

Testing Hypotheses

Comparing groups, and when a difference is more than luck

Muhtasim Munif Fahim, MSc

Stat.BD

Data Science Research Lab, Department of Statistics & Data Science,
University of Rajshahi

Session 4 of 14 • Theory

Last session put a margin on an estimate. This one asks whether a difference is real

Session 3 gave us

- ▶ probability, and the normal curve
- ▶ the standard error: how much an estimate wobbles
- ▶ the 95% confidence interval

Today we use the same standard error to judge a *difference*.

Today, in five acts

-
- | | |
|-----|--|
| I | The logic of a test, and the p -value |
| II | Comparing two groups: t -tests |
| III | Comparing three or more: ANOVA |
| IV | When the data are skewed: ranks |
| V | Reading the p -values in a published table |
-

The question behind every test

If nothing were going on, how surprising would these data be?

Class data throughout: **simulated** stand-in for the class survey (data/class-data.R, fixed seed). Not real students.



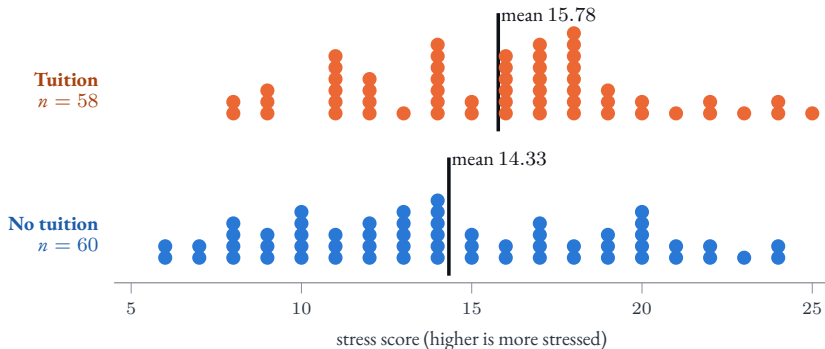
The test of this session

Given a comparison between groups, say what the null hypothesis is, choose a sensible test, read the p -value and the interval correctly, and say whether the difference matters.

- ▶ State a null and an alternative hypothesis for a research question
- ▶ Interpret a p -value correctly, and name two common misreadings
- ▶ Choose between a t -test, ANOVA and their rank-based alternatives
- ▶ Explain why a statistically significant result is not necessarily an important one



The question Session 3 left: is this gap luck?



Stress score, 5 to 25, from five questions. Tuition students average 15.78; the rest 14.33. A gap of 1.44 points.

Look at the spread inside each row. The gap between the bars is small next to it. Another class could easily reverse it. Or could it?

Simulated class data. The gap is computed from unrounded means; the rounded means differ by 1.45.



Act I

The logic of a test, and the p-value



A test starts by presuming nothing is going on

In a courtroom

- ▶ The accused is presumed innocent.
- ▶ The prosecution brings evidence.
- ▶ If the evidence would be very surprising for an innocent person, the court rejects innocence.
- ▶ Not guilty is not the same as proven innocent.

In a test

- ▶ The **null hypothesis** H_0 is presumed true.
- ▶ The data are the evidence.
- ▶ If the data would be very surprising under H_0 , we reject H_0 .
- ▶ Not rejecting is not the same as proving H_0 .

First example, from Session 1

H_0 : the average weeknight sleep of students like ours is 7 hours.

H_1 : it is not 7 hours.

The 7-hour figure: Watson NF et al., *Sleep* 2015; PMID 26039963 (seven hours or more for adults).



Writing the two hypotheses

Research question	H_0 (nothing going on)	H_1 (something is)
Is average weeknight sleep different from 7 hours?	$\mu = 7$	$\mu \neq 7$
Do tuition and non-tuition students differ in mean stress?	$\mu_{\text{tuition}} = \mu_{\text{none}}$	$\mu_{\text{tuition}} \neq \mu_{\text{none}}$
Does mean sleep differ across hall, mess and family home?	all three means equal	at least one differs

Hypotheses are about populations

H_0 says μ , not $\bar{x}$. Our sample mean is 6.80; we already know it is not 7.

Why H_0 is the dull one

“No difference” gives one exact value to test against.
“Some difference” does not.

Questions from the class survey. The means μ are for the population the class stands for.



The test statistic: how many standard errors from the null?

Session 3 asked how unusual a sample averaging 6.40 would be. Same question, now with our own mean.

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

Estimate minus the null value, divided by the standard error.

$$\bar{x} = 6.80, \mu_0 = 7, s = 0.93, n = 120$$

$$SE = 0.93 \div \sqrt{120} = 0.0849$$

$$t = (6.80 - 7) \div 0.0849$$

$$= -0.20 \div 0.0849 = -2.36$$

$$\text{Degrees of freedom: } n - 1 = 119$$

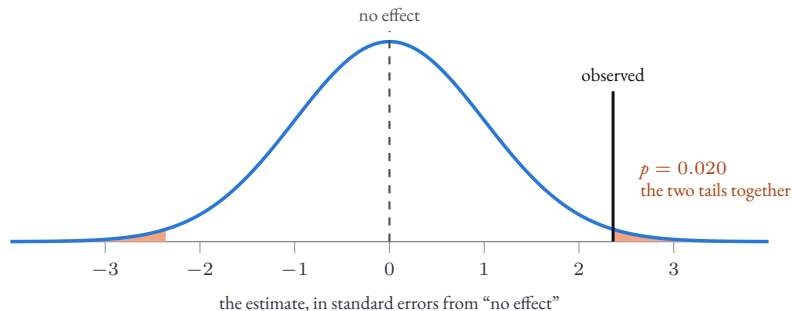
Read it in words

Our class sits 0.20 hours, 12 minutes, below 7 in raw terms, and 2.36 standard errors below it in the units a test uses.

Simulated class data. Hand arithmetic ours; software gives the same t .



The p-value: how often luck alone would do this



If the true average were 7 hours, samples of 120 would land 2.36 standard errors from 7, or further, in either direction, about 2% of the time.

The definition, in one line

The p -value is the chance of data at least this far from H_0 , if H_0 were true.

Simulated class data. p from the t curve with 119 df (software); the curve is drawn as the normal, which it nearly equals at this df.



Deciding: the significance level is set before the data arrive

Choose a threshold α in advance. By convention, $\alpha = 0.05$.

If	we say
$p < \alpha$	reject H_0 : “statistically significant”
$p \geq \alpha$	do not reject H_0

Our sleep test

$t = -2.36$, $df = 119$, $p = 0.020$

p is below 0.05, so reject H_0 .

The same decision another way

Critical value for 119 df: ± 1.980 .

$|-2.36| > 1.980$, so reject.

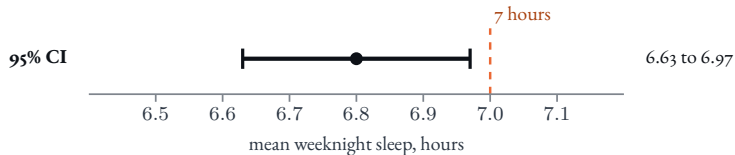
The sentence to write

Mean weeknight sleep was 6.80 hours, below the recommended 7 ($t = -2.36$, $df = 119$, $p = 0.020$).

Simulated class data. Test by software; critical value from the t table.



The test and the interval tell the same story



The rule. A 95% interval leaves out the null value exactly when the two-sided p is below 0.05. 7 is outside; $p = 0.020$.

What only the interval says. How far below 7: between about 2 and 22 minutes a night. The p -value cannot tell you that.

Simulated class data. Minutes are our arithmetic: $(7 - 6.97) \times 60$ and $(7 - 6.63) \times 60$.



The misreading	Why it is wrong
" $p = 0.020$: a 2% chance H_0 is true."	p is computed <i>assuming</i> H_0 is true.
" $p = 0.086$: there is no difference."	Weak evidence is not evidence of none.
"A smaller p means a bigger effect."	p mixes the size of the effect with n .
" $p < 0.05$, so it matters."	Importance is a judgement about size.

The ASA's statement, 2016

- ▶ p -values do not measure the probability that a hypothesis is true.
- ▶ Conclusions should not rest only on whether p passes a threshold.
- ▶ A p -value does not measure the size or the importance of an effect.

Wasserstein RL, Lazar NA. *Am Stat* 2016;70(2):129–133. doi:10.1080/00031305.2016.1154108. Three of its six principles, paraphrased.



Two ways to be wrong

	H_0 is true	H_0 is false
Reject H_0	Type I error (false alarm)	correct (power)
Do not reject	correct	Type II error (missed effect)

In court: jailing an innocent person, or freeing a guilty one.

α = chance of a Type I error when H_0 is true. We choose it: 0.05.

β = chance of a Type II error when a real effect exists.

Power = $1 - \beta$: the chance of detecting a real effect.

You cannot shrink both for free

Lowering α makes false alarms rarer and missed effects more common. Only more data reduces both.



One-sided or two-sided?

Two-sided: $H_1: \mu \neq 7$

Either tail counts. $p = 0.020$.

One-sided: $H_1: \mu < 7$

Only the lower tail counts. $p = 0.010$, half the two-sided value.

Choosing after looking

Picking the side after seeing which way the data went quietly doubles the false-alarm rate, from 5% to 10%.

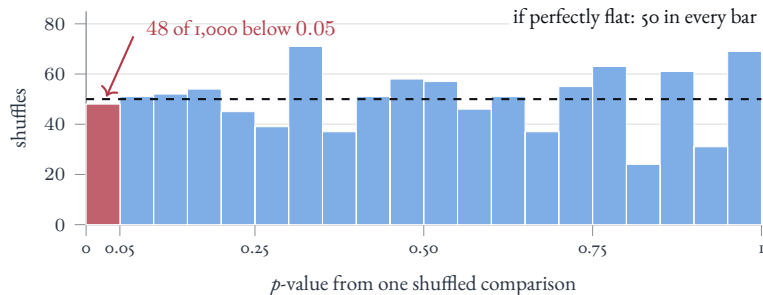
The default

Two-sided, unless a result in the other direction would be impossible or meaningless, and you said so before collecting the data.

Simulated class data. One-sided p is our arithmetic: half the two-sided value.



When nothing is going on, 5% of tests still say “significant”



We shuffled the tuition labels at random 1,000 times and ran the stress t -test each time. Shuffling destroys any real link, so every “difference” is luck. Still, 48 tests (4.8%) came out below 0.05.

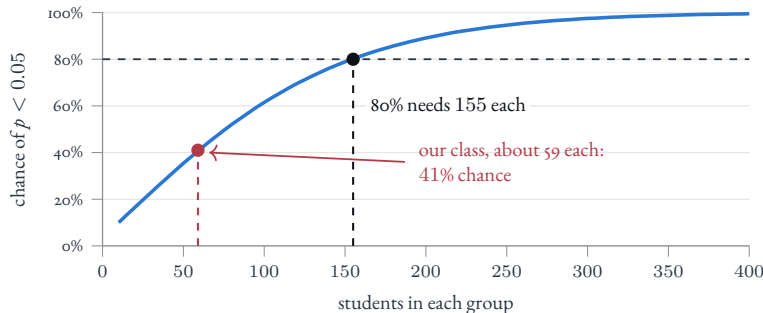
What α really is

The false-alarm rate you sign up for, test by test, when H_0 is true.

Simulated: 1,000 random shuffles of the **simulated** class data. The bars wobble because stress scores are whole numbers.



Power: would the test have noticed a real gap?



Suppose the true gap were exactly the size we saw ($d = 0.32$). A class our size would detect it only 41% of the time.

80% power needs 155 students per group; 90% needs 207. Power is planned *before* a study.

Our arithmetic (`R, power.t.test, two-sided,  $\alpha = 0.05$`), for a standardised gap of 0.32 from the **simulated** class data.



Many tests: the birthday problem returns

Run k tests on pure noise at $\alpha = 0.05$. The chance of at least one false alarm is $1 - 0.95^k$.

Tests	At least one false alarm
1	5.0%
3	14.3%
5	22.6%
10	40.1%
20	64.2%

Our arithmetic, assuming the tests are independent.

Same logic as Session 3

One test rarely fools you, as one pair rarely shares a birthday. Many tests give many chances.

Bonferroni: to keep the overall rate at 5% across 20 tests, use $0.05 \div 20 = 0.0025$ for each.



Ways to manufacture significance

- ▶ drop the values that “spoil” the result
- ▶ try many outcomes, report the one that worked
- ▶ keep collecting until p dips below 0.05
- ▶ slice into subgroups until one is significant

Session 2 called the first one misconduct. All four produce false alarms.

A field checks itself

100 psychology studies were repeated with larger samples.

97% → 36%

originals significant → repeats significant
The repeated effects were, on average, half the size.

Open Science Collaboration. *Science* 2015;349(6251):aac4716. PMID 26315443. Figures from its abstract.



Act II

Comparing two groups: t-tests

Two groups: the same logic, a different estimate

The question: do tuition and non-tuition students differ in mean stress?

$$H_0: \mu_{\text{tuition}} - \mu_{\text{none}} = 0$$

$$H_1: \mu_{\text{tuition}} - \mu_{\text{none}} \neq 0$$

The estimate: $\bar{x}_{\text{tuition}} - \bar{x}_{\text{none}} = 1.44$

	<i>n</i>	mean	SD
Tuition	58	15.78	4.14
No tuition	60	14.33	4.88

The recipe does not change

$$t = \frac{\text{estimate} - \text{null value}}{\text{standard error of the estimate}}. \text{ Only the standard error needs a new formula.}$$

Simulated class data. Two students with a missing stress item are left out.



The standard error of a difference

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Two sources of wobble, one per group. Their squares add.

$$\frac{4.14^2}{58} = 0.2955 \quad \frac{4.88^2}{60} = 0.3969$$

$$SE = \sqrt{0.2955 + 0.3969} = 0.8321$$

$$t = 1.44 \div 0.8321 = \mathbf{1.73}$$

SEs do not simply add

It is the *squares* that add. Adding the two SEs, $0.54 + 0.63$, would overstate the wobble.

Simulated class data. Hand arithmetic ours; software gives $t = 1.73$.



Reading the output: an interval that crosses zero

Welch two-sample t -test

$t = 1.73$, $df = 114.1$, $p = 0.086$
95% CI for the gap: -0.21 to 3.09

By hand: $1.44 \pm 1.981 \times 0.8321$
 $= 1.44 \pm 1.65$

Decision: p is above 0.05, so we do not reject H_0 .

The interval: the gap could be 0.21 points in the *other* direction, or as large as 3.09. Zero is inside.

Session 3's warning, in action

An interval for a difference that crosses zero is the same statement as $p > 0.05$.

Simulated class data. Test by software; the hand interval uses the t value for 114 df.



“Not significant” does not mean “no difference”

What $p = 0.086$ allows us to say

The data are compatible with no gap, and also with a gap of up to 3 points.

What it does not allow

“Tuition does not affect stress.” The study could not have shown that: its power for a gap this size was only 41%.

An honest sentence

Tuition students scored 1.44 points higher on average (95% CI -0.21 to 3.09 ; $p = 0.086$).
The study was too small to tell whether the gap is real.

A title worth remembering

“Absence of evidence is not evidence of absence.”

Altman DG, Bland JM. *BMJ* 1995;311:485. PMID 7647644. Class data **simulated**.



Welch or pooled? Why Welch is the default

	Pooled (Student)	Welch
Assumes equal SDs	yes	no
t	1.73	1.73
df	116	114.1
p	0.087	0.086

Here the SDs are 4.14 and 4.88 and the groups are nearly equal in size, so the two versions agree.

They part company when the groups differ in size *and* in spread. Then the pooled test can be badly wrong; Welch stays close to right.

Practical rule

Use Welch unless you have a reason not to. R uses it by default; SPSS prints both rows.

Simulated class data. Both tests by software.



What a t -test assumes, and how much it matters

Assumption	How to check	If it fails
Independent observations	The design: one answer per student, no clusters	Nothing in the test can fix it
Roughly normal in each group, <i>or</i> large groups	Histogram or Q-Q plot; the central limit theorem helps from about 30 per group	Rank-based tests (Act IV)
No extreme outliers	Box plot, Session 2's fences	Check for errors; report with and without

Formal normality tests mislead

Weeknight sleep fails the Shapiro–Wilk test ($p = 0.008$), mainly because answers come in half hours. With 120 students, the t -test is still fine.

Stress: skewness 0.09, close to symmetric. Screen time: skewness 0.67 and Shapiro–Wilk $p < 0.001$. Act IV is for screen time.

Simulated class data. Tests by software.



How big is the gap? Cohen's d

$$d = \frac{\text{difference in means}}{\text{pooled SD}}$$

The gap in standard deviations, so it can be compared across scales and studies.

Pooled SD = 4.53

$$d = 1.44 \div 4.53 = \mathbf{0.32}$$

$$\text{Sleep test: } d = -0.20 \div 0.93 = -0.22$$

$ d $	Cohen's rough label
0.2	small
0.5	medium
0.8	large

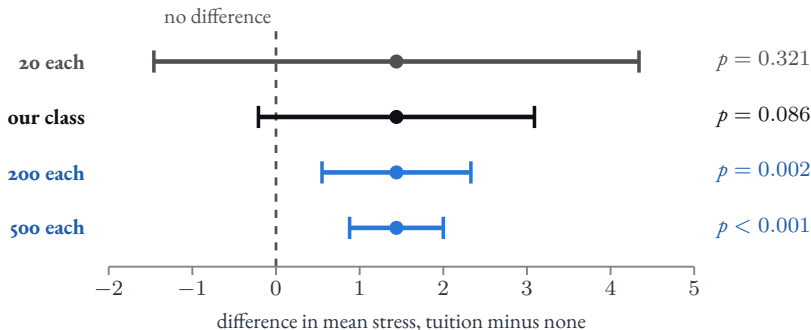
Labels are not verdicts

A “small” d can matter a great deal when many people are affected.

Cohen J. *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed., 1988. **Simulated** class data; d by hand is ours.



Same gap, different verdicts: significance depends on n



The gap and the SDs are fixed at ours. Only the number of students changes. The effect is exactly as large, or as small, in every row.

Hypothetical: our **simulated** gap and SDs with other group sizes. Our arithmetic (Welch).



A huge study can make a trivial gap significant

Imagine two groups of 20,000 students whose true weeknight sleep differs by 0.05 hours.

$$p < 0.001$$

highly significant

3 minutes

the size of the difference

No student would notice three minutes of sleep.

Significant, and unimportant.

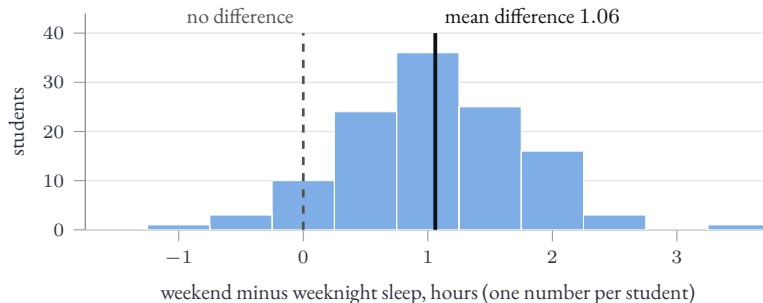
Two questions, always

Is it distinguishable from zero? Is it large enough to matter? Only the second needs knowledge of the subject.

Simulated: two groups of 20,000 drawn by `data/class-data.R`. Not real students.



Paired data: the same students, measured twice



Session 1 asked everyone about weeknight *and* weekend sleep. That gives one difference per student: 119 differences.

Mean difference 1.06 hours, SD 0.72. Only 4 students slept less at weekends; 10 slept the same.

Simulated class data. One student who skipped a sleep question is left out.



The paired t -test, by hand

H_0 : the mean difference is 0.

$$t = \frac{\bar{d}}{s_d / \sqrt{n}}$$

$$SE = 0.72 \div \sqrt{119} = 0.066$$

$$t = 1.06 \div 0.066 \approx 16$$

(software, from unrounded values: 15.99)

$$df = 118, p < 0.001$$

$$\begin{aligned} 95\% \text{ CI: } 1.06 \pm 1.98 \times 0.066 \\ = 0.93 \text{ to } 1.19 \text{ hours} \end{aligned}$$

$$\text{Effect size: } d_z = 1.06 \div 0.72 = 1.47$$

The sentence

Students slept 1.06 hours more at weekends (95% CI 0.93 to 1.19; paired $t = 15.99$, $df = 118$, $p < 0.001$).

Simulated class data. Hand arithmetic ours; the rounded SE gives 16.1, software 15.99.



The wrong test throws information away

	Paired	Unpaired (wrong)
t	15.99	8.27
95% CI	0.93 to 1.19	0.81 to 1.31

The unpaired test treats 119 weekend answers and 119 weeknight answers as if they came from 238 different students.

Weekend SD 1.05, weeknight SD 0.92, and the two go together: $r = 0.74$.

SD of the differences:

$$\sqrt{1.05^2 + 0.92^2 - 2 \times 0.74 \times 1.05 \times 0.92} \\ = \sqrt{0.5192} = 0.72$$

Pairing cancels each student's own baseline.

Pairing comes from the design, not the data

Same people twice, or matched pairs: paired test. Two separate groups: never.

Simulated class data. r is the correlation of Session 5. The SD formula is our arithmetic from the rounded values.



Which t -test?

Your data	Test	Our example
One group, compared with a fixed value	one-sample t	sleep against 7 hours
Two separate groups	Welch two-sample t	stress, tuition against none
Same people twice, or matched pairs	paired t	weekend against weeknight sleep
Three or more separate groups	ANOVA (Act III)	sleep by residence

One question decides it

Could the same student appear in both groups? If yes, the data are paired.

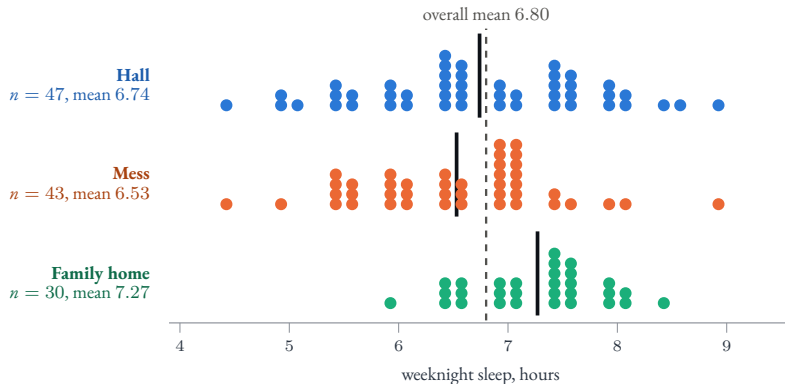
Examples from the **simulated** class data.



Act III

Comparing three or more groups: ANOVA

Three groups: does residence go with sleep?



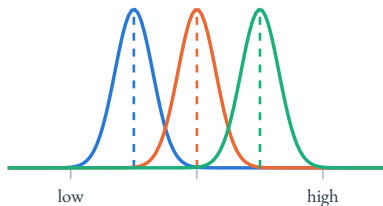
Three t -tests (hall–mess, hall–home, mess–home) would give three chances for a false alarm: $1 - 0.95^3 = 14.3\%$ if nothing is going on. One test of all three at once keeps it at 5%.

Simulated class data.

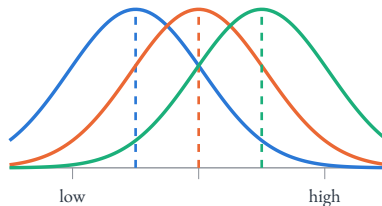


The idea of ANOVA: between-group spread against within-group spread

Tight groups: F is large



Wide groups: F is small



$$F = \frac{\text{variation between the group means}}{\text{variation within the groups}}$$

Same three means in both panels. On the left the groups barely overlap: the means are far apart *relative to the spread*. On the right the same gaps are lost in the noise.

Constructed for teaching: made-up means and spreads.



The ANOVA table

Source	Sum of squares	df	Mean square	F	p
Between residences	9.70	2	4.85	6.13	0.003
Within residences	92.50	117	0.791		
Total	102.20	119			

Mean square = sum of squares $\div$ df

$$9.70 \div 2 = 4.85$$

$$92.50 \div 117 = 0.791$$

$$F = 4.85 \div 0.791 = \mathbf{6.13}$$

$$\text{df between} = \text{groups} - 1 = 2$$

$$\text{df within} = n - \text{groups} = 117$$

$F(2, 117) = 6.13, p = 0.003$: reject H_0 . At least one residence group differs in mean sleep.

Simulated class data. ANOVA by software; the divisions are ours and reproduce it.



How much does residence explain? η^2

$$\eta^2 = \frac{SS \text{ between}}{SS \text{ total}} = \frac{9.70}{102.20} = \mathbf{0.095}$$

About 9.5% of the variation in weeknight sleep lies between residence groups. The other 90.5% is between students in the same group.

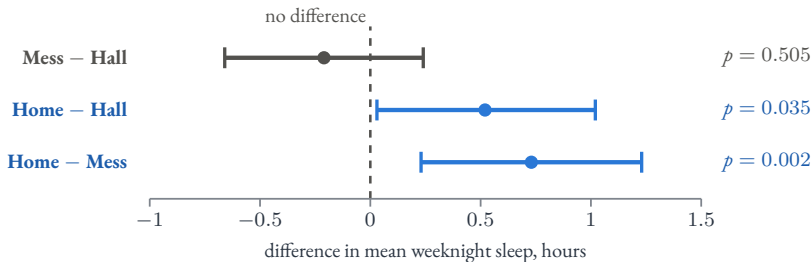
Simulated class data. η^2 by hand is ours.

Significant, and modest

$p = 0.003$ says the group means are unlikely to be equal. η^2 says residence is a small part of why students differ.



Which groups differ? Tukey's comparisons



Family-home students sleep about 0.52 hours more than hall students and 0.73 more than mess students. Hall and mess cannot be told apart.

Simulated class data. Tukey HSD by software; intervals and p adjusted for three comparisons.



Equal spreads?

SDs: 1.04, 0.88, 0.60. Largest over smallest: 1.7.

Brown–Forsythe test: $p = 0.020$. The spreads probably differ.

Normal residuals?

Shapiro–Wilk on the residuals: $p = 0.416$.

Welch's ANOVA, which does not assume equal spreads:

$F = 9.78$, df 2 and 77.8, $p < 0.001$.

The conclusion stands either way.

Rule of thumb

Equal-spread ANOVA copes when the largest SD is under about twice the smallest and the groups are similar in size. When in doubt, run Welch's version and see if anything changes.

Simulated class data. All tests by software.



Four checks, before any p -value

1. Plot every group: dots, box plots.
2. Look for outliers and impossible values.
3. Compare the spreads, not only the centres.
4. Ask whether a gap this size would matter.

In our data: 0.52 hours is about 31 minutes; 0.73 hours is about 44. Worth knowing for a student, and modest next to the spread inside each group (SDs from 0.60 to 1.04 hours).

A test answers only the question it was given

ANOVA compares means. It will not notice that one group is bimodal, that a group of 3 drives the result, or that the scale was misread.

Simulated class data. Minutes are our arithmetic.



ANOVA

Mean weeknight sleep differed by residence ($F(2, 117) = 6.13, p = 0.003, \eta^2 = 0.095$). Family-home students slept 0.52 hours more than hall students (95% CI 0.03 to 1.02, Tukey-adjusted).

Always include

- ▶ the estimate, in its units
- ▶ its 95% interval
- ▶ the test, its statistic and df
- ▶ the exact p , or $p < 0.001$

Avoid

- ▶ $p = 0.000$ (write $p < 0.001$)
- ▶ stars instead of numbers
- ▶ “ $p = \text{NS}$ ” with no estimate
- ▶ “proved”, “no effect”

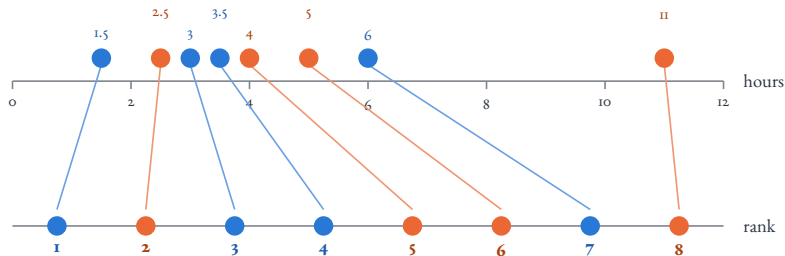
Simulated class data.



Act IV

When the data are skewed: ranks

Replace the values with their ranks



Tuition (orange): ranks 2, 5, 6, 8; rank sum 21

No tuition (blue): ranks 1, 3, 4, 7; rank sum 15

The 11-hour value becomes rank 8, the same rank it would have at 6.5 hours or 50. Ranks keep the order and ignore the distances, so one extreme value cannot drag the result.

Constructed for teaching: eight made-up screen-time values.



The Mann–Whitney test: two groups, skewed data

Screen time by tuition

	<i>n</i>	median, hours
Tuition	59	4.60
No tuition	58	4.15

$$W = 1807.5, p = 0.601$$

H_0 : neither group tends to have higher values than the other.

Rank all 117 students together; compare the ranks of the two groups. If tuition made no difference, the ranks would be shuffled evenly between them.

$p = 0.601$: no evidence either group spends more time on screens.

Not strictly a test of medians

It asks whether one group tends to be higher. Report medians beside it, but do not call it a “median test”.

Simulated class data. Test by software (normal approximation, ties corrected). Also called the Wilcoxon rank-sum test.



What W counts: a probability you can explain

Pair every tuition student with every non-tuition student: $59 \times 58 = 3,422$ pairs.

W counts the pairs in which the tuition student has more screen time, with ties counted as half.

$$1807.5 \div 3,422 = 0.53$$

Simulated class data. W by software (R); the division is our arithmetic.

Read it as

Pick one student from each group at random. The tuition student has more screen time about 53% of the time. No tendency at all would be 50%.



The Kruskal–Wallis test: three or more groups

Screen time by residence

	median, hours
Hall	4.10
Mess	4.45
Family home	4.65

$$H = 0.75, df = 2, p = 0.688$$

Simulated class data. Test by software.

The rank-based partner of one-way ANOVA. Rank everyone together; compare the average rank in each group.

$p = 0.688$: no evidence that screen time differs by residence.

Had it been significant, the next step would be pairwise rank tests with a Bonferroni or similar adjustment.



Mean-based or rank-based?

	t -test, ANOVA	Mann–Whitney, Kruskal–Wallis
Answers	Do the means differ?	Does one group tend to be higher?
Gives	a difference in means, with CI	a probability or rank summary
Sensitive to outliers	yes	hardly
Best when	roughly normal, or large groups	skewed, ordinal, or small with outliers

Choose by the question

If the mean is what matters (total hours, total cost), use the mean-based test, and with large groups it copes with skew.

Do not shop

Running both and reporting the one with the smaller p is another way to hunt for significance. Decide first.

A summary; details in the class notes.



Choosing a test: the map for Sessions 4 to 6

Question	Roughly normal, or large	Skewed, ordinal, or small
One mean against a value	one-sample t	Wilcoxon signed-rank
Two separate groups	Welch t	Mann–Whitney
Same people twice	paired t	Wilcoxon signed-rank
Three or more groups	one-way ANOVA	Kruskal–Wallis
Two categorical variables	chi-square test (Session 5)	
Two numeric variables	Pearson r (Session 5)	Spearman's ρ (Session 5)
Predict an outcome	regression (Session 6)	

Start from the question and the type of outcome, not from the software menu.

A summary; Session 5 turns it into a flowchart.



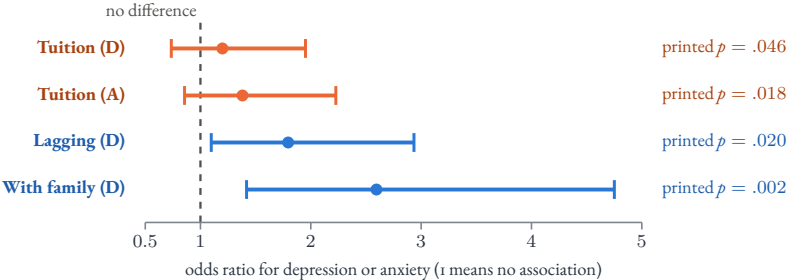
Act V

Reading the p-values in a published table



Stat.BD

Islam 2020, Table 3: an interval that includes 1



D: depression; A: anxiety. For a ratio, “no association” is 1, not 0. A 95% interval that includes 1 should come with $p > 0.05$. The two tuition rows include 1, yet print $p < 0.05$.

Islam MA et al. *PLoS One* 2020;15(8):e0238162, Table 3, as printed. PMID 32845928. CC BY 4.0.

Rebuilding a p -value from the odds ratio and its interval

Tuition, depression: OR 1.199, CI 0.736 to 1.952

1. $\ln(1.199) = 0.181$
2. $SE = (\ln 1.952 - \ln 0.736) \div (2 \times 1.96)$
 $= (0.669 + 0.307) \div 3.92 = 0.249$
3. $z = 0.181 \div 0.249 = 0.73$
4. Two-sided $p \approx 0.47$
(software, unrounded: 0.466). Printed: .046.

Row	Printed	Rebuilt
Tuition, depression	.046	0.466
Tuition, anxiety	.018	0.186
Age 21-24, depression	.082	0.828

The other 13 rows rebuild to their printed value.

Our reading, not the authors'

Each printed value looks like the rebuilt one with an extra 0 after the decimal point. We have not confirmed this with the authors.

Islam 2020, Table 3. The rebuilt p -values and the reading of the pattern are **ours**. Method: on the log scale the CI is estimate ± 1.96 SE.



What the paper concluded, and what the table supports

The paper

Its abstract reports the tuition odds ratio for depression as a finding. Its discussion calls tuition “a critical factor” in the rise of depression and anxiety.

The table

Both tuition intervals include 1. Rebuilt p : 0.466 and 0.186. The data are compatible with no association, and with a sizeable one.

Three more cautions

- ▶ 16 tests in one table: at 0.05, about one false alarm is expected even if nothing is going on.
- ▶ A snowball sample on Facebook (Session 3): the interval covers chance, not bias.
- ▶ Associations, not causes (Session 5).

The reader's habit

Check that each p agrees with its interval. When they disagree, trust neither until you know why.

Islam 2020, abstract and discussion; quotation from the discussion. Expected false alarms ($16 \times 0.05 = 0.8$) are our arithmetic.



Two rows that pass this check

Depression	OR (95% CI)	p
Lagging academically	1.795 (1.098–2.935)	.020
Living with family	2.595 (1.418–4.751)	.002

Rebuilt: 0.020 and 0.002. Interval and p agree.

Significant, but how large? The lagging interval runs from about 10% higher odds to nearly three times the odds. Precise enough to say “higher”, not to say “how much”.

Still not a cause: lagging academically may raise depression, or depression may cause students to fall behind.

Passing is not proof: Session 6 checks the family row against the paper’s own Table 2 counts, and there it fails.

Islam 2020, Table 3, as printed. Rebuilt p -values ours.



Activity: significant or not?

	Result	As reported
1	Islam: lagging academically, depression	OR 1.795 (1.098–2.935), $p = .020$
2	Islam: female, anxiety	OR 1.462 (0.866–2.468), $p = .155$
3	Islam: age over 24, depression	OR 2.886 (0.961–8.669), $p = .059$
4	Islam: tuition, anxiety	OR 1.381 (0.856–2.227), $p = .018$
5	Islam: exercise, depression	OR 0.880 (0.544–1.422), $p = .601$
6	Class: sleep by residence	$F = 6.13$, $p = 0.003$, $\eta^2 = 0.095$
7	Class: stress, tuition against none	gap 1.44 (–0.21 to 3.09), $p = 0.086$
8	Two groups of 20,000: sleep	gap 3 minutes, $p < 0.001$

In groups of three, for each row: (a) significant at 0.05? (b) do the p and the interval agree? (c) would you call it important, and what else would you need to know?

Rows 1–5: Islam 2020, Table 3, as printed. Rows 6–8: **simulated**. Answers in the class notes.



1. A test assumes nothing is going on, and asks how surprising the data would then be. That chance is the p -value.
2. A p -value is not the chance that H_0 is true, and it does not measure size or importance.
3. Two groups: Welch or paired t . Three or more: ANOVA, then Tukey. Skewed: rank-based tests.
4. Not significant does not mean no difference. Significant does not mean important. Report the estimate and its interval.
5. A p -value and its interval must agree. When they do not, something is wrong.



Session 5: Relationships between variables

The chi-square test for two categories, correlation for two numbers, and why association is not causation.

Before then

Read Chapter 4 of the class notes. Try the Part A exercises and the activity answers.

The residence and tuition figures use the **simulated** class data.

One question to think about

In our class, 74.4% of mess students give tuition, against 36.2% of hall students and 33.3% of those living at home. Is residence related to tuition, or could this be luck?



- ▶ Islam MA, Barna SD, Raihan H, Khan MNA, Hossain MT. Depression and anxiety among university students during the COVID-19 pandemic in Bangladesh: a web-based cross-sectional survey. *PLoS One* 2020;15(8):e0238162. PMID 32845928. PMC7449469. doi:10.1371/journal.pone.0238162. CC BY 4.0.
- ▶ Wasserstein RL, Lazar NA. The ASA statement on p -values: context, process, and purpose. *Am Stat* 2016;70(2):129–133. doi:10.1080/00031305.2016.1154108.
- ▶ Open Science Collaboration. Estimating the reproducibility of psychological science. *Science* 2015;349(6251):aac4716. PMID 26315443. doi:10.1126/science.aac4716.
- ▶ Altman DG, Bland JM. Absence of evidence is not evidence of absence. *BMJ* 1995;311(7003):485. PMID 7647644. doi:10.1136/bmj.311.7003.485.
- ▶ Watson NF, Badr MS, Belenky G, et al. Recommended amount of sleep for a healthy adult: a joint consensus statement of the American Academy of Sleep Medicine and Sleep Research Society. *Sleep* 2015;38(6):843–844. PMID 26039963. doi:10.5665/sleep.4716.
- ▶ Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum; 1988.
- ▶ Class data: **simulated** stand-in for the Stat.BD class survey, generated by data/class-data.R from a fixed seed. Not real students. Every figure computed from it, including the 1,000 shuffles, the power curve, the study of 20,000 per group and the class re-run at other sizes, is ours. So are the p -values rebuilt from Islam 2020, Table 3.
- ▶ Constructed examples, made up for teaching: the two panels of the F -ratio figure and the eight screen-time values in the ranks figure.

