ETC1000 Business and Economic Statistics

MONASH UNIVERSITY · UNDERGRADUATE LEVEL 1 · SEMESTER 2, 2026

Full guide + ask Sia about ETC1000

SIDE 1/2 DEFINITIONS → FORMULAS → DECISIONS → DEFENSIBLE SENTENCES **EXAM REVISION SHEET** Examination 40% · end-of-semester period

0 · Exam Map COURSE.JSON · BRIEF

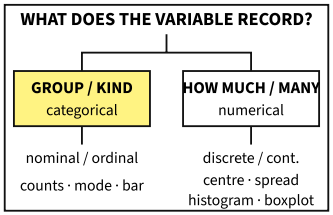
Exercises 20 Tests 20 Project 20 Examination 40

Exercises, completed during your allocated tutorial	20%
In-Class Tests, two tests worth 10 per cent each	20%
Group Project	20%
Examination	**40%**

Permitted materials: not confirmed for this offering in the course materials we hold — check the exam instructions on the LMS. *BRIEF.md*

★ The unit's repeated scoring move is **quantity + sentence**: compute the right object, name its denominator or scale, then state what it establishes and what it does not.

1 · Variable → Display → Summary CH 1–2



F1.1 redraw: classify first; every valid display and summary follows.

TYPE	TEST / EXAMPLE	USE
Nominal	no order; payment	counts, mode, bar
Ordinal	ordered bands	ordered bar, median band
Discrete	whole counts	frequency, histogram
Continuous	measured refinement	histogram, boxplot, mean

Digits do not decide type: postcode, customer ID and store number are nominal because an average is meaningless. A sorted bar chart compares lengths and usually beats a pie chart beyond about four categories.

2 · Contingency Tables CH 1

SAME CELL, CHANGING DENOMINATOR

```
marginal = row/column total ÷ grand total
joint = interior cell ÷ grand total
conditional = cell ÷ its row/column total
```

Independence: compute the *same* conditional probability inside each row. Equal conditionals → independent; unequal → relationship. Report the percentage-point gap. Signal words given, among, of those who demand a row or column denominator.

3 · Centre, Spread, Shape CH 2

MEASURE	PROMISE	TRAP
Mean	balance point; uses all values	pulled by extremes
Median	middle rank	ignores distance
Mode	most common	may tie / be absent
Range	max − min	two least typical cases
IQR	Q3 − Q1; middle half	tails omitted
SD	typical spread about mean	quote in original units

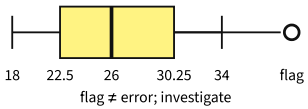
SAMPLE SUMMARIES

$\bar{x} = \Sigma x_i / n$

$s^2 = \Sigma(x_i - \bar{x})^2/(n-1)$ · $s = \sqrt{s^2}$

IQR = Q3-Q1 · fences = Q1-1.5IQR, Q3+1.5IQR

IQR = 30.25 − 22.5 = 7.75



F2.1 redraw: box, whiskers and the 1.5-IQR investigation flag.

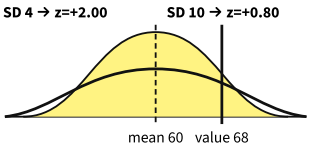
Skew: mean > median → long upper/right tail; mean < median → long lower/left tail. Prices, incomes and waiting times often need median + IQR. The mean remains the right centre when the question is about a total.

Why n−1? Deviations use the sample mean, which makes them unusually small; dividing sample squared deviations by n would systematically understate population variance. Use n−1 for a sample estimate.

Histogram rule: try two or three interval widths and retain one where the shape is stable. Too wide hides skew; too narrow reports sampling noise as structure. State the width in the caption.

Outlier workflow: flag → check recording and case context → report sensitivity with and without if influential. A flagged point may be the most informative case and is never deleted merely for being unusual.

4 · Standardise Before Comparing CH 2



F2.2 redraw: the same raw value has different standing in each group.

5 · Population → Interval CH 3

POPULATION LANGUAGE	SAMPLE LANGUAGE
population: every case the claim covers	sample: observed subset
parameter: fixed, usually unknown	statistic: known, varies by sample
bias: frame/process error	sampling variability: random noise

A larger sample shrinks noise, not bias. Selection, survivor and non-response bias remain. Ask how cases entered before asking how many there are. The law of large numbers stabilises sample summaries around population values; it cannot repair the wrong frame.

PROBLEM question → SAMPLE collect → MODEL fit → INFER claim?

Is the last arrow earned?

F5.1 redraw: inference is the last, conditional step.

SAMPLING DISTRIBUTION OF THE MEAN

centre = μ · SE(x̄) = σ/√n ≈ s/√n

quadruple n → halve SE

INTERVAL FOR A POPULATION MEAN

$\bar{x} \pm t^* \times s/\sqrt{n}$ · df = n-1

USE	REASON
Normal critical value	population σ known
t critical value	σ unknown; estimate with s; heavier tails

95% confidence: the *procedure* captures the parameter in about 95 of 100 repetitions. The realised interval either covers it or does not. An interval for a mean says nothing about where most individual observations lie.

FAILURE	WHAT WENT WRONG	DOES LARGER N FIX IT?
Selection bias	entry relates to measured quantity	No
Survivor bias	failed cases absent from frame	No
Non-response	respondents differ from non-respondents	No
Sampling noise	chance sample-to-sample movement	Yes, at √n rate

6 · Precision Levers CH 3

larger s	wider interval
larger n	narrower at √n rate
higher confidence	larger critical value → wider

CHECK *Write the SE first. If it did not shrink when n grew, you divided by n instead of √n. Then name the population supported by the sampling frame.*

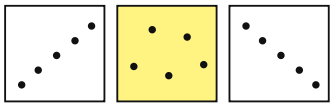
7 · Association → Regression CH 4

Scatter first: direction, form, strength and exceptions decide whether a straight-line summary is licensed. Put the explanatory variable on x and outcome on y. Swapping axes changes the fitted line because least squares measures vertical error; it does not change correlation.

JOINT MOVEMENT

sxy = $\Sigma(x_i - \bar{x})(y_i - \bar{y})/(n-1)$

$r = sxy/(s_x s_y)$, $-1 \le r \le 1$



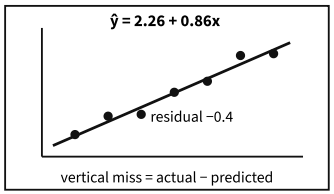
F4.2 redraw: r measures linear pattern, not any pattern.

LEAST SQUARES

$\hat{y} = b_0 + b_1 x$ · $b_1 = sxy/s_x^2$ · $b_0 = \bar{y} - b_1\bar{x}$

$e_i = y_i - \hat{y}_i$ · choose line minimising Σe_i^2

$R^2 = r^2$ in simple regression



F4.1 redraw: seven stores, fitted line and one residual.

Slope sentence: "A one-unit increase in x is associated with b_1 units of y." Name both units. Intercept only if x=0 is observed and meaningful. Residual curvature → wrong form; fan → changing spread; isolated large residual → influential exception.

R^2 is not a licence: it gives the share of sample outcome variation accounted for by the line; it proves neither linearity, causation nor validity outside the observed x-range.

12 · R + Wrangling Rules CH 7–8

VERB / OBJECT	CHANGE	CHECK
filter	keeps rows	condition first
select	keeps columns	variables retained
mutate	adds column	units / type
arrange	reorders	no data change
group	invisible label	expected groups
summarise	collapses rows; changes grain	carry count

8 · Hypothesis-Test Pipeline CH 5

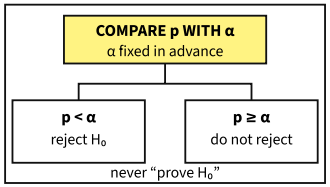
1. **Name the population parameter.**
2. Write H_0 as equality/no effect; H_1 from the question.
3. Fix α before seeing data; choose one- or two-sided from wording.
4. Compute test statistic = (estimate − null)/SE.
5. Compare p with α; decide; report effect, units and limit.

COMMON FORMS

$H_0: \mu = \mu_0$; $H_1: \mu \neq \mu_0$ (differs/changed)

$H_0: \beta_1 = 0$; $H_1: \beta_1 \neq 0$

"exceeds / improved / faster" → one-sided



F5.2 redraw: the decision is mechanical once p exists.

Correct p sentence: if H_0 were true, a result at least this extreme would occur with probability p. A large p means *insufficient evidence*, not that H_0 is true. Statistical significance does not certify commercial importance; read effect size and interval.

ERROR	MEANING	LEVER
Type I	reject true H_0; false alarm	rate = α
Type II	miss a real effect	power = 1−β
Trade	lower α usually raises β	larger n improves both

9 · Multiple Regression CH 6

COEFFICIENT SENTENCE

change in y for +1 unit of x_j, holding every other included variable fixed

Adding a variable changes the question and may shrink or reverse a coefficient because shared variation is reassigned. It does not make either estimate arithmetically wrong. Avoid causation from observational data and remember that omitted variables can still contaminate the model.

DEVICE	RULE	READ
k-category factor	k−1 dummies	each vs omitted baseline
Interaction	x×d	difference between slopes
Ordinary R^2	never falls when x added	not model chooser
Adjusted R^2	penalty per variable	compare specifications

10 · Classification CH 9

A model gives an estimated positive probability; a *threshold* turns it into a decision. Lower threshold → more positives, higher sensitivity and more false positives. Raise it when false action is costly; lower it when a missed positive is costlier. Moving the threshold trades errors; it does not improve the fitted model.

ACTUAL \ PREDICTED	positive	negative
positive	**48 TP** caught	**20 FN** missed
negative	**12 FP** false flag	**120 TN** cleared

acc .84 · sens .71 · spec .91

n=200 · rates use different denominators

F9.1 redraw: four counts before four rates.

RATES

accuracy=(TP+TN)/all

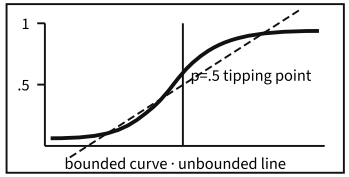
sensitivity=TP/(TP+FN) · specificity=TN/(TN+FP)

precision=TP/(TP+FP) · base rate=(TP+FN)/all

RATE	DENOMINATOR ASKS	LEAD WHEN...
Accuracy	all cases	classes not badly imbalanced
Sensitivity	actual positives	misses are costly
Specificity	actual negatives	false alarms are costly
Precision	predicted positives	a positive flag must be trusted

Accuracy must beat the base-rate benchmark. At 3% defaults, predicting none scores 0.97 and catches nothing. Fit on training data; judge on held-back data.

11 · Logistic Regression CH 10



F10.1 redraw: binary outcomes require a bounded probability curve.

SCALES

odds=p/(1-p) · log odds=ln[p/(1-p)]

log odds = $b_0 + b_1 x + \ldots$ · odds multiplier for +1 x = e^{b_1}

p = 1/(1+e⁻η), η=$b_0 + b_1 x + \ldots$

https://www.asksia.ai/cheatsheet/monash-etc1000

v2 · 2026-10-25

Forward this to your tutorial group

SHARED BY _____

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

Monash Semester 2 final assessment period, 2–18 November 2026 (check your personal exam timetable).

ETC1000 Business and Economic Statistics · Self-Test Drills

WORKED PATHS · JUDGEMENT CARDS · CLASSIC TRAPS · EXAM-MORNING RECALL

Full guide + ask Sia about ETC1000

SIDE 2/2 COVER THE ANSWER → WORK THE PATH → CHECK THE SENTENCE

SELF-TEST DRILLS

Five worked methods · judgement cards · morning plan

13 · Worked 1 – Probabilities

CH 1 WORKED

Data: 1,200 returns. Online 750, in store 450. Refunded: 300 online, 90 in store.

OUTCOME

	refunded	other	total
online	300	450	750
in store	90	360	450
total	390	810	1200

conditionals: .40 online vs .20 in store

Source-data redraw: choose the denominator from the wording.

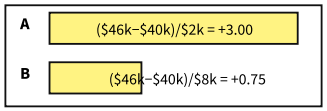
1. **Marginal refund:** (300+90)/1200 = **0.325**.
2. **Joint in-store and refunded:** 90/1200 = **0.075**.
3. **Independence:** P(refund|online)=300/750=**0.40**; P(refund|store)=90/450=**0.20**. Unequal → not independent in this sample.

TRAP *"Given/among" changes the denominator. "And/both" keeps the grand total. The claim concerns this sample, not customers in general.*

14 · Worked 2 – Standardise

CH 2 WORKED

Data: both branches average $40,000 weekly sales. SDs: A $2,000; B $8,000. Both post $46,000.



F2.2 values: equal sales; unequal standing inside each history.

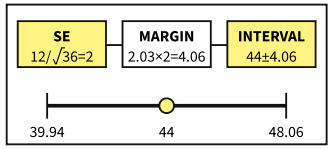
Answer: A's week is more unusual: +3 SD versus +0.75 SD. Ask A what changed and whether it can repeat; B is within normal variation and should focus on reducing variability. Raw $46,000 alone cannot rank the results.

IF YOU SEE	DO
same unit, different spread	standardise within each group
mean far above median	report right skew; prefer median for typical case
extreme observation	investigate; do not delete automatically
histogram changes with bins	try widths; keep stable shape; state width

15 · Worked 3 – Confidence Interval

CH 3 WORKED

Data: n=36 invoices, x̄=$44, s=$12. Approximate 95% interval; df=35, t*=2.03.



Worked path: SE → margin → interval.

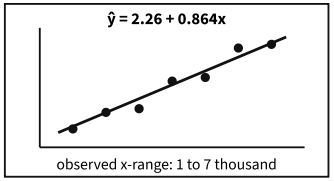
1. SE = 12/√36 = **$2**.
2. Margin = 2.03×2 = **$4.06**.
3. CI = 44±4.06 = **$39.94 to $48.06** (about $40–$48).
4. Sentence: the procedure brackets the population mean and succeeds in about 95 of 100 repetitions.

Qualify: this covers the *mean*, not individual invoices; it is only as defensible as the sampling frame. One month or one client can make a precise answer for the wrong population.

16 · Worked 4A – Fit the Line

CH 4 WORKED

Data: spend x=1…7 thousand; weekly sales y=3.2, 4.1, 4.4, 6.0, 6.3, 7.9, 8.1 thousand. x̄=4; ȳ≈5.71; deviation cross-products sum 24.2; squared x-deviations sum 28.



F4.1 values: seven stores and the least-squares line.

b_1=24.2/28=**0.864**; b_0=5.71−0.864×4≈**2.26**. Each extra $1,000 advertising is associated with about **$864** more weekly sales across these stores. It does not quite pay for itself on this evidence.

Refuse x=$30,000: the data cover $1,000–$7,000; the straight relationship may not continue. Also refuse causation: budgets were observed, not assigned.

24 · Classic Traps

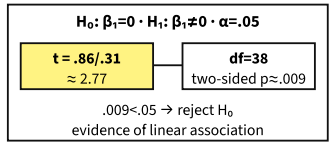
ALL CHAPTERS

WRONG MOVE	REPAIR
conditional ÷ grand total	divide inside stated row/column
large n fixes bias	only sampling noise shrinks
95% of individuals in CI	CI is for the population mean
r means cause	association; name third variables

17 · Worked 4B – Test the Slope

CH 5 WORKED

Data: n=40 stores; fitted slope 0.86 thousand sales per thousand spent; SE=0.31.



Worked path: population hypotheses → t → p → bounded conclusion.

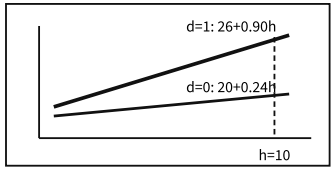
There is evidence of a linear association in the population represented by these stores. Estimated size: about **$860 more weekly sales per extra $1,000 spent**. No causal claim; no extrapolation beyond sampled spend.

BAD SENTENCE	REPAIR
"H_0 has a 0.9% chance."	p is probability of data this extreme if H_0 true.
"H_0 is proved" when p large.	insufficient evidence; real small effects may remain.
"Significant = important."	judge estimate and CI in business units.
choose one-sided after result	direction comes from question before data.

18 · Worked 4C – Interaction

CH 6 WORKED

Model: weekly spend = 20 + 0.24h + 6d + 0.66(hd), where member d=1.



Interaction redraw: dummy shifts intercept; product term tilts slope.

Non-member: d=0 → ŷ=20+0.24h. **Member:** d=1 → ŷ=26+0.90h. Slope difference 0.90−0.24=**0.66**, exactly the interaction. Member at h=10: 26+9=**$35**.

CHECK *If reconstructed slopes do not differ by b_3, you put the dummy coefficient into a slope. b_2 shifts the intercept; b_3 changes the slope.*

The member coefficient is a comparison at h=0, not a universal member premium. With an interaction present, every main effect is conditional on the other interacting variable being zero.

IF YOU SEE	WRITE
coefficient changes after adding z	it now compares cases alike on z; the question changed

19 · Worked 5A – Classifier

CH 9 WORKED

Held-back n=400: TP=30, FP=20, FN=50, TN=300; 20% churn.

PREDICTED

	churn	stay
churn	30 TP	50 FN
stay	20 FP	300 TN

acc .825 · sens .375 · precision .60
do-nothing accuracy = .80

Worked matrix: the headline barely beats the base-rate benchmark.

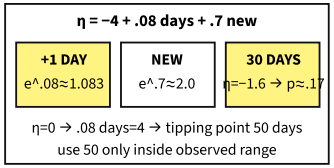
1. Accuracy=(30+300)/400=**0.825**.
2. Sensitivity=30/(30+50)=**0.375**.
3. Precision=30/(30+20)=**0.60**.
4. No-flags benchmark = non-churn share = **0.80**; gain is slight.

For retention, misses are expensive; catching under two fifths of churners is weak. Lower threshold → sensitivity rises, precision falls. Cheap contact relative to lost customer makes that trade defensible.

20 · Worked 5B – Logistic Model

CH 10 WORKED

Model: late-payment log odds
η=−4.0+0.08(days)+0.7(new), new=1.



Worked path: coefficient → odds multiplier → probability → tipping point.

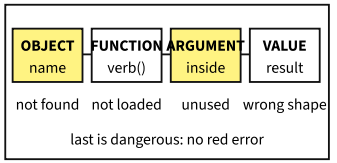
Each extra credit day multiplies late-payment odds by about **1.083**, holding customer status fixed; ten days multiplies odds by about **2.2**. A new customer has about **2.0×** the odds versus established on the same terms.

Established, 30 days: η=−4+2.4=−1.6 → p≈**0.17**. Tipping point: η=0 → days=**50**; beyond about 50 days, late payment becomes more likely, provided 50 is within the data's range.

DECISION CUE	MOVE
missed positive is expensive	lower threshold; sensitivity ↑, false positives ↑
false action is expensive	raise threshold; precision tends to ↑, sensitivity ↓
rare positive class	state base rate before accuracy
logistic coefficient requested	report sign, e^b odds multiplier, holding-fixed clause

21 · R Error Decoder

CH 7 WORKED



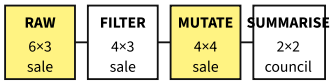
F7.1 redraw: match the message to the failed hop.

MESSAGE/RESULT	CAUSE → FIX
object not found	name not created/misspelt → run earlier line/check name
could not find function	package not loaded or verb misspelt
unused argument	argument name wrong
three times expected rows	non-unique join key multiplied matches

Read the first error, not the last. Predict result rows, columns and grain before running code. Check a joined row count against the input; an increase can expose duplicate keys.

22 · Wrangling Sequence Drill

CH 8 WORKED



only summarise changes what one row means
carry n beside every group statistic

F8.1 redraw: rows, columns and grain are different questions.

Faulty sequence: rank 313 suburbs by median price → keep top 20 → group by council → report council means. Those are means of each council's surviving expensive suburbs, not of its suburbs.

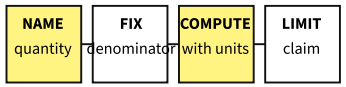
Repair: start with all 313 suburbs → group by council → summarise every suburb (include count) → arrange the council table. A council with one very expensive suburb no longer wins unfairly.

IF YOU SEE	DO / SAY
top within a subset	filter first, rank after
groups of unequal size	do not average their averages
missing after join	investigate; distinguish key mismatch from genuine absence
more rows after join	test key uniqueness
grouped output	state new grain and carry count

23 · Method Chooser

CH 1–10

IF THE QUESTION ASKS...	USE / SHOW	GUARDRAIL
category counts/shares	frequency, relative frequency, bar	no mean of labels
two categorical variables	contingency + conditionals	denominator follows "given"
typical skewed value	median + IQR	mean follows tail
compare unlike spreads	$z=(x-\bar{x})/s$	name reference group
population mean range	$\bar{x}\pm t^*s/\sqrt{n}$	procedure, not individuals
linear joint movement	scatter + r	r≈0 may hide curve
predict numerical y	least-squares line	residuals + x-range
generalise sample effect	hypothesis test + CI	population parameter
hold variables fixed	multiple regression	no omitted-variable immunity
group-specific slopes	dummy + interaction	b_3 = slope difference
predict a category	threshold + confusion rates	base rate + held-back data
binary probability model	logistic regression	coefficient is log-odds scale



probability · interval · slope · rate

F12.1 redraw: the reusable four-beat answer.

24b · Model + Data Traps

CH 5–10

WRONG MOVE	REPAIR
accuracy alone	base rate + sensitivity + precision
logistic b = probability change	e^b multiplies odds
join silently trusted	predict and verify row count
mean of group means	use original rows or weighted mean

25 · Exam-Morning Plan

STUDY STRATEGY

1. Recite: type → display; denominator; centre/spread/shape; z; SE/CI; scatter/r/line/residual/R^2; H_0/H_1/p/errors; holding-fixed/dummy/interaction; matrix/threshold; odds/log odds.
2. Redo the five numbers cold: **.325/.075; +3/+0.75; $39.94–$48.06; 0.864/2.26; .825/.375/.60; p≈.17 and 50 days.**
3. For every answer: **name → denominator/scale → compute with units → limit.**

Forward this to your tutorial group

SHARED BY _____

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

Monash Semester 2 final assessment period, 2–18 November 2026 (check your personal exam timetable).