STAT5002 Introduction to Statistics

UNIVERSITY OF SYDNEY · POSTGRADUATE · SEMESTER 2, 2026

Full guide + ask Sia about STAT5002

EXAM REVISION SHEET · SIDE 1/2

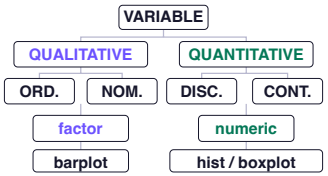
SIDE 1/2 DEFINITIONS → FORMULAS → DECISIONS → CHECKS | FINAL EXAM 60% · 2 HOURS · AI PROHIBITED | S2 2026 · course.json

Final Exam 60% | Quiz 20% | Assignment 10% | Weekly Online quizzes 10%

Final Exam 60% · Quiz 20% · Assignment 10% · Weekly Online quizzes 10%. Permitted materials: not confirmed for this offering in the course materials we hold — check the exam instructions on the LMS. *[BRIEF.md; course.json]*

1 · Start with the Variable

CH 1 · IDA



Classify the variable first; that determines the R storage check and the useful plot.

VARIABLES	FAST DISPLAY
1 qualitative	barplot(table(x))
2 qualitative	stacked / beside bars
1 quantitative	histogram, boxplot
2 quantitative	scatter plot
quantitative by group	boxplot(y ~ group)

Qualitative: nominal or ordered. *Quantitative*: counted (discrete) or measured (continuous). A numerical code can still be qualitative; inspect `str()`, then use `as.factor()` or `as.numeric()`. Replace missing-value codes by `NA`.

UNEQUAL HISTOGRAM INTERVALS

```
area = count ÷ total
height = area ÷ interval width
count = height × width × total
```

Area carries proportion; height carries crowding. With `right=F`, intervals include the left edge and exclude the right. Conditions act elementwise: `==, !=, &, |, !`; `sum(condition)` counts TRUE values.

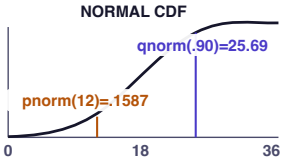
First-look sequence: establish the data source and research question → inspect dimensions and one row per subject → classify columns → repair storage and missing codes → select plots and numerical summaries → look for impossible values, outliers, skew and group differences. An anonymous ID is not an analysis variable.

R PATTERN	MEANING
table(x); barplot(...)	counts for a qualitative variable
hist(x, breaks=..., right=F)	left-closed classes
boxplot(y ~ group)	quantitative response across groups
sum(x>5 & x<=10)	count values in (5,10]
d[d$x>5 & d$g=="A",]	filter matching rows

TRAP + FIX *Digits do not make an ID quantitative: if averaging is meaningless, treat it as qualitative and use `as.factor()`. With `right=F`, a boundary enters the interval starting there; predict its bin first.*

3 · Normal Curve: Area or Cut-off?

CH 3



pnorm maps a cut-off to lower-tail area; qnorm maps lower-tail area back to a cut-off.

```
X ~ N(μ, σ²) · z=(x-μ)/σ
P(X<a)=pnorm(a, μ, σ)
P(X>a)=pnorm(a, μ, σ, lower.tail=F)
P(a<X<b)=pnorm(b,μ,σ)-pnorm(a,μ,σ)
Q_p=qnorm(p, μ, σ)=μ+σqnorm(p)
```

Second entry of N(μ,σ²) is variance; `pnorm/qnorm` expect SD. Total area = 1. Symmetry: P(Z<−a)=P(Z>a). The 68–95–99.7 rule places about 68%, 95%, 99.7% within 1, 2, 3 SDs.

QUESTION GIVES	QUESTION ASKS	USE
measurement	proportion / chance	pnorm
proportion	measurement / cut-off	qnorm

BAND / PERCENTILE	NORMAL SHORTCUT
μ±σ	about 68% inside; 16% in each tail
μ±2σ	about 95% inside; 2.5% each tail
μ±3σ	about 99.7% inside; 0.15% each tail
16th, 50th, 84th	approximately μ−σ, μ, μ+σ

Sketch before code: mark μ, the cut-off(s), and shade. One `pnorm` call returns one lower tail; an upper tail uses the complement, an interval subtracts two lower tails, and two outer tails must be added. For a standardised calculation, the area is unchanged when x becomes z.

4 · Correlation: Calculate, Then Distrust

CH 4

```
r = average(z_x z_y), -1≤r≤1 · cor(x,y)
```

Plot first. r measures only straight-line association. Its sign must match the visible slope. Adding a constant, multiplying by a positive constant, or swapping x/y leaves r unchanged; multiplying one variable by a negative constant flips its sign.

TRAP	CONSEQUENCE
one outlier	can dominate r
curved relation	r can be 0 despite exact dependence
group averages	ecological r tends to inflate
association	does not establish causation

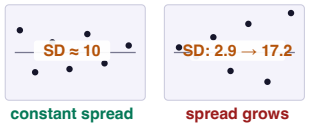
5 · Regression: Five Numbers → Line

CH 5

PREDICT Y FROM X

```
b = r·SD_y/SD_x · a = ȳ-b x̄ · ŷ = a + bx
e=y-ŷ · SSE=Σe² · SST=Σ(y-ȳ)²
R²=1-SSE/SST=r²
```

The line passes through (x̄,ȳ), minimises squared residuals, and moves only r SDs in y for each SD in x. Baseline prediction is always ȳ. Do not rearrange the y-on-x line to predict x; refit `lm(x ~ y)`.



An even residual band supports prediction; a fan shows changing spread and warns against it.

RESIDUAL PLOT	DECISION
random band around 0; equal strip spreads	linear fit, homoscedastic; prediction defensible
curve	line misses shape
fan / changing strip spread	heteroscedastic; do not trust prediction

CHECK *Residuals from the fitted line average zero. Positive residual = observed point above the line. R² is the share of y-variation removed relative to predicting ȳ.*

R OUTPUT DECODER

OBJECT / QUANTITY	READ AS
lm(y ~ x)	response first; predictor after ~
coef: (Intercept)	a, fitted y when x=0
coef: x	b, change in fitted y per x unit
fitted.values	$\hat{y}_i$ on the line
residuals	$y_i-\hat{y}_i$
cor(x,y)^2	R² for simple linear regression

Reverse prediction: y-on-x slope is rSD_y/SD_x; x-on-y slope is rSD_x/SD_y. Rearranging the first gives $SD_x/(rSD_y)$, which is different unless $|r|=1$. Extrapolation beyond the observed x-range also lacks data support.

12 · Chi-squared

CH 12

```
T=Σ(O-E)²/E ·
P=pchisq(T,df,lower.tail=F)
GOF: E_j=np_0j, df=k-1
independence: E_ij=row_i total×col_j total/n
df=(r-1)(c-1)
```

5b · Regression Answer Path

MARKS

1. Name response y and predictor x.
2. Compute b; its sign must equal r.
3. Compute a from the point of averages.
4. Substitute x; retain units.
5. Residual = observed − fitted; interpret sign.
6. Square r; write "% of variation in y explained".
7. Inspect residual pattern before endorsing prediction.

SSE versus SST: SST is the squared error from always predicting ȳ; SSE remains after using x. Thus 1−SSE/SST is improvement over the baseline. A large R^2 does not repair curvature, unequal spread, an influential point or a causal interpretation.

6 · Probability Rules

CH 6



Conditional probabilities are column heights; joint probabilities are the corresponding areas.

```
P(A^c)=1-P(A)
P(A|B)=P(A∩B)/P(B)
P(A∩B)=P(A)P(B|A)
P(A∪B)=P(A)+P(B)-P(A∩B)
```

CONDITION	MEANING	SHORTCUT
mutually exclusive	cannot both occur	P(A∩B)=0
independent	one does not change the other	P(A∩B)=P(A)P(B)

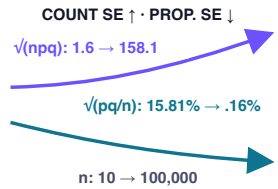
Exclusive positive-probability events are dependent. For "at least one", use 1 − P(none). Before multiplying, write the second chance as "given"; drop the condition only after independence is established. `sample(..., replace=FALSE)` changes the box after each draw; draws are dependent. Use `set.seed()` for reproducibility, put one experiment in a function, repeat with `replicate()`, and take `mean()` of logical results.

Set language: A∩B means both; A∪B means at least one; A^c means not A. Conditional probability restricts the denominator to B. Independence can be checked by P(A|B)=P(A) or P(A∩B)=P(A)P(B); one equality failure is enough to reject it.

PROMPT	FIRST MOVE
at least one success	1−P(no successes)

8 · Law of Averages + CLT

CH 8



As n grows, count SE rises as $\sqrt{n}$ while proportion SE falls as $1/\sqrt{n}$.

The sample average concentrates at μ because $\sigma/\sqrt{n}\to 0$. The sum's absolute error grows as $\sigma\sqrt{n}$; the proportion's error shrinks as $\sigma/\sqrt{n}$. The original ticket box does *not* become normal: the distribution of sums or averages does.

CLT FOR A SUM

```
z=(s-nμ)/(σ√n)
P(S≤s) ≈ pnorm(s, nμ, σ√n)
```

Approximation quality depends on box shape: a very skewed box needs more draws. Always identify sum versus average, write E and SE, then sketch the requested tail before calling `pnorm`.

Kerrich check: for 10,000 fair tosses, E(heads)=5,000 and SE(count)=0.5√10,000=50. The observed 5,067 is +67, or 1.34 SEs, from expectation. The count error is visibly large, but the proportion error is only 0.0067.

MULTIPLY N BY 10	EFFECT
SE of count	×√10 ≈ ×3.16
SE of proportion / average	÷√10 ≈ ÷3.16
shape of sums	usually closer to normal

9 · One Proportion: Interval + Z

CH 9

```
p̂=x/n · SE(p̂|p)=√[p(1-p)/n]
prediction interval (known p): p ± 1.96SE
Wilson CI:
binom.confint(x,n,method="wilson")
z=(p̂-p_0)/√[p_0(1-p_0)/n]
```

H_1	P-VALUE
$p<p_0$	pnorm(z)
$p>p_0$	pnorm(z, lower.tail=F)
$p\neq p_0$	2×pnorm(|z|, lower.tail=F)

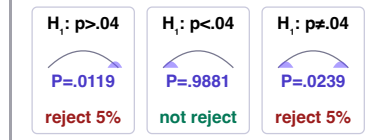
Wilson collects candidate p values whose prediction intervals contain $\hat{p}$; because SE depends on p, it is not centred exactly on $\hat{p}$. A 95% procedure captures the true p in 95% of repeated samples, not a 95% probability statement about this fixed interval.

10 · HATPC: Write the Whole Test

CH 9–12

1. **H**ypotheses: parameter and direction fixed before data.
2. **A**ssumptions: random / independent draws; distribution or large-sample condition; study design.
3. **T**est statistic: estimate minus null, divided by null SE.
4. **P**-value: assuming H_0, chance of evidence at least this strong in H_1's direction.
5. **C**onclusion: compare with α; reject or do not reject H_0; answer in context.

A large P-value does not prove or "accept" H_0. A two-sided level-α test and a (1−α) CI give the same decision about the null value.



The same z gives different P-values: choose the tail from H_1 before reading the data.

Tail logic: the statistic measures distance from H_0 in SEs. "Less" uses a lower tail; "greater" an upper tail; "different" both. The P-value is computed under H_0 and is not the probability that H_0 is true. α is the long-run false-alarm probability when H_0 is true.

11 · Choose the Mean Test

CH 10–11

DESIGN / INFORMATION	STATISTIC / R
one mean; σ known	Z=(x̄−μ₀)/(σ₀/√n)
one mean; σ unknown; near normal	T=(x̄−μ₀)/(s/√n), df=n−1; t.test(x,mu=)
2 independent; unequal spreads	Welch: t.test(x,y) default
2 independent; common SD justified	pooled: var.equal=T; df=m+n−2
same units twice / matched	differences d=x−y; paired=T; df=n−1
clear skew / doubtful normality	bootstrap T from centred surrogate box

TWO INDEPENDENT MEANS

```
SE(x̄-ȳ)=√(σ_x²/m+σ_y²/n)
pooled s_p=√([(m-1)s_x²+(n-1)s_y²]/(m+n-2))
T=(x̄-ȳ)/[s_p√(1/m+1/n)]
```

Student's t has fatter tails than N(0,1); use `pt/qt`, not `pnorm/qnorm`. QQ points near a line support a near-normal box. For paired data, argument order fixes difference sign.

Forward this to your tutorial group

V2 · 2026-10-25

Semester 2 formal examination period, from 16 November 2026 (the Exams Office publishes the exact date).

Check your exam instructions on the LMS for what you may bring; use this sheet to revise and self-test.

STAT5002 Introduction to Statistics

UNIVERSITY OF SYDNEY · POSTGRADUATE · SEMESTER 2, 2026

Full guide + ask Sia about STAT5002

EXAM REVISION SHEET · SIDE 2/2

SIDE 2/2 SELF-TEST DRILLS · ANSWER PATHS · JUDGEMENT CARDS · CLASSIC TRAPS

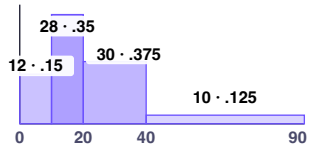
COVER → SOLVE → CHECK

Worked values from source/chapters.json

D1 · Summaries Drill CH 1–2 WORKED

200 parcels: intervals [0,1), [1,2), [2,5), [5,15); counts 40, 50, 80, 30.

INTERVAL	AREA	WIDTH	HEIGHT/KG
[0,1)	0.20	1	0.200
[1,2)	0.25	1	0.250
[2,5)	0.40	3	0.1333
[5,15)	0.15	10	0.015



Density histogram: area is proportion; height is proportion divided by interval width.

Answer path: areas add to 1. Most crowded = [1,2), tallest at 0.25/kg. Most parcels = [2,5), area 0.40. At least 2 kg = 0.40+0.15= 0.55.

TRAP *Comparing heights answers "most crowded", not "most subjects"; recover area as height × width.*

CENTRE-AND-SPREAD CHECKPOINT

Fares: 9, 11, 12, 12, 14, 15, 17, 30. $\bar{x}$=120/8=**15**; median=(12+14)/2=**13**. $\Sigma(x-15)^2$=300, so $s=\sqrt{300/7}$=**6.547**. z for 30=(30−15)/6.547=**2.29**.

IQR=15.5−11.75=3.75; UT=15.5+1.5(3.75)=**21.125**; 30 is an outlier. Mean>median and a right tail → report median 13, IQR 3.75 .

SELF-CHECK	REQUIRED STATEMENT
Σ deviations	must be 0
sample variance	300/7, not 300/8
z=2.29	30 is 2.29 SDs above mean
30>UT	flagged by 1.5IQR rule
right-skewed	prefer median + IQR

Reverse checks: data point = mean + SD×z, so 15+6.547(2.29)≈30. The lower threshold is 11.75−1.5(3.75)=6.125, so there is no low outlier. A flagged outlier is not automatically an error; investigate it before removal.

NORMAL-AREA CHECKPOINT

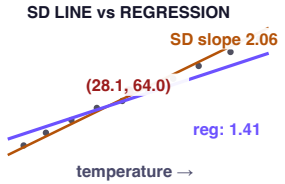
Bolt lengths follow N(40,0.5²) mm; reject below 39.2 or above 40.8. Limits standardise to −1.60 and +1.60. Each tail is 0.0548 by symmetry, so rejection probability is 0.0548+0.0548=0.1096 , about 11.0%.

```
pnorm(39.2,40,0.5) +
pnorm(40.8,40,0.5,lower.tail=F)
= pnorm(-1.6) +
pnorm(1.6,lower.tail=F)
```

D2 · Regression Drill CH 5 WORKED

Sleep x: $\bar{x}$=6.8 h, SD_x=1.1. **Reaction y:** $\bar{y}$=312 ms, SD_y=28. r=−0.55.

1. b=−0.55×28/1.1=**−14.0 ms/h**.
2. a=312−(−14)(6.8)=**407.2**.
3. Line: $\hat{y}$=407.2−14x.
4. x=5.5 → $\hat{y}$=407.2−77=**330.2 ms**.
5. R^2=(−0.55)²=**0.3025** → about 30% of reaction-time variation explained.
6. Observed y=340 → e=340−330.2=**+9.8 ms**: slower than predicted; point lies above line.



Both lines pass through the point of averages; regression slope is r times the SD-line slope.

CHECK *Slope sign matches r; line passes through (6.8,312); residual uses observed minus fitted; R^2 loses r's sign.*

Standard-unit audit: 5.5 h is (5.5−6.8)/1.1=−1.18 SDs from mean sleep. A regression prediction moves r times that many response SDs: (−0.55)(−1.18)=+0.65 SDs, or about 0.65×28=18.2 ms above 312 → 330.2 ms. This catches a swapped SD ratio.

Meaning of 30%: the line removes 30.25% of SST relative to the baseline $\bar{y}$=312; about 69.75% remains in residual variation. It does not mean 30% of people lie on the line, nor that sleep causes reaction time.

Judgement · Regression IF X → Y

IF YOU SEE...	DO...
predict x from y	refit lm(x ~ y)
curved scatter	do not summarise by r alone
fan residuals	flag heteroscedasticity; distrust prediction
one distant point	check influence on r and line

Also check the prediction lies within the observed x range. A good numerical fit can still be unsafe when extrapolating. A residual plot must show both random shape and roughly equal vertical spread; averaging residuals to zero is automatic and does not prove adequacy.

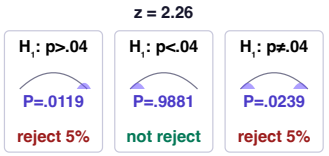
CORRELATION ARITHMETIC CHECKPOINT

D3 · One-Proportion Z Drill CH 9 WORKED

Clinic claim p_0=0.70; suspicion p<0.70. Random n=90 visits; x=54, so $\hat{p}$=0.600. Test at α=0.05.

HATPC

```
H₀:p=0.70 · H₁:p<0.70
SE₀=√(0.7×0.3/90)=0.0483
z=(0.600-0.700)/0.0483=-2.070
P=pnorm(-2.070)=0.0192
```



The same z gives different P-values: choose the tail from H_1 before reading the data.

Conclusion: 0.0192<0.05 → reject H_0. The data are evidence that fewer than 70% of patients are seen within 15 minutes.

TRAP *The null value p_0 belongs in the test SE. H_1 fixes the lower tail before the data are seen.*

Assumption line: the 90 visits are treated as random independent draws from a large population of visits; under H_0 the expected successes and failures are 63 and 27, so the normal approximation is reasonable. Name the population parameter before arithmetic.

Decision cross-check: a negative z supports only a lower alternative. If the question had asked "different", the two-sided P-value would be 2×0.0192=0.0384; if it had asked "greater", the upper-tail value would be large. Same data, different pre-specified question.

Fast Tail Decoder CH 3, 8–10

PHRASE	SHADE / R PATTERN
below, less, shorter	left tail; pnorm(z) / pt(t,df)
above, greater, longer	right tail; lower.tail=F
different, either direction	two tails; double matching tail
between a and b	lower tail at b minus lower tail at a
outside limits	add two separate tails

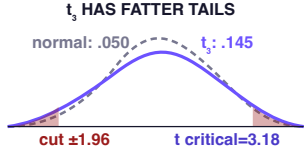
Never choose the tail from the observed sign. Choose from the research question and H_1.

INTERVAL AND TEST ARE NOT INTERCHANGEABLE INPUTS

D4 · Mean-Inference Drills CH 10–11 WORKED

Commute: n=16, $\bar{x}$=27.9 min, s=4.4 min; test μ>25. QQ plot roughly straight.

```
H₀:μ=25 · H₁:μ>25
SE=s/√n=4.4/4=1.1
t=(27.9-25)/1.1=2.636 · df=15
P=pt(2.636,15,lower.tail=F)=0.0093
```



Student's t has fatter tails: using pnorm for a small-sample T-statistic understates the P-value.

0.0093<0.05 → reject H_0; evidence mean commute exceeds 25 min. For a 95% CI, qt(0.975,15)=2.131:

```
27.9 ± 2.131×1.1 = (25.56, 30.24)
min
```

TRAP *Because σ is estimated by s, use t with n−1 df. A one-sided test still pairs here with the supplied two-sided 95% interval.*

Output route: `t.test(x, mu=25, alternative="greater")` should report t≈2.636, df=15 and the one-sided P-value. The ordinary 95% two-sided interval uses `qt(0.975,15)`; qt, not qnorm, supplies the multiplier because s is estimated.

CI logic: the null value 25 lies below (25.56,30.24), consistent with rejection by a two-sided 5% test. The one-sided P-value is smaller because only the upper tail counts. Always state minutes on the endpoints and the population mean being estimated.

Mean-Test Judgement Cards CH 10–11

IF YOU SEE...	DO...
population σ known	Z-test; normal critical value
σ estimated, one sample	one-sample T; df=n−1
2 independent groups; spreads differ	Welch default
same person/item twice	paired: test differences
skewed small sample	bootstrap T logic

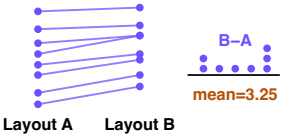
Bootstrap: use sample as surrogate box; resample; store T-statistics. If lower/upper T quantiles are ℓ,u, equal-tailed CI is [$\bar{x}$−uSE, $\bar{x}$−ℓSE]; skew can make it lopsided.

R FUNCTION	WHAT IT RETURNS
pt(t,df)	lower-tail t area
qt(p,df)	t cut-off with p below
t.test(x,mu=)	one-sample test + CI
qqnorm(x); qqline(x)	visual normality check

Paired-Design Branch INSIDE D4 · CH 11

Ten baristas use old and new machines. Differences old−new: **3, 3, 1, 5, 1, 5, 1, 3, 4, 1**.

```
d̄=2.70 s · s_d=1.636 · SE=1.636/
√10=0.517
T=2.70/0.517=5.218 · df=9
P=2×pt(5.218,9,lower.tail=F)=0.0006
CI=2.70±2.262×0.517=(1.53,3.87) s
```



Pairing removes between-person spread and tests the within-person differences.

Reject H_0: mean difference 0. Because old−new is positive, the new machine is faster by about 2.7 s on average.

TRAP *The unpaired method counts barista-to-barista variation as noise. Write differences first; swapping x,y flips the sign and reverses a one-sided alternative.*

Assumptions attach to differences: the ten baristas are the independent units; the distribution to inspect is the ten old−new differences, not the two raw columns separately. A dot plot shows no extreme outlier, so a one-sample T procedure on d is the intended model.

Interval conclusion: every plausible old−new mean in (1.53,3.87) is positive, so the new machine is faster. For `t.test(old,new,paired=T)`, R subtracts first argument minus second; reverse them and the interval becomes negative.

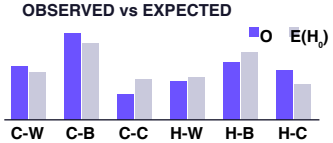
Independent Two-Sample Checklist CH 11

1. Confirm units in group X are independent of units in Y.
2. State parameter $\mu_x-\mu_y$ and argument order.
3. Inspect boxplots / spreads and distribution shape.
4. Use Welch by default; pool only if common SD is justified.
5. Read Welch's fractional df from R; do not force m+n−2.
6. For one-sided work, set `alternative=` to match the written H_1.

CLASSICAL / WELCH / PAIRED SEPARATOR

D5 · Chi-squared Independence Drill CH 12 WORKED

YEAR	R DAILY	NOT DAILY	TOTAL
First	18	42	60
Second	30	40	70
Third+	37	33	70
TOTAL	**85**	**115**	**200**



Expected cell counts come from row total × column total ÷ grand total.

H_0: year and daily R use independent. Expected counts: First 25.50,34.50; Second 29.75,40.25; Third+ 29.75,40.25. All exceed 5.

```
terms:
2.206+1.630+0.002+0.002+1.767+1.306
T=6.913 · df=(3-1)(2-1)=2
P=pchisq(6.913,2,lower.tail=F)=0.0315
```

0.0315<0.05 → reject H_0: evidence daily R use depends on year level; first-year daily use is lower than expected.

Arithmetic audit: expected rows sum to 60, 70, 70 and expected columns to 85, 115. Each contribution is non-negative; the largest gaps drive T. The 5% cut-off qchisq(0.95,2)=5.991, and 6.913 is beyond it, matching the P-value decision.

Goodness-of-fit branch: when one categorical variable is compared with fully stated p_0, use E=np_0 and df=k−1. For a two-way table, the margins are estimated from data and df=(r−1)(c−1). Small expected counts trigger simulation, not silent use of the curve.

Classic Traps + Exam Morning ALL CHAPTERS

TRAP	CORRECTION
density height = share	share is area
N(μ,σ²): use second entry as SD	take its square root
r=0 means no relation	only no linear relation
$\hat{y}-y$ residual	e=y−$\hat{y}$
exclusive = independent	positive-probability exclusive events are dependent
add SEs	add squared SEs
large P proves H_0	do not reject H_0
pairing ignored	analyse differences
χ^2 has two tails	upper tail only

1. Recite the method chooser and HATPC.
2. Redraw the four tail shapes; label `pnorm/pt`.
3. Check formula denominators, df and units.

Semester 2 formal examination period, from 16 November 2026 (the Exams Office publishes the exact date).

Check your exam instructions on the LMS for what you may bring; use this sheet to revise and self-test.