Data Analysis in R

Hypothesis Testing

Ken Stiller

5th 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

- Accuracy
  - Standard Errors
  - Confidence Intervals
- Hypothesis Testing
  - What is a hypothesis?
  - p-values
  - Type I and Type II errors
  - One and two-sided tests

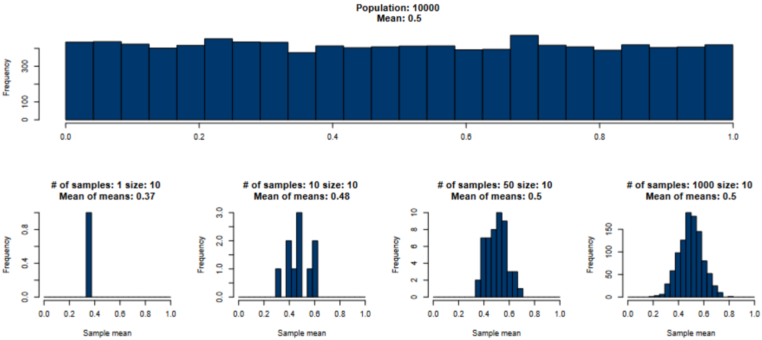
Table of Contents

Recap: Statistical Inference

- What we want to know: parameter $\theta \rightsquigarrow$ unobservable
- What you do observe: data
- We use data to compute an estimate of the parameter ($\hat{\theta}$)

- **But how good is $\hat{\theta}$ an estimate of θ?**
- Ideally, we want to know the estimation error $= \hat{\theta} - \theta$
- The problem remains unchanged: θ is unknown

Recap: Central Limit Theorem



Confidence Intervals

- We know the standard error and are aware of the Central Limit Theorem
- Thus, we can calculate specific ranges around the sample mean of which, if repeated over and over again, a certain share will contain the population mean. In other words, we can quantify how *confident* we are in our estimate.
- This is called a confidence interval

Confidence Intervals II

- An m-percent confidence interval establishes a boundary around the sample mean in which the true mean will lie m out of 100 times under repeated sampling
- Common values for m are 95 and 99 (sometimes 90)
- m is specified by choosing a significance level $\alpha : m = (1 - \alpha) * 100$
- Common significance levels are therefore 0.05 and 0.01 (sometimes 0.1)

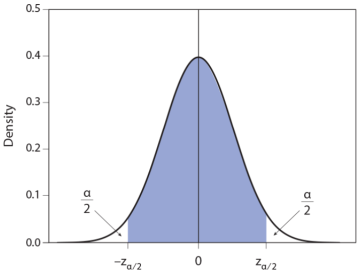
Confidence Intervals: Defining Boundaries

- In order to provide an interval estimate of the population mean μ we need to identify a lower bound (LB) and an upper bound (UB) such that P$(LB \leq \mu \leq UB)$
- Recall last week on probability: We can use our knowledge of the normal distribution to find this boundary
- After z-transformation of any normal distribution

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

  - Probability between -1 and 1 is 0.68
  - Probability between -1.96 and 1.96 is 0.95
  - Probability between -3 and 3 is 0.997
- $z_{\alpha/2}$ is the value associated with $(1 - \alpha) * 100\%$ coverage in the standard normal distribution

Example: Critical Values of Normal Distribution



- The lower and upper critical values, $-z_{\alpha/2}$ and $z_{\alpha/2}$, shown on horizontal axis
- Area under the density curve between these critical values (in blue) equals $1 - \alpha$

Confidence Intervals: Overview

- CI: boundaries in which μ will lie m-times out of a 100
- $(1-\alpha)*100\%$ confidence intervals:

$$CI_\alpha = [\overline{X} - z_{\alpha/2} * SE, \overline{X} + z_{\alpha/2} * SE]$$

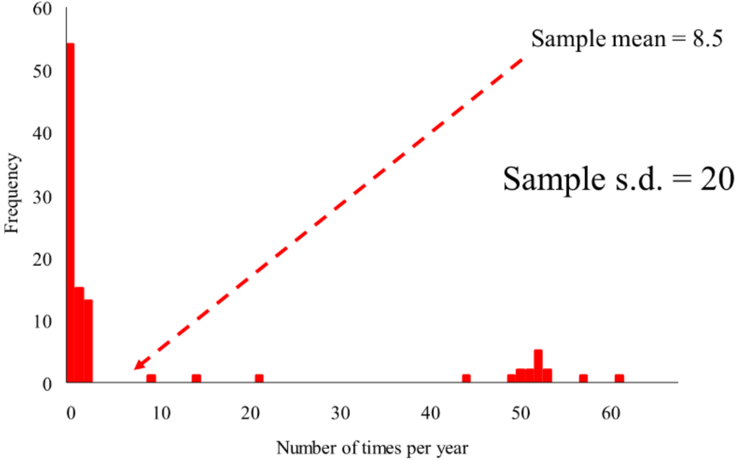
  where $z_{\alpha/2}$ is the critical value and α reflects our chosen significance level

- $P(Z > z_{\alpha/2}) = \alpha/2$ and $Z \sim \mathcal{N}(0,1)$
  1. $\alpha = 0.01$ gives $z_{\alpha/2} = 2.58$
  2. $\alpha = 0.05$ gives $z_{\alpha/2} = 1.96$
  3. $\alpha = 0.10$ gives $z_{\alpha/2} = 1.64$

Example: Confidence Intervals & Standard Error for Sample Mean

- How often do German people attend some form of religious worship?
- Take a random sample of 100 people from the German population and record how many times they attended a form of religious worship last year
  - Distribution is extremely skewed
  - Some went a lot, most went infrequently or not at all
- From that sample we get a sample mean and a sample standard deviation

Example: Confidence Intervals & Standard Error for Sample Mean II



Example: Confidence Intervals & Standard Error for Sample Mean III

- From the *sd* and *mean* of the sample we can calculate the standard error for the sample mean:

$$\frac{s}{\sqrt{n}} = \frac{20}{\sqrt{100}}$$

where $s =$ sample standard deviation

- From this we can calculate any confidence interval

$$\mathrm{CI}_\alpha = \left[\bar{X} - z_{\alpha/2} \times \text{ standard error}, \bar{X} + z_{\alpha/2} \times \text{ standard error } \right]$$

- Usually, we are interested in a 95% CI:

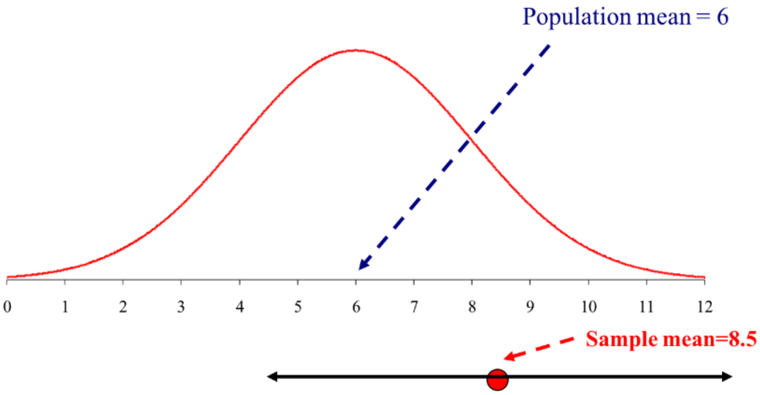
$$[8.5 - 1.96 \times 2, 8.5 + 1.96 \times 2] = [4.58, 12.42]$$

- In 95 out of a 100 times will the true mean lie within confidence intervals computed in this way

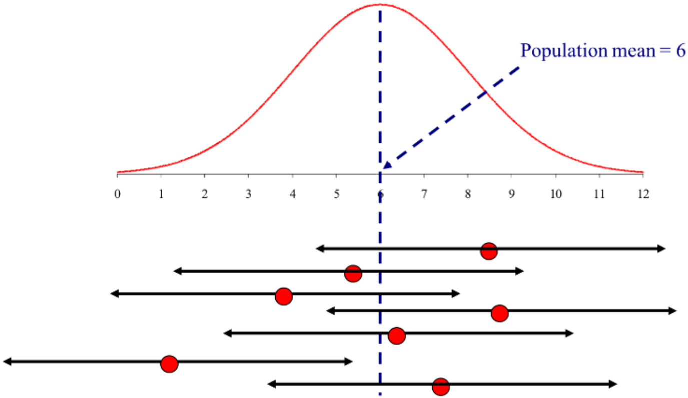
Example: Confidence Intervals & Standard Error for Sample Mean IV

- Suppose we know that the population mean for religious worship attendance in Germany is actually 6 times per year
  - Our particular sample is off by 2.5
  - The mean of all the possible sample means is equal to the population mean so the centre of the sampling distribution is 6

Example: Confidence Intervals & Standard Error for Sample Mean



Example: Confidence Intervals & Standard Error for Sample Mean II



Example: Confidence Intervals & Standard Error for Sample Mean VII

- Of the 7 samples, all the confidence intervals around the sample mean enclosed the actual true population mean apart from one
- If we repeated this lots of times, we would expect 95% of the confidence intervals to enclose the actual population mean
  - 95% because that's the level we set
  - If we had set 99%, the confidence intervals would be larger

Statistical Hypothesis Testing: Overview

1. Construct a null hypothesis (H_0) and its alternative (H_1)
2. Pick a test statistic T
3. Figure out the sampling distribution of T under H_0 (reference distribution)
   - For hypothesis tests regarding the mean, if sample size large, use the normal distribution
   - For other test statistics, you need to use other distributions
4. Is the observed value of T likely to occur under H_0?
   - **Yes** - Retain H_0
   - **No** - Reject H_0

What is a Hypothesis?

- Hypotheses = testable statements about the world
- Hypotheses = falsifiable
  - We test hypotheses by attempting to see if they could be false, rather than 'proving' them to be true
- Hypotheses come from:
  - Theory
  - Past empirical work
  - Common sense (?)
  - Anecdotal observations

Null and Alternative Hypotheses

- We choose between two conflicting statements, doing the following:

  1. The null hypothesis (H_0) is directly tested
     - This is a statement that the parameter we are interested in has a value similar to no effect (i.e., usually 0 for coefficients)
     - e.g. regarding ideology, old people are the same as young people
  2. Alternative (H_1) contradicts the null hypothesis
     - This is a statement that the parameter falls into a different set of values than those predicted by (H_0)
     - e.g. regarding ideology, old people are more right-wing than young people

- **Note that we actually 'test' the null hypothesis!**

Hypothesis Testing: Test Statistic

- In any statistical hypothesis test, a test statistic is computed from the data in order to test the null hypothesis.

$$T = \frac{\text{sample estimate} - \text{parameter value } \textit{under } H_0}{\text{standard error}}$$

- The larger T, the more the data contradict the null hypothesis
- For a given estimate, T becomes larger as the standard error decreases

Statistical Hypothesis Testing: Overview II

- Hypotheses - H_0: $\mu = \mu_0$ and $H_1 : \mu \neq \mu_0$
- Test statistic:

$$\text{z-score} = \frac{\bar{X} - \mu_0}{\text{standard error}} = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

- Under the null, by the central limit theorem

$$\text{z-score} \overset{\text{approx.}}{\sim} \mathcal{N}(0, 1)$$

- Is Z_{obs} unusual under the null?
  - Reject the null when $|Z_{obs}| > z_{\alpha/2}$
  - Retain the null when $|Z_{obs}| \leq z_{\alpha/2}$

Example: Exam Scores

- Suppose there's a standardised exam with marks ranging from 0-100
- Suppose further we know test scores are normally distributed with mean $\mu = 88$ and standard deviation $\sigma = 5$
- Now, in five tests cohorts receive test scores of $\overline{X} = 95$
  $H_0 : \mu = 88$ $H_1 : \mu \neq 88$

Example: Exam Scores II

- We know that the standard deviation of test in the population is 5. The sample size is 5 so we calculate the standard error as:

$$SE = \frac{5}{\sqrt{5}}$$

- Assuming H_0 was true, we know that the sampling distribution is

$$\mathcal{N}(88, (\frac{5}{\sqrt{5}})^2)$$

- Based on sampling distribution, how many standard deviations away is the observed mean from the hypothesized mean?

$$z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{95 - 88}{\frac{5}{\sqrt{5}}} = 5.82$$

- What is the probability of observing a z-value of more than 5.82 or below -5.82?

Example: Exam Scores III

The probability is smaller than 0.00001. What do we make of this?

p-Value

- Ok, so what's a p-value then?
- A p-value indicates the probability, under H_0, of observing a value of the test statistic at least as extreme as its observed value
- A smaller p-value presents stronger evidence against H_0
- The level of the test: $\Pr(\text{rejection}|H_0) = \alpha$
- A p-value less than α conventionally indicates statistical significance
  - Conventional values of α: 0.05 & 0.01

p-Value II

- A p-value is an *arbitrary* threshold that our test must meet
  - we might want to be 99% confident that we are correctly rejecting the null hypothesis
  - or we might make the judgement that p-values of e.g. 0.05 and below are **probably** good evidence the null hypothesis can be rejected

- Keep in mind: the p-value is **not** the probability that H_0 (H_1) is true (false)

Type I and Type II Errors

- Concern false rejection if the null is true (*type I error*)
- Two types of errors:

	Reject H_0	Retain H_0
H_0 is true	Type I error	Correct
H_0 is false	Correct	Type II error

- Type I error occurs when we reject H_0 even though it is true
  - Happens 5% of the time if we choose $\alpha = 0.05$
- Type II error occurs when we do not reject H_0 even though it is false
  - If $\alpha = 0.05$, sometimes a real difference won't be detected

Type I and Type II Errors

- There's a trade-off between the two types of error
  - What probability do you want to minimize? False positive or false negative?

One- or Two-Sided (Tailed) Tests

- In the example above, we were interested in the difference to the true value
  - one-sided alternative hypothesis: $H_1 : \mu > \mu_0$ or $\mu < \mu_0$
  - one-sided p-value= $\Pr(Z > Z_{obs})$ or $\Pr(Z < Z_{obs})$
- Convention is to use two-tailed tests
  - making it even more difficult to find results just due to chance
  - normally don't have very strong prior information about the difference

Significance Tests and CIs

- Note that our significance test looks similar to the CIs
- We could use a CI around the difference between the two sample means to 'test' the hypothesis that they are the same

- A 95% CI would just be $1.96 * SE$

- You can see this on first view if the 95% CI encloses zero
- CIs and significance tests are doing the same job, just presenting the information in a slightly different way

Binary Variables and Proportions

- We have been working with continuous variables and means
- This works for binary variables too, where the mean is just the proportion:
  - Population mean $= \mu =$ population proportion $= \pi$
  - Population standard deviation $= \sigma = \sqrt{\pi(1-\pi)}$
  - Sample proportion $= P$
  - Standard deviation $(P) = \frac{\sqrt{\pi(1-\pi)}}{\sqrt{n}}$
  - Standard error $(P) = \frac{\sqrt{P(1-P)}}{\sqrt{n}} = \sqrt{\frac{P(1-P)}{n}}$

Key Take Aways

- Knowing the shape of the sampling distribution, we can work out:
  - ranges around a sample mean that will enclose the population mean $X\%$ of the time
  - the probability that a null hypothesis about the population mean is 'true', given a particular sample mean
  - the probability that population means for different groups are different, given two sample means
  - all of the above for proportions
- Note that this allows us to make a **probabilistic** statement. Not more, not less.
- In expectation, a (non-negligible) share will be false positives!

The way ahead

- What if OLS assumptions are violated?
- When do we care?
- What can we do?
- What do we actually do?