Hypothesis testing: p-values

Data Analysis for Psychology in R 1

Semester 2, Week 2

Dr Umberto Noè

Department of Psychology
The University of Edinburgh

Learning objectives

1. Understand null and alternative hypotheses, and how to specify them for a given research question.

2. Understand the concept of and how to obtain a null distribution.

3. Understand statistical significance and how to calculate p-values from null distributions.

Part A

Introduction

Setting

- We cannot afford to collect data for the full population due to time and/or budget constraints

- Data collected for a random sample of size n

- We are interested in the population mean μ, but this is unknown as we cannot compute it

- Last week we learned how to:
    - obtain an estimate for the population mean $\rightarrow$ the sample mean $\bar{x}$
    - obtain a measure of precision of our estimate $\rightarrow$ the standard error $SE_{\bar{x}} = s/\sqrt{n}$
    - report the estimate along with the precision
    - compute and report a range of plausible values for the population mean, called **confidence interval** $\rightarrow$ $CI = \bar{x} \pm t^* \times SE_{\bar{x}}$

Where are we going?

- Are children exposed to pesticides more likely to develop ADHD (attention-deficit/hyperactivity disorder) that those who aren't?[1]
- Do students who eat breakfast achieve more than students who do not eat breakfast?
- Is the audience appreciation of shows appearing on Broadway lower than the audience appreciation of the touring version of the same show?
- If you want to remember something, should you take a nap or have some caffeine?

[1] Bouchard, M. F., Bellinger, D. C., Wright, R. O., & Weisskopf, M. G. (2010). Attention-deficit/hyperactivity disorder and urinary metabolites of organophosphate pesticides. Pediatrics, 125(6), e1270-e1277.

Where are we going?

- What do all of the previous questions have in common?
- Testing a claim about a population parameter!

Where are we going?

- Are children exposed to pesticides more likely to develop ADHD (attention-deficit/hyperactivity disorder) that those who aren't?
  - Is $p_{\text{exposed}} > p_{\text{not exposed}}$? where p is the proportion of all children diagnosed with ADHD. (Population proportion = p = parameter. Sample proportion = $\hat{p}$ = estimate).
- Do students who eat breakfast achieve more than students who do not eat breakfast?
  - Is $\mu_{\text{breakfast}} > \mu_{\text{no breakfast}}$? where μ is the mean achievement score.
- Is the audience appreciation of shows appearing on Broadway lower than the audience appreciation of the touring version of the same show?
  - Is $\mu_{\text{Broadway}} < \mu_{\text{Touring}}$? where μ is the mean audience appreciation score.
- If you want to remember something, should you take a nap or have some caffeine?
  - Is $\mu_{\text{nap}} \neq \mu_{\text{coffee}}$? where μ is the mean recall.

Where are we going?

- Many research hypotheses involve testing a claim about a population parameter.
- We will look at a widely applicable method (called **hypothesis test** or **test of significance**) that allows you to test an hypothesis about a population parameter.
- This method will allow you to answer many types of questions you may have about a population. All you have to do is
    - collect relevant sample data
    - perform a hypothesis test
    - report it correctly
- If you have a research question you are interested in, and you perform the steps above correctly, you may end up writing up your research results in your first journal paper after that!

Lecture example: Body temperature

- Today's recurring example will focus on answering the following research question:

  > Has the average body temperature for healthy humans changed from the long-thought 37 °C?

- We will use data comprising measurements on body temperature and pulse rate for a sample of $n = 50$ healthy subjects. Data link: https://uoepsy.github.io/data/BodyTemperatures.csv

```r
library(tidyverse)
tempsample <- read_csv('https://uoepsy.github.io/data/BodyTemperatures.csv')
glimpse(tempsample)  # n. rows and n. cols, variables, their type, and a preview
```

```
## Rows: 50
## Columns: 2
## $ BodyTemp <dbl> 36.44, 37.44, 37.22, 37.11, 36.67, 37.17, 37.22, 36.56, 36.00…
## $ Pulse    <dbl> 69, 77, 75, 84, 71, 76, 81, 77, 75, 81, NA, 78, 71, 80, 70, 7…
```

```r
tempsample <- tempsample %>%
    drop_na(BodyTemp)
```

Lecture example: Body temperature

```r
# both n. rows and n. cols
dim(tempsample)
```

```
## [1] 50  2
```

```r
# n. rows only: dim(tempsample)[1]
n <- nrow(tempsample)
n
```

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

```r
# sample mean
xbar <- mean(tempsample$BodyTemp)
xbar
```

```
## [1] 36.81
```

- The sample mean is $\bar{x}$ = 36.81 °C

Part B

Hypotheses and null distribution

Two hypotheses

- Let's start with an analogy from law. Consider a person who has been indicted for committing a crime and is being tried in a court.
- Based on the available evidence, the judge or jury will make one of two possible decisions:

  1. The person is not guilty.
  2. The person is guilty.

- Due to the principle of **presumption of innocence**, at the outset of the trial, the person is presumed not guilty.
  - "The person is not guilty" corresponds to what is called in statistics the **null hypothesis**, denoted H_0.
- The prosecutor's job is to prove that the person has committed the crime and, hence, is guilty.
  - "The person is guilty" corresponds to what is called in statistics the **alternative hypothesis**, denoted H_1.
- The evidence that the prosecutor needs to provide must be **beyond reasonable doubt**.

Two hypotheses

- In the beginning of the trial it is assumed that the person is not guilty.

- The null hypothesis H_0 is usually the hypothesis that is assumed to be true to begin with. It typically corresponds to "no change", "no effect", "no difference", "no relationship".

    - It involves the equality symbol $(=)$
    - The null hypothesis usually is the skeptical claim that nothing is different / nothing is happening.
    - Are we considering a (New! Improved!) possibly better method? The null hypothesis says, "Really? Convince me!" To convert a skeptic, we must pile up enough evidence against the null hypothesis that we can reasonably reject it.

- The alternative hypothesis is the claim that we wish to find evidence for. It is typically the hypothesis that embodies the research question of interest.

    - It involves the less than $(<)$ or greater than $(>)$ or not equal to $(\neq)$ symbols
    - If H_1 uses the symbol $<$, the test is called left-tailed or left-sided
    - If H_1 uses the symbol $>$, the test is called right-tailed or right-sided
    - If H_1 uses the symbol $\neq$, the test is called two-tailed or two-sided

Test of significance

- A **hypothesis test** (or **test of significance**) is a procedure for testing a claim about a population parameter (i.e. a property of a population).
- The test works by weighting the evidence **against** the null (and in favour of the alternative).
    - We want to be sure the sample data provide enough evidence against H_0 before rejecting it in favour of H_1.
- The evidence in statistics corresponds to the sample statistic (numerical summary of the sample data).
    - Informally, people say that the evidence corresponds to the sample data.
- The evidence provided must be **beyond reasonable doubt**.
    - If H_0 is true, it should be very unlikely for a random sample to give that value of the statistic.
      If a person is innocent, it should be very unlikely to pile up so much evidence against innocence.
    - If it were very likely for a random sample to give that value of the sample statistic when H_0 is true, then what we observed could just be a fluke due to random sampling rather than due to H_1.

Lecture example: Body temperature

> Has the average body temperature for healthy humans changed from the long-thought 37 °C?

- State the hypotheses using proper symbols for the population parameters.

$$H_0 : \mu = 37$$

$$H_1 : \mu \neq 37$$

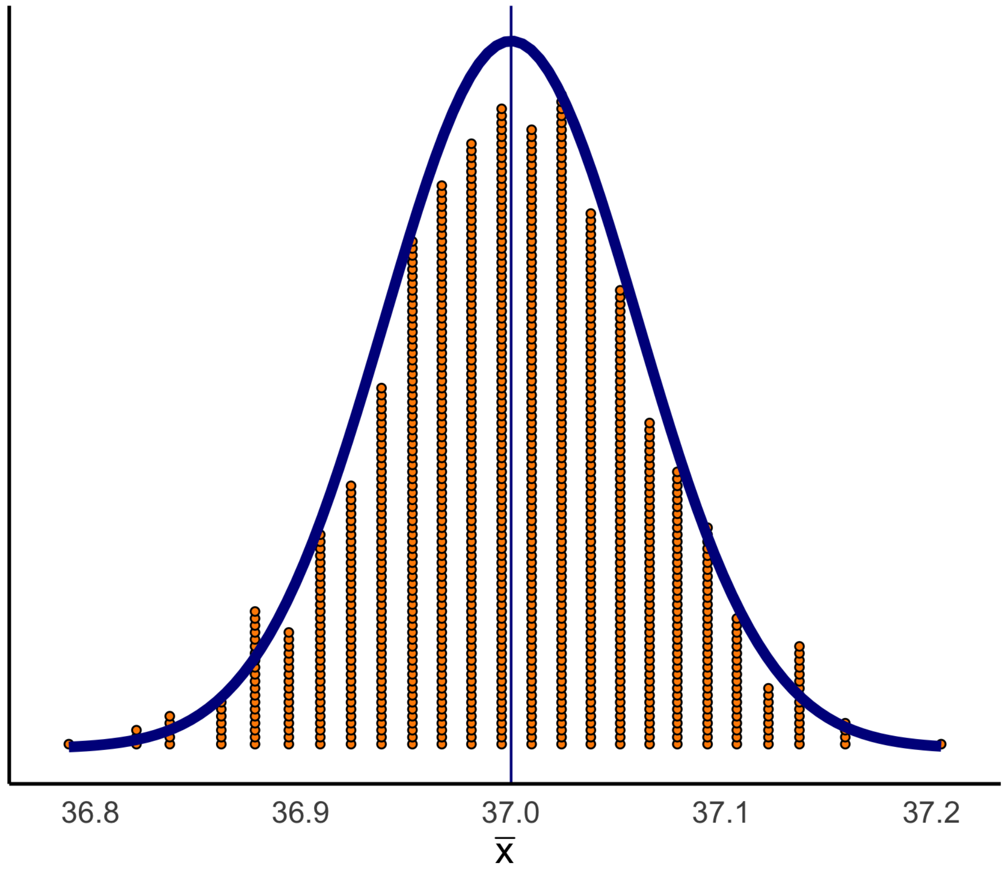
- From the sample data we can compute the sample mean, which is our estimate of μ

```
xbar <- mean(tempsample$BodyTemp)
xbar
```

```
## [1] 36.81
```

- $\overline{x} = 36.81$ °C, which differs from 37 °C
- Is this difference large enough to be really due to a systematic shift in the average body temperature of healthy humans?
- Or perhaps the population mean is truly = 37 °C, and the difference between 36.81 °C and 37 °C is simply due to random sampling?

Recap



Null distribution

- The sample mean varies from sample to sample, and all the possible values along with their probabilities form the sampling distribution:

$$\overline{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

- If the population mean μ was truly equal to 37, as the null hypothesis says, how would the sample means look?

- If $H_0 : \mu = 37$ is true, the sample mean would follow the distribution:

$$\overline{X} \sim N(37, \frac{\sigma}{\sqrt{n}})$$

- We can standardise it to obtain a distribution with mean = 0 and SD = 1 (**z-score**):

$$Z = \frac{\overline{X} - 37}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

Null distribution

- **However**, we cannot compute the population SD σ too...
- Estimate it with sample SD, denoted s. The distribution however becomes a $t(n-1)$
- When you standardise the sample mean using $SE_{\bar{x}} = s/\sqrt{n}$, you have the **t-statistic**:

$$\underbrace{t = \frac{\overline{X} - 37}{\frac{s}{\sqrt{n}}}}_{\textbf{t-statistic}} \sim t(n-1)$$

- The t-statistic is sometimes called the **t-score** (or t-scored sample mean, same thing)
- The distribution of the t-statistic, **assuming the null hypothesis to be true**, is called the **null distribution**.
    - It tells us which values of the t-statistic we would expect to see if H_0 were true.

Part C

t-statistic and p-value

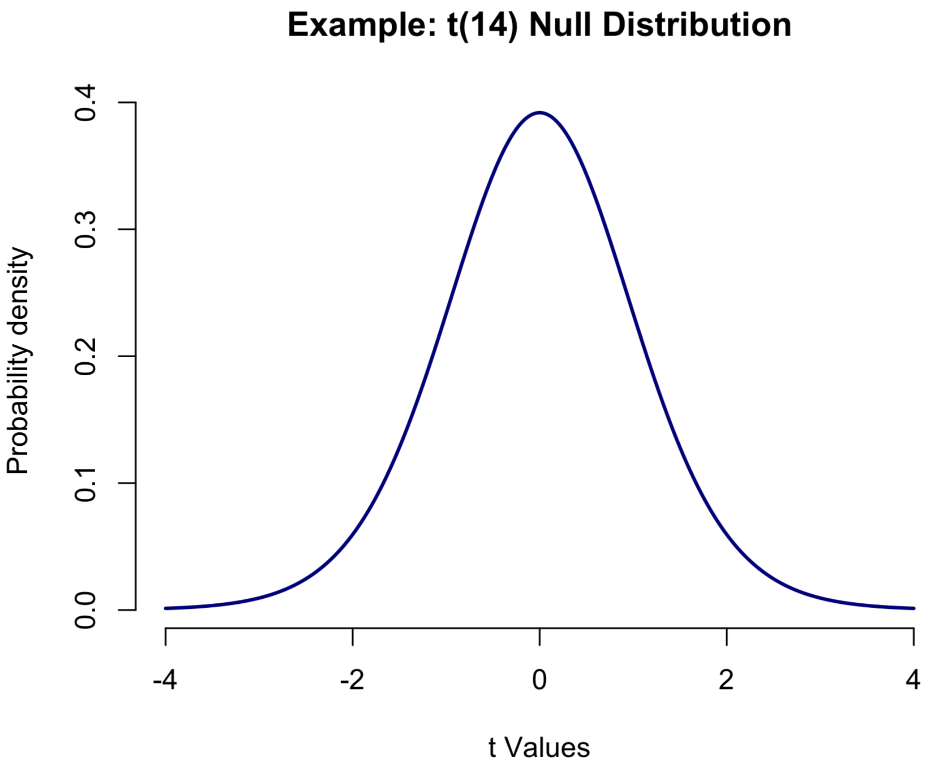
The t-statistic

- For $H_0 : \mu = \mu_0$ the t-statistic is:

$$t = \frac{\overline{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{\text{difference between sample and hypothesised mean}}{\text{variation in sample means due to random sampling}}$$

- The **t-statistic** measures how many standard errors away from μ_0 is our sample mean $\overline{x}$.
- It compares the difference between the sample and hypothesised mean, to the expected variation in the means due to random sampling.
- **Note**: The terms **t-score**, **t-statistic** and **t-value** are used as synonyms
- When referring to the t-statistic computed on the observed sample, people often say:
  - the observed value of the t-statistic
  - the observed t-value

Visually



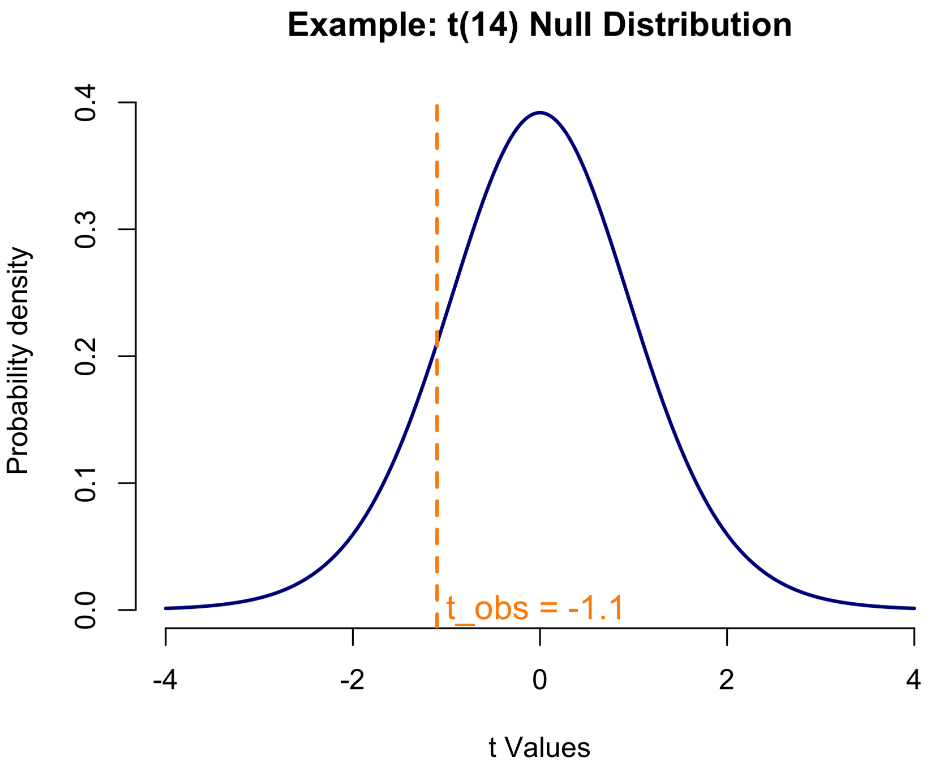
Consider $H_0 : \mu = \mu_0$

$$t = 0 \quad \text{when} \quad \frac{\overline{x} - \mu_0}{\frac{s}{\sqrt{n}}} = 0 \quad \text{when} \quad \bar{x} = \mu_0$$

Roughly speaking:

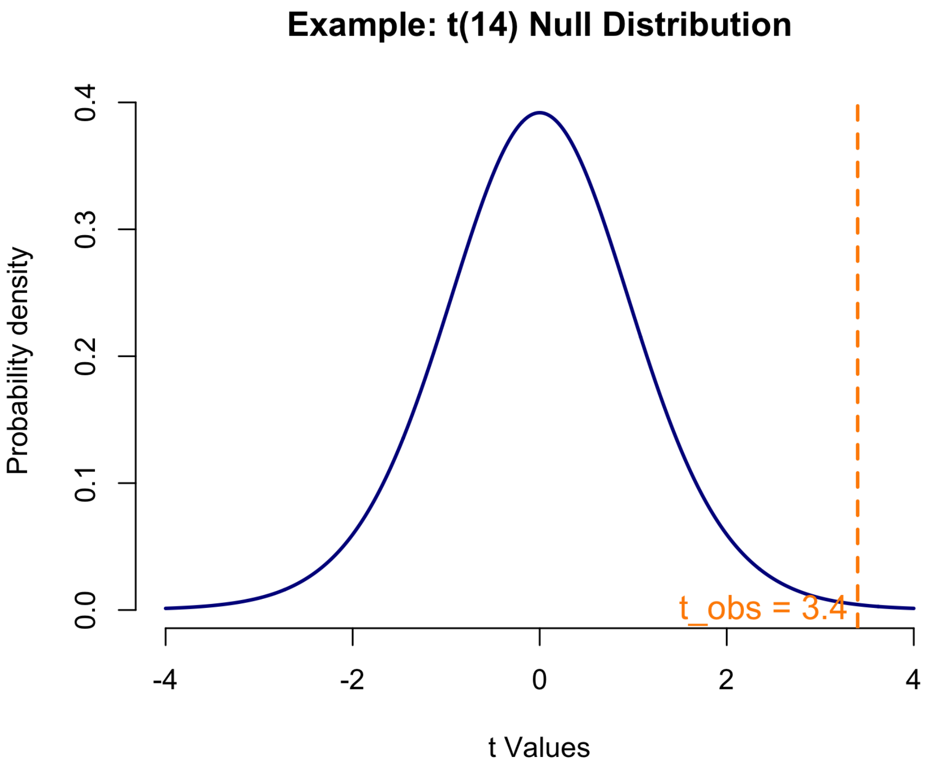
- We are very likely to see a t-score between -2 and 2 if in the population the mean is really μ_0 (37 in the Body Temperature example)
- We are very unlikely to see a t-score smaller than -2 or larger than 2 if in the population mean is really μ_0 (37 in the Body Temperature example)

Visually



- If our random sample leads to an observed t-value that has relatively high probability in the null distribution
  - There are many random samples leading to the same t-value when H_0 is true
  - Hence, it is very likely to obtain such t-value just from random sampling.

Visually



- If our sample leads to an observed t-value that has relatively low probability,
  - there are very few random samples leading to the same t-value when H_0 is true.
  - The observed t-value is **unlikely** to be obtained from random samples when H_0 is true. That surprisingly high or low t-value may be due to something else (our claim), rather than random sampling.

Evaluating how unlikely

- We need an objective criterion to evaluating how unlikely it is to see the observed t-value if H_0 is true.
- Just plotting a line on a graph can lead to very different conclusions based on the reader's perception of probability and their risk-aversion.

p-value

- In statistics, the evidence against the null hypothesis is provided by data (and not the prosecutor) and we use a probability to say how strong the evidence is.

- The probability that measures the strength of the evidence against a null hypothesis is called a **p-value**.

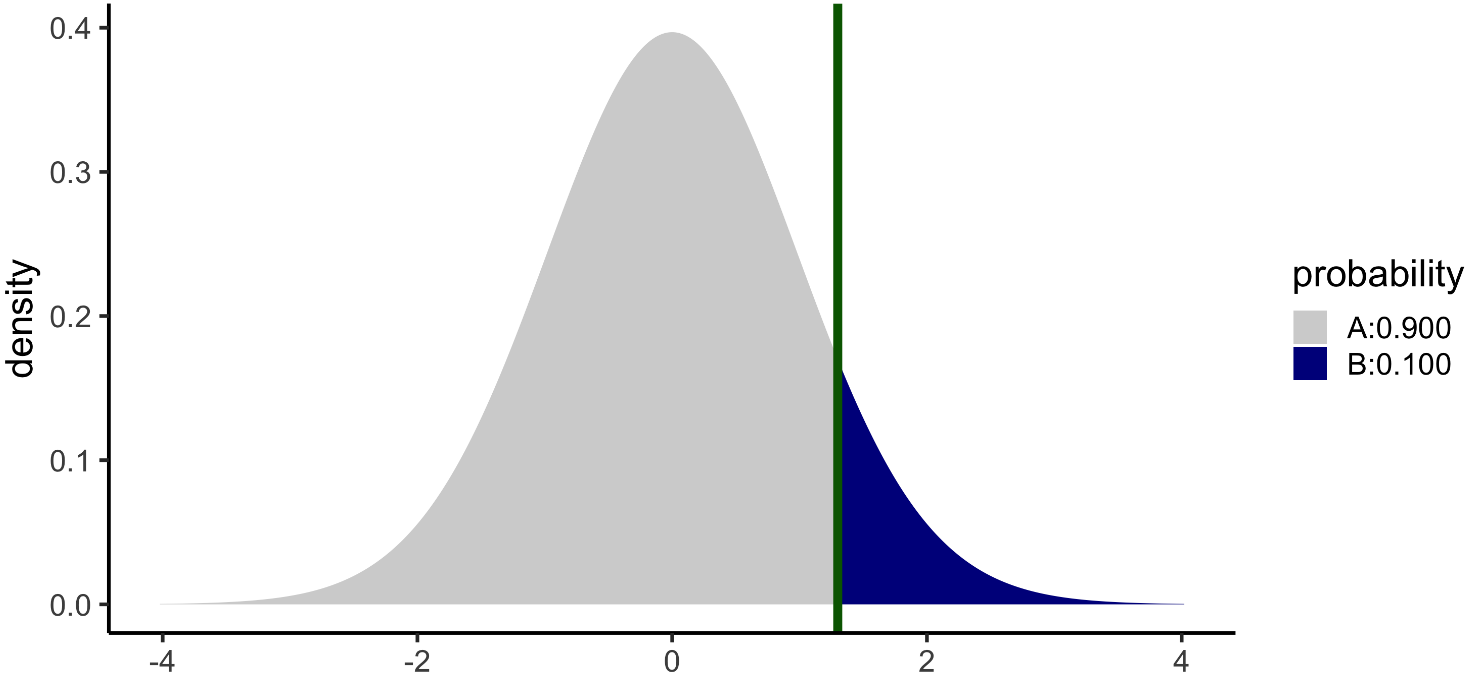
> **Definition**
> The p-value is the probability, computed assuming that H_0 is true, of obtaining a value of the t-statistic **at least as extreme as that observed**.

- Operationally, extreme corresponds to the direction specified by H_1.

  - If >, find the probability of larger t-scores than that observed
  - If < find the probability of smaller t-scores than that observed
  - If $\neq$ use both tails

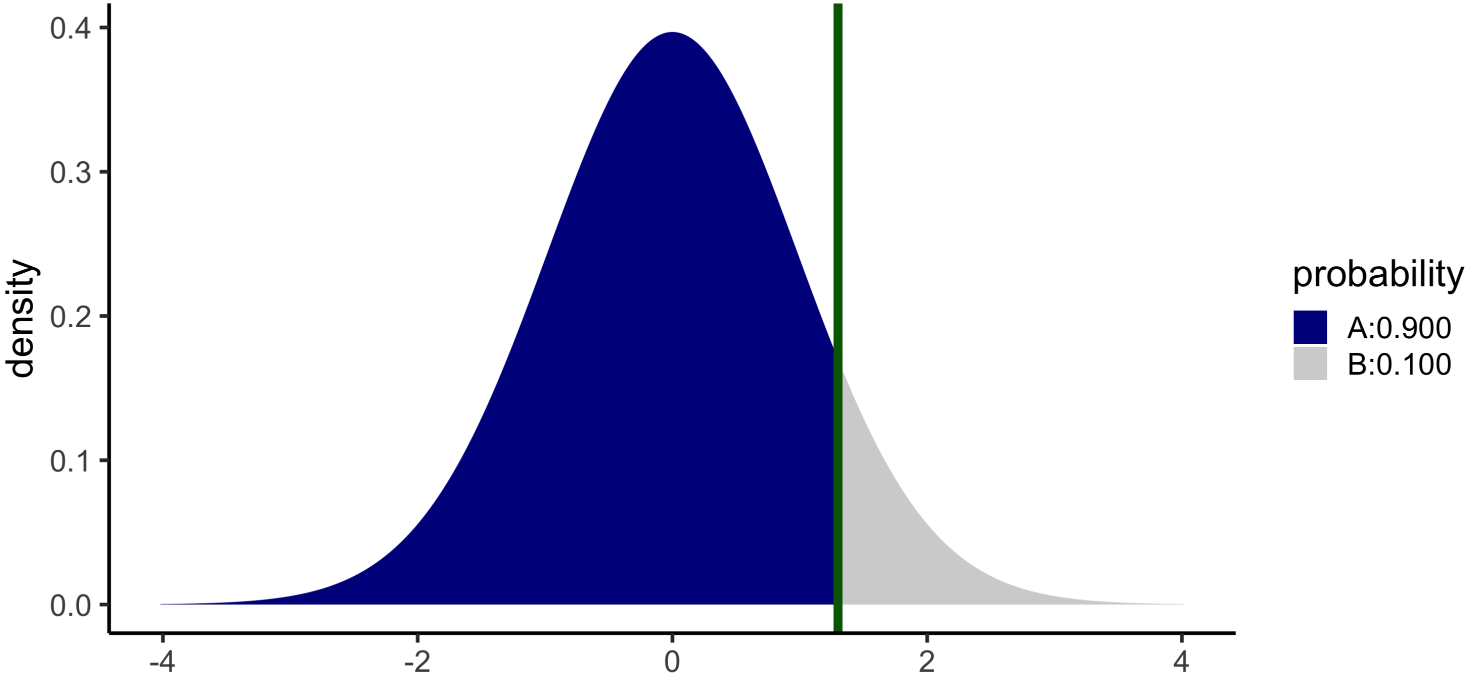
Visually: p-value

- If $H_1 : \mu > \mu_0$ and $t = 1.3$, **p-value = B**



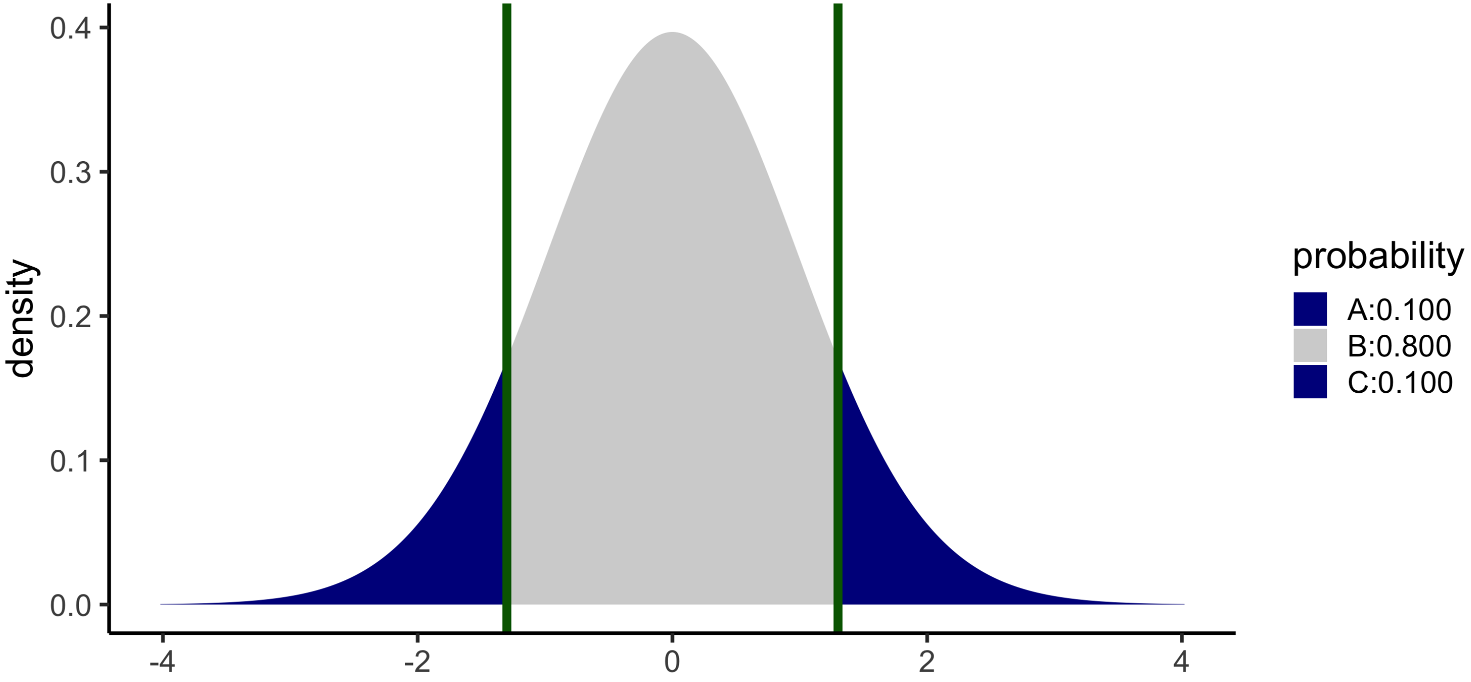
Visually: p-value

- If $H_1 : \mu < \mu_0$ and $t = 1.3$, **p-value = A**



Visually: p-value

- If $H_1 : \mu \neq \mu_0$ and $t = 1.3$, **p-value = A + C**



Body temperature example

- We have that $\bar{x} = 36.81$ °C. Let's compute the t-statistic, telling us how many SEs away from 37 °C the value 36.81 °C is.

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

mu0 <- 37  # null hypothesis value

tvalue <- (xbar - mu0) / SE
tvalue
```

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

The value of the t-statistic from the observed sample is

$$t = -3.141$$

Body temperature example

- Our alternative is $H_1 : \mu \neq 37$, so something is very different from that value either if it's (a) much bigger or (b) much smaller.
- The observed t-value is $t = -3.141$, so we compute the p-value as $P(T \leq -3.141) + P(T \geq +3.141)$
- If you drop the negative sign by using the absolute value $|t| = |-3.141| = 3.141$, you can write this as $P(T \leq -|t|) + P(T \geq +|t|)$.
- However, the t-distribution is symmetric, so those two probabilities will be the same.
- You can also compute it as $2 \cdot P(T \geq |t|)$.
- In R, the absolute value function is `abs()`

Body temperature example

```
tvalue
```

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

```
pvalue <- pt(-3.141, df = n-1) +
          pt(+3.141, df = n-1, lower.tail = FALSE)
pvalue
```

```
## [1] 0.002854
```

```
pvalue <- pt(-3.141, df = n-1) +
          (1 - pt(+3.141, df = n-1))
pvalue
```

```
## [1] 0.002854
```

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

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

Body temperature example

- We computed the probability of obtaining a t-score at least as extreme as the observed one when H_0 is true.
- The p-value is: $p = .003$

p-value

- The smaller the p-value, the stronger the evidence that the data provide against H_0.
- Small p-values are evidence against H_0, because they say that the observed result would be unlikely to occur if H_0 was true.
- Large p-values fail to provide sufficient evidence against H_0
- However, we need operational definition for *how small* a p-value should be to provide sufficient evidence against H_0. How small is small?

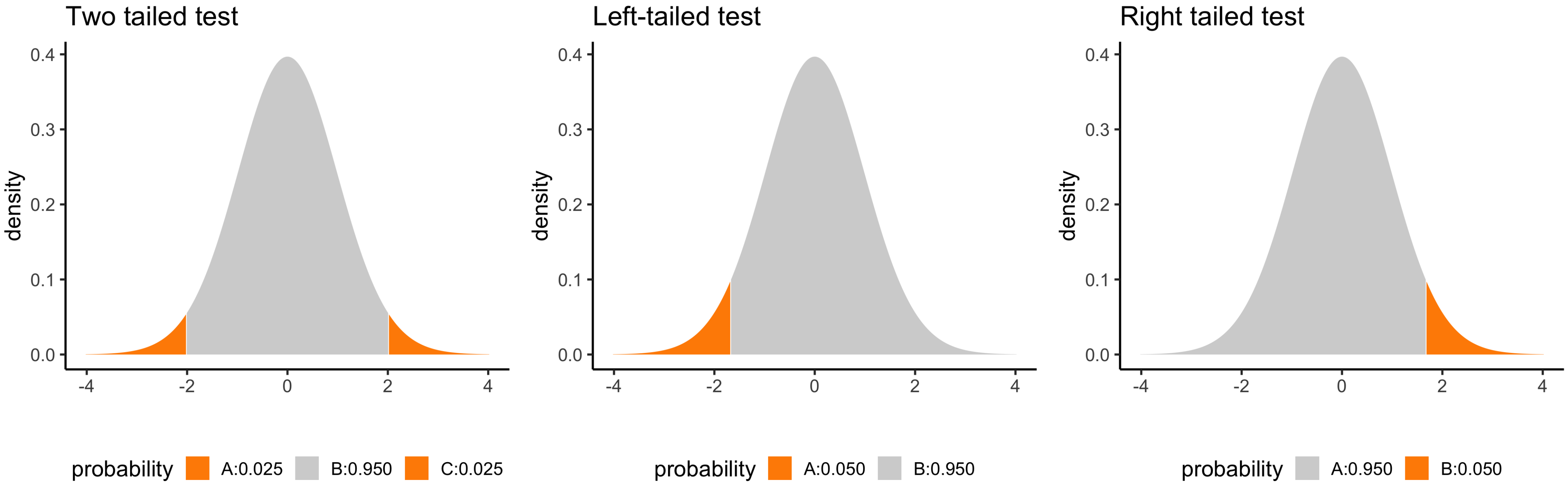
Part D

Significance level

Significance level

- We can compare a p-value with some fixed value (called **significance level** and denoted α) that is in common use as standard for evidence against H_0.
- The most common fixed values are $\alpha = 0.10$, $\alpha = 0.05$, and $\alpha = 0.01$.
- The value is chosen by the researcher (**you!**) once for all at the beginning of your study.
- It is important to clearly state the significance level at the start of your write-ups in every report or journal paper.
- If $p \leq 0.05$, there is no more than 1 chance in 20 that a sample would give evidence at least this strong just by chance when H_0 is actually true.
- If $p \leq 0.01$, we have a result that in the long run would happen no more than once per 100 samples when H_0 is true.

Visually: $\alpha = 0.05$



Statistical significance: interpretation

- If the p-value $\leq \alpha$, we say that the data are statistically significant at level α, and we reject H_0 in favour of H_1.
    - We say that the sample data provide significant evidence against H_0 and in favour of H_1.
- If the p-value $> \alpha$, we say that the data are **not** statistically significant at level α, and we do not reject H_0.
    - We say that the sample data do not provide sufficient evidence against H_0.
- "Significant" is a technical term in scientific research and it doesn't have the same meaning as in everyday English language.
    - It does **not** mean "important".
    - It means "unlikely to happen by random variations from sample to sample alone (assuming the null hypothesis is true)".

Guidelines for reporting strenght of evidence

The following table summarizes in words the strength of evidence that the sample results bring in favour of the alternative hypothesis for different p-values:

Approximate size of p-value	Loose interpretation
p-value > 0.1	little or no evidence against H_0
$0.05 <$ p-value ≤ 0.1	some evidence against H_0
$0.01 <$ p-value ≤ 0.05	strong evidence against H_0
p-value ≤ 0.01	very strong evidence against H_0

Reporting

- It is important to always report your conclusions in full, without hiding information to the reader.
- Restate your decision on whether you reject or fail to reject H_0 in simple non-technical terms, making sure to address the original claim, and provide the reader with a take-home message.
- Report test as follows: t(`df`) = `tvalue`, p = `pvalue`, `one/two`-sided.
  - t(49) = -3.14, p = .003, two-sided
- According to APA style, **don't** include the zero before the decimal place for p-values.
- Irrespectively of your α level, if your p-value is $\geq$ .001 it is good practice to report it **in full** but using proper rounding.
- Irrespectively of your α level, if your p-value is < .001 you can just report it as p < .001 as people don't really care about 5th or 6th decimal numbers.

Body temperature example

At the $\alpha = 0.05$ significance level, we performed a two-sided hypothesis test against the null hypothesis that the mean body temperature for all healthy humans is equal to 37 °C.
The sample results provide very strong evidence against the null hypothesis and in favour of the alternative one that the average body temperature differs from 37 °C; $t(49) = -3.14, p = .003$, two-sided.

Note

- Failing to find sufficient evidence against H_0 means only that the data are **consistent** with H_0, not that we have proven H_0 to be true.
- Example: not finding sufficient evidence that person is guilty doesn't necessarily prove they are innocent. They could have just hidden every single possible trace.

This week

- Attend the Live R programming lecture
- Read the lab worked example and complete your lab
- Complete any assigned readings
- Go to office hours if you have questions
- Complete the weekly quiz
    - Opens Monday at 9am
    - Closes Sunday at 5pm