Applied Statistical Analysis I

Statistical inference review

Elena Karagianni, PhD Candidate
karagiae@tcd.ie

September 24, 2025

Department of Political Science, Trinity College Dublin

Today's class

- Lecture Recap
- Exercises

Key concepts and terms

Key concepts and terms

- **Unit of analysis**: The observation described by a set of data. For example, voters,parties, bills, elections, voting decisions etc.

Key concepts and terms

- **Unit of analysis**: The observation described by a set of data. For example, voters,parties, bills, elections, voting decisions etc.
- **Variable**: A characteristic or attribute of the unit of analysis that can vary across observations.
    - "a characteristic that can vary in value among subjects in a sample or population" (Agresti and Finlay 2009, 11)

Key concepts and terms

- **Unit of analysis**: The observation described by a set of data. For example, voters,parties, bills, elections, voting decisions etc.
- **Variable**: A characteristic or attribute of the unit of analysis that can vary across observations.
    - "a characteristic that can vary in value among subjects in a sample or population" (Agresti and Finlay 2009, 11)
- **Dependent variable (DV)**: Also called the *outcome* or *response variable*,; Denoted as Y; the phenomenon we aim to explain.

Key concepts and terms

- **Unit of analysis**: The observation described by a set of data. For example, voters,parties, bills, elections, voting decisions etc.
- **Variable**: A characteristic or attribute of the unit of analysis that can vary across observations.
    - "a characteristic that can vary in value among subjects in a sample or population" (Agresti and Finlay 2009, 11)
- **Dependent variable (DV)**: Also called the *outcome* or *response variable*,; Denoted as Y; the phenomenon we aim to explain.
- **Independent variable (IV)**: Also called the *input*, *predictor*, or *covariate*; Denoted as X; used to explain variation in the DV.

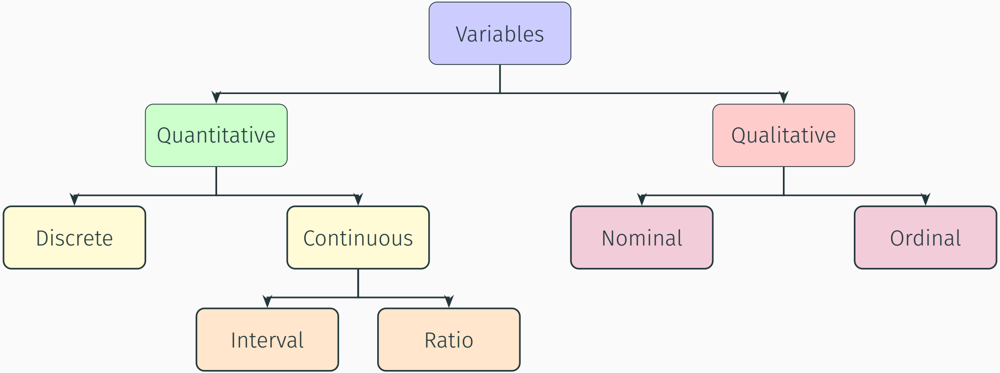
Key concepts and terms

- **Unit of analysis**: The observation described by a set of data. For example, voters,parties, bills, elections, voting decisions etc.
- **Variable**: A characteristic or attribute of the unit of analysis that can vary across observations.
    - "a characteristic that can vary in value among subjects in a sample or population" (Agresti and Finlay 2009, 11)
- **Dependent variable (DV)**: Also called the *outcome* or *response variable*,; Denoted as Y; the phenomenon we aim to explain.
- **Independent variable (IV)**: Also called the *input*, *predictor*, or *covariate*; Denoted as X; used to explain variation in the DV.
- What is variation? (Example: Age $\rightarrow$ Income)
- *Necessary terms for* ***regression analysis***.

Measurement

Refers to the way variables are quantified. (e.g., economic wealth measured as GDP).

Measurement



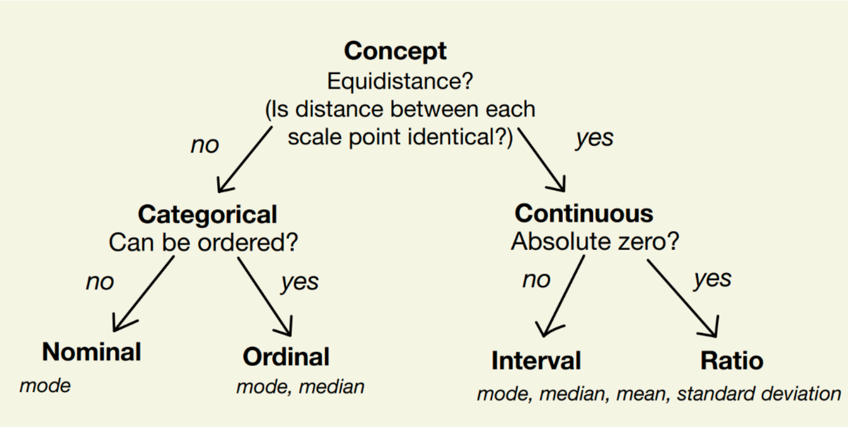


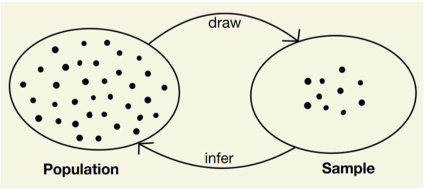
Figure 1: Kellstedt and Whitten 2018, Chap. 5

Population and Sample

- What is the relationship between population and sample?

Population and Sample

- **Population**: the total set of subjects of interest in a study
- **Sample**: the subset of the population on which the study collects data
- **Parameter**: numerical summary of the population
- **Statistic**: a numerical summary of the sample data



Descriptive and Inferential statistics

- **Descriptive statistics**: "summarize the information in a collection of data"

  Agresti and Finlay 2009, 4.

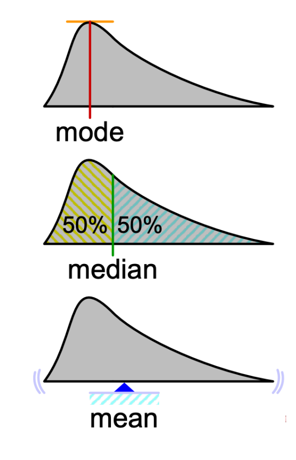
- **Inferential statistics**: "provide predictions about a population, based on data from a sample of that population" Agresti and Finlay 2009, 4.

Descriptive Statistics

Measures of Central Tendency and Variability

- Central tendency: mean, median, mode
- Variability: variance, standard deviation, range, IQR
- Visualization: boxplots

Descriptive Statistics: Central Tendency



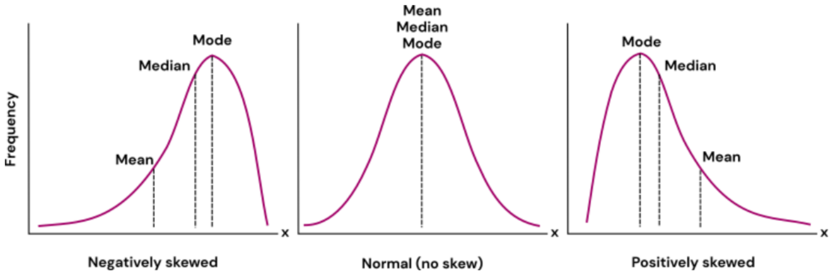
- Mode: Most frequently occurring value of X. Some distributions can have more than one mode.
- Median: Value of X that falls in the middle position when the observations are ordered from smallest to largest.
  - Median = 50th percentile = 2nd quartile
- Mean: Most common measure of central tendency.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Descriptive Statistics: Comparing Measures of Central Tendency

- In a perfectly symmetric distribution, e.g., normal distribution: **mode = median = mean**
- Not true when the distribution is non-symmetric:
  - right-skewed distribution (positive skew): median < mean
  - left-skewed distribution (negative skew): median > mean
- Mean is sensitive to outliers, while the median is more robust.

Descriptive Statistics: Comparing Measures of Central Tendency



Descriptive Statistics: Variability

- Sample Variance: Average of the square deviations from the mean:

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

  Why do we average by dividing by n-1? The sum of the deviations is always zero. Thus, the last deviation can be found once we know the other n-1. So we are not averaging n unrelated numbers. Only $n-1$ squared deviations vary freely, these are called *degrees of freedom* of the variance.

- (Sample) Standard Deviation: Square-root of (sample) variance:

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

Descriptive Statistics: Variability

- Range: Difference between largest and smallest measurement:

$$Range = x_{max} - x_{min}$$

- Interquartile Range (IQR):Difference between upper and lower quartiles (range of the middle 50% of the distribution):

$$IQR = x_{Q3} - x_{Q1} \tag{1}$$

Probability

Probability: Basic terminology

- What is a probability?
- What is a distribution?
- What is a probability distribution?

Probability: Basic terminology

- An experiment is a repeatable procedure for making an observation.
- An outcome is a possible results of such an experiment.
- The sample space (Ω) of an experiment is the set of all possible outcomes.
- An event is a subset of the sample space, i.e., any set of outcomes.
- The probability of an event is its long-run relative frequency.
    - If Pr(A) = 0.5, i.e., probability of event A is 0.5, then event A will occur approximately half of the time when the experiment is repeated infinitely often.
    - If the experiment is repeated many (finite) times, then the approximation as relative frequency (proportion) is expected to improve as the number of repetitions increases.
    - $$P(A) = \frac{\text{Number of elements in A}}{\text{Number of all elements}}$$

Probability: Distribution

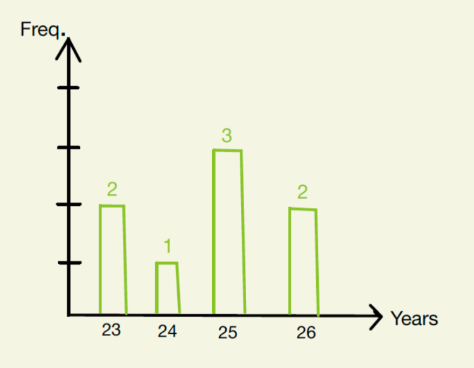


Figure 2: Example: Age of people in the room

It can take different shapes and, therefore, names: normal, binomial, t-distribution etc.

Probability: Probability Distribution

- Distributions of random variables are probability distributions if for all possible outcomes, it tells us the probabilities for these outcomes to occur.

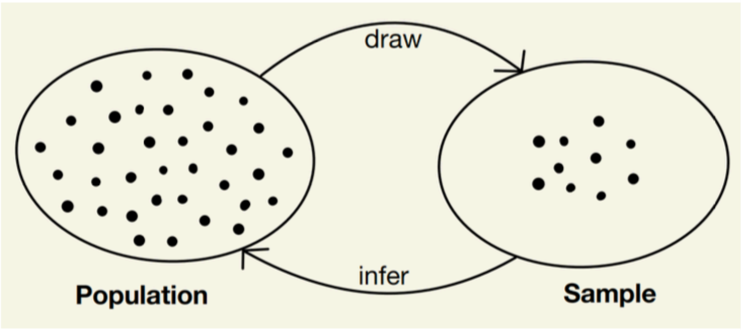
Probability: Probability Distribution

- Distributions of random variables are probability distributions if for all possible outcomes, it tells us the probabilities for these outcomes to occur.
- Probability distributions are analogous to frequency distributions, except that they are based on **probability theory** rather than observations in sample data.

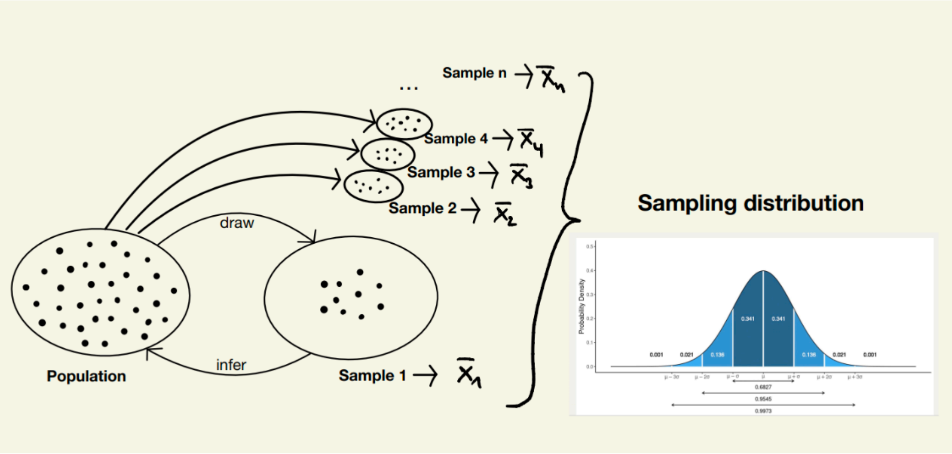
Probability: Probability Distribution

- Distributions of random variables are probability distributions if for all possible outcomes, it tells us the probabilities for these outcomes to occur.
- Probability distributions are analogous to frequency distributions, except that they are based on **probability theory** rather than observations in sample data.
- Definition by Agresti and Finlay (2009, 75): "lists the possible outcomes and their probabilities."

Recall the basic idea for empirical research:



Sampling Distribution



Sampling Distribution

- Sampling Distribution: "A sampling distribution of a statistic is the probability distribution that specifies probabilities for the possible values the statistic can take" (Agresti and Finlay, 2009, 87).
- For example: we have a population distribution with mean μ and variance σ^2 and we are interested in its mean.
- Repeatedly taking samples from that population and calculating the mean for each sample yields the sampling distribution of the mean.

Why is this important?

- The corresponding probability theory "helps us predict how close a statistic falls to the parameter it estimates" Agresti and Finlay 2009, 87 $\rightarrow$ how close is $\bar{x}$ to μ?
- Usually only one sample/estimate $\rightarrow$ Point estimate: "is a single number that is the best guess for the parameter value" Agresti and Finlay 2009, 107
- "If we repeatedly took samples, then in the long run, the mean of the sample means would equal the population mean μ".
- "The standard error describes how much $\bar{x}$ varies from sample to sample" $\rightarrow$ the SE is estimated based on the standard deviation:

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

Central Limit Theorem

What is the Central Limit Theorem?

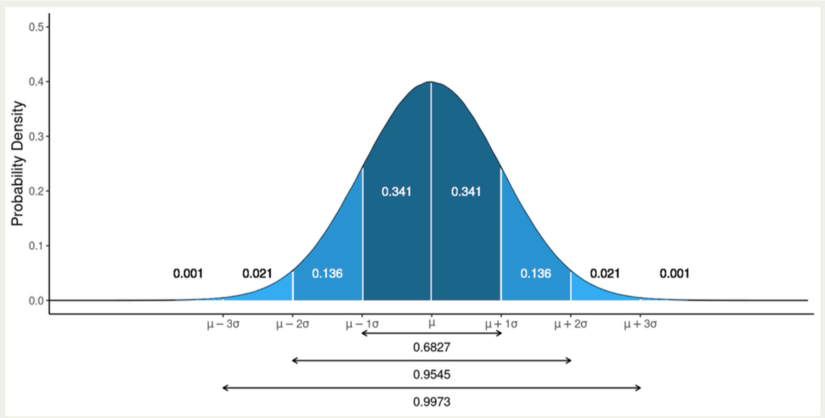
Central Limit Theorem

What is the Central Limit Theorem? $\rightarrow$ The sampling distribution of the statistic approaches a normal distribution with mean μ and variance σ^2/n as n increases.

- This hodls *regardless* of the shape of the original population distribution.
- Basis for application of statistics to many 'natural' phenomena (which are the sum of many unobserved random events).
- How? Take a sample, calculate its mean. Do the same thing again and again. The distribution of sample means will be normal even if the population distribution was not.
- If you repeatedly draw random samples from the same population, calculate the means and plot them, you get a histogram that approaches a bell-shaped curve.

Normal Distribution

- Continuous distribution that describes data clustered around the mean.
- Uniquely determined by its mean/median/mode μ and variance σ^2.
- Important for the **Central Limit Theorem**.



What are confidence intervals?

Confidence Intervals

What are confidence intervals?
“an interval of numbers around the point estimate that we believe contains the parameter value” $\rightarrow$ point estimate ± margin of error (Agresti and Finlay 2009, 110)

Confidence Intervals

What are confidence intervals?
"an interval of numbers around the point estimate that we believe contains the parameter value" $\rightarrow$ point estimate ± margin of error (Agresti and Finlay 2009, 110)

Let's explain this a bit more!

Confidence Intervals

- Our estimate of a population parameter varies across repeated samples, thus generating a *sampling distribution*.
- Instead of a point estimate, we should better get an interval estimate - **a range within the true parameter lies with some level of certainty**.
- We can construct confidence intervals using the standard error or the variance of our estimates.
- We call a CI a q% confidence interval if it is constructed that it **contains the true parameter at least q% of the time if we repeat the experiment a large number of times**.
- Check out this visualization: `https://rpsychologist.com/d3/ci/`
- Attention! This does not mean that there is a q% probability for the population parameter to lie inside the interval!