DATA2002 Data Analytics: Learning from Data

THE UNIVERSITY OF SYDNEY · SEMESTER 2, 2026

Full guide + ask Sia about DATA2002

SIDE 1 · RECALL — CLOSED-BOOK MEMORY SHEET · definitions → formulas → decisions — LEARN · RECITE · CHECK

1 · Exam map COURSE DATA · SAMPLE COVER

Final 60 | P1R | P2 10 | FE 9 | P1E | Git

Weights: Final Exam 60% · Group Project Report (Project 1) 10% · Individual Project Report (Project 2) 10% · Individual Feedback Exercise 9% · Group Project Presentation (Project 1) 5% · Individual GitHub Contribution (Project 1) 5% · Individual Early Feedback Task 1%.

PERMITTED MATERIALS — EXACT RULE

"Exam Conditions: This is a closed-book examination — no material permitted."
DATA2002 Semester 2 2026 exam sample paper, cover instructions (PDF_2026S2_DATA2x02_Exam_Sample_DW_V3)

Final: 2 hours, 60%. Released sample: no calculator; 30 equally weighted MCQs = ¾, one extended answer = ¼; key procedures follow the questions. Learn recognition, expressions, assumptions and interpretation.

ROUTE BEFORE CALCULATING

response? → groups? → paired?

count → χ^2 / exact; label → classify

measure → t/rank; none → PCA/cluster

Four questions from the guide: response, groups, pairing/design, goal.

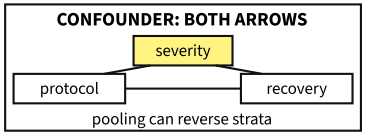
2 · Six-slot answer CH 3

WRITE	CHECK
1 level α	chosen before data
2 H_0 / H_1	population parameters
3 assumptions	design + distribution
4 statistic	value + null law + df
5 p-value	correct tail
6 conclusion	variables + population

Reject when $p \le \alpha$. Otherwise "insufficient evidence against H_0", never "H_0 is true".

EXAM MOVE *Write slots 1–3 before arithmetic; finish with variables and population.*

CONFOUNDER: BOTH ARROWS



Source redraw: a third variable must connect to both sides; compare strata before pooling.

4 · χ^2 counts CH 3 · CH 5

GOODNESS OF FIT

```
eᵢ = n pᵢ₀
χ² = Σ (yᵢ-eᵢ)²/eᵢ
df = k-1-q, q = parameters estimated from these data
```

R × C TABLE

```
eᵢⱼ = (row totalᵢ)(column totalⱼ)/n
χ² = ΣᵢΣⱼ (yᵢⱼ-eᵢⱼ)²/eᵢⱼ
df = (r-1)(c-1)
```

Upper tail only: χ^2 is non-negative distance. Check expected, not observed, counts. Every expected count ≥ 5; otherwise pool adjacent meaningful categories and recount df, or use exact/simulation.

SAMPLING SENTENCE	NAME / CONCLUSION
fixed n per group	homogeneity: distributions match?
one sample, two labels	independence: variables associated?

For a small 2×2 table, Fisher's exact test enumerates tables with fixed margins; no approximation means no expected-count condition. Larger sparse tables: simulated null with margins fixed.

5 · Two-by-two reflexes CH 4

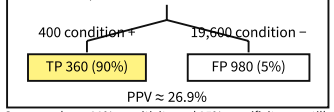
	TRUTH +	TRUTH −
TEST +	a = TP	b = FP
TEST −	c = FN	d = TN

CONDITION ON THE NAMED MARGIN

```
Prev=(a+c)/n · Accuracy=(a+d)/n
Sens=a/(a+c) · Spec=d/(b+d)
FNR=c/(a+c)=1-Sens ·
FPR=b/(b+d)=1-Spec
PPV=a/(a+b) · NPV=d/(c+d)
```

Sensitivity/specificity condition on truth and belong to the instrument. Predictive values condition on result and change with prevalence. For rare conditions, accuracy can flatter a useless classifier: compare with always predicting the majority class.

20,000 PEOPLE · PREV 2%



Source redraw: 90% sensitivity and 95% specificity can still yield low PPV at a 2% base rate.

6 · Risk versus odds CH 4

	OUTCOME +	OUTCOME −
EXPOSED	a	b
UNEXPOSED	c	d

MEASUREMENT → WHICH t?

one sample | two paired | independent

$\bar{x}$ vs μ_0 | $d_i = y_i - x_i$ | Welch default

normality applies to values / differences

Source redraw: relationship between observations decides the denominator and df.

8 · Pairing, intervals, power CH 6–7

PAIRED → ONE SAMPLE OF DIFFERENCES

```
dᵢ=yᵢ-xᵢ; t=(d̄-Δ₀)/(s_d/√n), df=n-1
n = pairs, not measurements
```

Pairing removes between-unit variation. Ignoring it can enlarge the denominator and p-value. Normal quantile plot the *differences*, not the two columns. Large samples make normality less consequential; small samples make it real.

TEST–INTERVAL IDENTITY

```
two-sided reject μ₀ ⇔ |x̄-μ₀| > c·s/√n
CI = x̄ ± c·s/√n
power(μ)=Pμ(reject H₀)=1-β(μ)
```

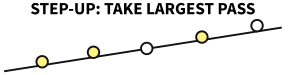
MOVE	POWER	COST / MEANING
|effect| ↑	↑	not controlled
n ↑	↑	time/money; only no statistical concession
spread ↓	↑	usually not controlled
α ↓	↓	Type I ↓, Type II ↑

Coverage belongs to the repeated procedure, not the realised interval. A 95% interval does not put a 95% probability on a fixed parameter. Sample-size answers always round up.

9 · Rank ladder CH 8

PROCEDURE	KEEPS	USE
t	values	normal model defensible
signed rank	sign + magnitude rank	paired, symmetric shape
sign	direction only	paired; strongest immunity
rank sum	pooled ranks	two independent groups

STEP-UP: TAKE LARGEST PASS



Source redraw: do not test each ordered p-value independently.

RANK TRAP *Discard zero differences before counting or ranking. Ties receive average ranks and force an approximation. Rank tests have no df: null laws are indexed by the usable sample size.*

10 · Resample correctly CH 9

QUESTION	PERMUTATION	BOOTSTRAP
Purpose	test / p-value	estimate / interval
Action	shuffle labels under H_0	sample with replacement
Size	preserve group sizes	same n as original
Assumption	exchangeability	sample represents population

Any statistic may be permuted; using a t statistic does *not* invoke a t distribution. Paired design: permute signs, not labels. Enumerate all reallocations when feasible for an exact p-value; otherwise simulate and report repetitions.

PERCENTILE BOOTSTRAP

```
repeat: draw n values with replacement
→ compute statistic
95% endpoints = 2.5th and 97.5th percentiles
```

Report resamples, seed and interval method. Resampling estimates sampling variability; it cannot repair selection bias or manufacture information in a small sample.

SAME DATA · DIFFERENT QUESTION

PERMUTE labels null distribution → p | RESAMPLE values sampling spread → CI

Source redraw: shuffling answers a test; replacement resampling answers uncertainty.

12 · ANOVA machinery CH 11–12

```
SS_total = SS_treatment + SS_residual
df: N-1 = (g-1)+(N-g)
MS = SS/df; F = MS_treatment/MS_residual
F ~ F(g-1, N-g) under H₀
```

One-way H_0: all group means equal; upper tail only. Rejecting says at least one differs, not which. Residual QQ plot checks normality; residuals vs fitted checks equal spread. Use the pooled residual variance for follow-ups.

FOLLOW-UP	VALID SCOPE
Bonferroni	contrasts fixed in advance
Tukey	all pairwise; exact if equal n

SS + df: TWO CHECKS

treatment | residual

SS parts add · df parts add

mean squares do not add

Source redraw: fill df first; sums of squares and df each reconcile to totals.

8b · Critical-value notation CH 7

TN(Γ) = LOWER-TAIL Γ QUANTILE	REJECT WHEN
H_1: $\mu<\mu_0$	$T \le tv(\alpha)$
H_1: $\mu>\mu_0$	$T \ge -tv(\alpha)$
H_1: $\mu\neq\mu_0$	$\lvert T\rvert \ge -tv(\alpha/2)$

For small α the unit's tv(α) is negative. Read the definition before choosing the sign.

13 · Second factor CH 12

DESIGN	STRUCTURE	CAN TEST INTERACTION?
blocked	one observation/cell; nuisance factor	no: interaction becomes error
two-factor replicated	>1 observation/cell	yes: genuine within-cell residual

```
df A=a-1; df B=b-1; df A×B=(a-1)(b-1)
df residual=ab(r-1); df total=abr-1
```

Read interaction first. Significant A×B → main-effect averages can describe neither condition; report simple effects. Parallel interaction-plot lines mean no interaction. Replication means repetition *within* cells.

14 · Linear model CH 13

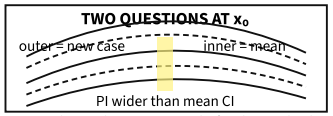
```
Yᵢ=β₀+β₁xᵢ+εᵢ; fitted line minimises Σeᵢ²
s²=SSE/(n-p-1); t=bⱼ/SE(bⱼ)
R²=1-SSE/SST
```

ASSUMPTION	CHECK / FAILURE
linear mean	residuals vs fitted: curve
constant spread	residuals vs fitted: funnel
normal errors	residual QQ: curvature/tail point
independence	residuals in collection order

Fit → inspect residuals → only then infer. Slope: change in *mean* response per unit x; in multiple regression add holding other variables constant. Intercept is interpretable only if x=0 is meaningful and near the observed range.

Mean-response CI answers "average at x_0". **Prediction interval** answers "new case at x_0" and is always wider because it also carries irreducible individual scatter.

TWO QUESTIONS AT x_0



Source redraw: only uncertainty in the fitted mean shrinks away with more data.

INTERVAL WORDING *Mean at x_0 → confidence interval; new case at x_0 → prediction interval. More data shrink fitted-mean uncertainty, not residual scatter, so prediction width has a floor.*

16 · Classification CH 15

```
η=β₀+β₁x₁+⋯; odds=e^η; p=e^η/(1+e^η)
log-odds neutral 0 → odds ratio neutral 1 → probability 0…1
```

Logistic coefficient inference uses **z, no df**. Thresholding probabilities is a separate decision. Printed confusion matrices put **predictions in rows, truth in columns**; transposition leaves accuracy unchanged but swaps sensitivity/specificity.

METHOD	HIGH-VALUE RULE
Tree	larger complexity constant demands more improvement → smaller tree
Forest	more trees, not deeper; sample predictors to decorrelate trees
k-nearest neighbours	k=1 is perfect in-sample; standardise distances; choose k out-of-sample

Before praising accuracy, compute majority-class accuracy. Lowering the classification threshold trades specificity for sensitivity.

ONE EFFECT · THREE SCALES

log-odds β, null 0 | odds e^{β}, null 1 | probability 0…1

Source redraw: never compare an odds ratio with zero.

17 · PCA and clustering CH 16

QUESTION	PCA	CLUSTERING
Works on	columns	rows
Object	variance directions	distance groups
Interpret	loadings name; scores locate	cut dendrogram / nearest centre

```
PC₁=a₁'x maximises variance subject to a₁'a₁=1
standardised data: Σ eigenvalues = p
maximum components = min(n-1,p)
```

Standardise by default: units otherwise dominate variance and distance. Keep components by stated cumulative threshold, elbow, or λ>1 convention; these may disagree. Hierarchical: no k first, deterministic. Partitioning: choose k first, fix seed, assigns new rows to centres.

IS THERE A TARGET?

YES: predict, hold data out | NO: structure, argue interpretation

Source redraw: unsupervised output is not a prediction score in disguise.

CLASSIFIER TRAP *Accuracy on fitting data is one minus resubstitution error and is optimistic. For kNN, k=1 is a degenerate in-sample optimum; standardise first, or the predictor with the largest numerical scale controls distance.*

Forward this to your tutorial group

V2 · 2026-10-24

Your exam is closed book: nothing printed goes in. Use this sheet to revise and self-test, not to carry.

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

DATA2002

Self-test drills · answer paths · classic traps

DATA ANALYTICS: LEARNING FROM DATA · THE UNIVERSITY OF SYDNEY

Full guide + ask Sia about DATA2002

SIDE 2 · DRILL — WORK THE PATH · route → expression → decision → sentence — **ANSWERS INCLUDED**

18 · Worked 1: base rate

CH 4 EXAMPLE

Prompt. Sensitivity 90%, specificity 95%, prevalence 2%, N=20,000. What fraction of positive results are correct?

1. Condition+: 0.02×20,000=400; condition−=19,600.
2. TP=0.90×400=360; FP=0.05×19,600=980.
3. PPV=360/(360+980)=360/1,340≈ 26.9%.

Sentence. High sensitivity and specificity do not guarantee high predictive value when the condition is rare; false positives from the large healthy group dominate.

READ THE DENOMINATOR

sens = TP / truth+ | PPV = TP / test+

360 / 400 = 90% | 360 / 1,340

same TP · different denominator

Diagnostic labels grade the test result first: true/false, then positive/negative.

19 · Denominator speed drill

CH 4

WORDING	DENOMINATOR	MEASURE
among condition+	TP+FN	sensitivity
among condition−	FP+TN	specificity
among test+	TP+FP	PPV
among test−	FN+TN	NPV
risk in exposed / unexposed	row totals	RR

COVER-AND-SAY *Which word restricts the population? That word fixes the denominator.*

20 · Study-design judgement cards

CH 2

If users choose a programme → observational; report association; name a confounder with both arrows. **If access is randomised →** causal comparison is licensed in expectation.

If millions reply from a biased frame → precision rises, accuracy does not. **If rates reverse after pooling →** compare subgroup mix and outcome rate: Simpson's paradox needs both.

If a third variable differs only by exposure → imbalance, not yet a confounder. It must also relate to outcome and not be downstream of exposure.

REPAIR THE CLAIM *Controlling for a measured confounder splits the sample but does not randomise it: unmeasured confounders remain, cells shrink and precision falls. Selection bias is a frame failure before contact; non-response bias occurs after invitation. More n fixes neither.*

21 · Worked 2: retrospective

CH 4 EXAMPLE

Prompt. Recruit 200 cases and 200 controls; 30% of cases and 15% of controls report exposure. Can risk be said to double?

	EXPOSED	NOT EXPOSED
Cases	60	140
Controls	30	170

1. Outcome totals were chosen: retrospective case–control design.
2. Risks of outcome are design constants; **RR unavailable**.
3. OR=(60×170)/(140×30)=10,200/4,200= 2.43.
4. Say odds of exposure were higher among cases; recall bias threatens the comparison. Do not say risk doubled.

WHICH MARGIN WAS FIXED?

start at exposure — RR + OR | start at outcome — OR only

design first · arithmetic second

Risk requires observed outcome proportions; an odds ratio survives fixed outcome totals.

22 · χ^2 repair drill

CH 3–5

Printed: χ^2=11.34, df=5, p=.045. Expected probabilities used one parameter estimated from these data.

1. Printout's df=k−1=5 → k=6 categories.
2. Correct df=6−1−1= 4.
3. Same χ^2 on smaller df has a smaller upper-tail p-value; rejection at 5% remains.

If expected counts include 3.1 and 4.2 in adjacent categories: pool before testing, recount k, then recompute df. Do not check observed counts.

IF YOU SEE	DO
small expected in 2×2	Fisher exact; perhaps continuity correction if borderline
small expected in larger table	pool meaningfully or simulate fixed margins
one sample cross-classified	independence
fixed group sample sizes	homogeneity

22b · Count-table self-checks

CH 3 · CH 5

CHECKPOINT	MUST EQUAL / IMPLY
GOF expected totals	$\Sigma e_i = n$
GOF raw gaps	$\Sigma(y_i - e_i) = 0$
r×c expected rows	their observed row totals

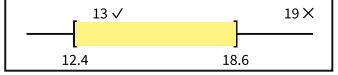
23 · Worked 3: interval logic

CH 7 EXAMPLE

Given: 95% CI for μ is 12.4 to 18.6.

1. Two-sided H_0:μ=19 at α=.05 → 19 is outside → reject.
2. H_0:μ=13 → 13 is inside → **do not reject**.
3. No new calculation: the interval is exactly the set of values not rejected by the matching two-sided test.
4. Wrong: "95% probability μ lies here." Correct: repeated intervals built this way cover fixed μ about 95% of the time.

CI = VALUES NOT REJECTED



Source redraw: test and interval at the same level cannot disagree.

24 · t / rank judgement cards

CH 6–8

If the same 18 trainees use two interfaces → paired; form 18 differences; df=17. Treating 36 times as independent gives df=34 and keeps between-person variation, usually making p too large.

If 9 differences include one zero → sign and signed-rank n=8. With six positive non-zero differences, sign statistic=6; $W_+ + W_- = 8\times9/2 =$ 36.

If rank output disagrees with hand statistic → check convention: software may print W_+ rather than min(W_+,W_-), or subtract the minimum possible rank sum. The p-value can match.

If normality fails → do not shop across tests. Choose rank or permutation from the design and diagnostic before viewing p-values. For paired normality, inspect differences.

25 · Power directions

CH 7

CHANGE	A	B	POWER
α .05 → .01	↓	↑	↓
n rises	same	↓	↑
spread rises	same	↑	↓
true gap rises	same	↓	↑

One-sided versus two-sided comes from the question *before* data. Choosing direction after seeing the sign silently inflates Type I error.

POWER-PLAN TRAP *A target power calculation relies on an assumed population spread from pilot or published data; if the true spread is larger, achieved power is lower. Round required n up. Retrospective power using the observed effect adds nothing.*

26 · Worked 4: ANOVA table

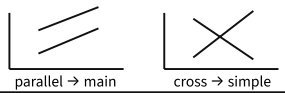
CH 11 EXAMPLE

Given: 4 groups × 7 observations; SS_treat=96.0; SS_total=312.0.

SOURCE	SS	DF	MS
Treatment	96	4−1=3	32
Residual	312−96=216	28−4=24	9
Total	312	27	—

F=32/9 on (3,24) df. Checks: 96+216=312 and 3+24=27. Mean squares do not add. A significant F says at least one mean differs; follow-up comparisons name where.

READ INTERACTION FIRST



Source redraw: interaction changes what a main-effect average can mean.

27 · Multi-test drill

CH 10 EXAMPLE

Eight independent null tests at α=.05:

```
P(any false +)=1-.95⁸=.337
Bonferroni threshold=.05/8=.00625
corrected FWER=1-.99375⁸ < .05
```

The strict inequality is the point: Bonferroni bounds FWER and is conservative. Do not claim it sets the realised rate exactly to .05.

Step-up trap: sort; find the *largest* rank meeting $p_{(i)} \le i\alpha/m$; take it and every smaller rank, even if an earlier point sits above its own threshold.

28 · Two-factor df drill

CH 12 EXAMPLE

3 levels × 4 levels × 2 replicates → N=24.

```
df A=2 · df B=3 · df A×B=6
df residual=24-12=12 · df total=23
check: 2+3+6+12=23
```

With one observation per cell, N=12 and model df consume all 11 total df: no residual variance, so no F ratios. More combinations do not substitute for repetition within cells.

FAILURE	RESPONSE
interaction significant	report factor effect within each level
blocked design	block row is housekeeping, not causal finding
rank-based blocked analysis	rank within each block

POST-HOC ROUTE *Match method to family: Bonferroni for fixed-in-advance contrasts; Tukey for all pairwise (exact at equal n); Scheffe for any contrast chosen after seeing means. Choosing whichever gives smallest p is data snooping.*

29 · Worked 5: eigenvalues

CH 16 EXAMPLE

Six standardised variables: λ = 3.10, 1.42, 0.71, 0.44, 0.20, 0.13.

1. Sum=6, because standardising gives total variance p=6.
2. First two share=(3.10+1.42)/6=4.52/6= 75.3%.
3. Cumulative: 51.7%, 75.3%, 87.2%, 94.5% → 90% rule keeps **4**.
4. Elbow after the steep early drop could keep about 2; different conventions may honestly disagree.

EIGENVALUES · TOTAL = 6



Source redraw: the highlighted first two explain 75.3%, not the 90% target.

30 · Classifier judgement card

CH 15 EXAMPLE

Reported: accuracy 94%; dataset 95% negative; sensitivity .02; specificity .99.

1. Always-negative baseline accuracy= 95%; model is worse than doing nothing.
2. Specificity .99 confirms it nearly always says negative.
3. Sensitivity .02 is decisive: it catches 2% of events.
4. If misses cost more, lower threshold to trade specificity for sensitivity; report probabilities so users can choose.

Logistic output: coefficient β is log-odds; e^{β} is an odds ratio. Use z, no df. Carry full precision through η → odds → probability.

NO RESPONSE LEAKAGE

train folds → select + tune → hold-out

repeat every fold → pool errors

Source redraw: selection once on all data leaves optimism untouched.

31b · Structure without a target

CH 16

OUTPUT CLUE	INTERPRETATION
one value per variable/component	loading; use signs and magnitudes to name direction
one value per observation/component	score; locate or plot observations
all similar positive loadings	overall size direction

32 · Method chooser

ALL CHAPTERS

SEE	ASK	ROUTE
one categorical count	stated/fitted proportions?	χ^2 GOF
two categorical variables	one sample or fixed groups?	independence / homogeneity
small 2×2 expected	fixed margins?	Fisher exact
one measurement group	normal mean?	one-sample t
two measurements	matched one-to-one?	paired t / Welch
non-normal paired	symmetry defensible?	signed rank / sign
non-normal independent	rank or shuffle?	rank sum / permutation
>2 means	one or two factors?	ANOVA
numeric target	explain or predict?	linear model + validation
binary target	probabilities or labels?	logistic/tree/kNN
no target	columns or rows?	PCA / clustering

ONE EXTENDED-ANSWER PIPELINE

design → method → working → sentence

assumptions → statistic + df → p / CI

then name variables, population, limits

never stop at "reject"

Answer structure earns marks even when the final arithmetic is unfinished.

33 · Classic reversals

HIGH-FREQUENCY TRAPS

WRONG REFLEX	CORRECT RECALL
large p proves H_0	insufficient evidence against
check observed counts	check expected counts
ratio interval vs 0	compare with 1
paired n = measurements	n = pairs
rank test has df	indexed by sample size
smaller α raises power	smaller α lowers power
F test names groups	only "not all equal"
read main effects first	interaction first
higher IC is better	lower is better
bigger tree constant	smaller tree

COVER → SAY → WRITE → CHECK

route — rule — work — claim

Retrieval practice: the check is useful only after an unaided attempt.

Forward this to your tutorial group

V2 · 2026-10-24

Your exam is closed book: nothing printed goes in. Use this sheet to revise and self-test, not to carry.

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