Hypothesis testing & Confidence intervals

Data Analysis for Psychology in R 1

Semester 2, Week 4

Dr Umberto Noè

Department of Psychology
The University of Edinburgh

Course Overview

Exploratory Data Analysis	Research design and data
	Describing categorical data
	Describing continuous data
	Describing relationships
	Functions
Probability	Probability theory
	Probability rules
	Random variables (discrete)
	Random variables (continuous)
	Sampling

Foundations of inference	Confidence intervals
	Hypothesis testing (p-values)
	Hypothesis testing (critical values)
	Hypothesis testing and confidence intervals
	Errors, power, effect size, assumptions
Common hypothesis tests	One sample t-test
	Independent samples t-test
	Paired samples t-test
	Chi-square tests
	Correlation

Learning objectives

1. Interpret a confidence interval as the plausible values of a parameter that would not be rejected in a two-sided hypothesis test.

2. Determine the decision for a two-sided hypothesis test from an appropriately constructed confidence interval.

3. Be able to explain the difference between a significant result and an important result.

Part A

Introduction

Population vs Sample

- We performed a study by selecting a sample from the population.

- The goal is to use the sample data to answer questions about the population data. For example:

    - What's the population mean? **Goal A:** estimating a parameter.

    - Could the population mean be zero? **Goal B:** testing a theory (hypothesis) about a parameter.

Confidence interval

Figure 1. Invisible lady walking her dog.

- We said that the SE of the mean measures how far $\bar{x}$ is likely to be from the population mean μ.

- Figure 1 is a drawing of an invisible lady walking her dog. The dog, which is visible, is on a leash. The leash is such that the dog is within $t^* \cdot SE$ of the lady 95% of the time.

- **In one picture** we can see the dog, but we would like to know where the lady is.

- Since the lady and the dog are usually within $t^* \cdot SE$ of each other, we can take the following interval as a range that would **typically** include the lady:

$$[\text{dog} - t^* \cdot SE, \ \text{dog} + t^* \cdot SE]$$

or, in short:

$$\text{dog} \pm t^* \cdot SE$$

- We could say that we are 95% confident that the lady is within this interval.

Confidence interval

- We have one picture, which corresponds to one sample.

- We would like to know the value of the population mean μ, which corresponds to the invisible lady, but we cannot directly observe it.

- We can however see the sample mean $\bar{x}$, corresponding to the dog.

- We use what we can see, $\bar{x}$, along with its standard error, as a way of constructing an interval that we hope will include what we cannot see, i.e. the population mean μ.

- The interval $\mathrm{dog} \pm t^* \cdot SE$ is a 95% confidence interval for the position of the lady.

- The interval $\bar{x} \pm t^* \cdot SE$ is a 95% confidence interval for the population mean μ.

  - Recall that the standard error of the mean is computed as $SE = s/\sqrt{n}$.
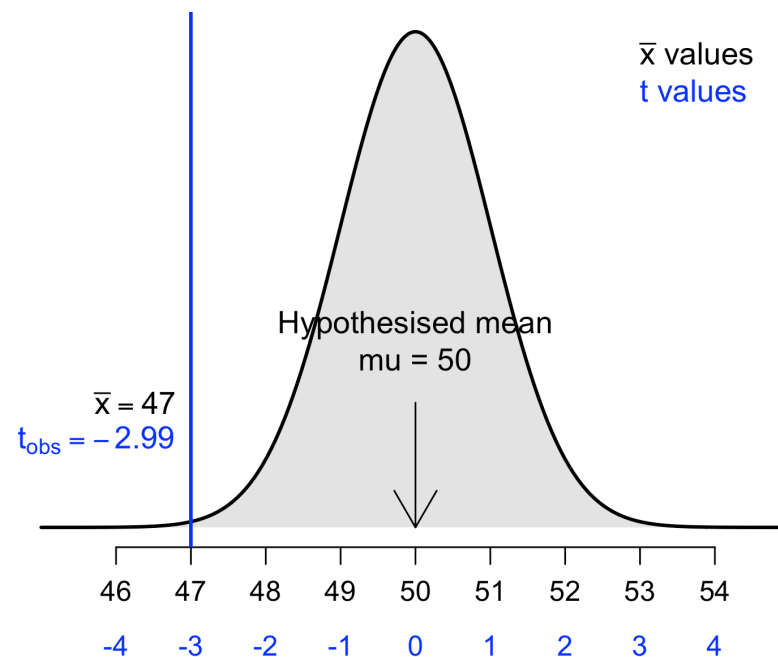
Hypothesis testing

Consider this example:

- Sample of $n = 30$, with mean $\bar{x} = 47$ and $s = 5.5$

- Want to test whether

$$H_0 : \mu = 50$$
$$H_1 : \mu \neq 50$$

- $SE = s/\sqrt{n} = 5.5/\sqrt{30} = 1.004$

- $t = t_{obs} = \frac{47-50}{1.004} = -2.99$

- The sample mean (47) is about 3 standard errors smaller than the claimed mean (50)

$\bar{x}$ values
t values

Hypothesised mean
mu = 50

$\bar{x} = 47$
$t_{obs} = -2.99$

46 47 48 49 50 51 52 53 54
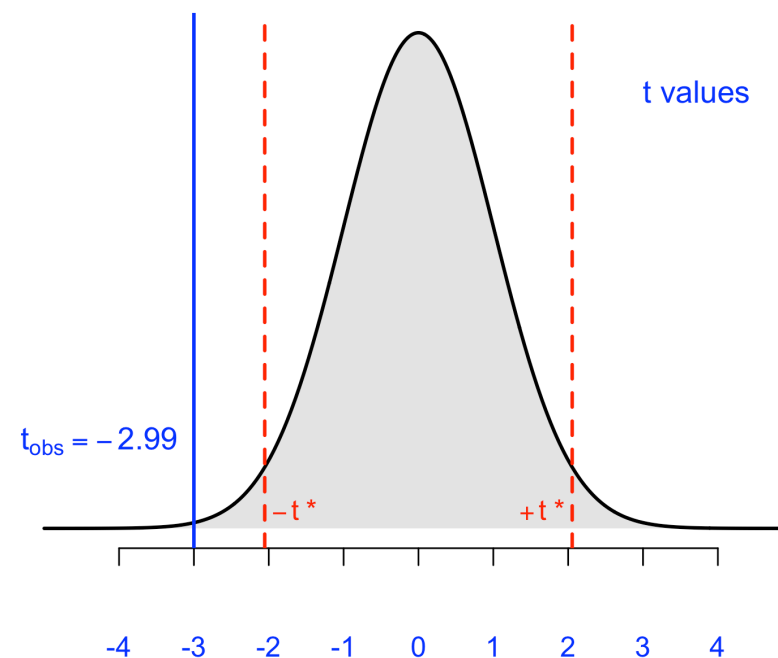
-4 -3 -2 -1 0 1 2 3 4

Hypothesis testing

- Critical values for a $t(29)$ distribution with $\alpha = 0.05$:

```
tstar <- qt(c(0.025, 0.975), df = 29)
tstar
```

```
## [1] -2.045  2.045
```

- As $t_{obs} \leq -t^*$, we reject H_0.

Part B

Hypothesis testing & Confidence intervals

Hypothesis testing

- Consider again the following example, where we wish to test the following claim at the $\alpha = 0.05$ significance level:

$$H_0 : \mu = \mu_0 \qquad \text{vs} \qquad H_1 : \mu \neq \mu_0$$

- We start by computing the observed value of the t-statistic for our sample:

$$t = t_{obs} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

P-value method

- Compute the p-value, i.e. probability of obtaining a test statistic at least as extreme as the observed one, when H_0 is true.

  ```
  2 * pt(abs(tobs), df = n-1, lower.tail = FALSE)
  ```

- Make a decision:

  - Reject H_0 if $p \leq \alpha$
  - Do not reject H_0 if $p > \alpha$

Critical value method

- Compute the lower and upper critical values $-t^*$ and t^*:

  ```
  qt(c(0.025, 0.975), df = n-1)
  ```

- Make a decision:

  - Reject H_0 if $t \leq -t^*$ or $t \geq +t^*$
  - Do not reject H_0 if $-t^* < t < +t^*$

Hypothesis testing

- **Do not reject** $H_0 : \mu = \mu_0$ **if:**

$$-t^* < t < +t^*$$

$$-t^* < \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} < +t^*$$

$$-t^* \cdot \frac{s}{\sqrt{n}} < \bar{x} - \mu_0 < +t^* \cdot \frac{s}{\sqrt{n}}$$

$$-\bar{x} - t^* \cdot \frac{s}{\sqrt{n}} < -\mu_0 < -\bar{x} + t^* \cdot \frac{s}{\sqrt{n}}$$

$$\bar{x} + t^* \cdot \frac{s}{\sqrt{n}} > \mu_0 > \bar{x} - t^* \cdot \frac{s}{\sqrt{n}}$$

$$\bar{x} - t^* \cdot \frac{s}{\sqrt{n}} < \mu_0 < \bar{x} + t^* \cdot \frac{s}{\sqrt{n}}$$

Hypothesis testing

- **Do not reject $H_0 : \mu = \mu_0$ when:**

$$\underbrace{\bar{x} - t^* \cdot \frac{s}{\sqrt{n}}}_{\text{lower limit of CI}} < \mu_0 < \underbrace{\bar{x} + t^* \cdot \frac{s}{\sqrt{n}}}_{\text{upper limit of CI}}$$

- **Reject $H_0 : \mu = \mu_0$ when either:**

$$\mu_0 \geq \underbrace{\bar{x} + t^* \cdot \frac{s}{\sqrt{n}}}_{\text{upper limit of CI}}$$

or

$$\mu_0 \leq \underbrace{\bar{x} - t^* \cdot \frac{s}{\sqrt{n}}}_{\text{lower limit of CI}}$$

- That is, we **do not reject H_0** when μ_0 lies within the CI:

$$\text{Do not reject } H_0 \text{ if } \left[\bar{x} - t^* \cdot \frac{s}{\sqrt{n}}, \ \bar{x} + t^* \cdot \frac{s}{\sqrt{n}}\right] \text{ includes } \mu_0$$

Part C

Significance vs Importance

Body temperature example

- Recall the body temperature study, where the goal was to assess whether the sample data provide evidence, at the 5% significance level, that the population mean body temperature for healthy humans changed from the long thought 37 °C.

- Data:

```
library(tidyverse)
tempsample <- read_csv('https://uoepsy.github.io/data/BodyTemperatures.csv')
tempsample <- tempsample |> drop_na(BodyTemp)
```

```
head(tempsample)
```

```
## # A tibble: 6 × 2
##    BodyTemp Pulse
##       <dbl> <dbl>
## 1      36.4    69
## 2      37.4    77
## 3      37.2    75
## 4      37.1    84
## 5      36.7    71
## 6      37.2    76
```

```
nrow(tempsample)
```

```
## [1] 50
```

Body temperature example

- Null and alternative hypotheses:

$$H_0 : \mu = 37$$
$$H_1 : \mu \neq 37$$

- Chosen significance level: $\alpha = 0.05$

- Test statistic:

$$t = \frac{\bar{x} - 37}{s/\sqrt{n}}$$

```
n <- nrow(tempsample)
xbar <- mean(tempsample$BodyTemp)
s <- sd(tempsample$BodyTemp)
se <- s / sqrt(n)

mu0 <- 37
tobs <- (xbar - mu0) / se
tobs
```

```
## [1] -3.141
```

Body temperature example

If you prefer to use tidyverse, you could equivalently do:

```
stats <- tempsample |>
    summarise(
        n = n(),
        xbar = mean(BodyTemp),
        s = sd(BodyTemp),
        se = s / sqrt(n),
        mu0 = 37,
        tobs = (xbar - mu0) / se
    )
stats
```

```
## # A tibble: 1 × 6
##       n  xbar      s      se   mu0   tobs
##   <int> <dbl>  <dbl>   <dbl> <dbl>  <dbl>
## 1    50 36.81 0.4252 0.06013    37 -3.141
```

Body temperature example

- Recall the value of the t-statistic for the observed sample:
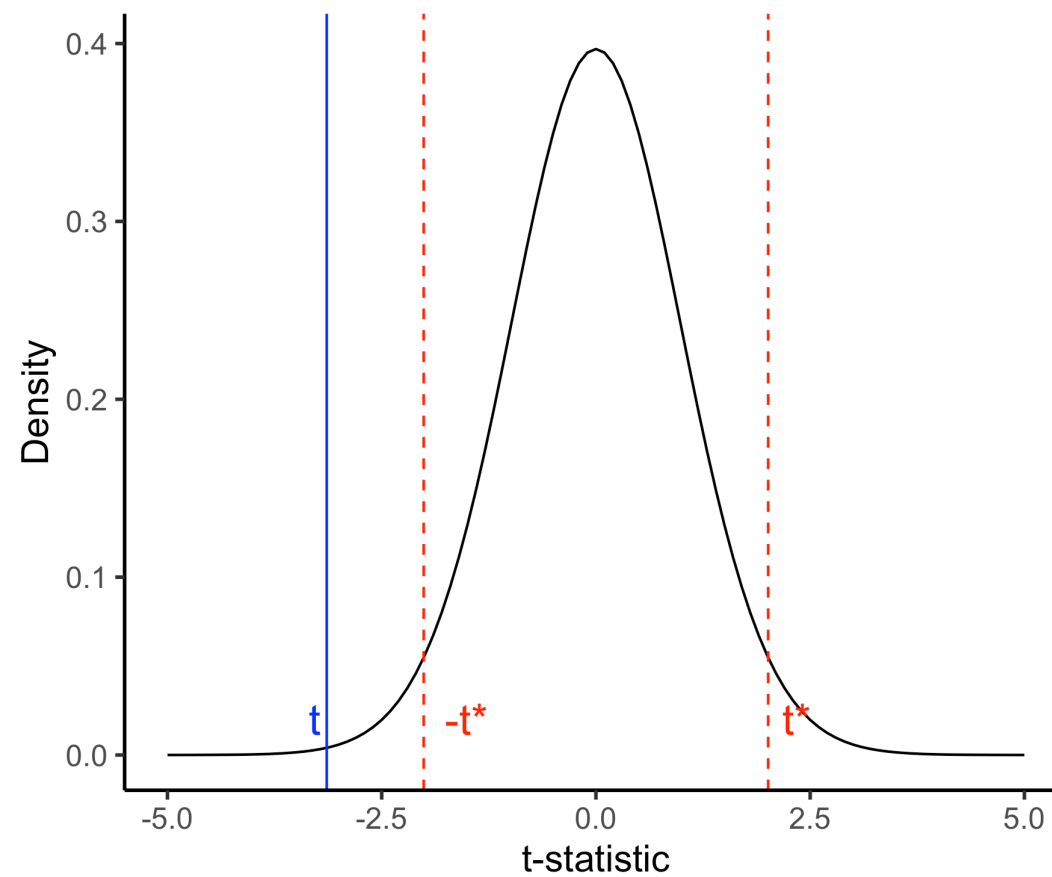
```
tobs
```

```
## [1] -3.141
```

- To make a decision, we compare the observed t-statistic, $t = -3.141$, to the critical values $\pm t^*$ from a $t(49)$ distribution.

```
tstar <- qt(c(0.025, 0.975), df = n - 1)
tstar
```

```
## [1] -2.01  2.01
```

- Two-sided alternative, so we ask: is the observed t **as or more extreme** than the two critical values?

  - Is $t \leq -t^*$ or $t \geq t^*$?

  - Yes, so reject H_0.

Body temperature example

Body temperature example

- If you were to compute the p-value, you would reach the same conclusion.

```
pvalue <- 2 * pt( abs(tobs), df = n - 1, lower.tail = FALSE)
pvalue
```

```
## [1] 0.002851
```

- We could say:

> Let μ denote the mean body temperature of all healthy humans. At the 5% significance level, we performed a one-sample t-test of $H_0 : \mu = 37$ against $H_1 : \mu \neq 37$. The sample data provided very strong evidence to reject the null hypothesis in favour of the alternative one that the mean body temperature of healthy humans is different from the long thought 37 °C; $t(49) = -3.14, p = .003$, two-sided.

Body temperature example

- At the $\alpha = 0.05$ significance level, the sample data provide very strong evidence to reject $H_0 : \mu = 37$ in favour of $H_1 : \mu \neq 37$.

- However, this doesn't give us any idea of what the actual population mean may be. It just tells us there is very strong evidence that it's not 37 °C.

> **Good practice**
> When you find a statistically significant result, it is good practice to follow-up the hypothesis test with a confidence interval, in order to provide the reader with an idea of what a plausible value of the population parameter may be.

```
tstar <- qt(c(0.025, 0.975), df = n - 1)
tstar
```

```
## [1] -2.01  2.01
```

```
xbar + tstar * se
```

```
## [1] 36.69 36.93
```

- We are 95% confident that the population mean body temperature for healthy humans is between 36.69 °C and 36.93 °C.

Hypothesis testing and Confidence intervals

- Notice that the 95% confidence interval [36.69, 36.93] does not include the hypothesised value in the null hypothesis (37).

  - As 37 is not within [36.69, 36.93], we reject the null hypothesis $(H_0 : \mu = 37)$ at the 5% significance level.

- The 95% CI gives us a range of plausible values for the population mean.

- As the claimed value 37 is not within the range, that is not a plausible value for μ. If you test whether the population mean could be equal to 37, you end up rejecting that value.

- **Generic fact:**
  The 95% CI gives you a range of values for the parameter that would **not** be rejected at the 5% significance level in a two-sided hypothesis test.

Body temperature example

- The p-value is .003, indicating that the sample data provide very strong evidence that the population mean is significantly different from 37 °C.

    - This is about **statistical significance**

    - That is, the result is *very unlikely* to come from a population where H_0 is true and the discrepancy between the sample and population mean is only due to random sampling.

- We are 95% confident that the population mean body temperature for healthy humans is between 36.69 °C and 36.93 °C.

    - The confidence interval tells us that perhaps the true population mean body temperature is somewhere in between 36.69 and 36.93, rather than 37.

    - Is this departure from 37 big enough to be of any practical impact for decision makers?

    - Pretty much no, it's negligible.

    - The result, while being significant, is not an **important** result.

    - Hence, statistical significance is different from practical significance (which we call importance), and a statistically significant result may not be of practical significance at all.

Significance vs Importance

- In scientific research, and in this course too, the word **significance** has a special meaning which is different from its everyday language meaning.

- It is a keyword reserved to mean **statistical significance**, i.e. the sample results being unlikely to be observed just because of random variation due to sampling when in fact H_0 is true in the population.

- This is **not** the same meaning as the sample results being **important** or **practically significant**.

Significance vs Importance

- You can have a significant result, but this doesn't necessarily mean that it's **important**.

  - It's good practice to always follow-up a significant result (i.e. when you reject the null hypothesis) with a confidence interval, to provide the reader with a measure of the "magnitude" of the departure from the hypothesised value $H_0 : \mu = \mu_0$.

  - In our case, the 95% confidence interval [36.69, 36.93] °C indicates that the departure from the currently accepted standard of 37 °C is not that big to be of much impact for decision makers.

  - This is an example of a statistically significant result that is not of practical importance.

Why both?

- If a confidence interval can also be used to perform a hypothesis test, why do we need both?

- The answer is that they both provide useful and complementary information.

  - A p-value gives you the **strength of evidence** that the sample data bring against the null hypothesis. This is not conveyed by a confidence interval.

  - The confidence interval gives you an idea of the "magnitude" of the population parameter, which the hypothesis test doesn't convey.

- In summary, report both if you have a significant result, i.e. if you reject the null hypothesis!

  - Always follow up a significant hypothesis test with a confidence interval.

  - *Side note*. If the hypothesis test is not significant, and you do not reject $H_0 : \mu = 0$ say, then you do have an idea about the magnitude of the population mean. In fact, by not rejecting the claim that $\mu = 0$ you are told that 0 is a plausible value for the population mean.

This week

Tasks

- Attend both lectures

- Attend your lab and work together on the lab tasks

    - Tip: read the worked example in advance!
    - Reminder: Formative Report C will be due 12 noon on 14 February 2025 (next week)

- Complete any lecture activities and/or readings

- Complete the weekly quiz

    - Opens Monday at 9am
    - Closes Sunday at 5pm

Support

- **Office hours**: for one-to-one support on course materials or assessments
  (see LEARN > Course information > Course contacts)

- **Piazza**: help each other on this peer-to-peer discussion forum

- **Student Adviser**: for general support while you are at university
  (find your student adviser on MyEd/Euclid)