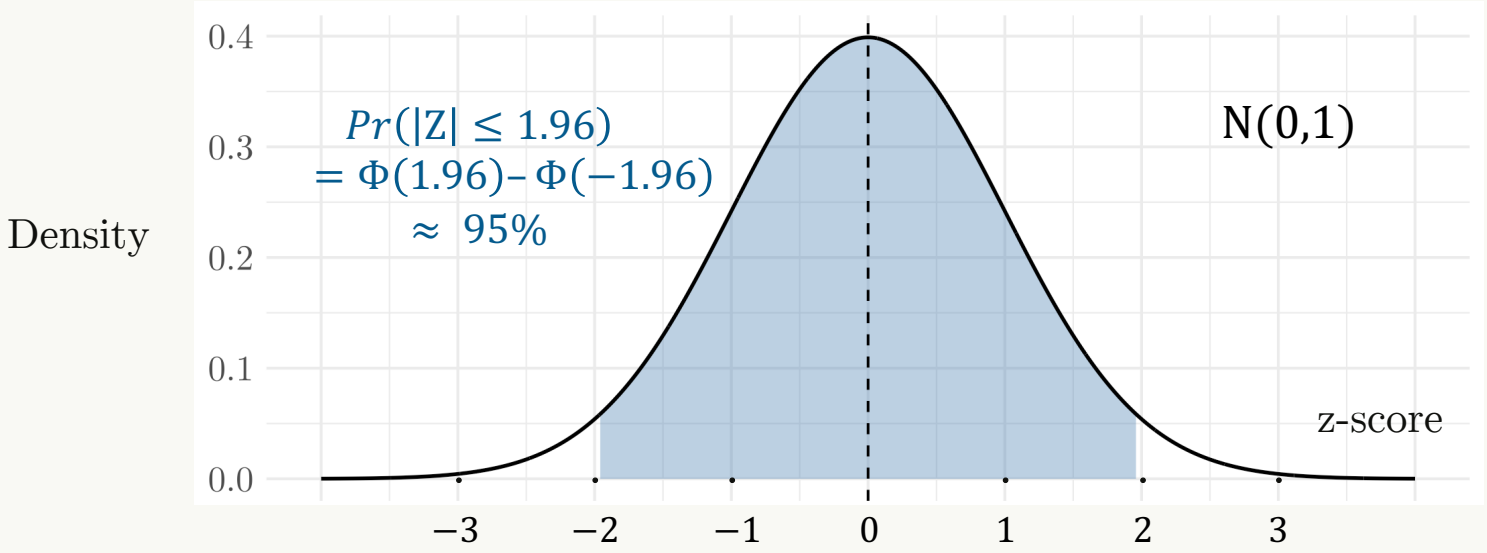
Symmetric around mean μ

larger σ = more spread

Remember: this is a **probability distribution** defined by the PDF, not an empirical distribution.

Probability | Standard normal distribution and z-scores

The case of $\mu = 0$ and variance $\sigma^2 = 1$ is referred to as the **standard normal distribution**.



The standard normal CDF will be a key ingredient for statistical inference!

Recap | Key takeaways from today

We study samples to learn about populations.
But what we can credibly conclude depends on how our data were collected and measured.

Describing data is a necessary first step.
Visualisation and summary stats help us understand distributions, variation, and anomalies.

Summary statistics compress information.
They are useful but can be misleading, so need to be interpreted in context, with caution.

Probability provides the language of uncertainty.
It allows us to reason about random variation and prepares us for statistical inference.

Homework | Formulate a research question

This course is about giving you the tools to answer the questions that you care about.

By 10am next Monday: Submit a research question related to your research interests that you think can be answered using analytical statistics.

This will **help you** start thinking about your summative.
This will **help me** better understand your what you want to use stats for.

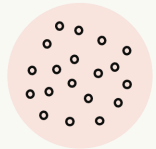
Note:
- You are not tied to the question you submit. It will not be graded.
- Consider feasibility (e.g. data access) but don't give it too much weight at this point.

Submit your question on **Canvas** ("Assignments" -> "Week 1 Homework").

Next week | Statistical inference and uncertainty

Next week, we will **bring together statistics and basic probability theory**

This will allow us to move away from just describing our sample,
towards **drawing conclusions about wider populations**.



📚 **Readings**: C4 RegOS and C3 ArtOS will be particularly useful.