Applied Statistical Analysis I

Introduction & stats review

Dr Redmond Scales

Department of Political Science, Trinity College Dublin

September 19, 2025

Today's Agenda

(2) Lecture recap

(3) Tutorial exercises

Regression analysis

What is a variable? What is regression analysis?

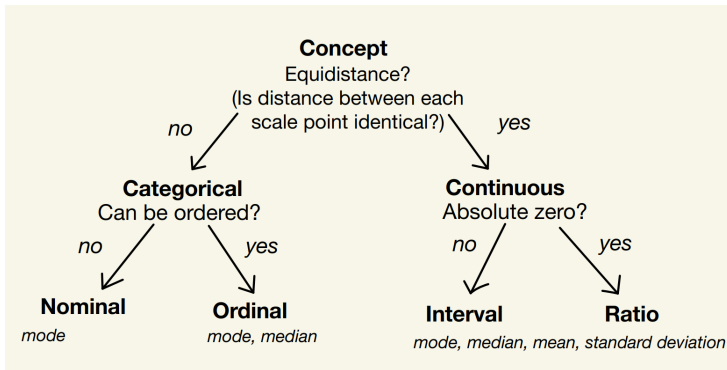## Regression analysis

What is regression analysis?

* <u>Variable:</u> "a characteristic that can vary in value among subjects in a sample or population" (Agresti and Finlay 2009, 11)

* <u>Dependent variable (DV):</u> outcome, response variable, Y, phenomenon to be explained.

* <u>Independent variable (IV):</u> input, explanatory variable, covariate, predictor, $X \rightarrow$ Explain variation in DV using the IV.

* What is variation? (Example: Age $\rightarrow$ Income)

Measurement Scales

How can we measure concepts? And why does it matter?

Regression
○○○●

Population, sample
○○○○

Description
○○○

Distributions
○○○○○○○○○○

CLT
○○○

CIs
○○

Exersices
○○○

Measurement Scales

How can we measure concepts? And why does it matter?

Concept
Equidistance?
(Is distance between each
scale point identical?)

no   *yes*

Categorical
Can be ordered?

no   *yes*

Continuous
Absolute zero?

no   *yes*

Nominal
mode

Ordinal
mode, median

Interval
mode, median, mean, standard deviation

Ratio

(Kellstedt and Whitten 2018, Chap. 5)

<u>Discrete:</u> finite set of possible values.
<u>Continuous:</u> infinite set of possible values.

Population, sample, parameter, variable

What is the relationship between population and sample?

Population, sample, parameter, statistic

What is the relationship between population and sample?

- Population: "the total set of subjects of interest in a study" (Agresti and Finlay 2009, 5).
- Parameter: "numerical summary of the population" (Agresti and Finlay 2009, 5).
- Sample: "the subset of the population on which the study collects data" (Agresti and Finlay 2009, 5).
- Statistic: "a numerical summary of the sample data" (Agresti and Finlay 2009, 5).
- Observation: single subject/unit, one row in dataset
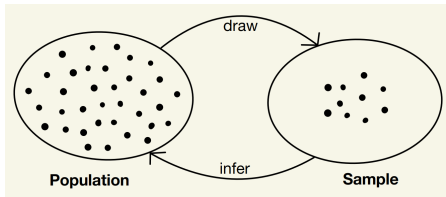
Inferential and descriptive statistics

What is the difference between inferential and descriptive statistics?

Inferential and descriptive statistics

What is the difference between inferential and descriptive statistics?

- <u>Descriptive statistics</u>: "summarize the information in a collection of data" (Agresti and Finlay 2009, 4).

- <u>Inferential statistics</u>: "provide predictions about a population, based on data from a sample of that population" (Agresti and Finlay 2009, 4).

Measures of central tendency and variability (dispersion)

How can we describe variables?

Measures of central tendency

How can we describe variables?

- <u>Mean</u>: $\bar{y} =$ Sum of all values divided by the number of observations, $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$

Measures of variability (dispersion)

How can we describe variables?

- <u>Variance</u>: $s^2(y) =$ Sum of squared deviations divided by number of observations (deviation is the difference between observed value and the mean, $y_i - \bar{y}$), $s^2 = \frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}$
- <u>Standard Deviation</u>: Return original units by taking square root, $s = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}}$
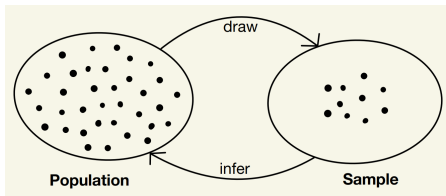
Probability

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

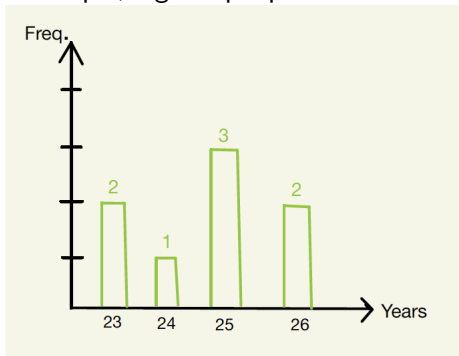## Probability

What is probability?

- Probability: "the probability that an observation has a particular outcome is the proportion of times that outcome would occur in a very long sequence of like observations" (Agresti and Finlay 2009, 73). $\rightarrow P(A) = \frac{Numer\ of\ elements\ in\ A}{Number\ of\ all\ elements}$
- Why do we need probability?

Distributions and probability distributions

What is a distribution?

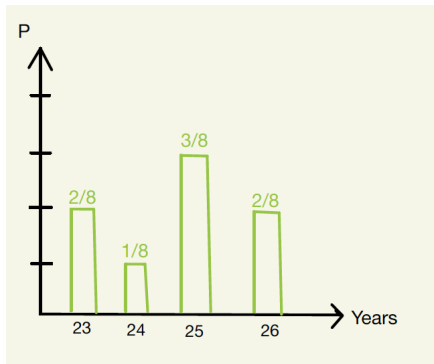Example, Age of people in the room.

* Different shapes, for example, binomial distribution, normal distribution, t-distribution...

Distributions and probability distributions

What is a probability distribution?

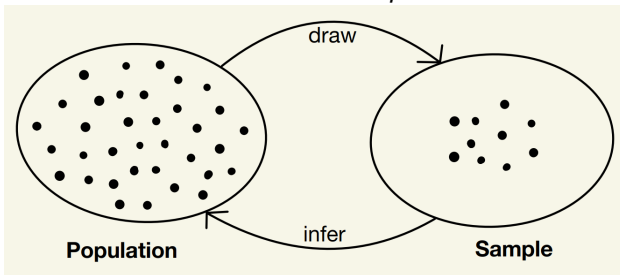- Probability distribution "lists the possible outcomes and their probabilities" (Agresti and Finlay 2009, 75).

Sampling distribution

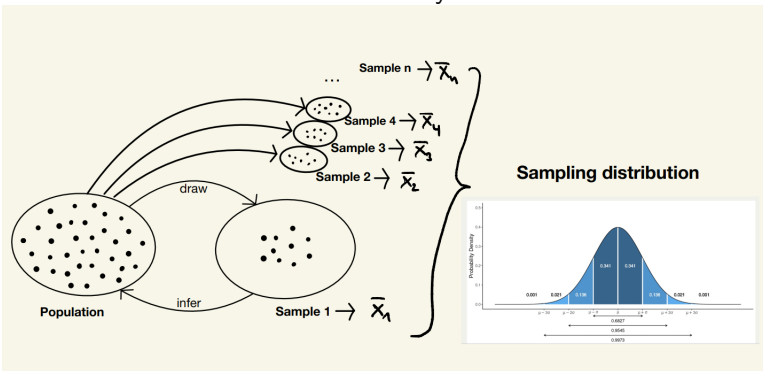*What is a sampling distribution? Why is this important?*

Sampling distribution

Recall the basic idea of empirical research

Sampling distribution

theoretically...

Sampling distribution

What is a 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).

- In other words, a probability distribution for a statistic rather than values of observations $\rightarrow$ What is the probability of $\bar{Y} = 0.5$, rather than what is the probability of $Y = 3$?

Sampling distribution

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 $\bar{y}$ to μ?

- Usually only one sample/one estimate $\rightarrow$ <u>Point estimate</u>: "is a single number that is the best guess for the parameter value" (Agresti and Finlay 2009, 107).

The sampling distribution of the mean, $\bar{y}$

- "If we repeatedly took samples, then in the long run, the mean of the sample means would equal the population mean μ" (Agresti and Finlay 2009, 90). $\rightarrow$ mean of the sampling distribution of $\bar{y}$ equals the population mean, hence, $\mu = \bar{y}$

- "The standard error describes how much $\bar{y}$ varies from sample to sample" (Agresti and Finlay 2009, 90). $\rightarrow$ standard error is estimated based on standard deviation, hence, $\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$

- *Why does this work?*

Central Limit Theorem (CLT)

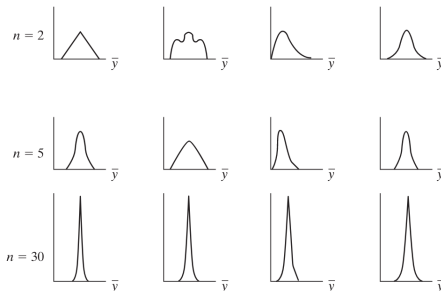*What is the Central Limit Theorem?*

Central Limit Theorem

What is the Central Limit Theorem?

- "For random sampling with a large sample size n, the sampling distribution of the sample mean $\bar{y}$ is approximately a normal distribution" (Agresti and Finlay 2009, 93). $\rightarrow$ regardless of the population distribution
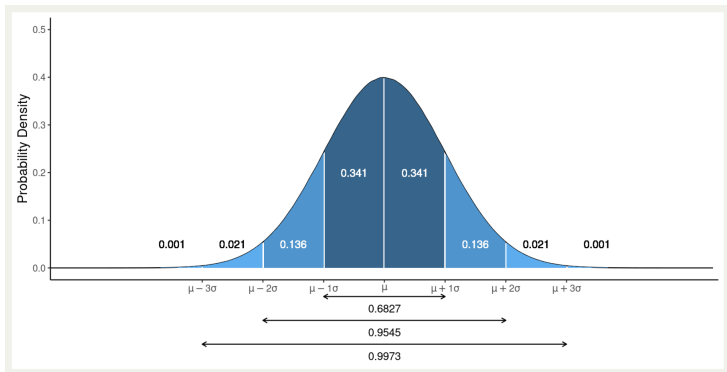

Population distributions

Sampling distributions of $\bar{y}$

$n = 2$

$n = 5$

$n = 30$

Central Limit Theorem

What is the Central Limit Theorem?

- "Knowing that the sampling distribution of $\bar{y}$ can be approximated by a normal distribution helps us to find probabilities for possible values of $\bar{y}$ (Agresti and Finlay 2009, 94). → key in inferential statistics

Confidence intervals

*What are confidence intervals?*

Confidence intervals

What are confidence intervals?

- <u>Confidence interval</u>: "an interval of numbers around the point estimate that we believe contains the parameter value" (Agresti and Finlay 2009, 110). → Point estimate $\pm$ Margin of error

- <u>Confidence level</u>: "The probability that this method produces an interval that contains the parameter" (usually 0.95, 0.99) (Agresti and Finlay 2009, 110).

- Margin of error = multiple of the standard error, $\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$ (Agresti and Finlay 2009, 117).

- For example, for 95% confidence level, the margin of error is $\pm 1.96 \sigma_{\bar{y}}$ (have a look at the normal distribution).
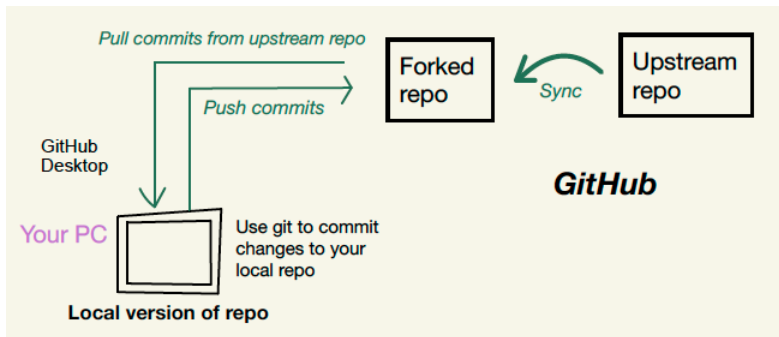
Software check

1. R and RStudio
2. LaTeX and TeXstudio/Overleaf
3. git, GitHub account and GitHub Desktop

git, GitHub workflow
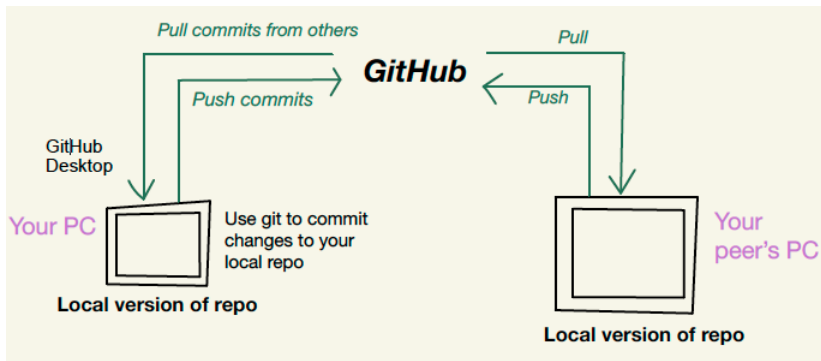
git, GitHub workflow

References I

📄 Agresti, Alan, and Barbara Finlay. 2009. *Statistical methods for the social sciences.* Essex: Pearson Prentice Hall.

📄 Kellstedt, Paul M., and Guy D. Whitten. 2018. *The fundamentals of political science research.* Cambridge: Cambridge University Press.