universität wien

BAK3: Introduction to Quantitative Methods

Week 10: Hypothesis Testing II

Leonardo Carella

universität wien

The Plan for today

- Statistics:
  - **Two-sample tests** for differences in means and proportions.
- Research Design:
  - The **concept of "Average Treatment Effect"** in experimental and quasi-experimental research designs: a very gentle introduction.
- Coding in R:
  - **Two-sample tests in R,** plus dot-and-whisker plots (if there's time).
  - The effects of the 'Ibiza affair': an example of quasi-experimental design.

universität wien

Hypothesis Testing: Recap

- Framework to evaluate the plausibility of **hypotheses about the population** from a sample estimate via **proof of contradiction**:
  - State the “null hypothesis”: the thing we want to disprove, or H_0.
  - Compare our sample estimate (e.g. our sample mean) to the parameter value that we are hypothesising under H_0.
  - *p*-value: summary of the evidence against the “null hypothesis”.

universität wien

Hypothesis Testing: Recap

- Last time:
  - **one-sample t-test** for sample means: I want to prove that my value is **different from** a certain value that I define as important.
  - It will compute a **t-statistic**: the difference between the observed mean and the population mean under the null hypothesis, expressed in number of standard errors.
  - Produces also a **p-value** that indicates how likely it is to get a sample mean as far from the population mean **under the null hypothesis** as the sample mean we actually observe.

universität wien

Hypothesis Testing: Recap

- Some things to remember:
  - A **low p-value** means means that the observed difference between the sample mean and the hypothesised population mean is unlikely to have occurred by random chance if the null hypothesis were true.
  - We reject the null at the 95% level when $p < 0.05$ (***a convention***).
  - For tests of **proportions**, we use a different test (chi-squared test), because we assume a binomial distribution of the population. Same interpretation of the p-value, but no t-statistic here.

universität wien

Two-Sample t-tests

- Test the **significance of a difference-in-means**.
- We have already encountered, informally, differences-in-means:
  - As “mean comparison” (week 6): “The unemployed support higher levels of welfare spending than employed people” or “incumbents receive more votes than non-incumbent candidates on average.”
  - Assignment 1 (week 4): “What is the average (mean) GDP per capita in countries rated “Free” by Freedom House and in countries rated something other than “Free” by Freedom House”?

universität wien

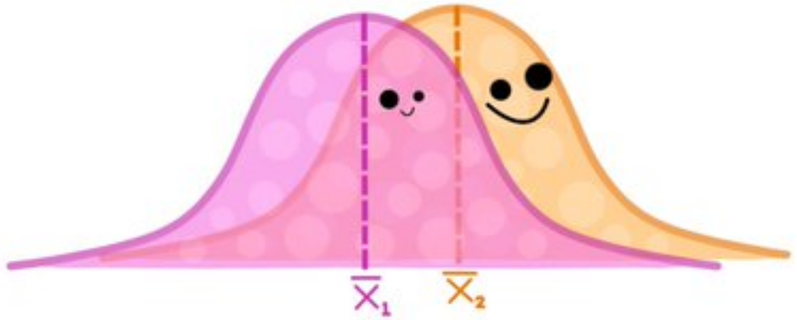
Two-Sample t-tests

- Test the **significance of a difference-in-means**.
- H_0: in the population there's **no difference** between the mean of group A and the mean of group B.
- H_a: in the population, group A has **a different mean** than group B.
- *p*-value: probability to obtain a difference between the mean of group A and the mean of group B **at least as large** as the one we observe in our sample, if group A and group B in the population belonged to distributions **with the same mean**.

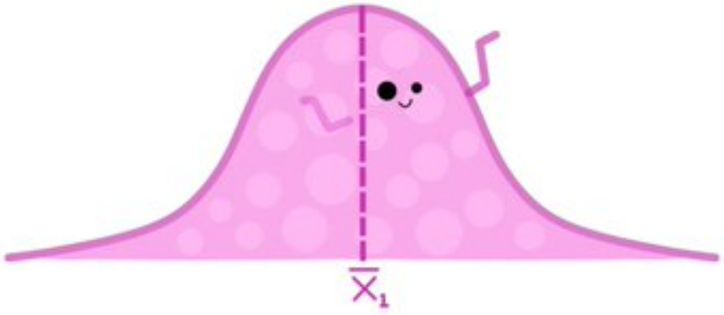
LET'S START HERE: if random samples are drawn from populations w/ the same mean...

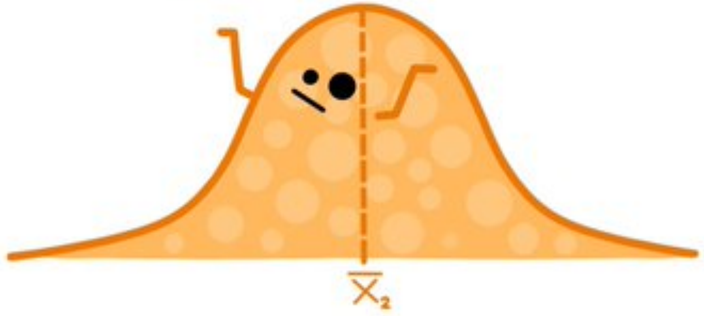
(i.e. the same population)

Then it is more likely that the 2 sample means will be close together...



...and it is less likely (but always possible!) that the sample means will be far apart.





@allison_horst

Illustration credits: Allison Horst

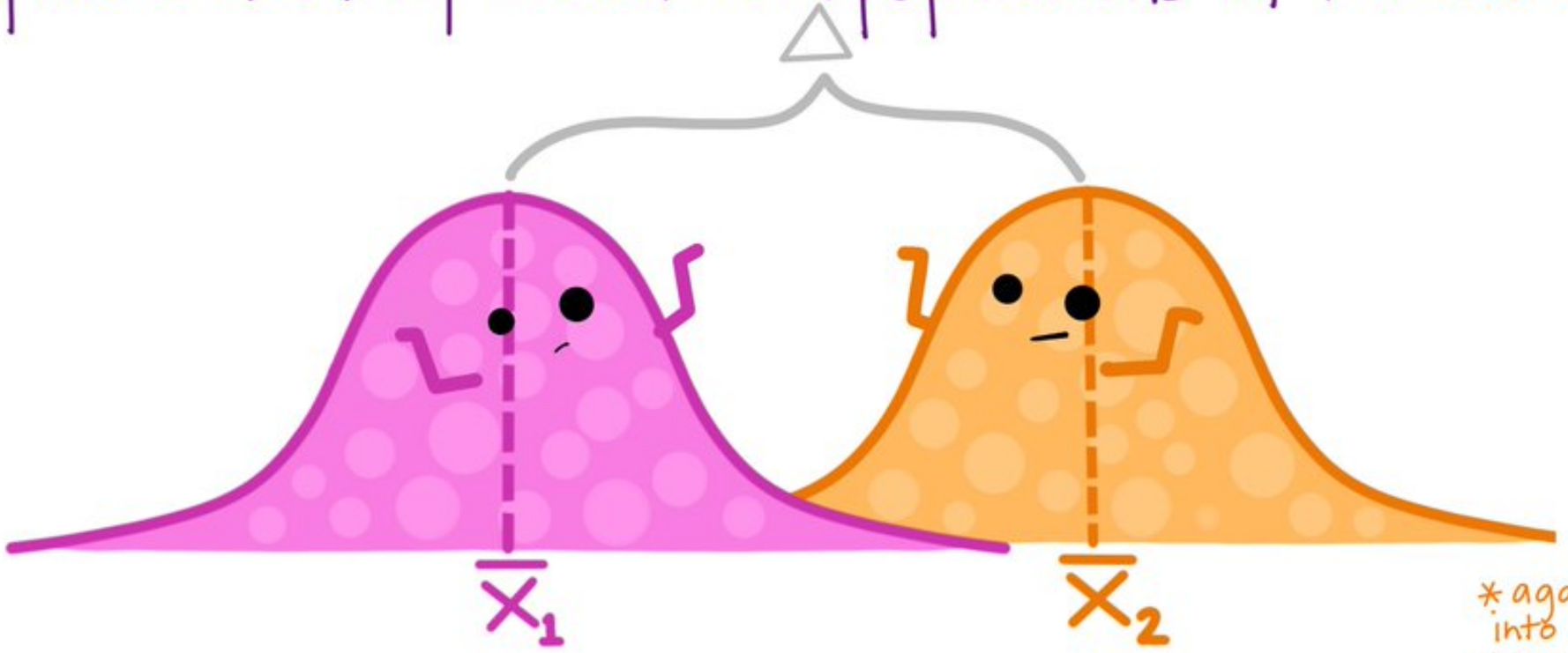
in OTHER WORDS... : The more different the sample means are,* the less likely it is they were drawn from populations w/ the same mean.

*(when taking into account sample spread + size, & assuming we've randomly sampled

Illustration credits: Allison Horst

So for our 2 random samples, we ask:

WHAT IS THE PROBABILITY OF GETTING 2 SAMPLE MEANS THAT ARE AT LEAST THIS DIFFERENT,* if they were actually drawn from populations w/ the same mean?

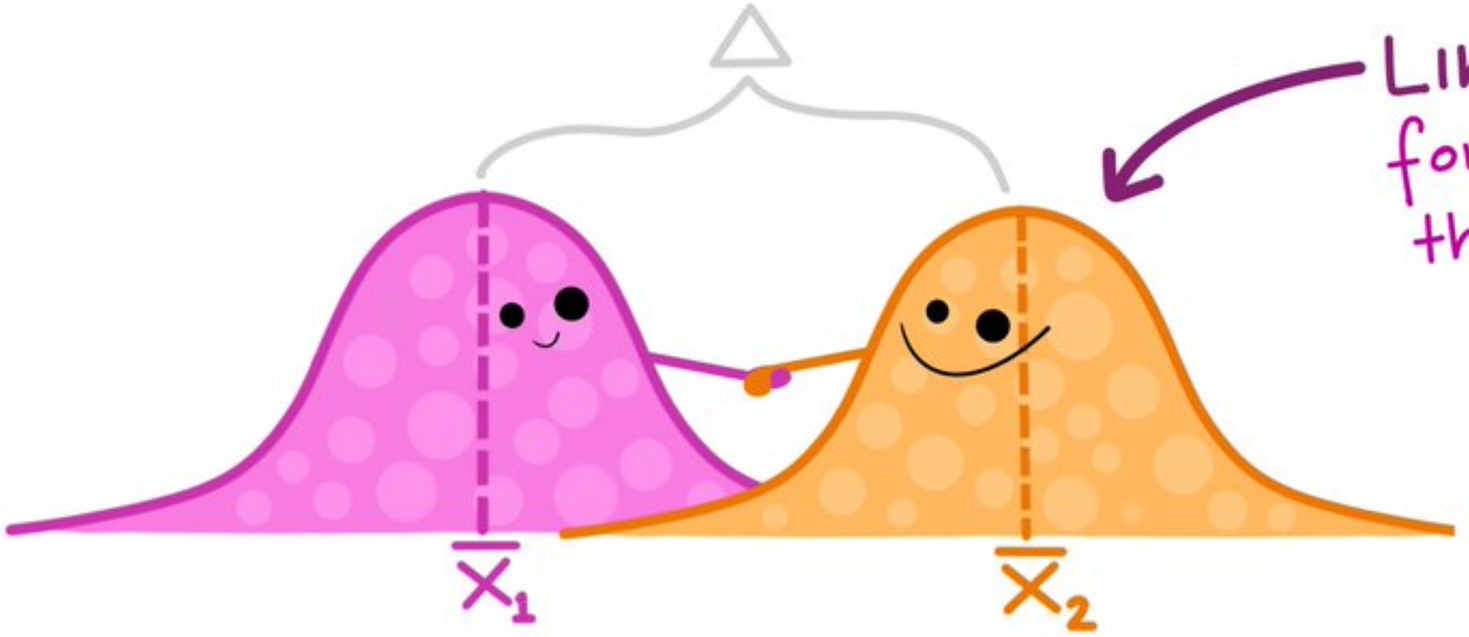


*again when taking into account sample spread & size, + assumptions...

Illustration credits: Allison Horst

That's our p-value!

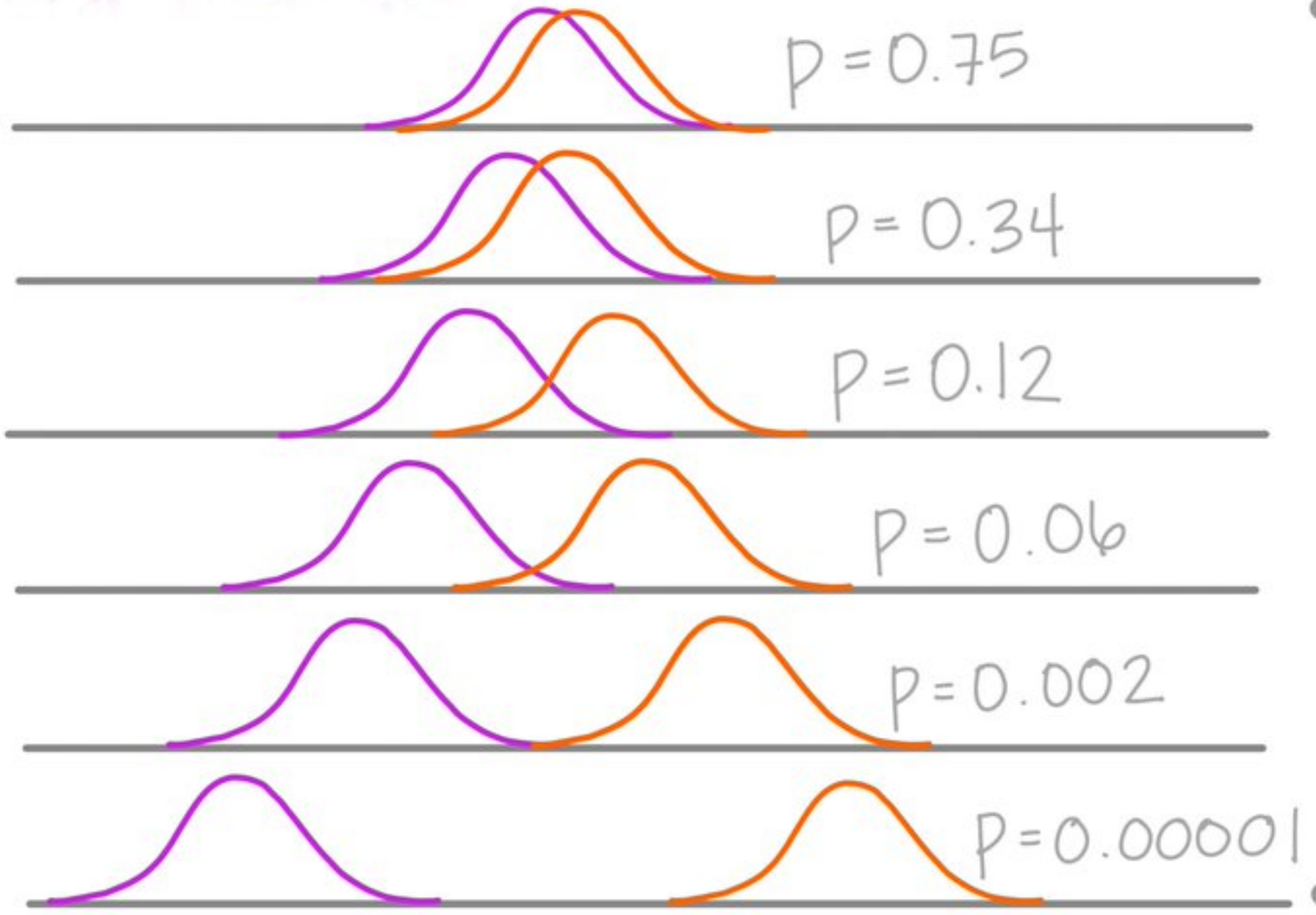
WHAT IS THE PROBABILITY OF GETTING 2 SAMPLE MEANS THAT ARE AT LEAST THIS DIFFERENT, if they were actually drawn from populations w/ the same mean?



LIKE: If a 2-sample t-test for these samples yields p=0.03, that means there is a 3% chance of getting means that are at least this different, if they're drawn from populations with the same mean.

Illustration credits: Allison Horst

P-VALUES, SCHEMATICALLY:



Higher p-values

(HIGHER PROBABILITY OF 2 SAMPLE MEANS BEING AT LEAST THIS DIFFERENT, IF DRAWN FROM POPULATIONS WITH THE SAME MEAN) = LESS EVIDENCE OF DIFFERENCES BETWEEN POPULATION MEANS

Lower p-values

(LOWER PROBABILITY OF 2 SAMPLE MEANS BEING AT LEAST THIS DIFFERENT, IF DRAWN FROM POPULATIONS WITH THE SAME MEAN) = MORE EVIDENCE OF DIFFERENCES BETWEEN POPULATION MEANS

Illustration credits: Allison Horst

Question:

WHEN DO WE HAVE ENOUGH EVIDENCE TO SAY THERE IS A SIGNIFICANT DIFFERENCE?

Answer:

WHEN OUR P-VALUE IS BELOW OUR SELECTED SIGNIFICANCE LEVEL (α), USUALLY (BUT NOT ALWAYS) = 0.05.

Which means:

IF THE PROBABILITY (p-value) OF FINDING AT LEAST OUR DIFFERENCE IN SAMPLE MEANS (IF THEY WERE DRAWN FROM POPULATIONS WITH THE SAME MEANS) IS LESS THAN 5%, THAT'S ENOUGH EVIDENCE FOR US TO DECIDE THEY ARE LIKELY FROM POPULATIONS WITH UNEQUAL MEANS.

Illustration credits: Allison Horst

universität wien

Two-sample t-test

- Example: We a sample of 220 **rural and urban** Austrians' support for a tax on high-emission vehicles (x) on a scale from 0 to 10.
  - Urban sample: $n_1 = 120$; $\bar{x}_1 = \mathbf{6.13}$; $s_1 = 1.79$.
  - Rural sample: $n_2 = 100$; $\bar{x}_2 = \mathbf{5.36}$; $s_2 = 1.83$.
- Difference-in-means: $\bar{x}_1 - \bar{x}_2 = \mathbf{6.13 - 5.36 = 0.77}$.
  - The null hypothesis: $\mu_1 = \mu_2$, or equivalently $\mu_1 - \mu_2 = 0$
  - The alternative hypothesis: $\mu_1 \neq \mu_2$, or equivalently $\mu_1 - \mu_2 \neq 0$

Two-sample t-test

- T-statistic: the difference between (1) the observed difference-in-means and (2) the difference-in-means under the null hypothesis (0), expressed in number of standard errors of the difference-in-means.

The observed difference-in-means

$$t = \frac{\bar{x}_1 - \bar{x}_2 - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

The population difference-in-means under the null hypothesis

formula for the standard error of the difference in means (no need to learn it!)

- In our case, $t = \frac{0.77}{0.245} = 3.15$

universität wien

Two-sample t-test

- What's the probability of observing a difference-in-means that is **3.15** standard errors away from 0 if the population difference-in-means is 0?
- **Very low!** We know from the CLT that the distribution of sample estimates (like the difference-in-means) is approximately normal.
- So about 95% of sample estimates will be within 1.96 standard errors of the population difference-in-means, 99% within 2.58 etc.
- In fact, **p = 0.002**. We reject the null. The difference in support for the tax between urban and rural residents is **statistically significant**.

universität wien

Two-sample t-test in R

```
> t.test(df$support ~ df$residence)

   Welch Two Sample t-test

data:  df$support by df$residence
t = 3.1545, df = 209.13, p-value = 0.001845
alternative hypothesis: true difference in means between
group Urban and group Rural is not equal to 0
95 percent confidence interval:
 0.2901702 1.2571825
sample estimates:
mean in group Urban mean in group Rural
           6.130883            5.357207
```

Two-sample test for proportion

- Remember our example from week 2 (Bertrand and Mullainathan, 2004)

	Call-Back Rate for White Names	*Call-Back Rate for African American Names*
Sample:		
All sent resumes	**10.06%** [2445]	**6.70%** [2445]

- Same story as last time:
- A t-test will usually do just about fine, but ***technically*** you should also make the assumption that the population distribution is binomial (only 0-1).
- The chi-squared test for equality of proportion does that.

universität wien

Why do we care?

- Difference-in-means (and differences-in-proportions) are descriptive statistics. **No causal interpretation** (even if $p < 0.00000001$)…
- …***Unless*** our groups are defined by the random assignment of the thing whose effect we care about. We usually talk about this "thing" as the **treatment**, borrowing a term from medical research.
- In the Bertrand and Mullainathan example: CVs are the same, **except for the racially distinctive name** of the applicant, which applies to some CVs (the treatment group), and not in others (the control group).

universität wien

ATEs

- In these cases, we can discuss the difference-in-means (or difference-in-proportions) in causal terms, as the **Average Treatment Effect**.

$$\text{ATE} = \bar{Y}(\text{treatment} = 1) - \bar{Y}(\text{treatment} = 0)$$

	Call-Back Rate for White Names	*Call-Back Rate for African American Names*
Sample:		
All sent resumes	**10.06%** [2445]	**6.70%** [2445]

universität wien

Experiments

- In a randomised experiment, people are assigned to the two groups by chance: a **treatment group**, and a **control group**.
- Otherwise, the groups start out (roughly) similar in all other respects.
- So any difference in their mean outcomes can be attributed to the treatment, and we can use the language of cause and effect:
- The difference-in-means between the mean of the outcome variable computed in the treatment and the mean of the outcome variable computed in control group is the **Average Treatment Effect**.

universität wien

Quasi-Experiments

- Often, we cannot randomly assign our independent variable:
  - **Impossible**: how do we randomly assign “a presidential system” to some countries but not to other?
  - **Unethical**: would you randomly assign some people to smoking cigarettes for a year to test whether smoke is bad for you?
  - Or just plain **expensive**. I’d love you to run your own experiment for the seminar paper, but that’s going to be several thousand euros each.

universität wien

Quasi-Experiments

- A lot of work in political science today is conducted with designs where the ‘treatment’ (= the independent variable whose effect on some outcome we care about) is assigned **‘as if random’**.
- If we can make a credible case that our groups are **similar, except for that one thing**, then we can apply the same logic as experiments, but nature, institutions, or timing do the “randomisation” for us.

universität wien

Sharp Cutoffs

- Sometimes access to something changes **abruptly** at a cutoff value:
  - Eligibility to vote on your 16th birthday.
  - Electoral thresholds: e.g. in Austria, if a party gets 3.99% of the vote, they are out of parliament; if they get 4.01 they are in.
- Units just below and just above the cutoff are often very similar, except for whether they receive the “treatment”: eligibility to vote or entry in parliament.
- We can compute the difference-in-means on some outcome (dependent) variable between units “just above” vs. “just below” the cutoff, and argue that it is the average effect of the treatment.

Carpenter, C., & Dobkin, C. (2009). The effect of alcohol consumption on mortality: regression discontinuity evidence from the minimum drinking age. *American Economic Journal: Applied Economics*, *1*(1), 164-182.

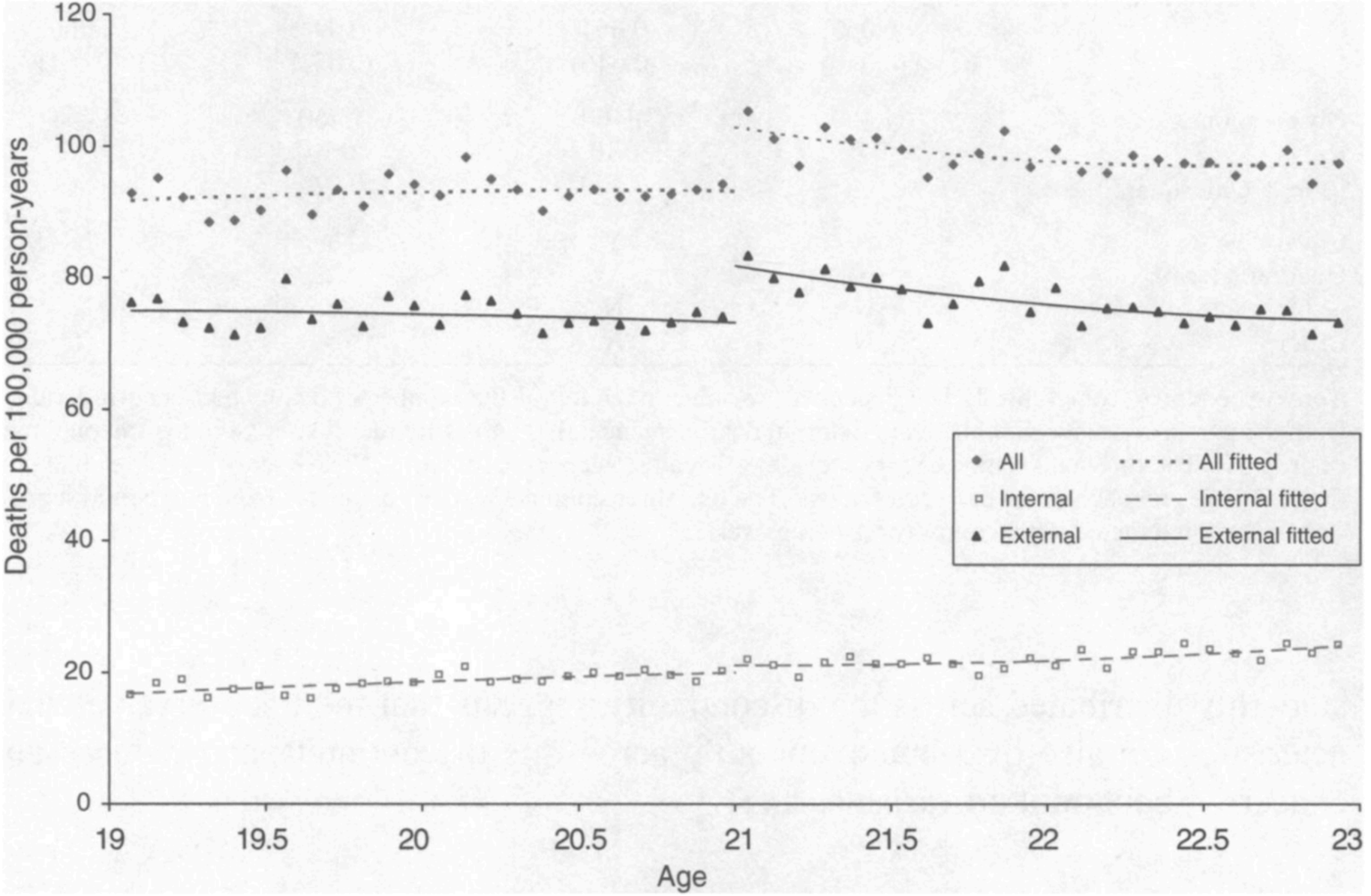


FIGURE 3. AGE PROFILE FOR DEATH RATES

Carpenter, C., & Dobkin, C. (2009). The effect of alcohol consumption on mortality: regression discontinuity evidence from the minimum drinking age. *American Economic Journal: Applied Economics*, *1*(1), 164-182.

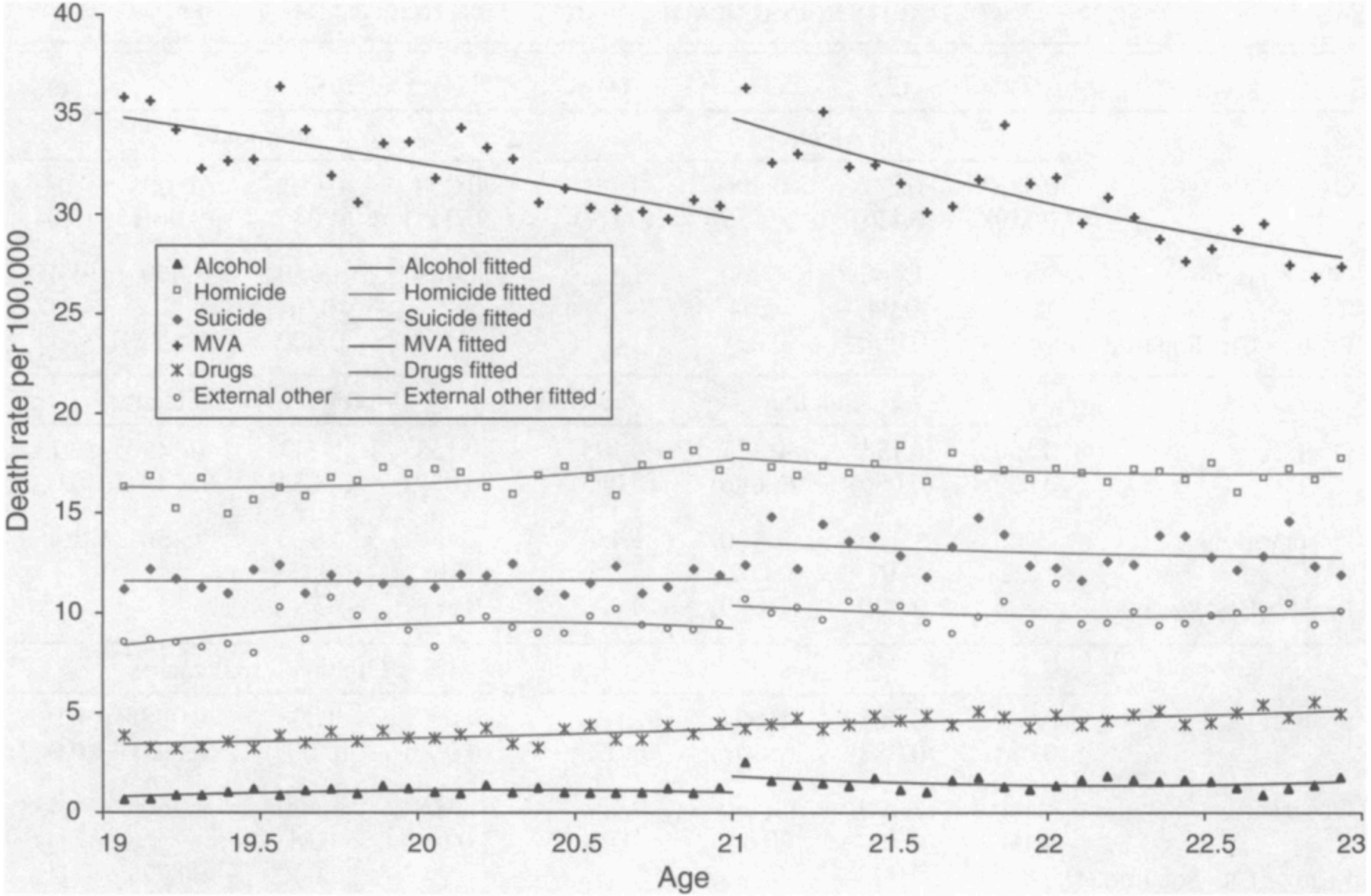


FIGURE 4. AGE PROFILES FOR DEATH RATES BY EXTERNAL CAUSE

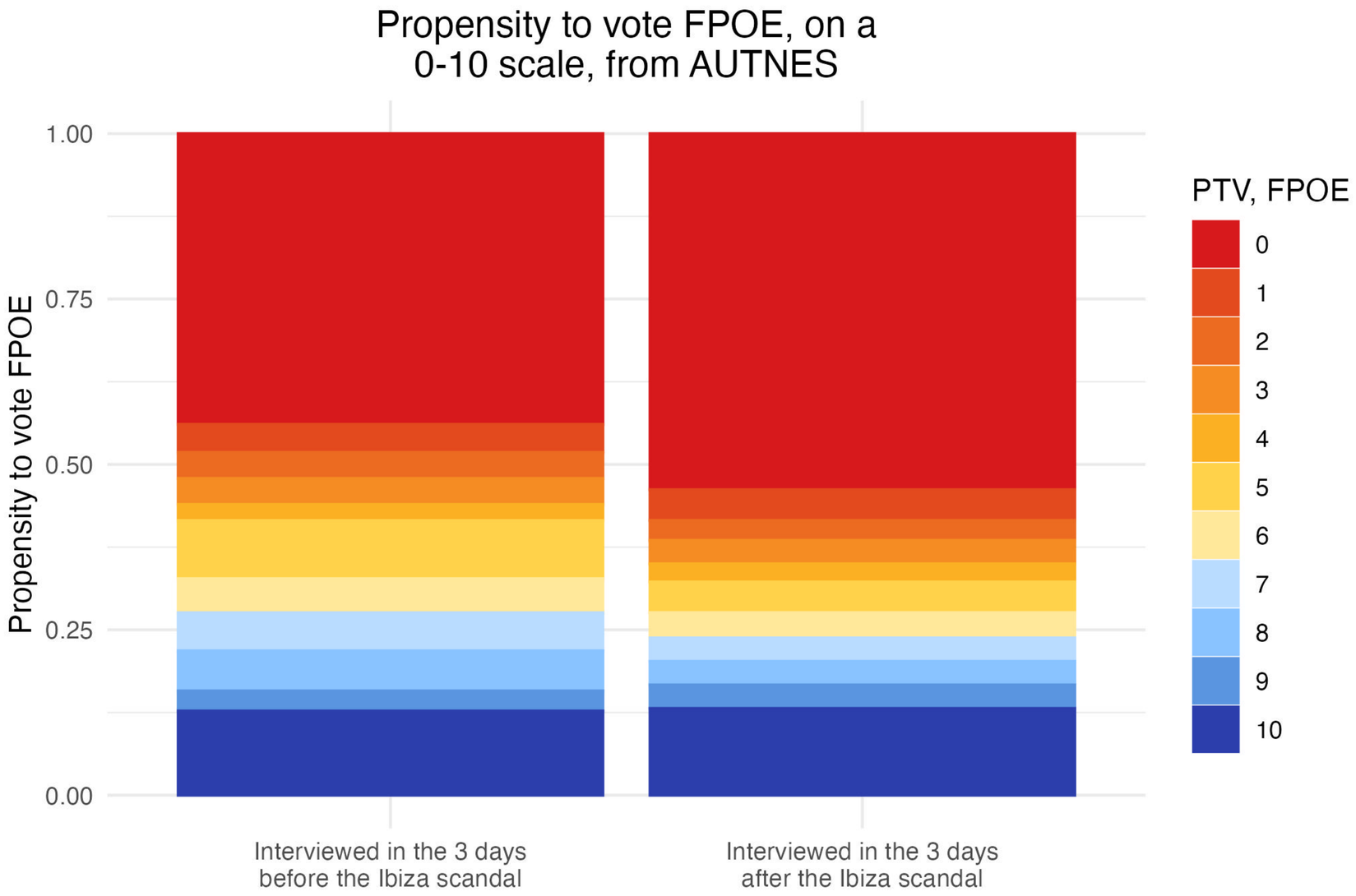
universität wien

Unexpected Events

- Imagine you're running a survey: it usually takes a week or two to get all the respondents you need to fill up your sample.
- Something **big** happens in the days of the fieldwork: a big scandal breaks out, a new policy is announced, there's a terrorist attack etc.).
  - Respondents interviewed just before the event form the **control group**.
  - Respondents interviewed just after the event form the **treated group**.
- If your respondents are **equally likely** to be interviewed at any point during your fieldwork, then the difference in means between "before" and "after" respondents estimates the **ATE of the event** on some variable of interest.

universität wien

Unexpected Events



universität wien

Summing Up…

- Two-sample tests tell us whether observed differences between two groups are significant, or may have been due to random chance.
- *p*-value summarises our evidence against the null that the two groups have the same mean in the population, i.e. that the difference-in-means is zero.
- Two-sample t-test for differences-in-means; two-sample chi-squared test for differences-in-proportions.
- When the independent variable is assigned “at random”, the difference-in-means between ‘treatment’ and ‘control’ can be interpreted as the Average Treatment Effect: i.e., ***in experiments and quasi-experiments***.