Methods in Human Geography

Quantitative Methods: Statistical Analysis I

Dr Justin van Dijk
j.t.vandijk@ucl.ac.uk

UCL

This week

Part I

- Population and samples.
- Testing hypotheses.

Part II

- Parametric tests.
- Non-parametric tests.
- Novel data sources.

Population and samples

Populations and samples

- Problem:

  We want to know something about a population but we do not have complete data.

- Solution:

  Use samples and draw statistical inferences from them about their ‘parent’ population.

Populations and samples

- Problem:

  With sample data, exact results will always vary from sample to sample.

- Solution:

  Use tests of statistical significance to assess levels of certainty with our sample.

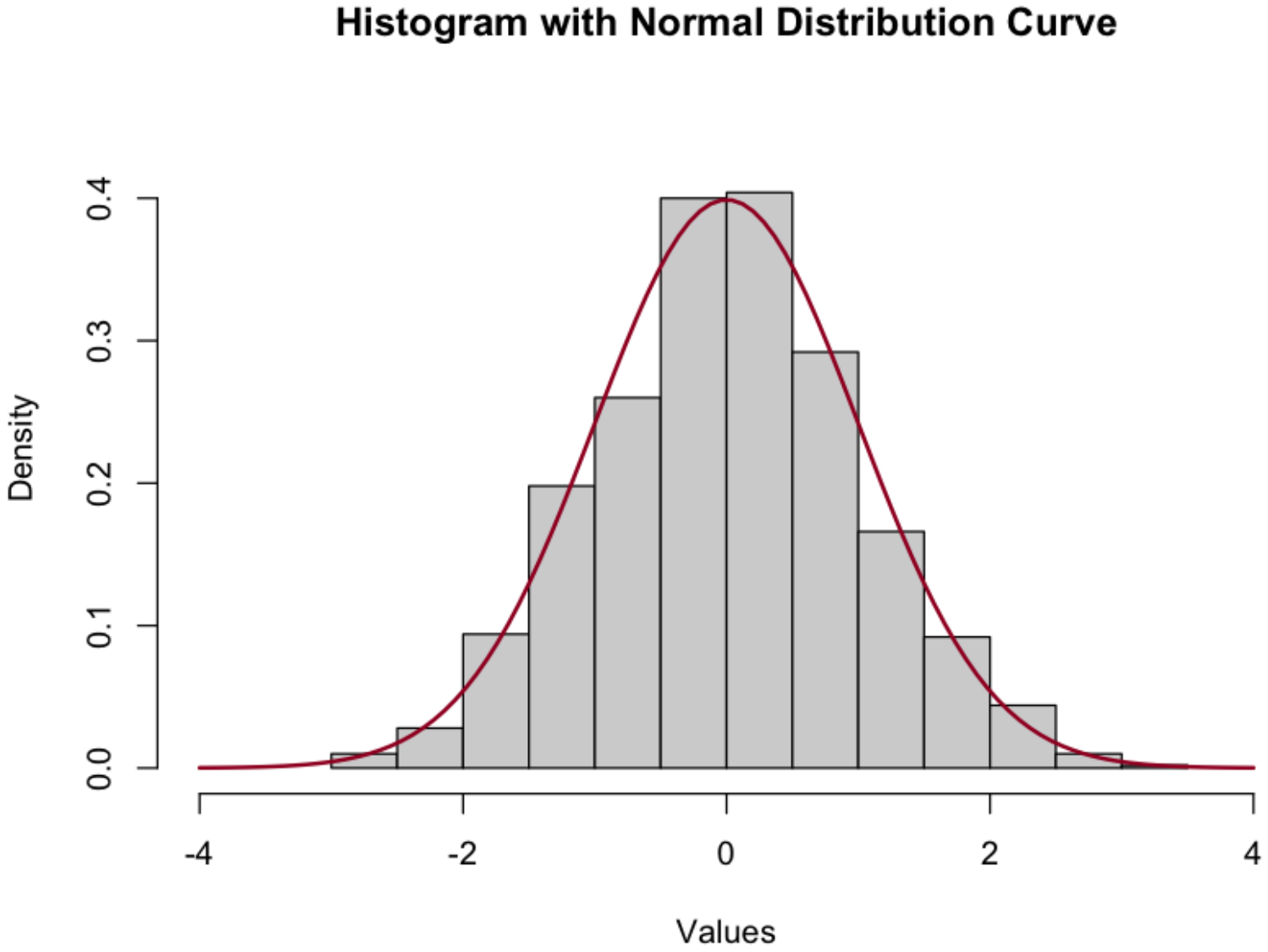
Statistical significance

- Harris (2016: 96): "*The basic idea of statistical testing is simple enough. It is used to test whether some level of difference (or association) between two or more variables has arisen by chance.*"
- How do we test for statistical significance?

Sampling distribution of means

- The sampling distribution of means is the probability distribution of all possible sample means obtained from a given population, assuming repeated sampling of a specific sample size.
- The mean of the sampling distribution of the means is equal to the population mean.

Sampling distribution of means

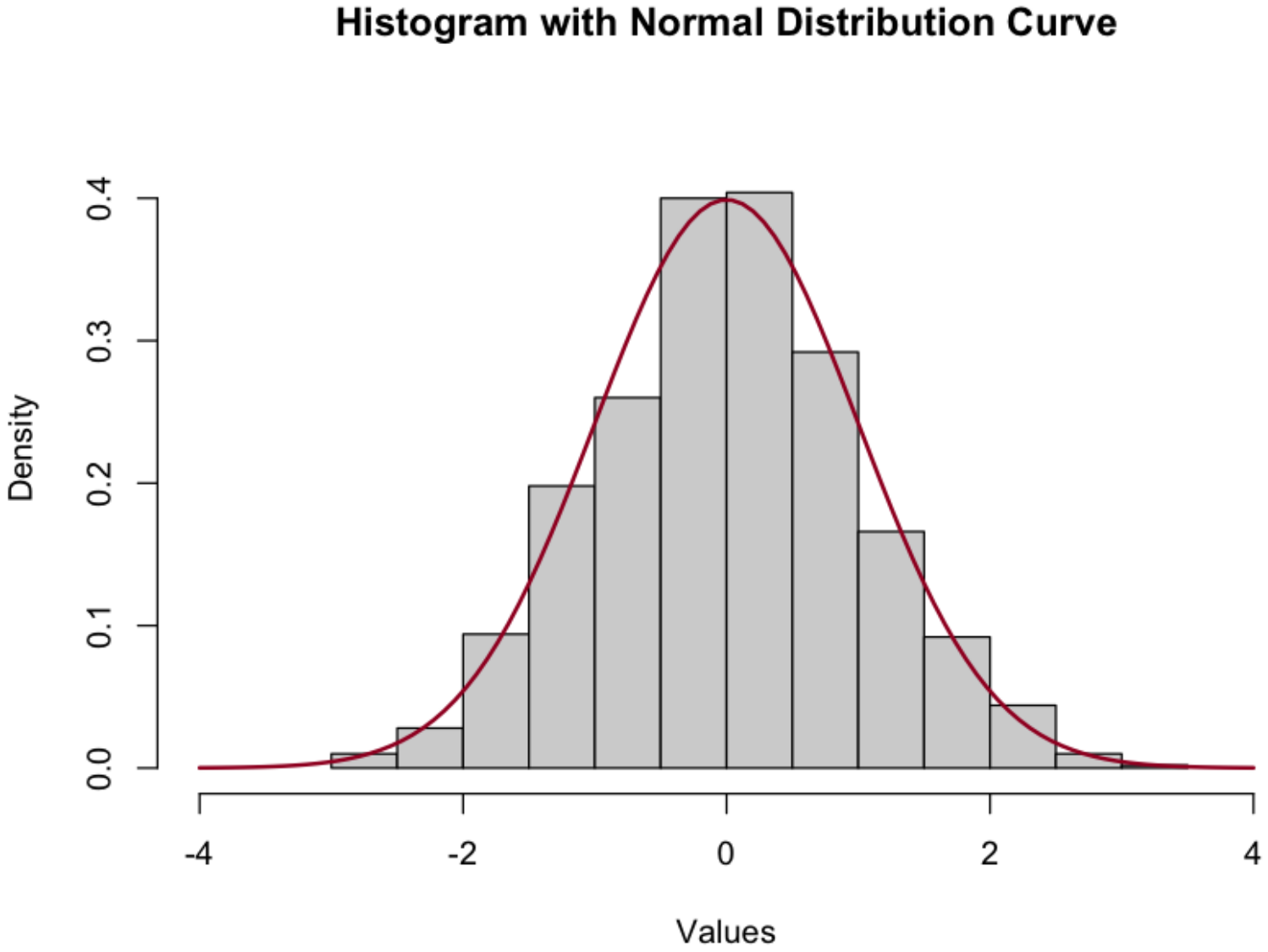


Sampling distribution of means

- The mean age of people in England and Wales in 2021 Census was 46.81.
- If we took four samples of 200 people, we may find the following means:

  $\bar{x}$ = 46.57; $\bar{x}$= 43.54; $\bar{x}$= 47.32; $\bar{x}$= 44.98.
- The average of these mean scores is 45.61.
- If we could take every possible sample, all those means would eventually average out to 46.81: the mean of means.

Sampling distribution of means



Sampling distribution of means

- The sampling distribution of the mean is theoretical and represents the distribution we would get if we could take every possible sample of a given size from the population and calculate their means.
- Even though we normally only collect one sample, the concept of the sampling distribution is useful because it allows us to predict the variability of that sample mean relative to the population mean.

Standard error

- The standard error of the mean (SEM) is the standard deviation of the distribution of the sample means:

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$$

- Unless population standard deviation is known, SEM is an estimate of the standard deviation of the entire sampling distribution based on our one sample.

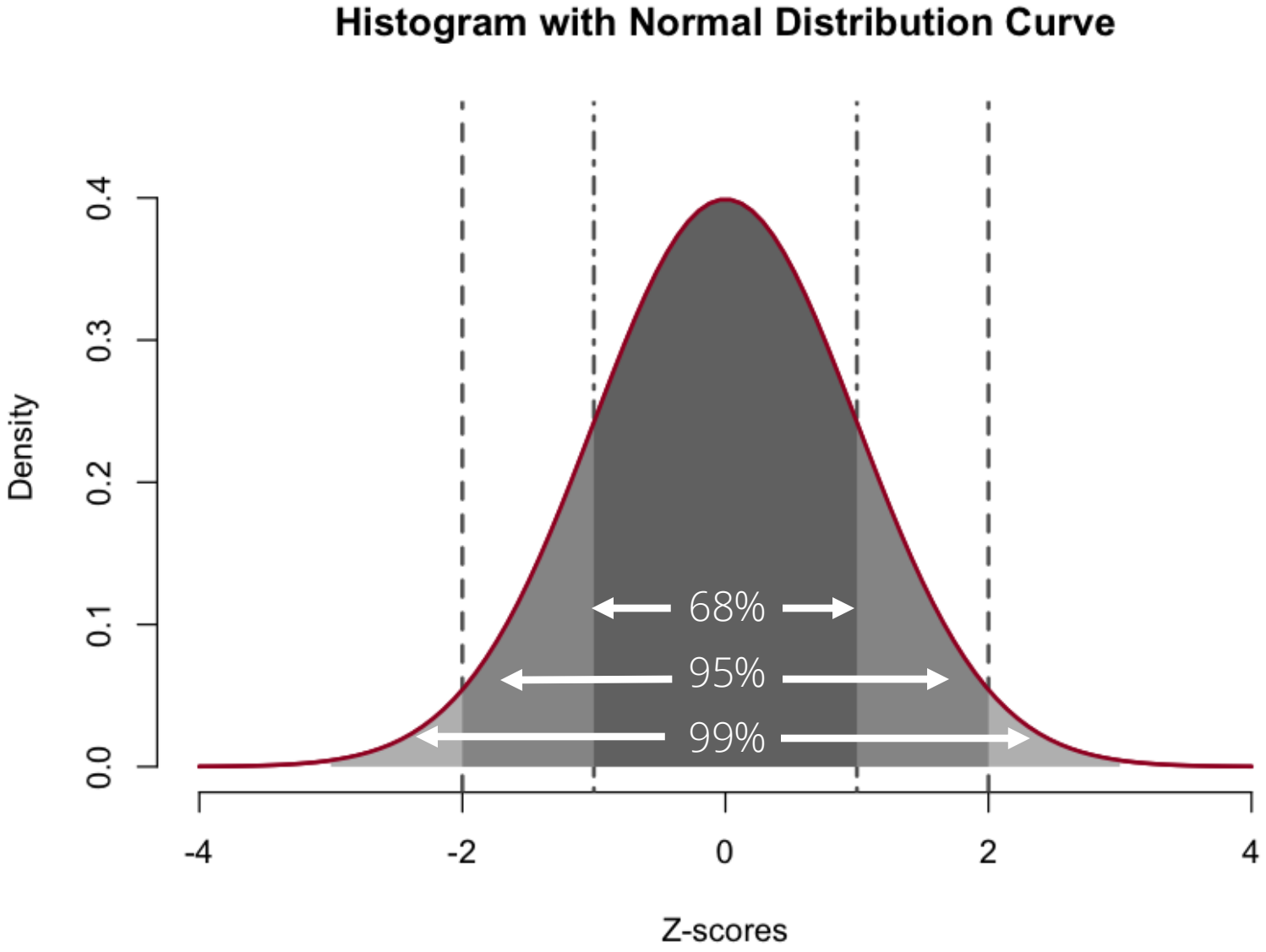
Standard error

- As sample size (n) increases, SEM decreases because you are averaging over more data points, making the sample mean more reliable.
- SEM helps construct confidence intervals, which estimate a range within which the true population mean is likely to lie.

Standard error

- Central Limit Theorem: for a large enough sample sizes the sampling distribution of the mean will approximate a normal distribution.
- With a sampling distribution that is normally distributed, we can use z-scores to estimate a confidence interval: the range of values within which we expect the true population parameter to fall.

Confidence intervals



Mentimeter

- Go to www.menti.com.
- Use code: 5422 7780

Hypotheses

Null and alternative hypothesis

- Harris (2016: 96): "*The basic idea of statistical testing is simple enough. It is used to test whether some level of difference (or association) between two or more variables has arisen by chance.*"
- Whether a value is significant is based on four concepts:
  - The null hypothesis.
  - The alternative hypothesis.
  - The alpha (α) value or alpha level.
  - The p-value.

Null and alternative hypothesis

- Hypotheses are predictive statements that are tested to answer our question.
- We cannot prove claims using empirical data: falsification.

Null and alternative hypothesis

- H_0 specifies that there is no relationship or pattern or change.
- H_A posits that there is a relationship or pattern or change.

Null and alternative hypothesis

- Theory:
  - Attending all GEOG0018 lectures leads to better final assessment performance.
- Method:
  - Draw a random sample of GEOG0018 students.
  - Classify students into two groups and calculate mean grades.

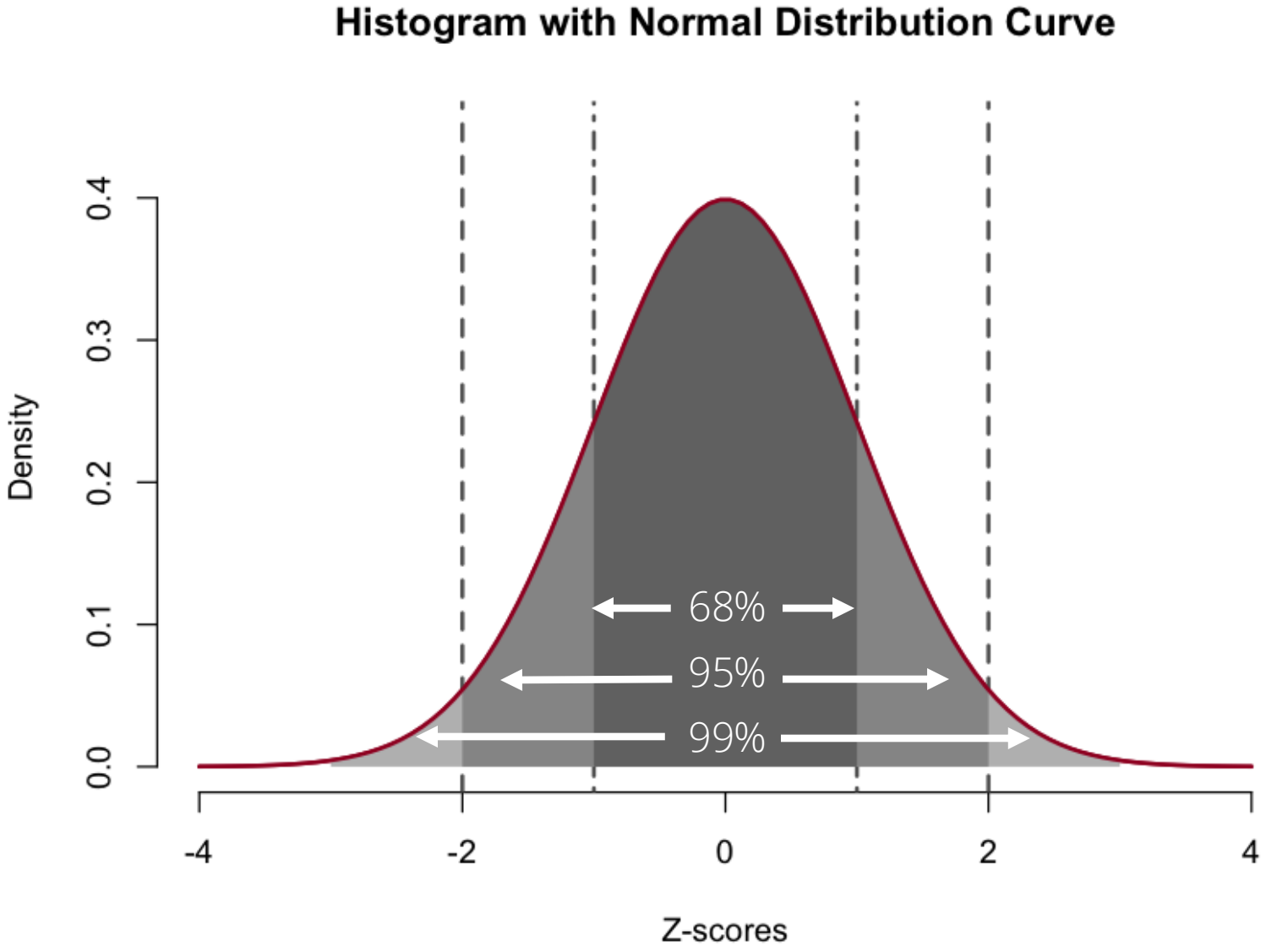
    Group A: Attended all lectures.

    Group B: Did not attend all lectures.
- Results:
  - The mean assessment grade of Group A is higher than that of Group B.

Null and alternative hypothesis

- H_0 Attending all GEOG0018 lectures does not lead to better final assessment performance: the difference that we observed is by chance.
- H_A Attending all GEOG0018 lectures leads to better final assessment performance: there is a genuine difference between the two groups.

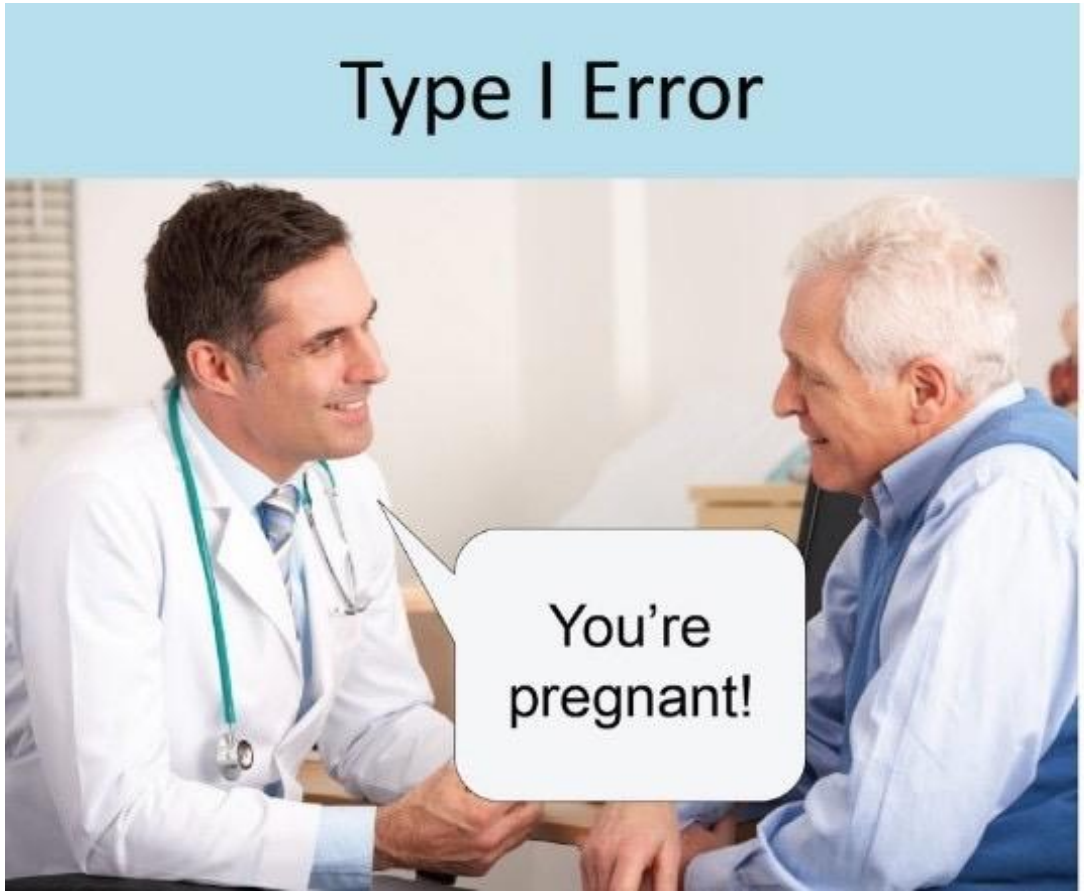
Hypothesis testing

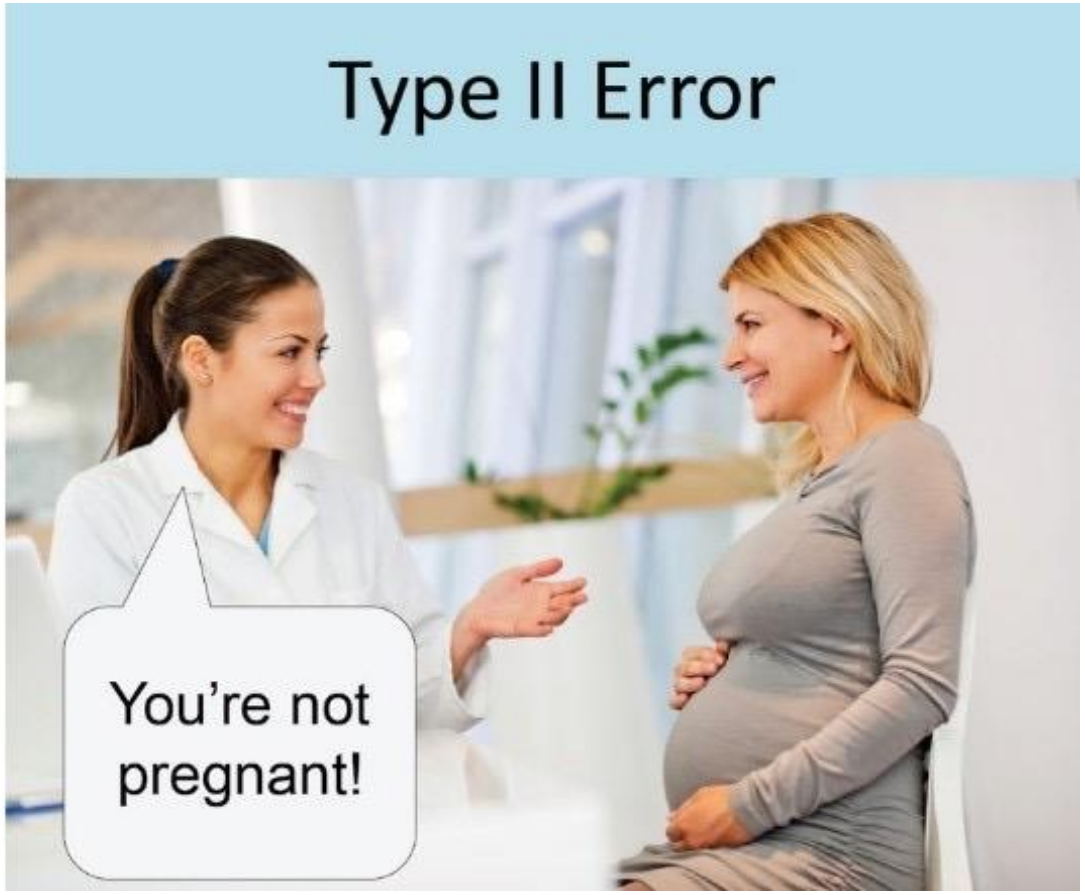


Different errors

- We set out to reject the null hypothesis in our research.
- Type I error: is the incorrect rejection of the null hypothesis (significance level).
- Type II error: is the incorrect acceptance of the null hypothesis.

Different errors





General test procedure

1. State H_0 and H_A with clear reference to the population.
2. Choose the level of Type I error to be tolerated (i.e. choose the significance level).
3. Estimate how likely it is that we would observe the sample outcome.
4. Compare this to the significance level – or use confidence intervals.

General test procedure

- A group of parents believe that the class sizes in primary schools in London are larger than elsewhere in the country.
- According to national figures class sizes have a mean of 25, standard deviation of 8.
- You select a random sample of n = 30 classes and find that class size $\overline{x}$ = 28.

General test procedure

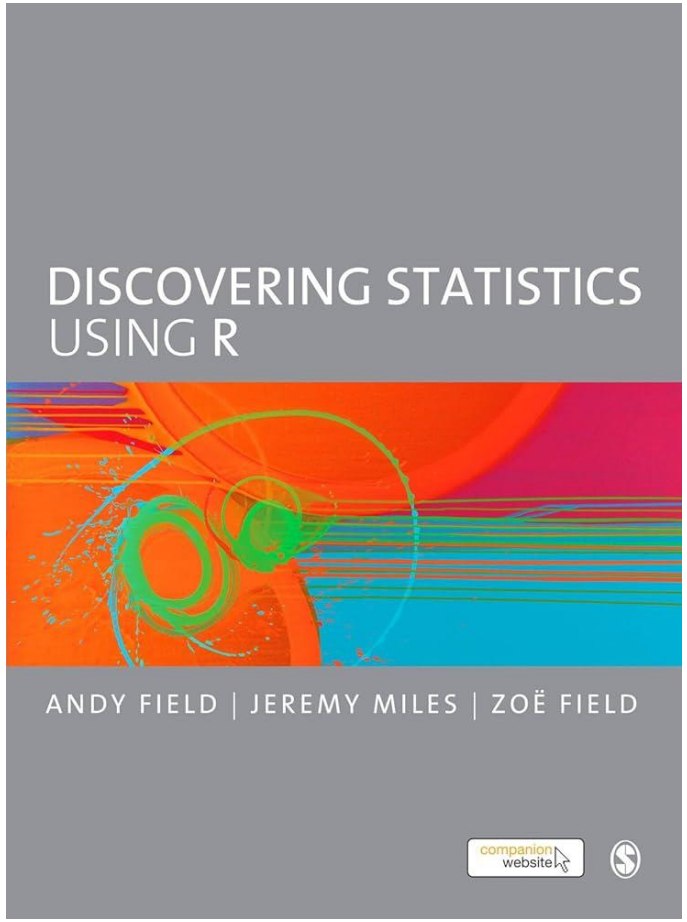
- H_0 Average class size in London is the same as the national average.
- H_A Average class size in London is different from the national average.
- Decide on the significance level. Example: 5% (α = 0.05).
  - Extract a p-value: 0.04.
  - Confidence intervals: (25.1, 30.9)
- Rejection of the null hypothesis.


boy, that escalated quickly

Break

Parametric tests

Suggested reading



Comparing groups

- Testing difference between group means.
- Parametric: certain assumptions are made (e.g. normal distribution).
- Comparing two groups: Two-sample t-test.
- Comparing multiple groups: One-way ANOVA.

Two-sample t-test

- A two-sample t-test allow us to determine if differences in means between two groups are statistically different.
- A two-sample t-test requires continuous data on one variable and nominal data on another (i.e. two categories).
- Samples can be paired (dependent) and non-paired (independent) samples.

Two-sample t-test

How does it work (simplified):

- The average score for each group is calculated.
- The test looks at how much the averages differ.
- If the difference between the averages is big enough compared to how spread out the scores are in each group, it suggests the group means are likely different.

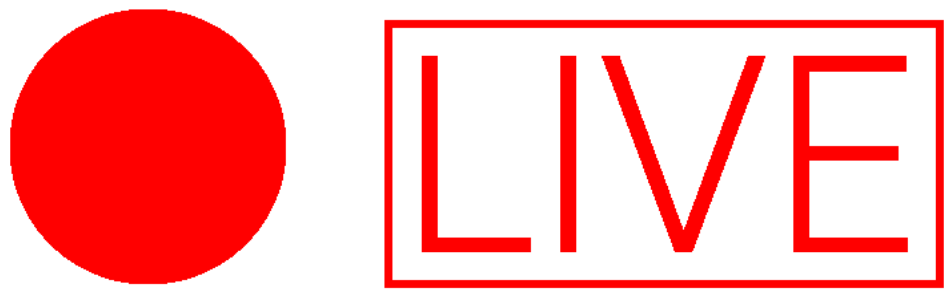
One-way ANOVA

- A One-way ANOVA is used to compare means across two or more groups to determine if at least one group mean is statistically different from the others.
- A One-way ANOVA requires continuous data on one variable and nominal data on another (i.e. two or more categories).
- A significant result indicates that at least one group mean differs from the others, but not which group(s).

One-way ANOVA

How does it work (simplified):

- The average score for each group is calculated.
- ANOVA examines the variance within each group and the variance between the groups to evaluate the overall differences.
- If the differences between the group averages are larger than what would be expected by random chance, it suggests that at least one group is different from the others.


LIVE

Non-parametric tests

Non-parametric tests

- Data are not normally distributed.
- Sample sizes are small.
- Data are captured on an ordinal or non-metric scale.
- Unequal variances.
- Comparing medians instead of means.

Comparing groups

- Testing difference between groups.
- Non-parametric: assumptions are not being met.
- Comparing two independent groups: Mann-Whitney U test.
- Comparing multiple groups: Kruskal-Wallis test.

Mann-Whitney U test

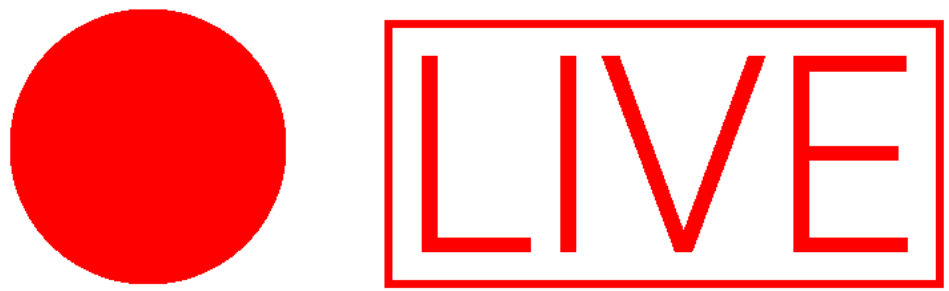
How does it work (simplified):

- The Mann-Whitney U-test checks if two independent groups come from different distributions.
- Combines data from both groups and ranks all values, then looks at the ranks within each group to see if one group generally has higher or lower ranks than the other.
- If the rank differences are large, it suggests the groups differ significantly.

Kruskal-Wallis test

How does it work (simplified):

- The Kruskal-Wallis test checks if three or more independent groups come from different distributions.
- Combines data from all groups and ranks all values, then compares the average ranks for each group.
- If the rank differences between groups are large, it suggests that at least one group is different from the others.


LIVE

Novel data sources

Traditionally, quantitative datasets in geography and the social sciences:

- Are collected for a specific purpose, following a careful study and design.
- Contain very detailed information on a particular topic.
- Are of high quality and of known provenance.

Novel data sources

Traditionally, quantitative datasets in geography and the social sciences are:

- Expensive to produce (e.g., Census, longitudinal surveys).
- Of limited spatial granularity (privacy preserving).
- Infrequently updated.

Novel data sources

- New sources of data: the digital exhaust.
- Diverse in quality and resolution.
- Arguably: higher spatial granularity, higher temporal granularity.

Novel data sources

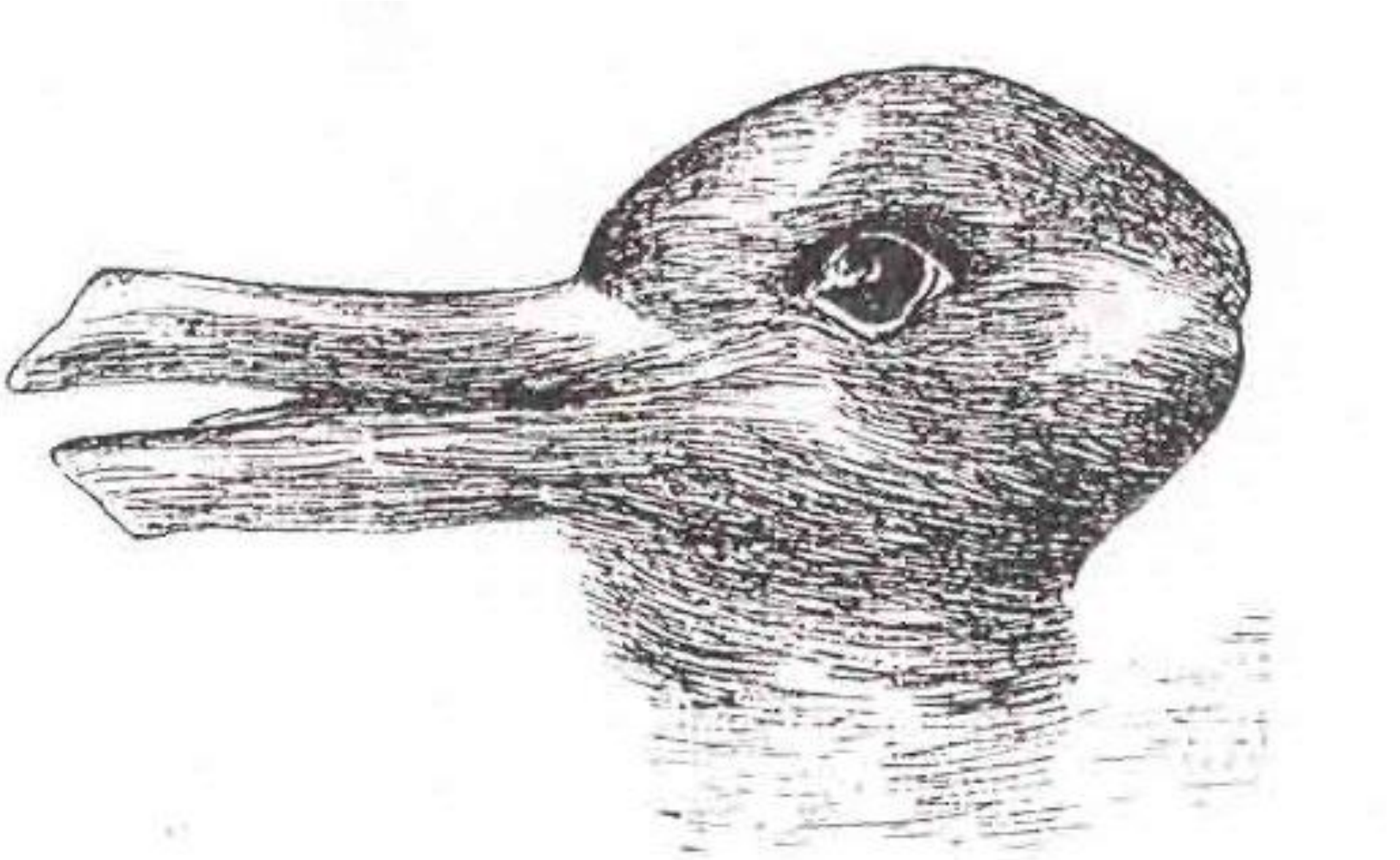
Kitchen 2014:

“Traditionally, data analysis techniques have been designed to extract insights from scarce, static, clean and poorly relational data sets, scientifically sampled and adhering to strict assumptions (such as independence, stationarity, and normality), and generated and analysed with a specific question in mind. The challenge of analysing Big Data is coping with abundance, exhaustivity and variety, timeliness and dynamism, messiness and uncertainty, high relationality, and the fact that much of what is generated has no specific question in mind or is a by-product of another activity.” (p.2)

Novel data sources

- Representation in new data sources raises new epistemological challenges.
  - What do we know and how can we know it.
  - What tools do we use to study the world.
- “Big Data analytics enables an entirely new epistemological approach for making sense of the world; rather than testing a theory by analysing relevant data, new data analytics seek to gain insights ‘born from the data.” (Kitchen 2014, p.2)

Novel data sources



Novel data sources

Summary

Summary

- Samples allow us to infer characteristics of a population.
- Known statistical distributions help us make these inferences.
- Parametric tests to compare groups: Two sample t-test, One-Way ANOVA.
- Non-parametric tests to compare groups: Mann-Whitney U test, Kruskal-Wallis test.
- Novel data sources come with new epistemological challenges.

Questions

Justin van Dijk

j.t.vandijk@ucl.ac.uk