Lecture 3: Hypothesis Testing

Paul Röttger

Applied Analytical Statistics

3rd of February 2026

Housekeeping | Tutorials

Tutorial attendance was very low last week.

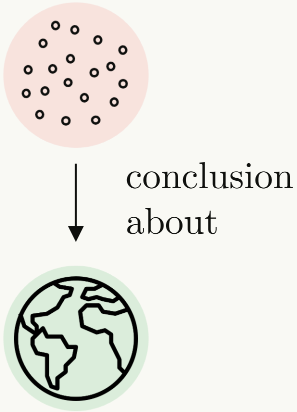
What can we do to make the live tutorial sessions more useful to you?

🤔

Plan for today | Hypothesis testing

Today we move from estimation to decision-making
We move from quantifying uncertainty to formally evaluating claims about populations.

1. **Hypotheses**: null and alternative claims about populations
2. **Null distributions**: sampling distributions under the null
3. **Test statistics**: standardised distance from the null
4. **Decision rules**: P-values and significance levels
5. **Common hypothesis tests**: means, proportions, categorical associations
6. **Bootstrap testing**: computational method, connection to CIs


conclusion
about

We will again finish with a class activity, focusing on the data for your summative.

Hypothesis testing | What is a hypothesis?

Statistical hypotheses are **claims** about population parameters.

→ fixed but unknown descriptors of our target population

We typically formulate two competing claims:

The **null hypothesis** H_0: a specific claim, often about the absence of an effect or relationship.

e.g.: H_0: % of UK adults using AI at least once per week $= 30\%$

The **alternative hypothesis** H_A: the complement of the null hypothesis.

e.g.: H_A: % of UK adults using AI at least once per week $\neq 30\%$

These claims are mutually exclusive and together capture all possible outcomes.

Hypothesis testing | Back to the sampling distribution

In week 2, we learned about the **sampling distribution**, which is the distribution of a statistic across repeated i.i.d. samples from the same population.

Sampling distributions are described by **fixed but unknown population parameters**, so we relied on large sample theory (LLN, CLT) to approximate these distributions.

Hypothesis testing **flips the perspective**. Instead of asking how the statistic varies under the true (unknown) population, we assume a null hypothesis that **fixes the parameter value**.

e.g.: H_0: $p = 30\%$

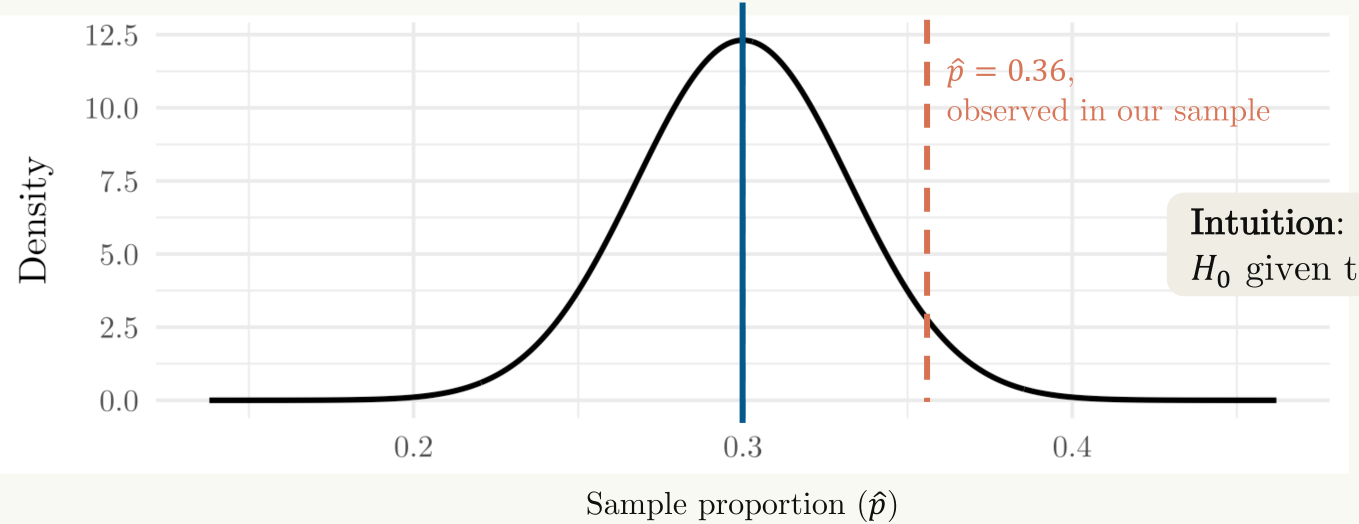
We then **study the sampling distribution of the statistic under the null hypothesis** and ask: How unusual is the statistic we observed in our sample, if the null were true?

We broadly make the same assumptions: observations are i.i.d. draws from same population.

Hypothesis testing | Sampling distribution “under the null”

Data: Self-reported AI usage data from a representative survey of 200 UK adults.

By the CLT: $\hat{p} \sim \mathrm{N}\left(p_0, \frac{p_0(1-p_0)}{n}\right)$ where $\boldsymbol{p_0} = \mathbf{30\%}$ under the null.



$\hat{p} = 0.36$, observed in our sample

Intuition: How plausible is H_0 given that we observed $\hat{p}$?

Hypothesis testing | Test statistics

A **test statistic** measures the distance between our **sample statistic** and the **null value**.

e.g. sample mean $\hat{p} = 0.36$,

General form: test statistic $= \dfrac{\text{estimate} - \text{null value}}{\text{standard error}}$

This produces a **standardised scale:** distance in standard errors from the null value.

In our AI usage example: $\hat{p} = 0.36$, $p_0 = 0.30$, $\text{n} = 200$

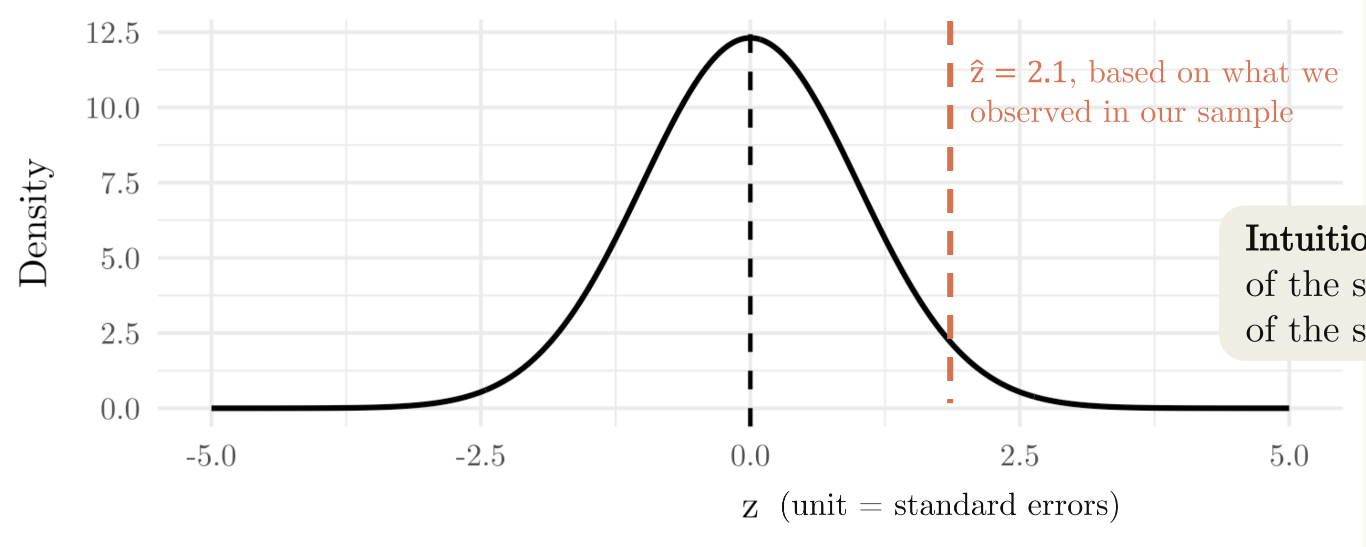
Then the standard error under the null: $SE_0(\hat{p}) = \sqrt{\dfrac{p_0(1-p_0)}{n}}$

And thus the z-statistic: $\text{z} = \dfrac{\hat{\text{p}} - \text{p}_0}{\text{SE}_0(\hat{\text{p}})} = \dfrac{0.36 - 0.30}{\sqrt{0.3(0.7)/200}} \approx 2.1$

🤔 Why do we use p_0 to calculate $\text{SE}_0(\hat{\text{p}})$?

Hypothesis testing | Null distribution of the test statistic

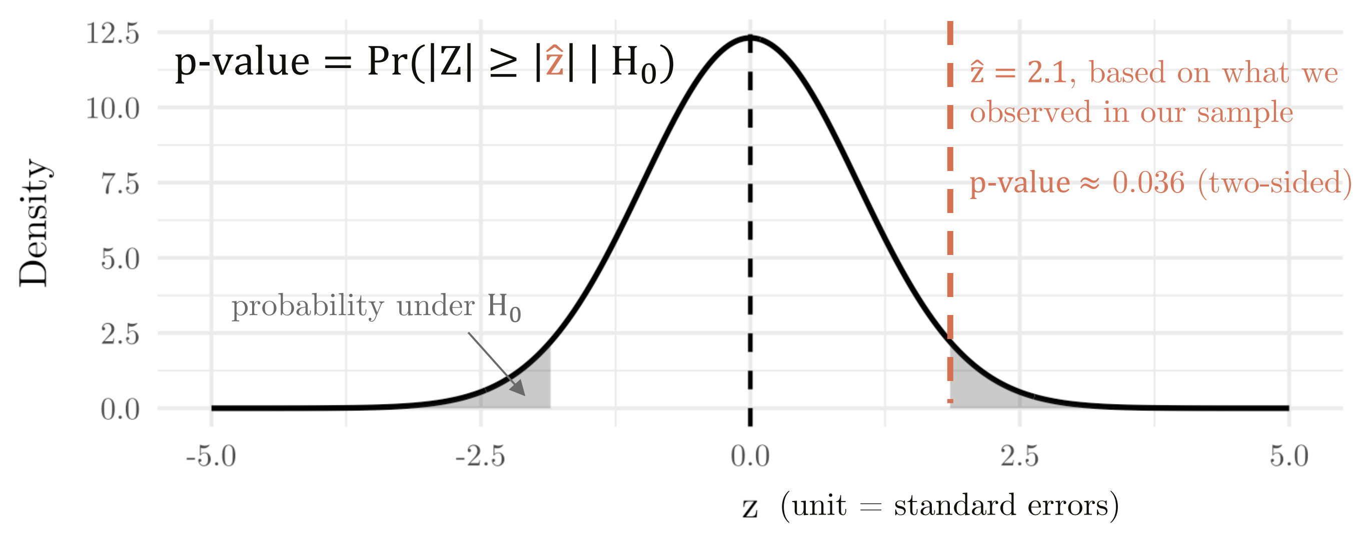
The **null distribution** is the sampling distribution of the test statistic when H_0 is true. The test statistic is a standardised version of our sample statistic (e.g. sample mean).



Intuition: Rescaled version of the sampling distribution of the sample statistic.

Hypothesis testing | P-values

The **p-value** is the probability, **under the null hypothesis**, of observing a test statistic **at least as extreme as the one we observed** in repeated i.i.d. draws from the same population.



Hypothesis testing | Significance levels and decision rules

The **significance level** α is a **pre-specified threshold** for how much evidence against H_0 we require to reject it. → design choice before observing the data

Decision rule:
If p-value $\leq \alpha$: **reject** H_0
If p-value $> \alpha$: **do not reject** H_0
→ Failing to reject H_0 does not mean H_0 is true!

α controls what **probability of falsely rejecting H_0** we are willing to accept.
→ false positive rate

$\alpha = 5\%$ is a historical convention, not some magic threshold!

Hypothesis testing | Decision errors

The true state of the world is fixed but unknown.
There are four possible outcomes to a hypothesis test.

	H_0 true	**H_0 false**
Reject H_0	Type I error (α) = False Positive	Correct decision
Do not reject H_0	Correct decision	Type II error (β) = False Negative

Power $= 1 - \beta$
(will cover later)

We control long-run Type I error rate directly by choosing a significance level α.
Lower α (stricter threshold) decreases risk of false positives, increases risk of false negatives.
Higher α (looser threshold) increases risk of false positives, decreases risk of false negatives.

Hypothesis testing | General template

1. Choose a test statistic.

Sample statistic (e.g. Δ in means) standardised using standard error.

2. State the hypotheses.

Null hypothesis $\mathrm{H_0}$ (data-generating assumption) and alternative hypothesis $\mathrm{H_A}$.

3. Derive the null distribution.

Sampling distribution of the test statistic under the assumption that $\mathrm{H_0}$ is true.

4. Compute the p-value.

Probability of observing a test statistic at least as extreme as the data, under $\mathrm{H_0}$.

5. Draw a conclusion

Compare p-value to significance level α. Reject or do not reject $\mathrm{H_0}$.

Hypothesis testing | Tests by data type

Continuous outcomes (means)

1. **One-sample t-test**
   → Is a population mean equal to a hypothesised value?
2. **Two-sample t-test (independent samples)**
   → Do two groups have different population means?
3. **Two-sample t-test (paired samples)**
   → Is the mean of the paired differences equal to zero?

Categorical outcomes (counts / proportions)

1. **Chi-squared test of independence**
   → Are two categorical variables associated?

All tests follow the same general template described on the previous slide.

Tests for continuous data | One-sample t-test

Data: Human quality ratings (0-100 scale) for LLM-generated answers to 50 questions.

We assume observations $X_1, \dots, X_n$ are i.i.d. draws from the same population.

average rating across all human ratings

The parameter of interest is the **population mean** $\boldsymbol{\mu}$.
The sample statistic is the **sample mean** $\overline{\boldsymbol{X}}$

average rating across our observed set of human ratings

Hypotheses:
$H_0: \mu = \mu_0$
$H_A: \mu \neq \mu_0$

🤔 This is for a **two-sided** test. What would H_0 and H_{A} be for a **one-sided** test?

Not normal!

Test statistic: $t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$ where under H_0 $t \sim t_{n-1}$.

Tests for continuous data | z-test vs. t-test

IF population SD σ were known, we could use $z = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$ where under H_0 by CLT $z \sim N(0,1)$.

→ **z-test**

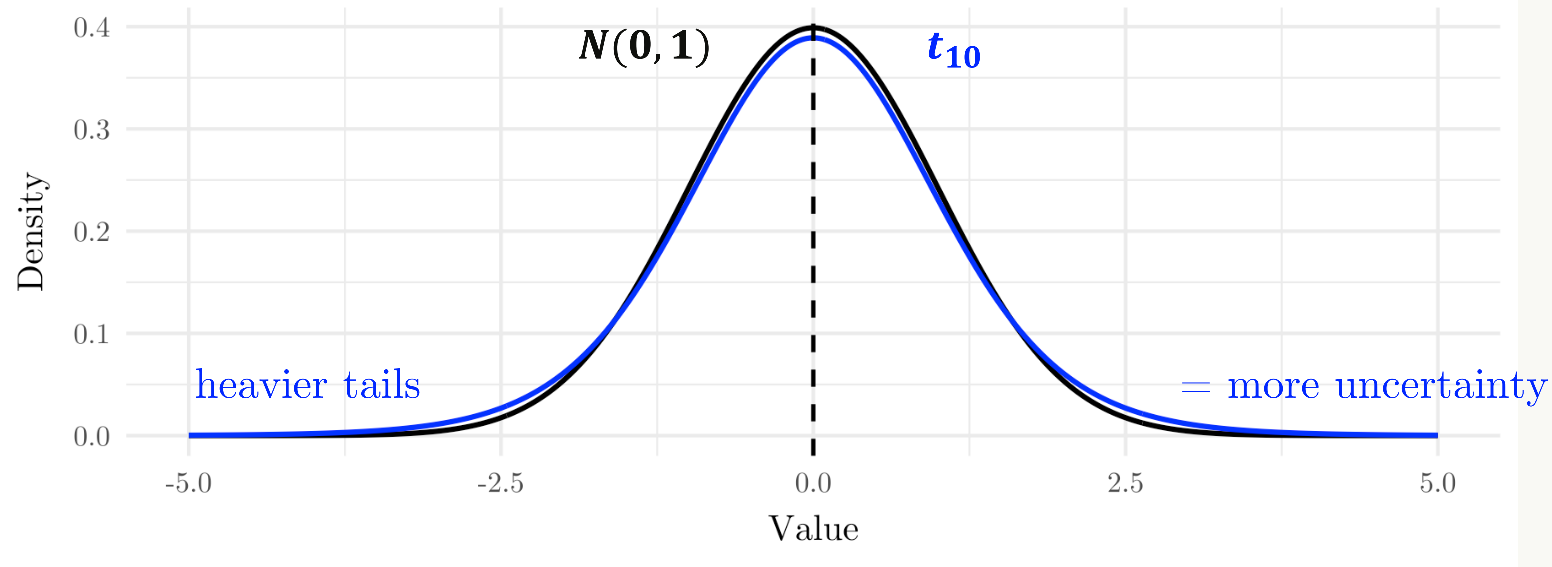
In practice, we **estimate variability from our sample**, replacing σ with sample SD s, to produce the estimated standard error $SE(\bar{X}) = \frac{s}{\sqrt{n}}$

This creates additional uncertainty, which we need to account for in inference:
Our new test statistic $t = \frac{\bar{X}-\mu_0}{s/\sqrt{n}}$ under H_0 follows a **t-distribution** $t \sim t_{n-1}$

→ **intuition**: distance in standard errors

degrees of freedom: how many independent pieces of information remain after estimating parameters from the data

Tests for continuous data | The t distribution



As $n \to \infty$, s becomes a precise estimate of σ. Therefore $\boldsymbol{t} \to N(0,1)$
This is why z-tests and t-tests give nearly identical results in large samples.

Tests for continuous data | Independent two-sample t-test

Data: Human ratings (0-100) for answers from LLM A vs. LLM B for 50 questions.

The parameter of interest is the **difference in population means** $\mu_1 - \mu_2$
The sample statistic is the **difference in sample means** $\bar{X}_1 - \bar{X}_2$

Hypotheses:
$H_0: \mu_1 - \mu_2 = 0$
$H_A: \mu_1 - \mu_2 \neq 0$

In practice, **we compare** two estimates, which creates additional uncertainty.

This is reflected in our two-sample t-statistic: $t = \frac{(\bar{X}_1 - \bar{X}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

→ "Welch's t-test" allows for unequal variance

Tests for cont. data | Deriving Welch's two-sample t-statistic

By independence and the CLT: $\overline{X}_1 - \overline{X}_2 \approx N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$

If σ_1^2 and σ_2^2 were known: $Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$

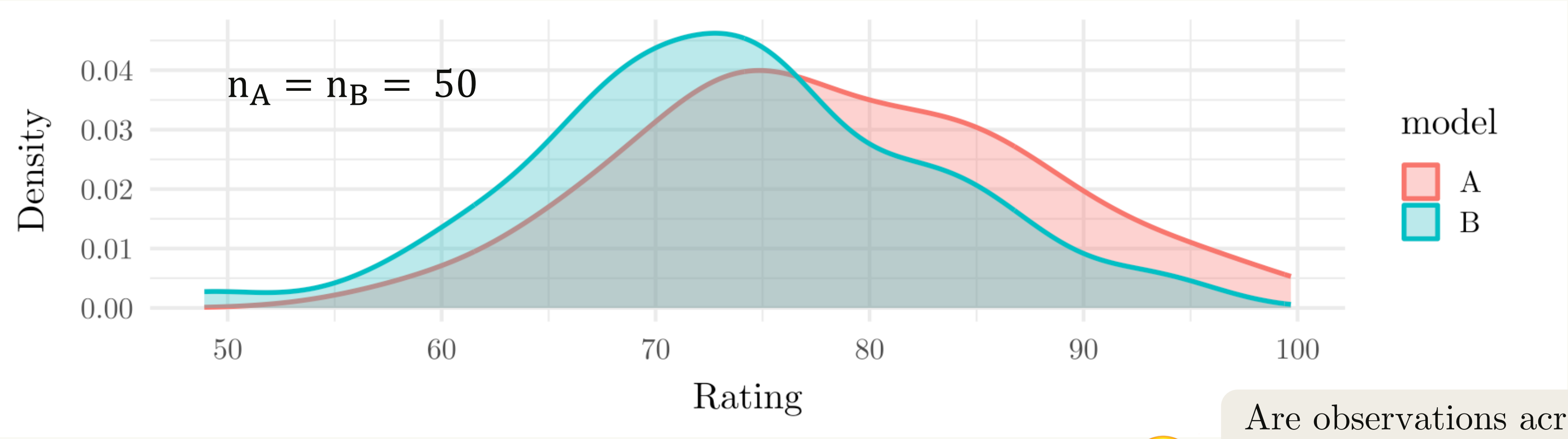
In practice, $SE(\overline{X}_1 - \overline{X}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

This yields the **two-sample t-statistic**: $t = \frac{(\overline{X}_1 - \overline{X}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t_{df}$

degrees of freedom given by
computational approximation

Tests for continuous data | Example of two-sample t-test

Data: Human ratings (0-100) for answers from LLM A vs. LLM B for 50 questions.



🤔 Are observations across groups really independent?

$\overline{X}_A = 78.34, s_A^2 = 85.72$ $\quad$ $\overline{X}_B = 73.46, s_B^2 = 81.98$

$t = (\overline{X}_A - \overline{X}_B)/\sqrt{s_A^2/n_A + s_B^2/n_B} = 2.66$ where under H_0 $t \sim t_{97.95}$ so that $p = 0.009 < \alpha$

Tests for continuous data | Pooled standard error

Welch's t-test allows for samples with unequal variance. **This should be our default.** However, it is sometimes convenient to assume equal population variance: $\sigma_1^2 = \sigma_2^2$

We can then combine information from both groups to estimate a **single shared variance**:

$$s_{\text{pooled}}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Standard error becomes: $SE_{\text{pooled}}(\bar{X}_1 - \bar{X}_2) = \sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$

The test statistic becomes: $t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2}$

Tests for continuous data | Paired two-sample t-test

Data: Human ratings (0-100) for answers from LLM A vs. LLM B for the same 50 questions.

For **paired data**, we observe linked observations across two groups (X_{i1}, X_{i2}) for $i = 1, \dots, n$.
We can make use of this structure to run a **more powerful single-sample t-test**.

Let $D_i = X_{i1} - X_{i2}$ describe within-pair differences.
The parameter of interest is now the **population mean of these differences** μ_D

Our hypotheses, as in the one-sample test: $H_0: \mu_D = 0$, $H_A: \mu_D \neq 0\ (two-sided)$

Test statistic: $t = \frac{\bar{D}-0}{s_D/\sqrt{n}}$ where under H_0 $t \sim t_{n-1}$.

🤔 Are we right to use a paired t-test for the example above?

Pairing removes between-unit variation, which reduces uncertainty.

Tests for categorical data | Goodness of fit test

Data: Counts of users reporting their primary social media platform.

We observe **one categorical variable** with K categories, counting $O_1, \dots, O_K$.

We want to test **goodness of fit**: observed distribution vs. hypothesised distribution

Hypotheses: → from other data source (e.g. Ofcom data on platform usage)

H_0: Category probabilities equal specified values $(p_1, \dots, p_\mathrm{k})$

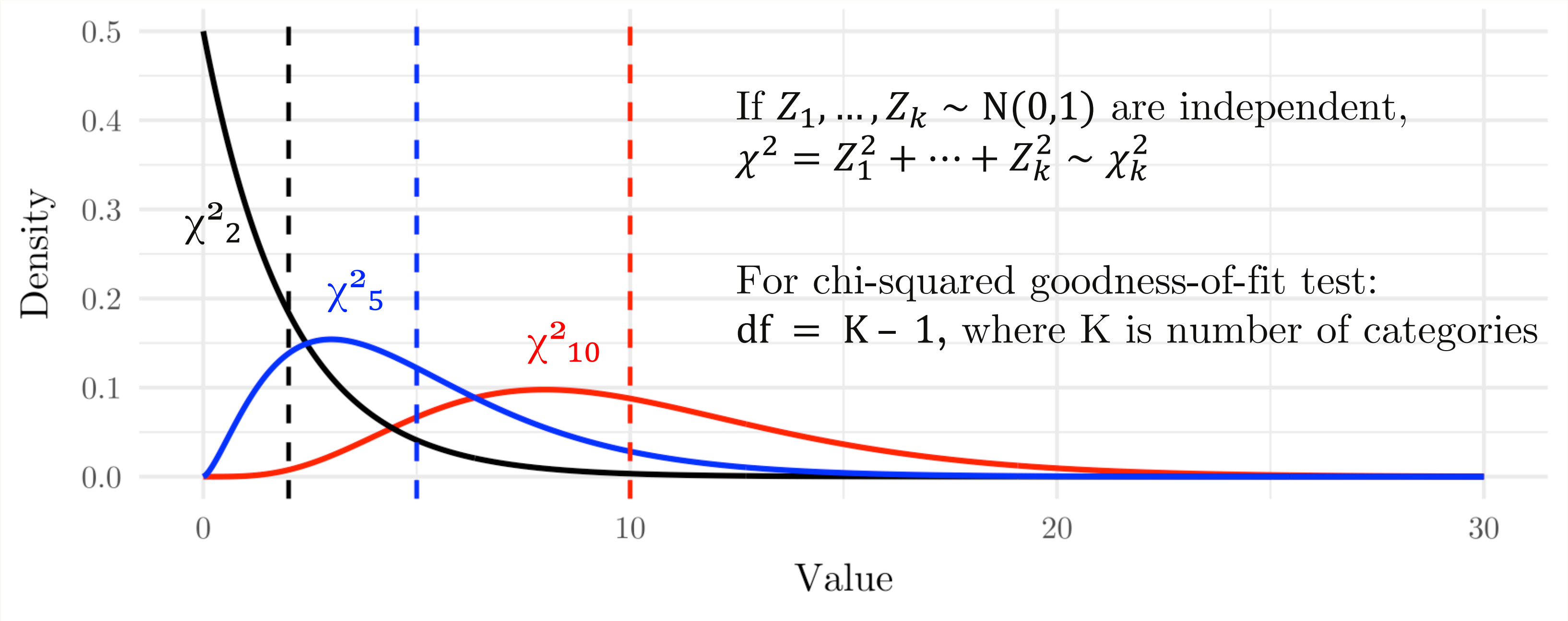
H_A: At least one category probability differs

Under H_0, the **expected count** in category k is $E_k = n \cdot p_k$

The chi-squared test statistic $\chi^2 = \sum_{k=1}^{K} \frac{(\mathrm{O}_k - E_k)^2}{E_k}$ where under H_0 $\chi^2 \sim \chi^2_{K-1}$

Tests for categorical data | The chi-squared distribution

χ^2 distribution describes the null distribution of **sums of squared, standardised deviations**.



If $Z_1, \dots, Z_k \sim \mathrm{N}(0,1)$ are independent,
$\chi^2 = Z_1^2 + \cdots + Z_k^2 \sim \chi_k^2$

For chi-squared goodness-of-fit test:
df = K – 1, where K is number of categories

Tests for categorical data | Contingency tables

Data: Tweets in 4 languages labelled for hate speech (Tonneau et al., 2024).

We often want to study the relationship between **two categorical variables**.

Language	**Hateful**	**Offensive**	**Neither**	**Total**
English	87	1,816	28,097	30,000
French	220	1,318	28,462	30,000
Spanish	177	2,208	27,615	30,000
Turkish	177	1,320	28,503	30,000
Total	661	6,662	112,677	**120,000**

🤔

Does language predict label?

Is the label distribution the same across languages?

Formally, we test whether our two variables are **statistically independent**: H_0: $X \perp\!\!\!\perp Y$

Tests for categorical data | Observed vs. expected counts

As for goodness-of-fit, we compute **expected cell counts** $\boldsymbol{E_{ij}}$.

Under H_0, these are now determined by marginals: $\mathrm{E_{ij}} = \dfrac{(\text{row total}_\text{i})(\text{column total}_\text{j})}{\text{n}}$

Language	Hateful	Offensive	Neither	Total
English	87	1,816	28,097	30,000
French	220	1,318	28,462	30,000
Spanish	177	2,208	27,615	30,000
Turkish	177	1,320	28,503	30,000
Total	661	6,662	112,677	**120,000**

$$\mathrm{E_{English,Hateful}} = \frac{30\text{k}*661}{120\text{k}} = 165.25$$

$$\mathrm{O_{English,Hateful}} = 87$$

🤔 Are the discrepancies **across all cells combined** too large to attribute to chance?

Tests for categorical data | Chi-squared test of independence

Test statistic $\chi^2 = \sum_{\mathrm{i}} \sum_{\mathrm{j}} \frac{(\mathrm{O_{ij}} - E_{\mathrm{ij}})^{\mathbf{2}}}{E_{\mathrm{ij}}}$

→ Aggregates discrepancy between observed and expected counts across the table.
→ Penalises large deviations relative to expected counts.

Under H_0: $\chi^2 \sim \chi^2_{(\boldsymbol{R}-1)(\boldsymbol{C}-1)}$ where $\boldsymbol{R}$ = No. of row categories, $\boldsymbol{C}$ = No. of column categories

→ Degrees of freedom reflect how many cell counts are free to vary
→ Marginal totals impose constraints

In our **example** (previous slide): $\chi^2_{\text{obs}} = 409.44$, where $\chi^2_{\text{obs}} \sim \chi^2_6$
$\mathrm{p} = \Pr(\chi^2 \geq \chi^2_{\text{obs}}" \mid H_0) \;=\; 0.000$, so that $\mathrm{p} < \alpha$

→ Reject H_0, conclude highly significant association between language and label.

Hypothesis testing | Connection to confidence intervals

Last week, we computed **confidence intervals** around sample statistics to express:
“Which parameter values are plausible, given our data?”

This week, we used **hypothesis tests** to answer questions like:
“Is this specific parameter value plausible, given our data?”

→ same inferential logic,
differently framed

We can **invert a hypothesis test to produce a confidence interval**:

- Fix a significance level, e.g. $\alpha = 5\%$.
- A parameter value μ_0 is not rejected if $\frac{\hat{\mu}-\mu_0}{\mathrm{SE}(\hat{\mu})}$ is not too large in absolute value.
- The set of all μ_0 values not rejected forms a $(1 - \alpha)$ confidence interval.

For two-sided tests, we can reject H_0 at 95% confidence if μ_0 lies outside the 95% CI!

🤔 Why does this not work for one-sided tests?

Bootstrap | Application to hypothesis testing

Last week, we used the bootstrap to **approximate sampling distributions around the observed statistic**, which helped us construct confidence intervals.

For hypothesis testing, we use the bootstrap to **approximate the null distribution**, i.e. the sampling distribution of the test statistic when H_0 is true.

- We treat our sample as a stand-in for the population **under H_0**.
- **We modify the data so that H_0 is true by construction.**
- We repeatedly draw samples with replacement.
- We recompute the test statistic for each resample.
- We use the resulting “bootstrap null distribution" of the statistic as a stand-in for the unobserved sampling distribution **under H_0**.

Bootstrap | Example of two-sample t-test

Data: Human ratings (0-100) for answers from LLM A vs. LLM B for 50 questions.

H_0: $\mu_A = \mu_B$

1. Compute observed difference in means, our sample statistic: $\bar{X}_A - \bar{X}_B$
2. Compute observed means in each group and overall: $\bar{X}_A$, $\bar{X}_B$, $\bar{X}_{A \cup B}$
3. **Shift both groups to a common mean** while preserving within-group variability:
   For each observation, subtract the corresponding group mean and add overall mean $\bar{X}_{A \cup B}$
4. Resample from the shifted data, calculating $\bar{X}_A - \bar{X}_B$ for each resulting bootstrap sample.
5. Compute the bootstrap p-value as the % of bootstrap sample statistics that are larger in absolute value than your observed sample statistic.

OR: compute 95% bootstrap CI around statistic, reject H_0 at $\alpha = 5\%$ if 0 lies outside CI.

Hypothesis testing | Correcting for multiple comparisons

When running many hypothesis tests, **some tests will be significant by chance alone**.
This is because we set $\alpha = 0.05$ for each test. Running multiple tests inflates Type I error.

To mitigate this risk and **control overall error rate**, we can **adjust the per-test α**.

Bonferroni adjustment

If we run m tests, use $\alpha_{\text{adjusted}} = \frac{\alpha}{m}$. Reject H_0 only if $p \leq \alpha/m$.

Holm-Bonferroni adjustment

Sort p-values from smallest to largest: $p_{(1)}, \ldots, p_{(m)}$. Compare $p_{(k)} \leq \frac{\alpha}{m-k+1}$.

Hypothesis testing | Practical vs. statistical significance

⚠ Results can be **statistically significant** but **practically insignificant**!

Statistical significance asks: “Is this result **distinguishable from random variation**?”
This depends on sample size, variability, significance level.

Practical significance asks: “Is this result **large enough to matter in the real world?**”
This depends on context, domain knowledge, ...

Always keep this in mind when interpreting the results of statistical analyses.
In a very large sample, almost any two variables have a statistically significant association.

Hypothesis testing | Effect sizes

To help gauge **practical significance**, we can compute **effect sizes**.

Raw effects

Directly reporting difference in means, proportions, etc.

Standardised effects

Cohen's d: $d = \frac{\bar{X}_1 - \bar{X}_2}{s_{\text{pooled}}}$ which takes values between 0 and 1.

Rule of thumb: $d \approx 0.2$ is small, $d \approx 0.5$ is medium, $d \approx 0.8$ is large.

→ other effect sizes: odds ratios for proportions, Cramer's V for categorical data, ...

Recap | Key takeaways from week 3

Hypothesis testing evaluates sample evidence against a null model.
We compute how compatible our observed statistic is with a specific population claim.

Inference proceeds via sampling distributions under the null.
Test statistics and p-values are defined assuming the null hypothesis is true.

Significance levels formalise decision-making under uncertainty.
Choosing α controls long-run Type I error rate, not the probability that H_0 is true.

Statistical significance does not equal practical significance.
Effect sizes and context determine whether results matter in the real world.

Class activity | Group assignment based on your RQs

G1: **Language, Communication, and Bias in AI & Media** – Caleb Agoha, Noha Mahgoub, Yunjia Qi

G2: **AI, Generative Models, and Evaluation** – Max Davy, Howard Leong, Audrey Yip

G3: **Media, Platforms, and Audience Response** – Sophie Bair, Charlotte Peart, Michi Wong

G4: **Po** **Institutions** – Celikhan Baylan, Grahm Gaydos, Caleb Tan

G5: **So** **Adoption** – Min Jung, Mia Kussman, Isaac Backer

G6: **Health, Medicine, and Neuroscience** – Amelia Mercado, Laura Wegner, Ines Trichard

G7: **Education, Labor, and Socioeconomic Outcomes** – Rehmat Arora, Yilin Qian, Yue Zhang

G8: **Culture, Mobility, Lifestyle** – Teo Canmetin, Alena Tsvetkova, Fucheng Wang, Nesma Hammouda

Same groups as last week!

Everyone not named: please get together in groups of 3.

Class activity | Overview

Please access the Week 3 Class Activity Google Doc on Canvas.

Last week we focused on research questions and conceptual challenges.

Today, we focus on the **data** that you plan to use to answer your research question.
Data is **crucial for the feasibility of your summative**. NO DATA = NO PROJECT.

Looking for data sources? Check out suggestions on the summative description → Canvas

---

In the next class, we will do a **peer feedback session** on your **half-page summative proposal** to prepare you for the submission deadline on **February 13th** (next week Friday!).