Data Analysis in R

Probability & Uncertainty

Ken Stiller

13th December 2024

Syllabus: Data Analysis in R

1. Introduction
2. Causality & Basics of Statistics
3. Sampling & Measurement
4. Prediction
5. Multivariate Regression
6. **Probability & Uncertainty**
7. Hypothesis Testing
8. Assumptions & Limits of OLS
9. Interactions & Non-Linear Effects

Plan for Today

- Our goal is statistical inference - making statements about populations
- We use probability theory to make such statements based on samples -i.e. to construct inferential statistics
- Probability Distributions
  - Random variables & probability distributions
  - Normal distribution
  - Sampling distributions
  - Standard error

Table of Contents

Probability: The Basics

- Experiment:
  1. Flipping a coin
  2. Rolling a die
  3. Voting in a referendum

- Sample space Ω: all possible outcomes of the experiment
  1. head, tail
  2. 1,2,3,4,5,6
  3. abstain, Leave, Remain
- Event: any subset of outcomes in the sample space
  1. head, tail, head or tail, etc.
  2. 1, even number, odd number, does not exceed 3, etc.
  3. do not abstain, do not vote leave, etc

Basic Concepts for Inference

- Probability distribution
- Normal distribution
- z-scores
- Sampling distribution

Random Variables and Probability Distributions

- Random variables assign numbers to events
  1. Coin flip: head = 1 and tail = 0
  2. Voting: vote = 1 and not vote = 0
  3. Survey response: strongly agree = 4, agree = 3, disagree = 2, and strongly disagree = 1

- Random variables can be discrete or continuous

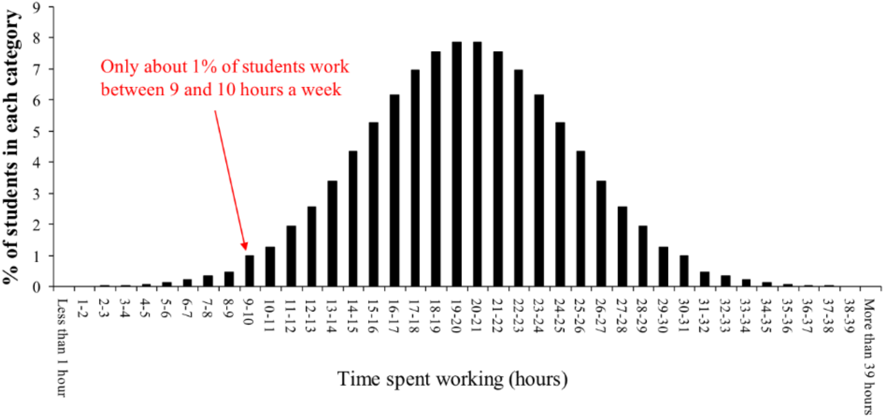
- A probability distribution indicates the probability of an event that a random variable takes a certain value
  1. $P(\text{coin})$: $P(\text{coin} = 1)$, $P(\text{coin} = 0)$
  2. $P(\text{survey})$: $P(\text{survey} = 4)$, $P(\text{survey} = 3)$ etc

Probability Distributions I

- Take a continuous variable, like hours worked by students per week
  - Working hours as random variable: worked 1 hour = 1, worked 15 hours = 15, etc
  - Imagine a population where the mean = 20, and standard deviation = 5

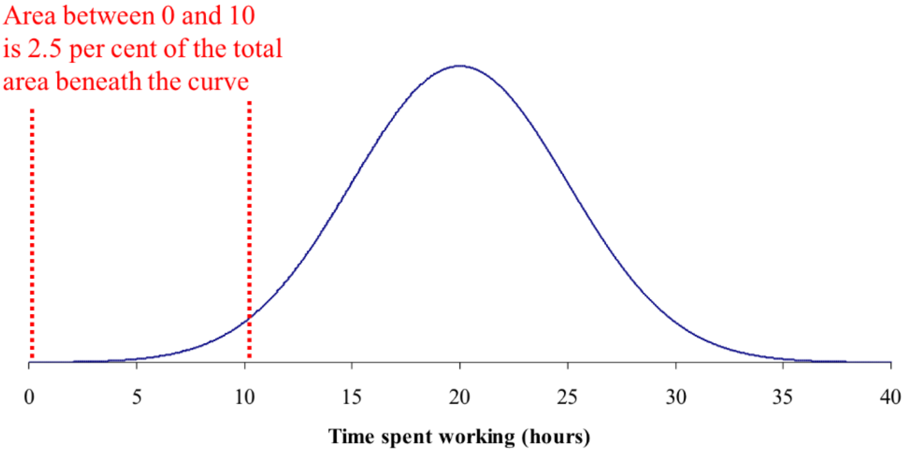
- What about the *distribution* of students?
  - First, just think of students being in certain categories
    - e.g. working $<$ 1 hour, or working 1-2 hours a week
  - Thus, we have a discrete interval level variable
  - A barplot can represent the percentage of students for each number of hours worked

Probability Distributions II



Only about 1% of students work between 9 and 10 hours a week

Probability Distributions III



Normal Distribution

- Normal Distributions are uni-modal (bell shaped) and symmetrical
  - Mode, median, mean at the same point
  - The distribution above the mean is the *same* as the distribution below the mean
  - $\neq$ income distribution, which has a skewed distribution

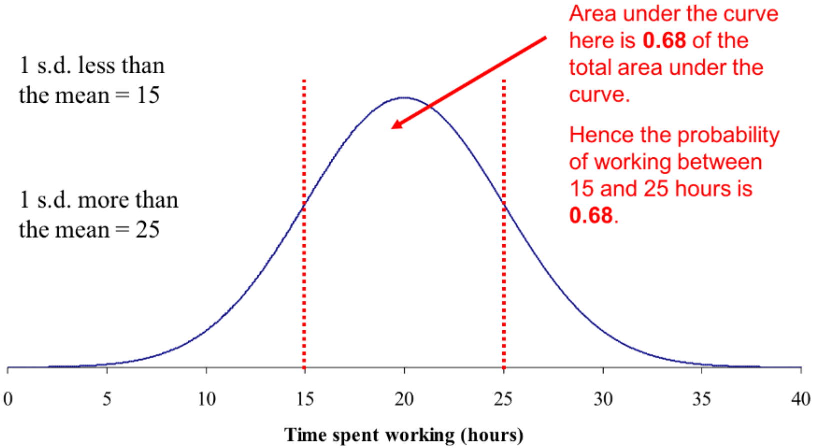
Z-Scores

- The **z-score** for a value x_i of a variable is the number of standard deviations that x_i falls from the mean of x ($\overline{x}$)

$$z = \frac{x_i - \overline{x}}{sd_x}$$

- For any normal distribution, the probability of falling within z standard deviations of the mean is the same, regardless of the distribution's standard deviation
  - For 1 s.d. (or a z-value of 1) the probability is 0.68
  - For 2 s.d. the probability is 0.954
  - For 3 s.d. the probability is pretty close to 1 (0.9975)

Normal Distribution: Example



Area under the curve here is **0.68** of the total area under the curve.

Hence the probability of working between 15 and 25 hours is **0.68**.

Normal Distribution - Example

- For any value of z there is a corresponding probability
  - Most stats book have [used to have?] z tables in their front/back covers
- So: If we pick a student out of our population of a normal distribution we could work out how likely it would be that they worked more than a particular number of hours:
  - e.g., $P(> 25) =?$

Z-Scores - Applied

- **Recall:** The **z-score** for a value x_i of a variable is the number of standard deviations that x_i falls from the mean of x ($\overline{x}$)

$$z = \frac{x_i - \overline{x}}{sd_x}$$

- $P(> 25) =?$ For our example: $\overline{x} = 20$, $sd_x = 5$

$$z = \frac{25 - 20}{5} = 1$$

- Then look at Z value on normal distribution table
- $P(> 25) = 1 - P(\leq 25) \approx 0.16$

Normal Distribution - Parameters

- The particular shape of a normal distribution is defined by its mean and standard deviation
  - These are called the parameters of the normal distribution
  - A particular normal distribution can be represented by the following notation: $N(\mu, \delta^2)$

- To describe the distribution of student work hours from the previous example, we can use the following notation:
  - $N(20, 25)$ with $\delta^2 = 5^2 = 25$

Sampling

- Sampling is the process by which we select a portion of observations from all the possible observations in the population

- **Our aim is estimating what we do not observe from what we do observe**

- What we want to know: Population mean (θ) - which is unobserved [unobservable]
- What we do observe: our sample data
- Our best take at the parameter of interest then is to compute an estimate of the mean ($\hat{\theta}$) based on the sample

Law of Large Numbers

- As the sample size increases, the sample average of a random variable approaches its expected value

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \longrightarrow \mathbb{E}(X)$$

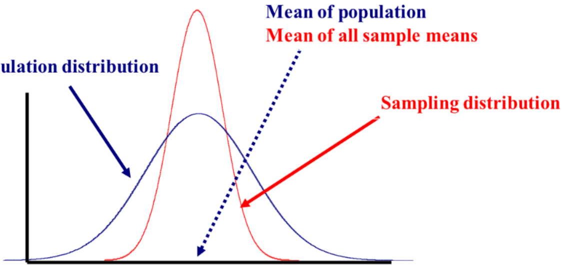
- Example:
  1. You flip a coin 10-times and count the number of heads
  2. What's your guess at the number of heads?
  3. You probably wouldn't be sure about any single round
  4. Repeat many times and compute mean - here your guess probably isn't far off
- If we have lots of sample means then the average will be the same as the population mean
  - In technical language, the sample mean is an unbiased estimator of the population mean

Sampling Distributions I

- If we took lots of samples, we would get a distribution of sample means - i.e., the sampling distribution
  - The sampling distribution of a statistic (in this case the mean of our sample) is the probability distribution that specifies probabilities for the possible values the statistic can take
- This sampling distribution (the distribution of sample means) is normally distributed
- If we took lots of samples then the distribution of the sample means would be centred around the population mean

Sampling Distributions II

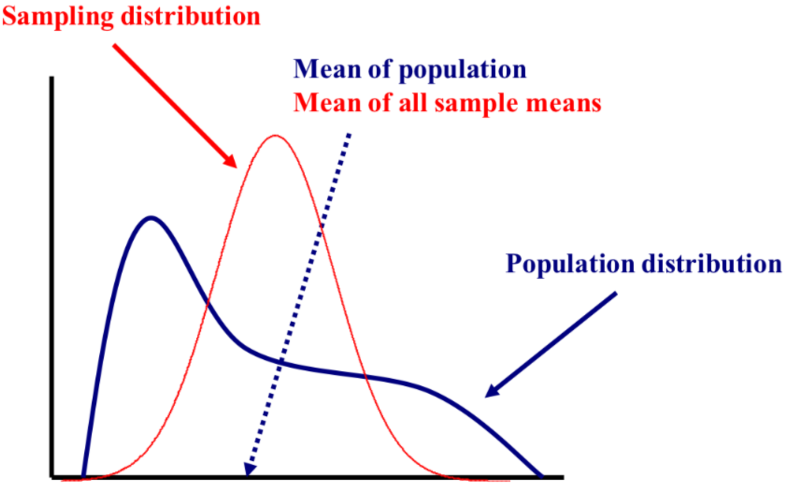
- If we took lots of samples, there would be a normal distribution of their means, centred around the population mean



Sampling Distributions & Central Limit Theorem

- X does not have to be normally distributed for distribution of $\overline{X_n}$ to be normal!
- If the sample size is large enough, the distribution of sample means (what is called the sampling distribution) is approximately normal
  - This is true regardless of the shape of the population distribution
- As n (the sample size) increases, the sampling distribution looks more and more like a normal distribution
  - This is what the *central limit theorem* described
- If we took lots of samples then the distribution of the sample means would be centred around the population mean

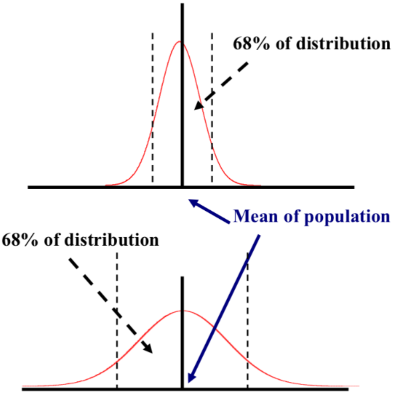
Sampling Distributions III



Uncertainty

- Some sampling distributions are tighter than others...

- The top sampling distribution is ’better’ for estimating the population mean as more of the sample means lie near the population mean



Uncertainty II

- Sampling distributions that are tightly clustered will give us a more accurate estimate on average than those that are more dispersed
  - The standard deviation of a sampling distribution is called a standard error to distinguish it from the standard deviation of a population or sample
  - A large standard error reflects a 'short and wide spread' sampling distribution and a low standard error reflects a 'tall and tight' sampling distribution (thus a less uncertain estimate)
- We need to estimate our sampling distribution's standard error
  - How though?

Some Helpful Notation

- Population mean = $\mathbb{E}(X) = \mu$
- Population standard deviation = σ

- Sample observation = X
- Sample mean = $\overline{X}$
- Sample standard deviation = s
- Sample size = n

Standard Error

- Let's say we know for a single sample:
  - Sample mean: $\overline{X} = 450$
  - Sample standard deviation: $s = 150$
  - Sample size: $n = 2500$
- But we want to know the standard deviation of the sampling distribution, so we can see what the typical deviation from the population mean will be

Standard Error II

- Fortunately, we know:
    - Standard error $(\overline{X}) = \frac{\sigma}{\sqrt{n}}$
    - We don't know σ, but we do know s
    - Estimated standard error $(\overline{X}) = \frac{s}{\sqrt{n}}$
- The standard error is an estimate of how far any sample mean 'typically' deviates from the population mean

Standard Error III

- For the sample:
  - Standard error $(\overline{X}) = \frac{s}{\sqrt{n}} = \frac{150}{\sqrt{2500}} = \frac{150}{50} = 3$
- So, the ’typical’ deviation of a sample mean from the unknown population mean would be 3, if we repeatedly sampled the population

Standard Error IV

- Standard error $(\overline{X}) = \frac{s}{\sqrt{n}}$
- The formula for standard errors entails that:

  1. As the n of the sample increases, the sampling distribution gets tighter
     - The bigger the sample the better it is at estimating the population mean
  2. As the distribution of the population becomes tighter, the sampling distribution also gets tighter
     - If a population is dispersed it will you will be less likely to sample observations near the mean

Key Take Aways

- We want to make *inferences* about the real world, yet have to work with samples
- Probability theory provides a foundation that allows us to make such statements
- We use properties of sampling distributions to relate our sample data to (unknown) population parameters
- Unless we know the entire population, we can only make *probabilistic statements* about the population
- This doesn't mean we can't make meaningful statement - but we keep in mind that they always entail some degree of uncertainty
- Making uncertainty explicit is usually the preferable option
- That's why adding confidence intervals to plots is insightful - it tells us about substantive findings and the degree of certainty of a model.