



CLASS NOTES • CHAPTER 4

Testing Hypotheses

Comparing groups, and when a difference is more than luck

Statistics & Data Analysis Fundamentals
Stat.BD

Mentor: Muhtasim Munif Fahim, MSc

CHAPTER 4

Testing Hypotheses

A difference between two groups in a sample can be real or can be luck. This chapter builds the test that tells them apart, and shows what the test cannot say.

In the simulated class, the 58 students who give private tuition score 15.78 on average on a stress scale that runs from 5 to 25. The 60 who do not score 14.33. Worked from the unrounded means, the gap is 1.44 points. Chapter 3 ended on the obvious question: is that gap real, or could it come from the luck of which students happened to be surveyed? Another class of the same size could easily show a gap of 2 points, or of none, or one the other way round. The tool that answers the question is the hypothesis test, and its result is usually reported as a p -value. By the end of section 2 the answer for tuition will be “not proven either way”, and the reasons for that answer matter more than the answer itself.

The chapter builds the test in five steps. Section 1 sets out its logic on the simplest case, one average compared with a fixed value, and names the ways a test can mislead. Section 2 compares two groups, first separate groups and then the same students measured twice. Section 3 compares three or more groups at once. Section 4 handles skewed data by replacing values with their ranks. Section 5 applies the chapter’s main check to a published table from a Bangladeshi survey, and finds three p -values that disagree with their own intervals.

The data are the class survey of Chapter 1. Until the real answers are coded, every figure drawn from it comes from a **simulated** stand-in of 120 students, generated by a program from a fixed seed. They are not real students. A few examples are constructed on purpose, with made-up numbers, and are labelled as constructed where they appear.

What this chapter establishes

1. The logic of a test: presume nothing is going on, then ask how surprising the data would be.
2. How to write a null and an alternative hypothesis, and why both are about the population.
3. How to compute a test statistic, and what a p -value is and is not.
4. Why a 95% interval and a two-sided test at 0.05 always agree.
5. The two ways a test can be wrong, what power is, and why a small study can miss a real difference.
6. Why many tests produce false alarms, and how a Bonferroni correction limits them.
7. How to compare two separate groups with the Welch t -test, and two measurements on the same people with the paired t -test.
8. How to measure the size of a difference with Cohen’s d , and why significance depends on the sample size.

9. How one-way ANOVA compares three or more groups, how η^2 measures what grouping explains, and how Tukey's method finds which groups differ.
10. When to use rank-based tests, and what the Mann–Whitney statistic counts.
11. How to check a published p -value against its own interval.

Notation

H_0, H_1	the null hypothesis (nothing going on) and the alternative
μ_0	the value of the mean under the null hypothesis
t, df	a t statistic, and its degrees of freedom
p	a p -value
α	the significance level, set in advance; usually 0.05
β , power	the chance of missing a real effect, and $1 - \beta$
$\bar{x}_1, s_1, n_1$	the mean, SD and size of group 1; likewise for group 2
SE	a standard error (Chapter 3)
s_p	the pooled SD of two groups
d, d_z	Cohen's d for two groups, and for paired differences
$\bar{d}, s_d$	the mean and SD of paired differences
SS, MS	a sum of squares, and a mean square (SS divided by its df)
F	the ANOVA statistic: between-group MS over within-group MS
η^2	eta squared: the share of variation between groups
W, H	the Mann–Whitney and Kruskal–Wallis statistics
OR	an odds ratio (Chapter 6 explains it)

Contents

What this chapter establishes	1
Notation	2
1 The logic of a test	5
1.1 A test presumes nothing is going on	5
1.2 Writing the two hypotheses	5
1.3 The test statistic	6
1.4 The p -value	7
1.5 Deciding: the significance level	7
1.6 The test and the interval agree	8
1.7 What a p -value is not	8
1.8 Two ways to be wrong	9
1.9 Power: would the test have noticed?	10
1.10 One-sided or two-sided?	10
1.11 When nothing is going on	10
1.12 Many tests: the birthday problem returns	11
1.13 Hunting for significance	12
2 Comparing two groups	12
2.1 The same logic, a different estimate	12
2.2 The standard error of a difference	13
2.3 Reading the output: an interval that crosses zero	14
2.4 “Not significant” does not mean “no difference”	14
2.5 Welch or pooled?	14
2.6 What a t -test assumes	15
2.7 How big is the gap? Cohen’s d	16
2.8 Significance depends on the sample size	16
2.9 A huge study can make a trivial gap significant	17
2.10 Paired data: the same students, measured twice	17
2.11 The wrong test throws information away	18
2.12 Which t -test?	19

3	Comparing three or more groups: ANOVA	20
3.1	Three groups, one test	20
3.2	The idea: spread between against spread within	20
3.3	The ANOVA table	21
3.4	How much does residence explain? η^2	22
3.5	Which groups differ? Tukey's comparisons	22
3.6	Checking the assumptions	23
3.7	Look before testing	24
3.8	Reporting a test	24
4	When the data are skewed: rank-based tests	25
4.1	Replace the values with their ranks	25
4.2	The Mann–Whitney test	25
4.3	What W counts	26
4.4	The Kruskal–Wallis test	26
4.5	Mean-based or rank-based?	26
4.6	Choosing a test: the map	27
5	Reading the p-values in a published table	28
5.1	An interval that includes 1	28
5.2	Rebuilding a p -value from an interval	29
5.3	What the paper concluded	29
5.4	Rows that pass this check	30
6	Testing a hypothesis: the procedure	30
7	Chapter summary	31
8	Key terms	32
9	Exercises	33
10	Solutions	37
11	Our arithmetic and constructed numbers	42
12	References and further reading	42
13	For students in the Stat.BD course	43

1 The logic of a test

1.1 A test presumes nothing is going on

A criminal court does not ask whether the accused is innocent. It presumes innocence, hears the evidence, and asks whether that evidence would be very surprising if the accused were innocent. If it would, the court rejects innocence and convicts. If it would not, the verdict is “not guilty”, which is a statement about the evidence and not a finding of innocence. A statistical test follows the same steps (Table 1).

Table 1: The logic of a court and the logic of a test.

In a court	In a test
The accused is presumed innocent.	The <i>null hypothesis</i> H_0 is presumed true.
The prosecution brings evidence.	The data are the evidence.
If the evidence would be very surprising for an innocent person, the court rejects innocence.	If the data would be very surprising under H_0 , the test rejects H_0 .
“Not guilty” is not the same as “proven innocent”.	Not rejecting H_0 is not the same as proving it.

The last row is the one most often forgotten, and section 2.4 returns to it. The first example is the question Chapter 1 raised and Chapter 3 answered with an interval: is the average weeknight sleep of students like these different from the seven hours recommended for adults (reference 2)?

1.2 Writing the two hypotheses

Definition 1.1 • Null and alternative hypotheses

The **null hypothesis** H_0 states that nothing is going on: no difference, no effect, or a mean equal to a named value. The **alternative hypothesis** H_1 states that something is. Both are statements about the *population*, never about the sample.

Table 2 writes the pair for the three questions this chapter tests.

The null is always the dull hypothesis, and there is a practical reason. “No difference” pins down one exact value, so it is possible to work out what the data should look like if it is true. “Some difference” names no value, and so gives nothing to compute with.

Hypotheses are about populations

Writing $H_0: \bar{x} = 7$ is a common examination mistake. The sample mean is already known: it is 6.80, and it is not 7. A hypothesis is a claim about the population mean μ , which cannot be seen.

Table 2: Three research questions from the class survey, each as a pair of hypotheses. The means μ are for the population the class stands for.

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 are equal	at least one differs

1.3 The test statistic

Chapter 3 asked how unusual a sample of 120 averaging 6.40 hours would be if the true average were 6.80, and answered by counting standard errors. A test asks the same question of the class's own mean, with the null value in place of the truth. The count of standard errors is the test statistic:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \quad (1.1)$$

Test statistic

a number that measures how far the data sit from H_0 , usually in standard errors

$\bar{x}$, s , n the sample mean, SD and size

μ_0 the value the null hypothesis names; here 7 hours

$s/\sqrt{n}$ the standard error of the mean (Chapter 3)

In words: the estimate, minus what the null says, divided by the standard error. Every test in this chapter has a statistic of this shape, and only the standard error changes from one test to the next.

Example 1.1 • Is average sleep seven hours?

The class has mean $\bar{x} = 6.80$ hours, SD $s = 0.93$ and $n = 120$. The null value is $\mu_0 = 7$.

Step	What	Working
1	standard error	$0.93 \div \sqrt{120} = 0.0849$
2	estimate minus null	$6.80 - 7 = -0.20$
3	test statistic	$t = -0.20 \div 0.0849 = -2.36$
4	degrees of freedom	$n - 1 = 119$

The class sits -0.20 hours below 7 in raw terms, which is 408 minutes against 420 on average, and 2.36 standard errors below 7 in the units a test uses. Software gives the same t .

Chapter 2 promised that degrees of freedom would return in every test. They return here. Because the sample SD stands in for the unknown population SD, the reference curve is the t distribution of Chapter 3 with $n - 1 = 119$ degrees of freedom. With that many it is almost exactly the normal curve.

1.4 The p -value

A t of -2.36 is a distance. The test turns it into a chance. Imagine the null were true, so that the average of all students like these really were 7 hours. Samples of 120 would then scatter around 7, and their t values would follow the curve in Figure 1. The shaded tails hold the samples that land at least 2.36 standard errors from 7, in either direction. They make up about 2% of the curve.

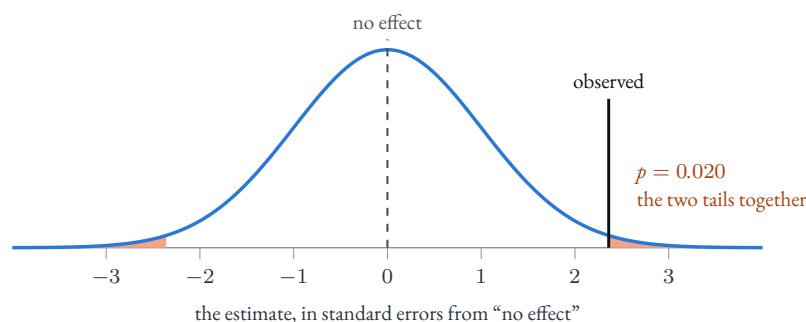


Figure 1: The p -value for the sleep test: the share of samples at least 2.36 standard errors from 7 hours, if 7 were the truth. The curve is drawn as the normal, which the t curve with 119 df almost exactly matches. **Simulated** class data.

Definition 1.2 • p -value

The **p -value** is the probability of data at least as far from H_0 as the data observed, *calculated on the assumption that H_0 is true*.

For the sleep test, $p = 0.020$. Both tails count because the alternative said “not 7”, and a class sleeping far more than 7 hours would have been just as much evidence against the null as one sleeping far less. The condition in the definition, “if H_0 were true”, is the part that is easy to lose. Every p -value is computed in an imaginary world where the null holds, and every misreading in section 1.7 comes from forgetting that.

1.5 Deciding: the significance level

A p -value is a continuous measure of surprise. A decision needs a cut-off, and the cut-off must be chosen before the data arrive. It is the significance level, written α , and by convention it is 0.05.

If	The decision
$p < \alpha$	reject H_0 ; the result is “statistically significant”
$p \geq \alpha$	do not reject H_0

For the sleep test, $p = 0.020$ is below 0.05, so the test rejects the null: the evidence says the average is not 7 hours. The same decision can be reached a second way. The value of t at which p would be exactly 0.05 is the critical value, and for 119 df it is ± 1.980 , the 95% multiplier of Chapter 3. Since 2.36 is larger than 1.980, the test rejects. The two routes always agree, because the critical value is defined as the point where p equals α .

Significance level

α , the p -value below which H_0 will be rejected, fixed in advance

Critical value

the value of the test statistic at which p equals α ; beyond it, H_0 is rejected

Why 0.05? Convention, not science. It means accepting that when nothing is going on, the test will be fooled by luck about one time in twenty. Some fields use 0.01, and some have argued for 0.005. Whichever level is used, it is set first and not adjusted afterwards.

The sentence to write

Mean weeknight sleep was 6.80 hours, below the recommended 7 ($t = -2.36$, $df = 119$, $p = 0.020$). The estimate comes first, because it answers the research question. The test supports it, in brackets.

1.6 The test and the interval agree

Chapter 3 computed a 95% interval for the same mean: 6.63 to 6.97 hours. It leaves out 7 (Figure 2). The test rejects 7. That is not a coincidence.

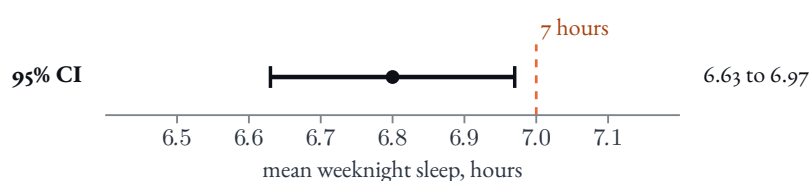


Figure 2: The 95% interval for mean weeknight sleep, against the null value of 7 hours. **Simulated** class data.

The interval rule

A 95% confidence interval leaves out a value exactly when a two-sided test of that value gives $p < 0.05$. The same holds for a 99% interval and a test at 0.01.

The rule works because both are built from the same three numbers: the estimate, its standard error and the critical value 1.980. The interval is the estimate plus or minus $1.980 \times 0.0849 = 0.17$; the test rejects when the estimate is more than that distance from the null value. A 99% interval uses the larger multiplier 2.618, giving 6.80 ± 0.22 , from 6.58 to 7.02. That interval *includes* 7, so a test at the 1% level would not reject, which agrees with $p = 0.020$ being above 0.01.

The interval says one thing the test cannot. It gives the size of the shortfall: between about 2 and 22 minutes a night. The p -value says only that the shortfall is probably not zero. This is why many journals ask for intervals alongside p -values, and some ask for intervals instead.

1.7 What a p -value is not

The p -value may be the most misread number in science. Table 3 lists four common misreadings.

The first misreading is the swapped conditional of Chapter 3. $P(\text{data} \mid H_0)$ and $P(H_0 \mid \text{data})$ share their ingredients, yet they can differ enormously. The exam-hall exercise of Chapter 3 showed the same thing: the share of cheaters carrying phones differs from the share of phone carriers who cheat.

Table 3: Four misreadings of a p -value. The p -values are from this chapter's **simulated** class data.

The misreading	Why it is wrong
" $p = 0.020$, so there is a 2% chance that H_0 is true."	The p -value is computed <i>assuming</i> H_0 is true. It is the chance of the data given the hypothesis, not of the hypothesis given the data.
" $p = 0.086$, so there is no difference."	Weak evidence against a difference is not evidence of no difference (section 2.4).
"A smaller p means a bigger effect."	A p -value mixes the size of the effect with the sample size (section 2.8).
" $p < 0.05$, so it matters."	Importance is a judgement about size, made with knowledge of the subject (section 2.9).

THE AMERICAN STATISTICAL ASSOCIATION'S STATEMENT, 2016

In 2016 the American Statistical Association issued its first formal statement on a single statistical method, because misreadings of p -values had become so widespread (reference 3). It set out six principles. Three of them, paraphrased: p -values do not measure the probability that a hypothesis is true; scientific conclusions should not rest only on whether a p -value passes a threshold; and a p -value does not measure the size or the importance of an effect.

1.8 Two ways to be wrong

Every decision a test makes can be wrong in one of two ways (Table 4). A Type I error is a false alarm: rejecting a null that is true. In a court, it is jailing an innocent person.

A Type II error is a missed effect: failing to reject a null that is false. In a court, it is freeing a guilty one.

Type I error

rejecting H_0 when it is true: a false alarm

Type II error

failing to reject H_0 when it is false: a missed effect

Table 4: The four outcomes of a test.

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

The chance of a Type I error, when the null is true, is α , and it is chosen: 0.05. The chance of a Type II error, when a real effect of a given size exists, is written β . Its complement is the power of the test, $1 - \beta$: the chance of detecting a real effect of that size.

Both errors cannot be shrunk for free

Lowering α makes false alarms rarer and missed effects more common. A court that demands more evidence convicts fewer innocent people and frees more guilty ones. The only way to reduce both errors at once is better evidence, which in statistics means more data.

Power

the chance that a test rejects H_0 when a real effect of a stated size exists; $1 - \beta$

1.9 Power: would the test have noticed?

Power depends on three things: the size of the real effect, the spread of the data, and the sample size. Figure 3 shows it for the tuition question of section 2. Suppose the true gap in stress were exactly the gap the class shows, which is 0.32 standard deviations (the d of section 2.7). A study with groups the size of the class's would detect it only 41% of the time: worse than a coin toss.

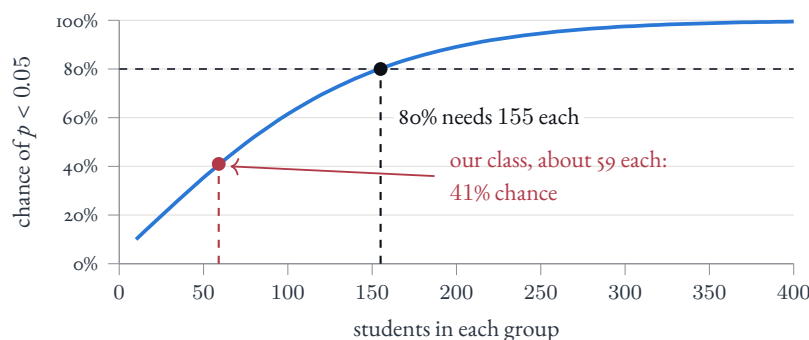


Figure 3: Power to detect a standardised gap of 0.32 with a two-sided test at 0.05, by the number of students in each group. Our arithmetic (R, `power.t.test`), for a gap taken from the **simulated** class data.

To reach 80% power, the conventional target, a study would need 155 students in each group, and 90% would need 207. Power is planned before a study, in the same way as the sample sizes of Chapter 3. Research proposals and ethics committees ask for exactly this calculation. Working it out after a non-significant result, from the effect that was observed, adds little: section 2.4 explains what to report instead.

1.10 One-sided or two-sided?

The sleep test was two-sided: its alternative was “not 7”, and both tails of Figure 1 counted. A one-sided test states the direction in advance, for example $H_1: \mu < 7$, and counts only the lower tail. For the class, the one-sided p would be half the two-sided one: $p = 0.010$.

Halving the p -value is exactly what makes one-sided tests tempting, and exactly what makes them dangerous. A researcher who looks at the data, sees which way they point, and then chooses that side is running a two-sided test at the 10% level while reporting 5%. The false-alarm rate has doubled without anyone saying so.

The default is two-sided

Use a one-sided test only when a result in the other direction would be impossible or meaningless, and when that was stated before the data were collected. Almost all published work uses two-sided tests, and an examiner will ask for the justification of a one-sided one.

Two-sided test

a test whose alternative allows a difference in either direction; both tails count

1.11 When nothing is going on

What does $\alpha = 0.05$ look like in practice? The class data allow a direct demonstration. Take the tuition labels, shuffle them at random among the students, and run the stress test again.

Shuffling destroys any real link between tuition and stress, so the null hypothesis is true by construction, and any “difference” is pure luck. Figure 4 shows the p -values from 1,000 such shuffles. They spread evenly from 0 to 1, and 48 of them, 4.8%, fell below 0.05. The expected number was 50.

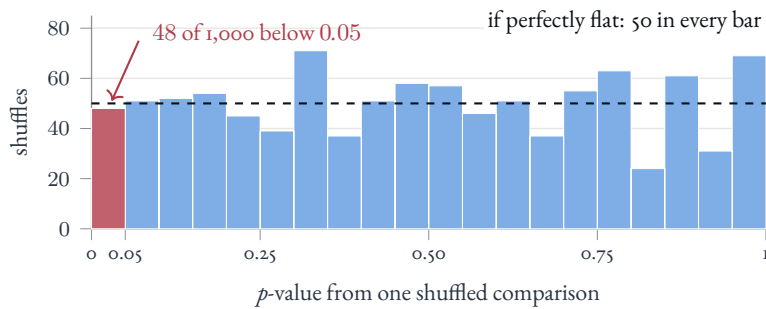


Figure 4: The p -values from 1,000 random shuffles of the tuition labels. The null is true in every shuffle, yet about one test in twenty is “significant”. The bars are uneven because stress scores are whole numbers. **Simulated** data: shuffles of the **simulated** class.

What α is

The significance level is the false-alarm rate accepted for each test when the null is true. The false alarms in Figure 4 look exactly like real findings. Nothing in a single result marks it as one of them.

1.12 Many tests: the birthday problem returns

Chapter 3 showed that 23 people make 253 pairs, which is why a shared birthday among them is more likely than not. Tests behave the same way. One test rarely produces a false alarm; many tests give many chances. If k independent tests are run on data where nothing is going on, each at $\alpha = 0.05$, the chance of at least one false alarm is

$$P(\text{at least one false alarm}) = 1 - 0.95^k \quad (1.2)$$

by the complement and multiplication rules of Chapter 3. This chance is called the familywise error rate, and Table 5 shows how fast it grows.

Familywise error rate
the chance of at least one false alarm across a family of tests

Table 5: The chance of at least one false alarm among k independent tests on pure noise, each at 0.05. Our arithmetic, from equation (1.2).

Tests k	At least one false alarm	Bonferroni level $0.05/k$
1	5.0%	0.05
3	14.3%	0.0167
5	22.6%	0.01
10	40.1%	0.005
20	64.2%	0.0025

With twenty tests on noise, a false alarm is close to a two-in-three chance. The simplest repair is the Bonferroni correction: divide α by the number of tests. To keep the overall rate at 5% across 20 tests, each must reach $0.05 \div 20 = 0.0025$. The correction is strict and costs

Bonferroni correction
testing each of k tests at α/k , so that the familywise rate stays at or below α

power, but it is honest. Section 3.5 meets a better tool for the common case of comparing every pair of groups.

A reader's rule

When a paper reports many tests and highlights one, count the tests.

1.13 Hunting for significance

The many-tests problem can also be hidden. Each practice below runs many tests in private and reports only the one that worked:

- dropping the values that “spoil” the result;
- measuring many outcomes, and reporting the one that reached significance;
- collecting data until p dips below 0.05, then stopping;
- slicing the sample into subgroups until one of them is significant.

Chapter 2 described the first as a recognised form of research misconduct, and promised to return to it. All four produce false alarms at far more than the stated 5%, even when nobody means to cheat. Together they are known as p -hacking.

p -hacking

any practice that tries many analyses and reports the one that gave $p < 0.05$

A FIELD CHECKS ITSELF

The Open Science Collaboration repeated 100 studies published in three psychology journals, with larger samples and the original materials where available (reference 4). According to the paper's abstract, 97% of the original studies had reported statistically significant results, but only 36% of the repeats did, and the repeated effects were on average half the size of the originals. Not all of the gap is p -hacking: some is chance, and some is the tendency of journals to publish only significant results. But it shows what happens when a field rewards $p < 0.05$ above everything else.

The cure is to decide the analysis before seeing the data, to write it down, and to report every test that was run. Chapter 2 met the same idea as a data-cleaning log.

2 Comparing two groups

2.1 The same logic, a different estimate

The chapter's opening question compares two groups: students who give private tuition and students who do not. Figure 5 shows every stress score in both groups. The two clouds overlap almost completely, and the tuition group sits a little to the right.

	n	Mean stress	SD
Tuition	58	15.78	4.14
No tuition	60	14.33	4.88

The hypotheses concern the population difference: $H_0: \mu_{\text{tuition}} - \mu_{\text{none}} = 0$ against

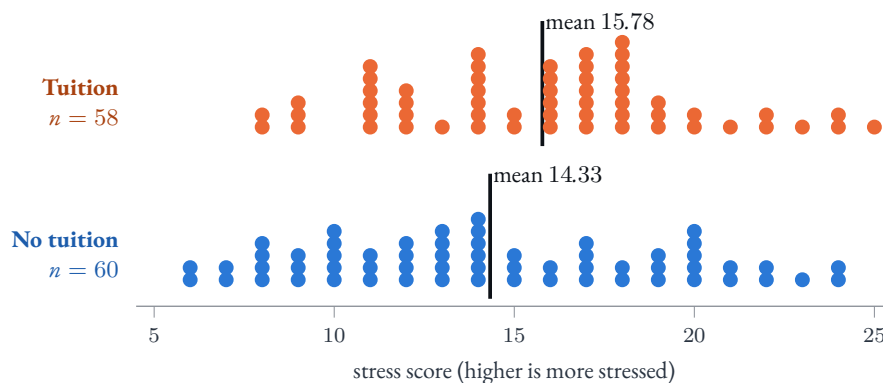


Figure 5: Stress scores by tuition, one dot per student, with each group's mean marked. Two students who skipped a stress item have no score and are left out. **Simulated** class data.

$H_1: \mu_{\text{tuition}} - \mu_{\text{none}} \neq 0$. The estimate is the difference between the two sample means, 1.44 points. It is worked from the unrounded means, so it can differ in the last digit from a subtraction of the rounded ones. The recipe of equation (1.1) does not change: the estimate, minus what the null says, divided by the standard error of the estimate. Only the standard error needs a new formula.

2.2 The standard error of a difference

Each group's mean wobbles from sample to sample, with its own standard error. The difference between the two means wobbles more than either, because it carries both wobbles at once. The two do not simply add. Their squares, the variances, do:

$$SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad (2.1)$$

s_1, n_1 the SD and size of group 1 (tuition)

s_2, n_2 the SD and size of group 2 (no tuition)

Example 2.1 • The tuition gap in stress, by hand

Step	What	Working
1	tuition: s_1^2/n_1	$4.14^2 \div 58 = 17.14 \div 58 = 0.2955$
2	no tuition: s_2^2/n_2	$4.88^2 \div 60 = 23.81 \div 60 = 0.3969$
3	standard error	$\sqrt{0.2955 + 0.3969} = 0.8321$
4	test statistic	$t = 1.44 \div 0.8321 = 1.73$

The gap is 1.73 standard errors from zero. The sleep test of Example 1.1 gave 2.36.

Standard errors do not add

The two groups' own standard errors are $4.14 \div \sqrt{58} = 0.54$ and $4.88 \div \sqrt{60} = 0.63$. Adding them gives 1.17, which overstates the wobble of the difference; the correct value is 0.8321. Square, add, then take the square root.

2.3 Reading the output: an interval that crosses zero

Software runs the Welch t -test on the same data and reports:

t	1.73
degrees of freedom	114.1
p	0.086
95% CI for the difference	−0.21 to 3.09

The t matches the hand value. The degrees of freedom are not a whole number: Welch’s version works them out from the two SDs and the two group sizes, and the result here, 114.1, is close to the $58 + 60 - 2 = 116$ of the older pooled version. The interval can be rebuilt by hand in the usual way, with the t multiplier for that many degrees of freedom, 1.981:

$$1.44 \pm 1.981 \times 0.8321 = 1.44 \pm 1.65, \quad \text{from } -0.21 \text{ to } 3.09.$$

The decision: $p = 0.086$ is above 0.05, so the test does not reject the null. The interval says the same thing more usefully. The population gap could be slightly negative (the interval starts at -0.21), could be zero, or could be as large as 3.09 points. Zero is inside the interval, and a 95% interval for a difference that contains zero is the same statement as $p > 0.05$ (section 1.6).

2.4 “Not significant” does not mean “no difference”

A tempting summary of the tuition result is “tuition does not affect stress”. The data do not support it. They are compatible with no gap, and equally compatible with a gap of up to 3.09 points, which on a scale from 5 to 25 would matter. Section 1.9 showed that a study of this size had only a 41% chance of detecting a gap of the size observed, even if that gap were real. A test with such low power will often fail to reject a false null, so its failure says little.

ABSENCE OF EVIDENCE IS NOT EVIDENCE OF ABSENCE

In a one-page note in the *BMJ*, Altman and Bland warned that a non-significant result is too often reported as proof that there is no difference (reference 5). They pointed out that the confidence interval shows what the data can and cannot rule out. Their title has become a proverb in medical statistics, and the note is worth reading in full.

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 non-significant result is inconclusive, not negative, unless the interval is narrow enough to rule out every difference that would matter. That happens only in large studies.

2.5 Welch or pooled?

The older version of the two-sample test, often called Student’s t -test after the pen name of W. S. Gosset, assumes that the two populations spread equally. It combines the two SDs into one pooled SD, weighting each group by its degrees of freedom:

Welch t -test

a two-sample t -test that does not assume the two groups have the same SD; the default in R

Pooled SD

a single SD for two groups, weighting each group’s variance by its degrees of freedom



$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \quad (2.2)$$

For the tuition data this gives

$$s_p^2 = \frac{(58 - 1) \times 17.14 + (60 - 1) \times 23.81}{116} = \frac{976.98 + 1404.79}{116} = 20.53,$$

$$s_p = \sqrt{20.53} = 4.53.$$

Welch's version, of section 2.3, drops the assumption of equal spread and pays for it with fractional degrees of freedom. Table 6 sets the two side by side.

Table 6: The pooled and Welch versions of the two-sample t -test on the tuition data. Both by software. **Simulated** class data.

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

Here the two agree closely, because the groups are nearly equal in size and the SDs are similar. They part company when the groups differ in size *and* in spread. If the smaller group has the larger SD, the pooled test rejects a true null far more often than 5% of the time. Welch's version stays close to 5% either way, and loses almost nothing when the SDs are in fact equal.

The practical rule

Use Welch's test unless there is a specific reason not to. R uses it by default. SPSS prints both rows, with Levene's test for equal spread above them; read the row labelled "equal variances not assumed".

2.6 What a t -test assumes

Table 7 lists the three assumptions, in order of importance.

Table 7: The assumptions of a t -test, how to check them, and what to do if they fail.

Assumption	How to check	If it fails
Independent observations	From the design: one answer per student, no clusters of friends or repeated people	Nothing in the test can repair it
Roughly normal in each group, <i>or</i> large groups	Histogram or Q-Q plot; from about 30 per group the central limit theorem of Chapter 3 helps	A rank-based test (section 4)
No extreme outliers	Box plot, with the fences of Chapter 2	Check for errors; report the result with and without them

The first assumption comes from how the data were collected, and no calculation can check it. Twenty answers from one group of friends are not twenty independent answers. The second matters less than students tend to fear, because the test is about means, and Chapter 3 showed that the mean of a large sample is close to normal even when the data are not.

The Shapiro–Wilk test checks normality formally, and it is tempting to let it decide.

Formal normality tests mislead

In large samples the Shapiro–Wilk test flags tiny, harmless departures; in small samples, where normality matters most, it misses large ones. Weeknight sleep in the class fails it ($p = 0.008$), mainly because the answers come in half hours, yet with 120 students the t -test of section 1.3 is entirely safe. Look at the histogram instead.

In the class data, stress is close to symmetric (skewness 0.09), so a t -test suits it. Screen time is right-skewed (skewness 0.67, Shapiro–Wilk $p < 0.001$), which is why section 4 uses it for the rank-based tests.

2.7 How big is the gap? Cohen’s d

A p -value says whether a gap can be told apart from zero. It does not say how big the gap is. For that, the gap is expressed in standard deviations, which makes it comparable across scales and across studies. The result is an effect size, and the most common one for two means is Cohen’s d :

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s_p} \quad (2.3)$$

For the tuition gap, $d = 1.44 \div 4.53 = 0.32$: tuition students sit 0.32 SDs higher on average. For the sleep test of section 1.3, with one group, the gap is divided by that group’s SD: $d = -0.20 \div 0.93 = -0.22$, so the class sits 0.22 SDs below seven hours.

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

These labels come from Cohen’s book on power analysis (reference 6). They are quoted here as they are widely reported; for these notes they have not been checked against the book itself. Cohen offered them as rough guides for when nothing better is known, not as verdicts. A “small” shift in the average blood pressure of a whole population, or in the grades of a whole cohort, can matter more than a “large” shift in a few people.

Two different questions

The p -value asks whether the gap can be told apart from zero. The effect size, with its interval, asks how big the gap is. A report needs both.

2.8 Significance depends on the sample size

Figure 6 keeps the tuition gap and the two SDs exactly as they are, and changes only the number of students in each group.

Shapiro–Wilk test

a test whose null hypothesis is that the data come from a normal distribution

Effect size

a measure of how large a difference or association is, separate from the p -value

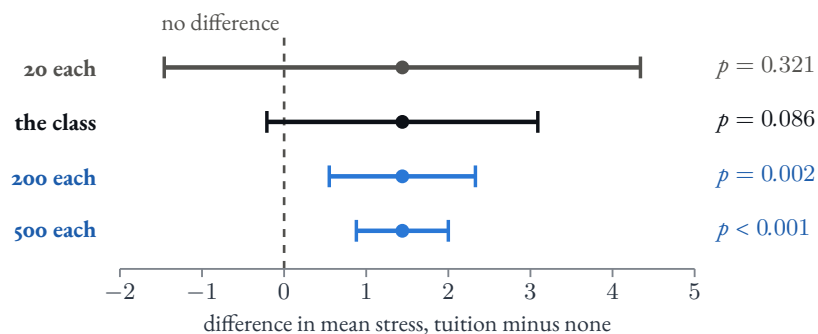


Figure 6: The same gap of 1.44 points, with the same SDs, at four group sizes. Only the width of the interval changes. Hypothetical: the **simulated** class's gap and SDs with other group sizes. Our arithmetic (Welch).

With 20 students in each group the gap is nowhere near significant. With 200 in each, $p = 0.002$; with 500, $p < 0.001$. The effect did not grow. The interval narrowed around it. A p -value is therefore partly a measurement of the sample size, which is the third misreading of Table 3.

2.9 A huge study can make a trivial gap significant

The same fact has a mirror image. The simulation program drew two groups of 20,000 students, and their average weeknight sleep differs by 0.05 hours. The test gives $p < 0.001$: highly significant. The difference is 3 minutes a night, which no student would notice.

Large surveys and studies of hospital records meet this constantly: nearly every difference is significant, and the useful question becomes which differences are large enough to act on. That is a judgement about sleep, or blood pressure, or grades, and not about statistics.

Two questions, always

Can the difference be told apart from zero? Is it large enough to matter? The p -value answers only the first. The second needs knowledge of the subject.

2.10 Paired data: the same students, measured twice

The class survey asked every student about weeknight *and* weekend sleep. The same student answers both questions, so the two answers are linked: a short sleeper on weeknights tends to be a short sleeper at weekends too. Data like these are paired. Treating them as two separate groups would throw away the link.

The paired analysis turns two columns into one. For each student, subtract weeknight sleep from weekend sleep. One student skipped a sleep question, which leaves 119 differences (Figure 7). Their mean is $\bar{d} = 1.06$ hours and their SD is $s_d = 0.72$. Only 4 students slept less at weekends, and 10 slept the same.

The question is now a one-sample question: is the mean difference zero? The paired t -test is the one-sample test of section 1.3, applied to the differences, with a null value of 0:

Paired data

two measurements on the same individuals, or on matched pairs; each pair gives one difference

Paired t -test

a one-sample t -test applied to the differences within pairs

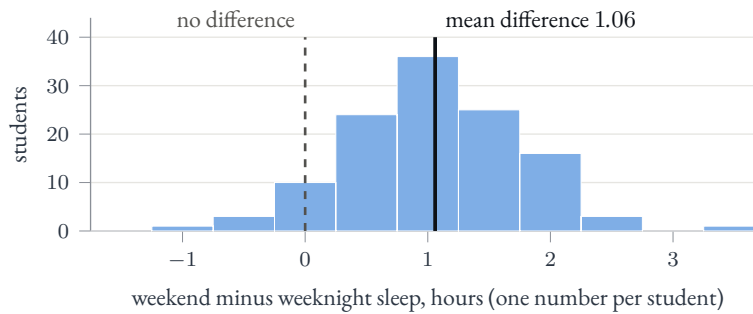


Figure 7: Weekend minus weeknight sleep, one difference per student. Most of the histogram lies to the right of zero. **Simulated** class data.

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} \quad (2.4)$$

Example 2.2 • Weekend against weeknight sleep

Step	What	Working
1	standard error	$0.72 \div \sqrt{119} = 0.066$
2	test statistic	$t = 1.06 \div 0.066 = 16.1$
3	degrees of freedom	$119 - 1 = 118$
4	95% interval	$1.06 \pm 1.980 \times 0.066 = 1.06 \pm 0.13$

Software, working from unrounded values, gives $t = 15.99$, $p < 0.001$ and a 95% interval from 0.93 to 1.19 hours. The hand t is slightly off only because the SE was rounded to three decimals; the interval reproduces exactly.

The mean difference is about 16.1 standard errors from zero. Students slept about 64 minutes more at weekends, give or take about 8 minutes. The effect size for paired data divides the mean difference by the SD of the differences: $d_z = 1.06 \div 0.72 = 1.47$, which is large by any label. A pattern like this is a common sign of sleep debt built up on weeknights, but a single survey cannot show that.

The sentence

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

2.11 The wrong test throws information away

Suppose the same data were analysed as two separate groups of 119, as if the weekend answers and the weeknight answers came from 238 different students. Weekend sleep has SD 1.05 and weeknight sleep has SD 0.92, so equation (2.1) gives

$$SE = \sqrt{\frac{1.05^2}{119} + \frac{0.92^2}{119}} = 0.128, \quad t = 1.06 \div 0.128 = 8.28.$$

Software gives $t = 8.27$. Table 8 compares the two analyses.

Table 8: The weekend sleep gain analysed correctly, as paired data, and wrongly, as two separate groups. Both by software. **Simulated** class data.

	Paired (correct)	Unpaired (wrong)
Standard error	0.066	0.128
t	15.99	8.27
95% CI, hours	0.93 to 1.19	0.81 to 1.31

Both analyses find a difference here, because the effect is large. But the unpaired interval is much wider, and with a smaller effect the wrong test could miss it altogether. The reason is that students differ a great deal in how much they sleep overall. The paired test subtracts each student's own baseline; the unpaired test leaves those differences in as noise.

The SD of the differences shows the mechanism. It depends on the two SDs and on how strongly the two answers go together, measured by the correlation r that Chapter 5 introduces. For these data $r = 0.74$, and

$$\begin{aligned}
 s_d &= \sqrt{s_{\text{we}}^2 + s_{\text{wd}}^2 - 2r s_{\text{we}} s_{\text{wd}}} \\
 &= \sqrt{1.05^2 + 0.92^2 - 2 \times 0.74 \times 1.05 \times 0.92} = \sqrt{0.5192} = 0.72.
 \end{aligned}$$

The larger r is, the more the last term removes, and the more pairing gains. This formula is stated here only to show why pairing helps; the paired test itself never needs it, since the differences can be computed directly.

Pairing comes from the design

The same people measured twice, or pairs matched on purpose, call for a paired test. Two separate groups never do. Whether data are paired is decided by how they were collected, not by looking at the numbers.

2.12 Which t -test?

Table 9: Choosing a t -test. Examples from the **simulated** class data.

The data	Test	Example in this chapter
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
The same people twice, or matched pairs	paired t	weekend against weeknight sleep
Three or more separate groups	one-way ANOVA (section 3)	sleep by residence

One question usually decides between the second and third rows: could the same student appear in both groups? If so, the data are paired.

3 Comparing three or more groups: ANOVA

3.1 Three groups, one test

Students in the class live in a residential hall, in a private mess, or at their family home. Does average weeknight sleep differ between the three? Figure 8 shows the data. Family-home students seem to sleep a little more; hall and mess students look much the same.

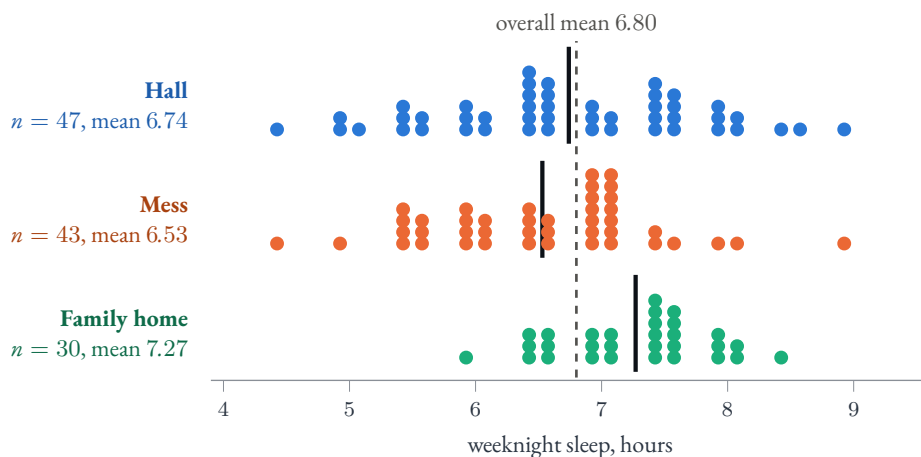


Figure 8: Weeknight sleep by residence, one dot per student. The bars mark the group means and the dashed line the overall mean. **Simulated** class data.

Residence	<i>n</i>	Mean sleep, hours	SD
Hall	47	6.74	1.04
Mess	43	6.53	0.88
Family home	30	7.27	0.60
All	120	6.80	

One approach would be three *t*-tests: hall against mess, hall against home, mess against home. Section 1.12 explains the problem. Three tests at 0.05 on data where nothing is going on give a $1 - 0.95^3 = 14.3\%$ chance of at least one false alarm. One test of all three means at once keeps the rate at 5%. That test is the analysis of variance, or ANOVA. With one grouping variable, as here, it is called one-way ANOVA. Its hypotheses are

$$H_0: \mu_{\text{hall}} = \mu_{\text{mess}} = \mu_{\text{home}} \quad H_1: \text{at least one mean differs from the others.}$$

Analysis of variance

ANOVA: a test of whether several group means are equal, comparing the spread between groups with the spread within them

3.2 The idea: spread between against spread within

The name sounds odd for a test about means. It is a test about means that works by comparing two kinds of spread. Figure 9 shows the idea on constructed data.

Both panels have the same three means. What differs is the spread inside each group. On the left the means are far apart *relative to* that spread; on the right they are not. ANOVA turns



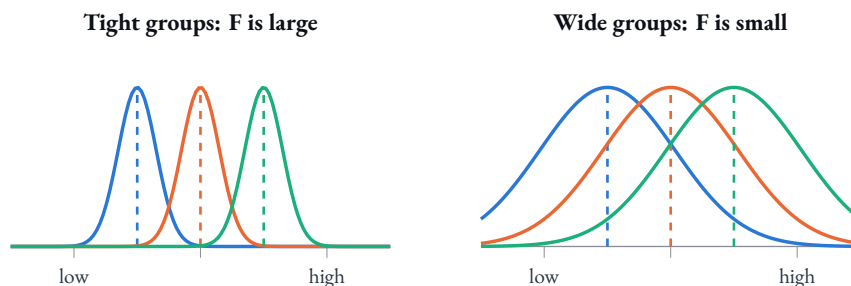


Figure 9: The same three group means in both panels. On the left, each group is tight, and the gaps between the means stand out. On the right, each group is wide, and the same gaps could easily be luck. **Constructed** for teaching: made-up means and spreads.

this comparison into a single ratio, the F statistic:

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

If the null is true, the group means differ only by luck, and the two kinds of variation estimate the same thing: F is close to 1. A large F says the means are further apart than the scatter inside the groups would explain.

F statistic

the ratio of the variance between group means to the variance within groups

3.3 The ANOVA table

Software presents the calculation as an ANOVA table (Table 10). Each row has a sum of squares, its degrees of freedom, and a mean square.

Sum of squares

a total of squared deviations; ANOVA splits it into a between-groups and a within-groups part

Table 10: One-way ANOVA of weeknight sleep by residence. By software; the divisions reproduce it. **Simulated** class data.

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			

The sums of squares measure spread. The between-groups sum of squares (SSB) measures how far each group mean sits from the overall mean, counted once for every student in the group:

$$SSB = \sum_j n_j (\bar{x}_j - \bar{x})^2 \quad (3.1)$$

$n_j, \bar{x}_j$ the size and mean of group j

$\bar{x}$ the overall mean of all 120 students

The within-groups sum of squares (SSW) adds up the squared deviations of every student from their own group's mean. It equals $\sum_j (n_j - 1)s_j^2$, the same weighting as the pooled SD of equation (2.2). The two parts add to the total sum of squares, the squared deviations of every student from the overall mean: $9.70 + 92.50 = 102.20$.



Example 3.1 • The between-groups sum of squares, by hand

With the means to three decimals, and the overall mean 6.800:

Group	n_j	$\bar{x}_j$	$\bar{x}_j - \bar{x}$	$n_j(\bar{x}_j - \bar{x})^2$
Hall	47	6.745	-0.055	0.14
Mess	43	6.535	-0.265	3.02
Family home	30	7.267	0.467	6.54
SSB				9.70

The family-home group contributes most, because its mean sits furthest from the overall mean. SSW is best left to software: working it from SDs rounded to two decimals gives a value noticeably off 92.50.

A sum of squares grows with the number of values in it, so it is divided by its degrees of freedom to give a mean square, which is a variance. The degrees of freedom are the number of groups minus one between groups, and the number of students minus the number of groups within them:

$$MSB = \frac{9.70}{2} = 4.85, \quad MSW = \frac{92.50}{117} = 0.791, \quad F = \frac{4.85}{0.791} = 6.13.$$

The between-group variance is 6.13 times the within-group variance, well above 1. The F statistic has two sets of degrees of freedom, one for the top of the ratio and one for the bottom, and is written $F(2, 117) = 6.13$. Its $p = 0.003$ comes from the F distribution with those degrees of freedom. The test rejects the null: at least one residence group differs in mean sleep.

3.4 How much does residence explain? η^2

The sums of squares give an effect size directly. Eta squared is the share of the total variation that lies between groups:

$$\eta^2 = \frac{SSB}{SST} = \frac{9.70}{102.20} = 0.095 \quad (3.2)$$

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

Significant, and modest

$p = 0.003$ says the three group means are unlikely all to be equal. $\eta^2 = 0.095$ says residence is a small part of why students differ in sleep. Both belong in the report.

3.5 Which groups differ? Tukey's comparisons

ANOVA says that at least one mean differs, not which. Comparing every pair of groups brings back the many-tests problem, and section 1.12 promised a better tool than Bonferroni for this case. Tukey's method, also called the honestly significant difference, gives an interval and a p -value for every pair of groups, adjusted together so that the familywise false-alarm

Mean square

a sum of squares divided by its degrees of freedom; a variance

Eta squared

η^2 , the share of the total variation that lies between groups

Tukey's method

Tukey's honestly significant difference: intervals and p -values for every pair of groups, adjusted so the familywise rate stays at 5%

rate stays at 5%. It is less strict than Bonferroni, because it uses the structure of all-pairs comparisons rather than treating them as unrelated tests.

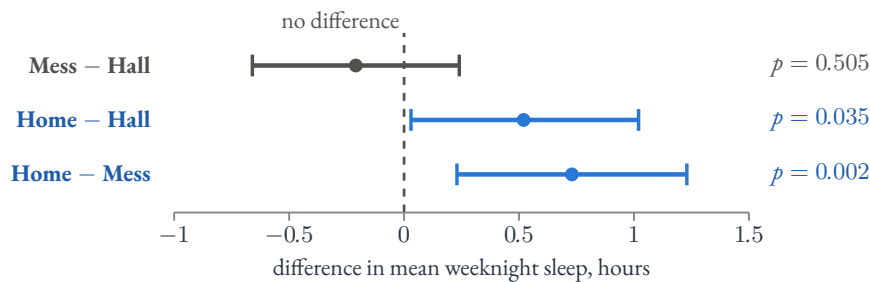


Figure 10: Tukey's 95% intervals for the three pairwise differences in mean sleep, adjusted for three comparisons. Intervals that clear zero are in blue. **Simulated** class data; Tukey HSD by software.

Two intervals in Figure 10 clear zero and one crosses it. Family-home students sleep about 0.52 hours more than hall students, about 31 minutes, and 0.73 hours more than mess students, about 44 minutes. Hall and mess students cannot be told apart. The home-against-hall interval comes very close to zero: its lower limit, 0.03 hours, is about 2 minutes. That difference is significant, but only just, and the data are compatible with a gap of a couple of minutes.

3.6 Checking the assumptions

ANOVA makes the assumptions of the t -test (Table 7): independent students, and roughly normal data in each group or large groups. The classic version also assumes that the groups spread equally, which is the pooled assumption of section 2.5 extended to three groups.

- **Equal spreads?** The SDs are 1.04, 0.88 and 0.60; the largest is 1.7 times the smallest. The Brown–Forsythe test, a version of Levene's test, gives $p = 0.020$: the spreads probably differ. Family-home students are noticeably more alike in their sleep.
- **Normal residuals?** The residuals are each student's distance from their own group mean. Shapiro–Wilk on the residuals gives $p = 0.416$, with no sign of a problem. The raw sleep values failed the same test (section 2.6). Subtracting each group's own mean breaks up the half-hour lumps, which is partly why the residuals pass.

Unequal spreads have the same remedy as in section 2.5. **Welch's ANOVA** does not assume equal spreads. It gives $F = 9.78$ on 2 and 77.8 degrees of freedom, $p < 0.001$, so the conclusion stands either way.

The right attitude to assumptions

Not “did it pass?” but “does the answer change if the assumption is relaxed?” A rough guide: 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 compare.

3.7 Look before testing

Every test in this chapter summarises the data into one number, and a summary can hide what matters. Four checks come before any p -value:

1. Plot every group, with dots or box plots.
2. Look for outliers and impossible values, as in Chapter 2.
3. Compare the spreads, not only the centres.
4. Ask whether a gap of the size seen would matter.

In the class data the Tukey gaps, about 31 and 44 minutes, are worth knowing about, and modest next to the spread inside each group, whose SDs run from 0.60 to 1.04 hours. A sentence claiming that residence strongly affects sleep would overstate the result.

A test answers only the question it was given

ANOVA compares means. It will not notice that one group has two clusters, that a group of three students drives the result, or that a question was misread. Chapter 5 shows a famous set of four datasets with identical summary statistics and completely different shapes.

3.8 Reporting a test

The report of a test is a sentence, with the estimate first and the test in support.

The ANOVA sentence

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).

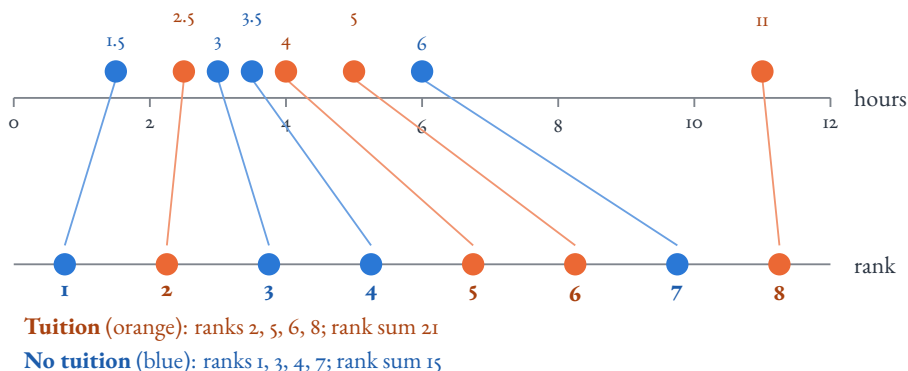
Always include	Avoid
the estimate, in its units	$p = 0.000$; write $p < 0.001$
its 95% interval	stars in place of numbers
the test, its statistic and df	“NS” with no estimate
the exact p , or $p < 0.001$	“proved”, “no effect”

SPSS prints “Sig. .000” when a p -value is very small. A p -value is never exactly zero: the printout has rounded it to three decimals, and it should be reported as $p < 0.001$.

4 When the data are skewed: rank-based tests

4.1 Replace the values with their ranks

Screen time is right-skewed: most students report a few hours, and a few report many (Chapter 2). A mean is pulled by the long tail, and so is a t -test. One remedy is to throw away the distances and keep only the order. Line all the values up from smallest to largest and number them 1, 2, 3, and so on. Each number is a rank. From then on only the ranks are used.



Rank

a value's position when all values are sorted from smallest to largest; tied values share the average of their positions

Figure 11: Eight screen-time values from two groups, replaced by their ranks. The 11-hour value becomes rank 8, as it would at 6.5 hours or 50. **Constructed** for teaching: eight made-up values.

Figure 11 shows eight constructed values. The tuition group's mean, 5.6 hours against 3.5, is pulled up by one heavy user at 11 hours. Its ranks simply say "mostly higher". The 11-hour value gets rank 8, and it would get rank 8 at 6.5 hours or at 50, so one extreme value cannot drag the result. The ranks 1 to 8 add to 36. If the groups did not differ, each would expect half, 18. The tuition group's ranks, 2, 5, 6 and 8, add to 21.

4.2 The Mann–Whitney test

The Mann–Whitney test ranks all the values from both groups together and asks whether one group's ranks run higher than chance would give. Its null hypothesis is that neither group tends to have higher values than the other. The same test is also called the Wilcoxon rank-sum test: R calls it `wilcox.test` and reports a statistic W , and SPSS calls it Mann–Whitney U . They are the same test.

Mann–Whitney test

a rank-based test of whether one of two groups tends to have higher values; also called the Wilcoxon rank-sum test

Screen time	n	Median, hours
Tuition	59	4.60
No tuition	58	4.15

For screen time by tuition in the class, all 117 students with an answer are ranked together, and the test gives $W = 1807.5$, $p = 0.601$. There is no evidence that either group spends more time on screens.



4.3 What W counts

The statistic W has a meaning that can be explained to anyone. Pair every tuition student with every non-tuition student. There are $59 \times 58 = 3,422$ such pairs. W counts the pairs in which the tuition student has more screen time, counting ties as a half. Dividing,

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

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%. A value of 70% or 80% would be a large difference, whatever the medians said. Some fields report this share directly as an effect size, under names such as the probability of superiority.

In the constructed example of Figure 11, the tuition group's rank sum is 21. The smallest rank sum a group of four could have is $1 + 2 + 3 + 4 = 10$, and W is the difference: $W = 21 - 10 = 11$. There are $4 \times 4 = 16$ pairs, so the tuition value is higher in $11 \div 16 = 0.69$ of them. Counting the pairs directly gives the same 11.

Not strictly a test of medians

The Mann–Whitney test asks whether one group tends to be higher, which is not the same as asking whether the medians differ. Report the medians beside it, but do not call it a median test.

4.4 The Kruskal–Wallis test

The Kruskal–Wallis test is the rank-based partner of one-way ANOVA. It ranks everyone together and compares the average rank in each group. Its statistic H is compared with the chi-square distribution, which Chapter 5 introduces, on the number of groups minus one degrees of freedom.

Kruskal–Wallis test
the rank-based partner of one-way ANOVA, for three or more groups

Residence	Median screen time, hours
Hall	4.10
Mess	4.45
Family home	4.65

For screen time by residence, $H = 0.75$ on 2 df, $p = 0.688$: 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.

For paired data there is a third rank-based test. The **Wilcoxon signed-rank test** ranks the sizes of the paired differences and compares the ranks of the positive differences with those of the negative ones. It is the rank-based partner of the paired t -test, and also of the one-sample t -test.

4.5 Mean-based or rank-based?

Choose by the question. If the mean is what matters, such as total hours or total cost, the mean-based test answers it, and with large groups it copes with skew. A mean-based test also gives a difference in hours with an interval, which is often what a reader wants. Ordinal data,



Table 11: Tests based on means and tests based on ranks.

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

such as the answer to a single Likert item, have an order and nothing else, so rank-based tests fit them naturally. Chapter 6 returns to Likert scales.

Do not shop

Running both a mean-based and a rank-based test and reporting the one with the smaller p is another way to hunt for significance (section 1.13). Decide first.

4.6 Choosing a test: the map

Table 12 gathers the tests of this chapter, with the ones Chapters 5 and 6 add. Start from the question and the type of outcome, not from the software menu.

Table 12: Choosing a test. Chapter 5 turns this table into a flowchart.

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
The same people twice	paired t	Wilcoxon signed-rank
Three or more groups	one-way ANOVA	Kruskal–Wallis
Two categorical variables	chi-square test (Chapter 5)	
Two numerical variables	Pearson r (Chapter 5)	Spearman's ρ (Chapter 5)
Predicting an outcome	regression (Chapter 6)	

5 Reading the p -values in a published table

5.1 An interval that includes 1

Chapters 1 and 3 used a survey of 476 Bangladeshi university students during the COVID-19 lockdown (reference 1). Its Table 3 reports which student characteristics went with depression and anxiety, as odds ratios, each with a 95% interval and a p -value. Chapter 6 explains odds ratios properly. For now one fact is enough: an odds ratio of 1 means no association, above 1 means higher odds of the outcome, and below 1 means lower odds. For a ratio, “nothing going on” is 1, not 0.

The interval rule of section 1.6 therefore reads: a 95% interval that includes 1 should come with $p > 0.05$, and one that excludes 1 should come with $p < 0.05$. Table 13 lists eight of the sixteen rows of Table 3, as printed.

Odds ratio

a ratio comparing the odds of an outcome in two groups; 1 means no association

Table 13: Eight rows of Islam 2020, Table 3, as printed, with p -values rebuilt from each row's own interval. D: depression; A: anxiety. The rebuilt p -values are **ours**.

Row	OR	95% CI	Printed p	Rebuilt p
Gives tuition (D)	1.199	0.736–1.952	.046	0.466
Gives tuition (A)	1.381	0.856–2.227	.018	0.186
Age 21–24 (D)	1.059	0.632–1.775	.082	0.828
Lagging academically (D)	1.795	1.098–2.935	.020	0.020
Living with family (D)	2.595	1.418–4.751	.002	0.002
Female (A)	1.462	0.866–2.468	.155	0.155
Age over 24 (D)	2.886	0.961–8.669	.059	0.059
Regular exercise (D)	0.880	0.544–1.422	.601	0.602

The first two rows break the rule. Both tuition intervals include 1, yet both print $p < 0.05$. Figure 12 shows them beside two rows that keep it.

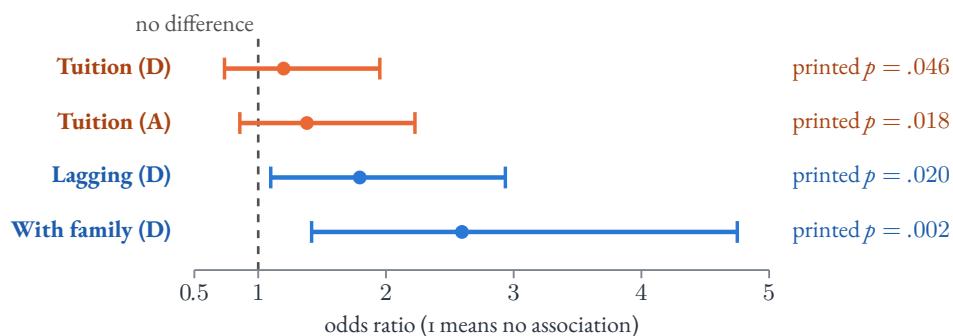


Figure 12: Four rows of Islam 2020, Table 3, as printed. The two orange intervals include 1 but print $p < 0.05$; the two blue intervals exclude 1 and print $p < 0.05$. Islam MA et al. *PLoS One* 2020, CC BY 4.0.

5.2 Rebuilding a p -value from an interval

Any ratio reported with a 95% interval carries its own p -value, and the p -value can be recovered with a calculator. The trick is to work on the log scale. There the interval is symmetric: the log of the odds ratio, plus or minus 1.96 standard errors. So the width of the interval on the log scale gives the SE, and the SE gives z and then p .

$$SE = \frac{\ln(\text{upper}) - \ln(\text{lower})}{2 \times 1.96}, \quad z = \frac{\ln(\text{OR})}{SE} \quad (5.1)$$

Here $\ln$ is the natural logarithm, the $\ln$ key on a calculator. The null value 1 becomes $\ln 1 = 0$ on the log scale, which is why z has no subtraction on top.

Example 5.1 • Tuition and depression, rebuilt

Printed: OR 1.199, 95% CI 0.736 to 1.952, $p = .046$.

Step	What	Working
1	log of the OR	$\ln 1.199 = 0.181$
2	logs of the limits	$\ln 1.952 = 0.669$; $\ln 0.736 = -0.307$
3	standard error	$(0.669 - (-0.307)) \div 3.92 = 0.249$
4	z	$0.181 \div 0.249 = 0.73$
5	two-sided p	$2 \times P(Z > 0.73) = 0.465$

Software, working from unrounded values, gives 0.466. The printed value is .046. A z of 0.73 is nowhere near the 1.96 needed for $p < 0.05$. The rebuilding is our arithmetic; the paper prints only the odds ratio, the interval and .046.

The same method rebuilds thirteen of the sixteen rows of Table 3 to their printed p -values, within rounding, which suggests the method is sound. It fails on three rows: the two tuition rows and age 21–24, whose printed values are .046, .018 and .082 against rebuilt values of 0.466, 0.186 and 0.828.

Our reading, not the authors'

Each of the three printed values looks like the rebuilt value with an extra zero after the decimal point: 0.466 printed as .046, and so on. That is our guess at what went wrong, and it has not been confirmed with the authors. What is certain is that the printed p -values disagree with their own intervals.

5.3 What the paper concluded

The paper's abstract reports the tuition odds ratio for depression as a finding, and its discussion calls private tuition "a critical factor" in the depression and anxiety of students. Table 3 does not support that. Both tuition intervals include 1, and the rebuilt p -values are 0.466 and 0.186. The data are compatible with no association, and also with a sizeable one: this is section 2.4 seen from the other side, a result reported as significant that the table does not show to be.

Three further cautions apply to the whole table.

- **Many tests.** Table 3 holds sixteen tests. At 0.05, if nothing at all were going on, about $16 \times 0.05 = 0.8$ false alarms would be expected.
- **The sample.** The survey recruited through Facebook, and Chapter 3 showed that an interval covers the luck of sampling, not the bias of who chose to answer.
- **Association, not cause.** A one-off survey cannot show which way an effect runs. Chapter 5 takes this up.

This is not a gotcha. Misprints happen in good journals, and most of the table rebuilds correctly. The point is the habit.

The reader's habit

Check that each p -value agrees with its interval. It takes a minute. When they disagree, trust neither until the reason is known.

5.4 Rows that pass this check

Two depression rows pass the interval check with clear significance: lagging academically, OR 1.795 (1.098 to 2.935), and living with family, OR 2.595 (1.418 to 4.751). Both rebuild to their printed p -values.

Even these need care. 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 higher. And the direction of cause is unknown. Falling behind may raise depression, or depression may make students fall behind.

Passing one check is not passing every check

Chapter 6 checks the living-with-family row against the counts the paper prints in its Table 2, and there it fails. Only 14 students did not live with family, and the counts give an odds ratio of about 0.77, below 1, with an interval from 0.17 to 3.52. One of the two tables must be wrong for this row, and it is not possible to tell which. Until the authors explain it, the row should not be cited.

6 Testing a hypothesis: the procedure

1. **State the question** as a null and an alternative hypothesis about the population, before seeing the data. Two-sided unless there is a stated reason.
2. **Fix α** , usually 0.05, and count the tests that will be run.
3. **Look at the data:** plot every group, check for impossible values and outliers, compare the spreads.
4. **Choose the test** from the design and the type of outcome (Table 12): paired or separate, two groups or more, means or ranks.
5. **Compute the estimate and its interval:** a difference in means, or a mean difference, with its 95% interval.
6. **Compute the test:** the statistic, its degrees of freedom and p .

7. **Check the assumptions**, and whether relaxing them changes the answer.
8. **Report** the estimate, its interval, the test and p in one sentence, with an effect size. Then judge whether the size matters.

7 Chapter summary

1. A test presumes the null hypothesis, that nothing is going on, and asks how surprising the data would be if it were true. The p -value is that chance.
2. A p -value is not the chance that H_0 is true, and it does not measure the size or the importance of an effect.
3. A 95% interval excludes a value exactly when a two-sided test of that value gives $p < 0.05$. The interval also gives the size.
4. A test can raise a false alarm (Type I, rate α) or miss a real effect (Type II). Power is the chance of detecting an effect of a stated size; small studies have little.
5. Many tests give many chances for a false alarm. Count the tests; a Bonferroni correction divides α by their number.
6. Two separate groups: the Welch t -test. The same people twice: the paired t -test, which is a one-sample test on the differences.
7. Not significant does not mean no difference. Significant does not mean important. Report the estimate, its interval and an effect size such as d .
8. Three or more groups: one-way ANOVA compares spread between groups with spread within them. η^2 gives the share explained, and Tukey's method finds which groups differ.
9. Skewed or ordinal data: rank-based tests. The Mann–Whitney statistic counts the pairs in which one group's member is higher.
10. A published p -value can be checked against its own interval. When they disagree, something is wrong.

8 Key terms

Alternative hypothesis (H_1)

The claim that something is going on; the complement of H_0 .

Analysis of variance (ANOVA)

A test of whether several group means are equal, using the ratio F .

Bonferroni correction

Testing each of k tests at α/k .

Cohen's d

A difference in means divided by the pooled SD, equation (2.3).

Critical value

The value of the test statistic at which $p = \alpha$.

Effect size

A measure of how large a difference or association is, such as d or η^2 .

Eta squared (η^2)

The share of total variation lying between groups, equation (3.2).

Familywise error rate

The chance of at least one false alarm across a family of tests, equation (1.2).

Kruskal–Wallis test

The rank-based partner of one-way ANOVA.

Mann–Whitney (Wilcoxon rank-sum) test

A rank-based test of whether one of two groups tends to be higher.

Null hypothesis (H_0)

The claim that nothing is going on, stated about the population.

p -value

The chance of data at least as far from H_0 as those observed, if H_0 were true.

p -hacking

Trying many analyses and reporting the one that gave $p < 0.05$.

Paired t -test

A one-sample t -test on the differences within pairs, equation (2.4).

Power

The chance of rejecting H_0 when a real effect of a stated size exists.

Significance level (α)

The p -value below which H_0 is rejected, fixed in advance.

Test statistic

How far the data sit from H_0 , usually in standard errors, equation (1.1).

Tukey's method

Intervals for every pair of group means, adjusted together.

Type I error

Rejecting a true H_0 : a false alarm.

Type II error

Failing to reject a false H_0 : a missed effect.

Welch t -test

A two-sample t -test that does not assume equal SDs, equation (2.1).

Wilcoxon signed-rank test

The rank-based partner of the one-sample and paired t -tests.

9 Exercises

Exercises in Part A need no calculation. Parts B to E need a calculator; where a critical value or a tail area is needed, it is given. Numbers not attributed to a paper or to the class data are constructed for teaching.

Part A. Ideas.

- Write a null and an alternative hypothesis for each question, in words and in symbols. (a) Is the average CGPA of students in one department different from 3.00? (b) Do male and female students differ in average daily phone use? (c) Does a study-skills workshop change the exam scores of the students who attend it, measured before and after? (d) Does average monthly spending differ across four faculties?
- A study reports $p = 0.03$ for a difference between two groups. Say whether each statement is right or wrong, and why. (a) There is a 3% chance that the null hypothesis is true. (b) If the null hypothesis were true, data at least this far from it would turn up about 3% of the time. (c) The difference is large. (d) At $\alpha = 0.05$, the null hypothesis is rejected. (e) If the study were repeated, it would give $p < 0.05$ with probability 0.97.
- A student is accused of copying in an examination, and a disciplinary committee hears the case. Using the court analogy of section 1.1: (a) what plays the part of the null hypothesis? (b) What would a Type I error be, and what would a Type II error be? (c) The committee finds the case “not proven”. Does that show the student did not copy?
- A university tests whether a new attendance policy raised average attendance. Describe, in words about attendance, (a) a Type I error and (b) a Type II error. (c) Which would be more costly if the policy is expensive to run, and which if it is cheap?
- Three studies each report a 95% interval for the difference in mean exam score, new teaching method minus old: (i) 0.4 to 2.8; (ii) -0.3 to 1.9; (iii) -2.5 to -0.1 . For each, (a) is the difference significant at 0.05? (b) Which method does the interval favour, if either? (c) What can be said about the size of the difference?
- Paired or separate? Say which, and why. (a) Blood pressure of 30 students measured the day before an examination and the day after. (b) Sleep of 40 hall students and 40 mess students. (c) Grip strength of the left and right hands of 25 students. (d) Examination marks of one class in Statistics and in English. (e) Monthly spending of 50 husbands and their 50 wives.
- A national survey of 12,000 students finds that those who use a certain study app score 0.4 points higher, on average, on a 100-point examination, with $p < 0.001$. A newspaper headline says: “App dramatically improves exam results.” Comment on the headline, using at least two ideas from this chapter.

Part B. Calculations.

- 8.** Twenty-five students in one department have mean CGPA 3.12 and SD 0.30. (a) Test whether the department's average differs from 3.00: find the SE, t and df. (b) The critical value for 24 df is 2.064. What is the decision at 0.05? (c) Find the 95% interval, and check that it agrees with the decision.
- 9.** Suppose the class survey had reached only 30 students, with the same mean weeknight sleep, 6.80 hours, and the same SD, 0.93. (a) Find the SE and the t statistic for a test against 7 hours. (b) The critical value for 29 df is 2.045. What is the decision? (c) The average shortfall is exactly the same as in section 1.3. Why has the conclusion changed?
- 10.** Two groups of students take a reading test. Group A: $n = 16$, mean 72, SD 8. Group B: $n = 20$, mean 67, SD 10. (a) Find the standard error of the difference with equation (2.1). (b) Find t . (c) Software gives Welch's df as 34.0, for which the 95% multiplier is 2.032. Find the 95% interval for the difference and state the decision.
- 11.** Eight students sit a practice test before and after a revision workshop.

Student	1	2	3	4	5	6	7	8
Before	52	60	45	70	58	63	49	55
After	56	63	47	75	60	68	50	59

- (a) Find the eight differences, after minus before, and their mean and SD. (b) Carry out a paired t -test; the critical value for 7 df is 2.365. (c) Find the 95% interval for the mean gain, and d_z .
- 12.** Fifteen students are taught by one of three methods, five to each, and then take a short test.

Method A	5	7	9	6	8
Method B	7	9	11	8	10
Method C	3	5	7	4	6

- (a) Find the three group means and the overall mean. (b) Find SSB with equation (3.1), and SSW by adding the squared deviations of each score from its own group mean. (c) Complete an ANOVA table. The critical value of F for 2 and 12 df is 3.89. (d) Find η^2 .
- 13.** For the two groups of Exercise 10, find the pooled SD with equation (2.2), and Cohen's d . Is the gap small, medium or large on Cohen's rough labels, and does the test of Exercise 10 contradict that?
- 14.** A student survey asks 40 questions and tests each one for a difference between men and women, at 0.05. Suppose there are no real differences at all. (a) How many "significant" results should be expected? (b) Treating the tests as independent, what is the chance of at least one? (c) What level would a Bonferroni correction use for each test?
- 15.** Group X has values 3, 7 and 9; group Y has values 2, 4, 5 and 6. (a) Rank all seven values together. (b) Find the rank sum of group X, and W by subtracting $1 + 2 + 3 = 6$. (c) Check W by counting the pairs, one from each group, in which the X value is higher. (d) Express W as a share of all pairs. (e) Software gives an exact $p = 0.40$. Why can such small groups never give $p < 0.05$ with this test?

Part C. The class data.

16. Some students say they aim for eight hours of sleep at weekends. For the 119 students with both sleep answers, weekend sleep has mean 7.87 hours and SD 1.05. (a) Write the hypotheses for a test against 8 hours. (b) Find the SE and t by hand. (c) Software gives $p = 0.178$. State the decision, and check it against the 95% interval, 7.68 to 8.06 hours.

17. Does private tuition take time away from self-study? Students who give tuition report 2.53 hours of self-study on a typical class day, and those who do not report 2.71. The Welch test gives $t = -0.81$, $p = 0.420$, and a 95% interval for the difference, tuition minus none, of -0.62 to 0.26 hours. (a) State the decision. (b) A classmate writes: “Tuition has no effect on study time.” Rewrite the sentence so that it says what the data show.

18. Repeat the decision of section 1.5 for the sleep test at $\alpha = 0.01$ instead of 0.05. Check your answer against the 99% interval of section 1.6.

19. A student looks at the class data, sees that the mean sleep is below 7 hours, and then runs a one-sided test with $H_1: \mu < 7$. They report $p = 0.010$. What is wrong, and what should they have reported?

20. The pooled test of Table 6 uses the standard error $s_p \sqrt{1/n_1 + 1/n_2}$. Using $s_p = 4.53$ from section 2.5, find this SE and the pooled t for the tuition gap in stress. Compare them with the Welch SE of Example 2.1 and the pooled t in Table 6.

Part D. Activities.

21. The class activity: for each row below, (a) is the result significant at 0.05? (b) Do the p -value and the interval agree? (c) Would you call it important, and what else would you need to know?

	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, $p < 0.001$

Rows 1 to 5 are from Islam 2020, Table 3, as printed; rows 6 to 8 use **simulated** data.

22. Twenty practice questions on choosing a test. For each, name the test from Table 12. Some need a test from Chapter 5 or 6; name it anyway.

- (a) Is the average height of male students at one university different from the national average of 165 cm?
- (b) Do hall and mess students differ in average monthly spending? The spending data are roughly symmetric, with about 60 students in each group.
- (c) Twenty students rate their stress before and after the examination week, on a scale

whose scores are roughly normal.

- (d) Screen time, heavily skewed, is compared between 12 first-year and 15 final-year students.
- (e) Average CGPA is compared across five faculties, with about 40 students in each.
- (f) Is gender related to whether a student gives tuition (yes or no)?
- (g) Monthly family income, heavily skewed, is compared across four divisions.
- (h) The same 30 students' typing speeds are measured on two keyboards.
- (i) Students answer one Likert item, from “strongly disagree” to “strongly agree”, and the answers of science and arts students are compared.
- (j) Do students who sleep more also have higher CGPAs?
- (k) Pairs of twins, one twin in each pair given coaching, sit the same test.
- (l) Weight loss is compared across three diets, with about 25 people on each and roughly normal data.
- (m) A researcher wants to predict CGPA from study hours, sleep and attendance together.
- (n) Ten patients rate their pain before and after a treatment; the differences are very skewed.
- (o) Is the average daily step count of office workers different from 10,000? The data are roughly normal, from 80 workers.
- (p) Mean reaction time is compared across four age groups, 30 people in each.
- (q) Is the district a student comes from related to their choice of faculty?
- (r) Do smokers and non-smokers differ in mean lung capacity? The smokers are fewer and more variable.
- (s) Household income, heavily skewed, is compared across hall, mess and family-home students.
- (t) The attendance of the same 40 students is recorded in the first month and in the last month of a term.

Part E. Challenge.

23. Rebuild the p -value of the “age over 24” row of Islam 2020, Table 3: OR 2.886, 95% CI 0.961 to 8.669, printed $p = .059$. Follow Example 5.1. The tail area you need is $P(Z > 1.89) = 0.0294$. Does the rebuilt value agree with the printed one? What should a reader conclude about age?

24. Figure 4 shows 1,000 shuffles of the tuition labels, of which 48 gave $p < 0.05$. (a) How many would you expect below 0.01? (b) The expected number below 0.05 was 50. Using the standard error of a proportion from Chapter 3, show that 48 is well within the range that luck allows.

25. A student plans to repeat the tuition study with 20 students in each group. (a) Use Figure 3 to estimate the power to detect a gap of $d = 0.32$. (b) How many students in each group would give 80% power? (c) The study with 20 in each group finds $p = 0.4$. What should the student conclude?

- 26.** Islam 2020, Table 3, holds sixteen tests. (a) Treating them as independent, what is the chance of at least one false alarm if nothing is going on? (b) What is the Bonferroni level for sixteen tests? (c) Which of the eight printed p -values in Table 13 pass it? (d) What does section 5.4 say about the row that does?
- 27.** Analyse the data of Exercise 11 wrongly, as two separate groups of eight. The SDs are 8.02 before and 9.13 after. (a) Find the SE of the difference with equation (2.1), and t . (b) Software gives $p = 0.462$. Compare this with the paired result, and explain the difference.

10 Solutions

Part A.

- 1.** (a) $H_0: \mu = 3.00$; $H_1: \mu \neq 3.00$, where μ is the department's average CGPA. (b) $H_0: \mu_{\text{male}} = \mu_{\text{female}}$; $H_1: \mu_{\text{male}} \neq \mu_{\text{female}}$. (c) With μ_d the average change, after minus before: $H_0: \mu_d = 0$; $H_1: \mu_d \neq 0$. The data are paired. (d) H_0 : all four faculty means are equal; H_1 : at least one differs. Every hypothesis is about the population. (Exercise 1.)
- 2.** (a) Wrong. The p -value is computed assuming H_0 is true; it is the chance of the data given the hypothesis, not of the hypothesis given the data. (b) Right: this is the definition. (c) Wrong. A p -value mixes the size of the effect with the sample size (section 2.8). (d) Right: $0.03 < 0.05$. (e) Wrong. The chance that a repeat reaches $p < 0.05$ depends on the true effect and the size of the repeat, neither of which the p -value gives. Section 1.13 describes a field in which most repeats of significant results failed. (Exercise 2.)
- 3.** (a) That the student did not copy: the presumption of innocence. (b) A Type I error would be finding a student guilty who did not copy. A Type II error would be clearing a student who did. (c) No. "Not proven" says the evidence was not strong enough to reject innocence. It is not a finding of innocence, just as not rejecting H_0 does not prove it. (Exercise 3.)
- 4.** (a) Concluding that the policy raised attendance when it did not. (b) Finding no evidence of a rise when the policy did raise attendance. (c) If the policy is expensive, a Type I error is costly: money is spent on a policy that does nothing. If it is cheap, a Type II error is the costlier one: a policy that works, at little cost, is dropped. The costs of the two errors should shape the choice of α and the size of the study. (Exercise 4.)
- 5.** (i) Zero lies outside, so the difference is significant at 0.05, favouring the new method. The gain is somewhere between 0.4 and 2.8 points, which may be small or moderate. (ii) Zero lies inside: not significant. The interval runs from a small disadvantage of 0.3 points to a gain of 1.9, so the result is inconclusive, not a finding of no difference. (iii) Zero lies outside: significant, but favouring the *old* method, since the new method scores lower. The loss could be as small as 0.1 points, which would be trivial, or as large as 2.5. (Exercise 5.)

6. (a) Paired: the same students, measured twice. (b) Separate: different students in each group. (c) Paired: two hands of the same student. (d) Paired: the same students in both subjects. (e) Paired: each husband is matched with his wife. (Exercise 6.)

7. With 12,000 students, even a tiny gap gives a tiny p -value (sections 2.8 and 2.9). A gap of 0.4 points out of 100 is far too small to matter, so “dramatically” confuses significance with importance. The survey also shows only an association: students who choose to use a study app may be more motivated to begin with, so the app need not cause any of the gap. Chapter 5 takes this up. (Exercise 7.)

Part B.

8. (a) $SE = 0.30 \div \sqrt{25} = 0.06$; $t = (3.12 - 3.00) \div 0.06 = 2.00$; $df = 24$. (b) 2.00 is less than 2.064, so the test does not reject at 0.05; software gives $p = 0.057$. (c) $3.12 \pm 2.064 \times 0.06 = 3.12 \pm 0.124$, from 2.996 to 3.244. The interval just includes 3.00, which agrees with not rejecting. A result this close to the line is best reported with its interval, not as a bare “not significant”. (Exercise 8.)

9. (a) $SE = 0.93 \div \sqrt{30} = 0.170$; $t = -0.20 \div 0.170 = -1.18$. (b) The size of t is less than 2.045, so the test does not reject; software gives $p = 0.249$. (c) With a quarter of the students, the SE is twice as large, because $\sqrt{120/30} = 2$. The same shortfall is then fewer standard errors from 7. The truth has not changed; the smaller study simply has less power to detect it. (Exercise 9.)

10. (a) $SE = \sqrt{8^2/16 + 10^2/20} = \sqrt{4 + 5} = 3$. (b) $t = (72 - 67) \div 3 = 1.67$. (c) $5 \pm 2.032 \times 3 = 5 \pm 6.10$, from -1.10 to 11.10 points. Zero is inside, so the test does not reject at 0.05; software gives $p = 0.105$. (Exercise 10.)

11. (a) The differences are 4, 3, 2, 5, 2, 5, 1 and 4. Their mean is $26 \div 8 = 3.25$. The squared deviations from 3.25 add to 15.5, so the SD is $\sqrt{15.5 \div 7} = 1.488$. (b) $SE = 1.488 \div \sqrt{8} = 0.526$; $t = 3.25 \div 0.526 = 6.18$ on 7 df. This is well beyond 2.365, so the test rejects; software gives $p < 0.001$. (c) $3.25 \pm 2.365 \times 0.526 = 3.25 \pm 1.24$, from 2.01 to 4.49 marks. $d_z = 3.25 \div 1.488 = 2.18$. Every one of the eight students improved. (Exercise 11.)

12. (a) The means are 7, 9 and 5; the overall mean is 7. (b) $SSB = 5 \times (0^2 + 2^2 + (-2)^2) = 40$. In each group the deviations from the group mean are $-2, 0, 2, -1$ and 1 , whose squares add to 10, so $SSW = 3 \times 10 = 30$. (c)

Source	SS	df	MS	F
Between methods	40	2	20.0	8.00
Within methods	30	12	2.5	
Total	70	14		

$F = 8.00$ is above 3.89, so the test rejects; software gives $p = 0.006$. (d) $\eta^2 = 40 \div 70 = 0.57$. The method explains a large share of the variation here, but with only five students per method the estimate is imprecise. (Exercise 12.)

13. $s_p^2 = (15 \times 64 + 19 \times 100) \div 34 = 2860 \div 34 = 84.12$, so $s_p = 9.17$ and $d = 5 \div 9.17 = 0.55$: about medium on Cohen's labels. There is no contradiction. The gap may well be medium-sized, but with 16 and 20 students the study cannot tell it from zero; its interval, -1.10 to 11.10 , covers everything from a small reverse gap to a large one. (Exercise 13.)

14. (a) $40 \times 0.05 = 2$. (b) $1 - 0.95^{40} = 87.1\%$. (c) $0.05 \div 40 = 0.00125$. (Exercise 14.)

15. (a) In order: 2 (Y) rank 1, 3 (X) rank 2, 4 (Y) rank 3, 5 (Y) rank 4, 6 (Y) rank 5, 7 (X) rank 6, 9 (X) rank 7. (b) Group X's ranks are 2, 6 and 7, adding to 15, so $W = 15 - 6 = 9$. (c) 3 beats 2; 7 beats all four Y values; 9 beats all four. That is $1 + 4 + 4 = 9$ pairs. (d) There are $3 \times 4 = 12$ pairs, and $9 \div 12 = 0.75$. (e) There are only 35 ways to split seven ranks into a group of 3 and a group of 4. The most extreme split, with X holding the three highest ranks or the three lowest, has a chance of $1/35$ each, so the smallest possible two-sided p is $2 \div 35 = 0.057$. With groups this small no result can reach 0.05. (Exercise 15.)

Part C.

16. (a) $H_0: \mu = 8$; $H_1: \mu \neq 8$, where μ is average weekend sleep. (b) $SE = 1.05 \div \sqrt{119} = 0.096$; $t = (7.87 - 8) \div 0.096 = -0.13 \div 0.096 = -1.35$ on 118 df. Software, from unrounded values, gives $t = -1.36$. (c) $p = 0.178$ is above 0.05, so the test does not reject. The interval, 7.68 to 8.06 hours, includes 8, which agrees. There is no evidence that average weekend sleep differs from eight hours. (Exercise 16.)

17. (a) $p = 0.420$ is above 0.05, so the test does not reject. (b) For example: "Tuition students reported slightly less self-study on average (-0.18 hours a day; 95% CI -0.62 to 0.26 ; $p = 0.420$). The data are compatible with no difference, and with any difference from -0.62 to 0.26 hours a day." (Exercise 17.)

18. $p = 0.020$ is above 0.01, so at $\alpha = 0.01$ the test does not reject. The 99% interval, 6.58 to 7.02 hours, includes 7, which agrees. The same data lead to different decisions at different levels, which is why α is fixed before the data are seen. (Exercise 18.)

19. The direction was chosen after looking at the data. Doing that doubles the false-alarm rate without saying so (section 1.10). The honest result is the two-sided $p = 0.020$. At 0.05 the decision happens to be the same, but reporting $p = 0.010$ overstates the evidence by a factor of two. (Exercise 19.)

20. $SE = 4.53 \times \sqrt{1/58 + 1/60} = 0.834$, and $t = 1.44 \div 0.834 = 1.73$. The Welch SE was 0.8321, almost the same, because the groups are nearly equal in size; the hand t matches the pooled t of Table 6. (Exercise 20.)

Part D.

21.**Row 1**

(a) Significant. (b) They agree: the interval excludes 1, and the rebuilt p is 0.020. (c) The odds are higher, but by anything from about 10% to nearly three times. And the direction of cause is unknown.

Row 2

(a) Not significant. (b) They agree: the interval includes 1, and the rebuilt p is 0.155. (c) Inconclusive. The data are compatible with 13% lower odds and with about two and a half times the odds. This is not evidence of no difference.

Row 3

(a) Not significant, only just. (b) They agree; rebuilt p 0.059. (c) The interval is very wide, from slightly lower odds to nearly nine times the odds. Not significant does not mean no effect; a wide interval like this usually signals a small group.

Row 4

(a) The printed p says significant. (b) They disagree: the interval includes 1, and the rebuilt p is 0.186 (section 5.2). (c) Trust neither until the reason is known. Since thirteen rows of the table rebuild correctly, the interval is the more likely to be right.

Row 5

(a) Not significant. (b) They agree; rebuilt p 0.602, within rounding of .601. (c) The interval runs from 46% lower odds to 42% higher odds, so the data cannot say whether exercise goes with less depression or more.

Row 6

(a) Significant. (b) No interval is given for F , but the Tukey intervals of Figure 10 agree: two of the three clear zero. (c) Modest: residence accounts for about 9.5% of the variation in sleep, and the gaps are about 31 and 44 minutes. **Simulated data.**

Row 7

(a) Not significant. (b) They agree: the interval includes 0. (c) Inconclusive, not negative. The study had only 41% power for a gap of this size.

Row 8

(a) Significant. (b) No interval is given; with 20,000 in each group it would be very narrow. (c) Not important: 3 minutes of sleep is too small to notice. Significant and trivial.

(Exercise 21.)

22. (a) One-sample t . (b) Welch t . (c) Paired t . (d) Mann–Whitney: skewed, and small groups. (e) One-way ANOVA. (f) Chi-square test (Chapter 5). (g) Kruskal–Wallis. (h) Paired t . (i) Mann–Whitney: a single Likert item is ordinal. (j) Correlation, Pearson's r or Spearman's ρ (Chapter 5). (k) Paired t : twins are matched pairs. (l) One-way ANOVA. (m) Regression (Chapter 6). (n) Wilcoxon signed-rank. (o) One-sample t . (p) One-way ANOVA. (q) Chi-square test (Chapter 5). (r) Welch t : unequal sizes and unequal spreads are exactly where Welch's version is needed. (s) Kruskal–Wallis. (t) Paired t , or Wilcoxon signed-rank if the differences are very skewed. (Exercise 22.)

Part E.

23. $\ln 2.886 = 1.060$; $\ln 8.669 = 2.160$; $\ln 0.961 = -0.040$. $SE = (2.160 - (-0.040)) \div 3.92 = 0.561$; $z = 1.060 \div 0.561 = 1.89$; $p = 2 \times 0.0294 = 0.059$. This agrees with the printed .059. The row is not significant, but the interval, 0.961 to 8.669, is compatible with slightly lower odds and with nearly nine times the odds. The paper cannot say whether age matters; it cannot say that it does not. The rebuilding is our arithmetic. (Exercise 23.)

24. (a) About 1% of 1,000, which is 10. (b) The SE of a proportion of 0.05 from 1,000 tries is $\sqrt{0.05 \times 0.95 \div 1000} = 0.0069$, about 0.7 percentage points. Two SEs either side of 5% run from about 3.6% to 6.4%, and 4.8% lies well inside. The shortfall from 50 is luck. (Exercise 24.)

25. (a) About 17%. (b) 155 in each group. (c) Very little. With power this low, a real gap of this size would be missed most of the time, so $p = 0.4$ is inconclusive. The student should report the estimate and its interval, and should have planned the sample size with a power calculation before collecting any data. (Exercise 25.)

26. (a) $1 - 0.95^{16} = 56.0\%$. The tests use the same students and are not truly independent, so this is a rough guide. (b) $0.05 \div 16 = 0.0031$. (c) Only .002, for living with family; lagging academically, at .020, does not pass. (d) It passes the interval check, but Chapter 6 finds that it contradicts the paper's own Table 2 counts. The one row that survives the correction is the row that should not be cited until the authors explain it. (Exercise 26.)

27. (a) $SE = \sqrt{8.02^2/8 + 9.13^2/8} = \sqrt{8.04 + 10.42} = 4.30$; $t = 3.25 \div 4.30 = 0.76$. (b) The paired test gave $p < 0.001$; the unpaired one gives $p = 0.462$. The students differ widely in ability, with SDs of 8 to 9 marks, and that variation swamps a gain of about 3 marks. Pairing subtracts each student's own baseline and leaves only the gain, which is consistent from student to student. (Exercise 27.)

II Our arithmetic and constructed numbers

Every number below was computed for these notes. None is printed by the papers cited, although the inputs from real papers are.

Value	How	Inputs from
Rebuilt p -values in Table 13	from each row's interval, equation (5.1)	Islam 2020, Table 3
SE 0.249, z 0.73	Example 5.1	Islam 2020, Table 3
The stray-zero reading	our guess at the three misprints, unconfirmed	Islam 2020, Table 3
$16 \times 0.05 = 0.8$; 56.0%; 0.0031	expected false alarms, the familywise rate and the Bonferroni level for sixteen tests	Islam 2020, Table 3
Odds ratio 0.77 (0.17 to 3.52)	crude odds ratio from the counts; worked in Chapter 6	Islam 2020, Table 2
2 to 22 minutes	how far the interval lies below seven hours	Watson 2015; the simulated class
Table 5; power in Figure 3	equation (1.2); R's <code>power.t.test</code>	none; the simulated gap
t and F critical values	standard tables, rounded to three decimals	none

Simulated, not real students: every figure from the class data, including the stress, sleep, screen-time and study-hour comparisons, the 1,000 shuffles of Figure 4, the two groups of 20,000 of section 2.9, and the other group sizes of Figure 6. All are generated by data/class-data.R from a fixed seed, and the class figures will be replaced by the real survey.

Constructed for teaching, and not from any study: the two panels of Figure 9; the eight screen-time values of Figure 11; the class of 30 in Exercise 9, which reuses the simulated mean and SD; and every scenario and number in Exercises 1 to 7, 8, 10 to 15, 22 and 27.

12 References and further reading

- 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. Licensed CC BY 4.0; free to read and share.
- 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. PMC4434546. doi:10.5665/sleep.4716. The PubMed abstract does not state the seven-hour figure, and the full text was not available to us; the figure is as summarised by the US Centers for Disease Control and Prevention.
- Wasserstein RL, Lazar NA. The ASA statement on p -values: context, process, and purpose. *The American Statistician* 2016;70(2):129–133. doi:10.1080/00031305.2016.1154108. Free to read. Not indexed in PubMed.

4. Open Science Collaboration. Estimating the reproducibility of psychological science. *Science* 2015;349(6251):aac4716. PMID 26315443. doi:10.1126/science.aac4716. The figures quoted are from the abstract.
5. Altman DG, Bland JM. Absence of evidence is not evidence of absence. *BMJ* 1995;311(7003):485. PMID 7647644. PMC2550545. doi:10.1136/bmj.311.7003.485. One page, free to read.
6. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum; 1988. A book, with no PMID. The benchmarks of section 2.7 are quoted as widely reported and were not checked against the book for these notes.
7. Class data: a **simulated** stand-in for the Stat.BD class survey, generated by data/class-data.R from the fixed seed 20261005. Not real students. No DOI or PMID.

13 For students in the Stat.BD course

Before the next session. Read this chapter, and work the Part A exercises (Exercises 1 to 7) and the activity answers of Exercise 21. Exercise 22, the twenty practice questions on choosing a test, is the best single preparation for choosing a test in practice. The rest are for practice at any point in the course. Bring a calculator.

What the next session does with this chapter. Every test in this chapter compared a number across groups. Session 5 turns to two categories at once. In the 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 that pattern be luck? None of this chapter's tests fits, because both variables are categories. The chi-square test answers it, with the same logic of a null hypothesis and a p -value. Session 5 also introduces correlation, and shows why an association is not a cause.

Questions this chapter left open, and where each is settled.

Question	Where it is settled
Is residence related to tuition?	Chapter 5, the chi-square test
What is the correlation r used in section 2.11?	Chapter 5, correlation
Does a significant association mean a cause?	Chapter 5
What is an odds ratio, and how is it estimated?	Chapter 6, logistic regression
Is the living-with-family row of Islam 2020 sound?	Chapter 6, a check against Table 2
How are Likert items combined into a scale?	Chapter 6
Running these tests in software	Session 7 onward