STATS100 Concepts in Statistics

DEPARTMENT OF STATISTICS · UNIVERSITY OF AUCKLAND · UNDERGRADUATE STAGE 1

Full guide + ask Sia about STATS100

SIDE 1/2 · DEFINITIONS → FORMULAS → DECISION RULES → CLAIM LIMITS · EXAM REVISION SHEET · S2 2026 · final examination 50% · source-grounded

1 · Exam map

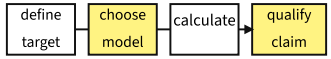
COURSE.JSON · BRIEF

Assignments 30% | Quizzes | Final Examination 50%

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

COMPONENT	PUBLISHED STRUCTURE	WEIGHT
Assignments	Best 10 of 11; 3% each retained	30%
Quizzes	Two quizzes; 10% each	20%
Final Examination	Minimum exam mark 45%	50%

Two separate pass checks: at least 50% overall and at least 45% in the final examination. Treat the component minimum and course total as separate conditions.



population · variables

process · uncertainty

Course-wide route: identify the target, represent it, calculate transparently, then enforce the design and uncertainty boundary.

The difficult move is translating a visual or simulated pattern into a parameter claim while keeping **sample, population, null model, uncertainty and practical meaning** distinct. A numerical answer without its denominator, unit, model or target population is unfinished.

1b · Universal answer frame

COURSE STUDY STRATEGY

1. **Identify:** population, observational units, variables and data-generating process.
2. **Represent:** descriptive display, chance/simulation model or inferential model.
3. **Compute:** write the relationship, substitute with units/signs, and preserve the denominator.
4. **Check:** inverse operation, boundary case, expected direction or independent calculation.
5. **Conclude:** answer the question, then name uncertainty, design and ethical limits.

SIA → *Recite the five verbs before every drill: identify, represent, compute, check, conclude.*

1c · Denominator & claim check

W2–3 · COURSE OVERVIEW

RATE	REQUIRED DENOMINATOR
overall proportion	all eligible sampled observations
subgroup proportion	eligible observations in that subgroup
misclassification	evaluated cases with a reference category

2 · Time series & change

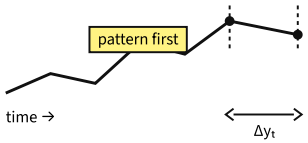
W1

Time series = measurements indexed and ordered by time, so temporal dependence and change stay visible. *Time difference* = change across a defined lag. *Temporal pattern* = trend, cycle, seasonal structure, dependence or unusual event.

ONE-PERIOD CHANGE

$\Delta y_t = y_t - y_{t-1}$

Bites when: the prompt asks what changed from the preceding time point. Preserve the original unit and interpret the sign.



Source-derived time-series map: read level and structure before turning adjacent observations into a difference.

SEE	DO	DO NOT CLAIM
ordered plot	describe level, direction, cycles/seasonality and unusual events	independence
before/after	compute $y_t - y_{t-s}$; name sign and unit	cause
pattern shift	locate when it starts and compare with surrounding structure	mechanism from timing alone

Boundary: a before–after difference describes movement; it does not establish why the movement occurred or make successive observations independent. Check that the lag is the one the question actually defines.

3 · Proportions & classification

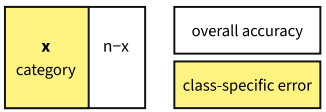
W2–3

Sample proportion = count in a defined category ÷ eligible sampled observations. A *classification rule* assigns categories from measured features; *misclassification rate* is the fraction whose predicted category differs from the reference category.

SAMPLE PROPORTION

$\hat{p} = x/n$

Bites when: counting a category. The numerator must be contained in the denominator on the same observation basis.



overall accuracy

class-specific error

Source-derived classification map: define the eligible whole, then inspect class-specific error as well as overall accuracy.

4 · Chance & null models

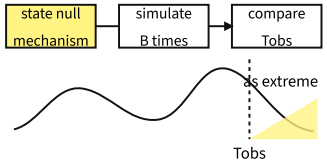
W4

Chance model specifies random conditions. A *null model* is the benchmark of no effect, no association or another stated reference. Its repeated statistics form a *simulation distribution*.

ADJUSTED SIMULATION P-VALUE

$\hat{p}_{sim} = (1 + \#\{T_\beta \text{ at least as extreme as Tobs}\})/(B + 1)$

Bites when: a finite null simulation is used to judge extremeness.



Source-derived null route: generate the reference distribution under the stated mechanism, then count results at least as extreme.

STEP	QUESTION TO ANSWER
Null	What chance process generates outcomes?
Statistic	What single quantity measures departure?
Tail	Which outcomes count as at least as extreme?
Decision	Is Tobs unusual under this model?

The +1 adjustment includes the observed statistic and prevents zero from a finite run. A p-value is not P(the null is true). Larger B improves simulation stability; it cannot repair a null mechanism that misrepresents the design.

5 · Sampling variation & SE

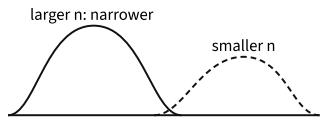
W5–6

Sampling distribution = the distribution of a statistic over repeated samples drawn by the same design. Its standard deviation is the *standard error*.

PROPORTION STANDARD ERROR

$SE(\hat{p}) = \sqrt{[p(1-p)/n]}$

Bites when: quantifying repeated-sample variability of a proportion under the stated model.



SE falls at $\sqrt{n}$ rate

Source-derived sampling map: repeated-sample spread narrows with larger n, at the square-root rate.

IF N CHANGES	SE RESPONSE
n multiplied by 4	SE divided by 2
n multiplied by k	SE divided by $\sqrt{k}$
same n, p near 0.5	larger p(1−p) than near 0 or 1

6 · Confidence intervals & claims

W5–6

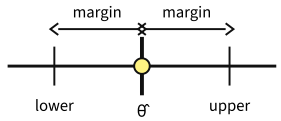
Point estimate targets a population parameter with one statistic. *Margin of error* is the symmetric half-width added to and subtracted from the estimate.

GENERIC INTERVAL

interval = $\hat{\theta} \pm c^*SE(\hat{\theta})$

margin = $c^*SE(\hat{\theta})$

Bites when: converting an estimate and its sampling error into a bounded population claim.



Source-derived interval map: the critical multiplier and SE determine the half-width around the point estimate.

Boundary: confidence is the long-run coverage of the interval procedure, not a posterior probability assigned to a fixed realised parameter. Narrow intervals do not correct selection bias.

7 · Data sources, missingness & ethics

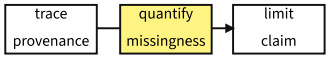
W7

Data provenance documents origin, collection, processing and custody. *Data ethics* covers consent, privacy, representation, fairness, stewardship and consequences.

MISSINGNESS RATE

$r_{miss} = n_{missing}/n_{eligible}$

Bites when: auditing whether expected values are absent. Define eligibility before counting missing values.



all | subgroup concentration? systematic relation to outcome?

Source-derived data-quality route: overall missingness is only the start; inspect concentration and mechanism.

AUDIT	ASK
Provenance	who collected what, when, how, for whom?
Eligibility	who should have had the value observed?
Pattern	does absence concentrate by subgroup or outcome?
Ethics	what consent, privacy, representation or consequence changes the defensible claim?

Boundary: a low overall missing fraction does not prove missing values are harmless when absence is systematically related to the outcome. Recalculate relevant subgroup rates.

8 · Relationships & linear models

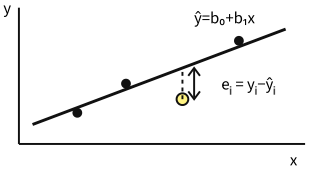
W8

An *explanatory variable* describes conditional variation in an outcome. The *fitted line* is the estimated linear mean response; a *residual* is the observed vertical departure from that line.

FITTED LINE AND RESIDUAL

$\hat{y}_i = b_0 + b_1 x_i \quad \cdot \quad e_i = y_i - \hat{y}_i$

Bites when: predicting within the observed x range, interpreting slope, or diagnosing model fit.



Source-derived linear-model map: prediction lies on the line; residual is the signed vertical gap to the observation.

OBJECT	INTERPRETATION / CHECK
b_0	fitted outcome at x = 0, only if that reference is meaningful
b_1	fitted change in outcome per one-unit increase in x
$\hat{y}$	conditional prediction at the stated x
e	positive if observed y is above the fitted line
residual pattern	structure suggests the linear representation is incomplete

Boundary: a fitted association does not establish causality. Prediction beyond the observed explanatory range is extrapolation and lacks direct support.

9 · Normal models & standardisation

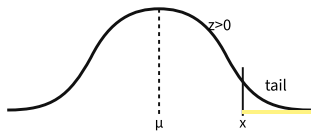
W9

A *normal distribution* is a symmetric bell-shaped model determined by μ and σ. A *standard score* expresses signed distance from the reference mean in standard-deviation units.

STANDARD SCORE

$z = (x - \mu)/\sigma$

Bites when: comparing relative location or using a defensible normal model for a tail probability.



Source-derived standardisation map: subtract the reference centre, scale by σ, then match the requested tail direction.

Z	MEANING
0	x equals the reference mean

10 · Tests, design & communication

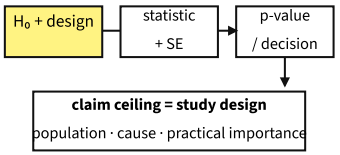
W10–12

A *test statistic* scales sample departure from a null benchmark. *Study design* covers selection, assignment, measurement and control of alternative explanations. *Uncertainty communication* reports variability, assumptions, limitations and plausible ranges.

STANDARDISED TEST STATISTIC

$z = (\hat{\theta} - \theta_0)/SE(\hat{\theta})$

Bites when: measuring how many standard errors the estimate lies from the null value.



Source-derived testing map: statistic and p-value answer compatibility with H_0; design controls the scope of the conclusion.

LAYER	WHAT EARNS THE CHECK
Question	state target parameter and comparison
Null	give benchmark θ_0 and chance mechanism
Evidence	show departure relative to SE; identify tail
Decision	describe compatibility with H_0, not probability H_0 is true
Meaning	separate statistical evidence from practical importance
Scope	respect selection, assignment, measurement and alternatives

INTERVAL PROMPT	COMPLETE RESPONSE
construct	state $\hat{\theta}$, SE, c^*, margin, endpoints
interpret	name target parameter and sampling procedure
compare	separate effect size from precision
critique	ask whether design justifies the population claim

Boundary: statistical significance does not supply practical importance, causal identification or freedom from selective analysis.

11 · Which representation?

ALL TOPICS

IF THE PROMPT ASKS...	START WITH...	FINISH WITH...
change over ordered time	time plot + Δy_t	pattern and non-causal limit
part of an eligible whole	$\hat{p} = x/n$	named denominator
chance extremeness	null simulation	tail count + model limit
repeated-sample spread	sampling distribution/SE	random-error statement
plausible parameter range	$\hat{\theta} \pm c^*SE$	procedure-based claim

Semester Two examination period, 29 October – 16 November 2026 (check your personal exam timetable).

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

STATS100 Concepts in Statistics

SELF-TEST DRILLS · JUDGEMENT CARDS · CLASSIC TRAPS · EXAM-MORNING PLAN

Full guide + ask Sia about STATS100

SIDE 2/2 WORKED PATHS → IF YOU SEE X, DO Y → TRAPS → RAPID RECALL

EXAM REVISION SHEET

S2 2026 · answer, then check the boundary

12 · Worked drill 1: p̂ + interval

COURSE WORKED EXAMPLE · 5 MARKS

Question. In an AskSia-authored random sample, 84 of 120 students prefer option A. Estimate the proportion and form an approximate 95% interval using two standard errors.

1 · ESTIMATE

$\hat{p} = 84/120 = 0.70$

2 · STANDARD ERROR

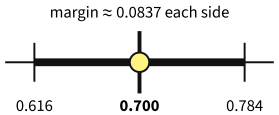
$SE = \sqrt{[0.70(0.30)/120]} \approx 0.0418$

3 · MARGIN

$2SE \approx 2(0.0418) = 0.0837$

4 · ENDPOINTS

$0.70 \pm 0.0837 \approx 0.616$ to 0.784



The verified arithmetic: centre 0.70, SE ≈ 0.0418 and two-SE margin ≈ 0.0837.

MARK PATH	CHECK
1	84/120 = 0.70
1	SE ≈ 0.0418
1	margin ≈ 0.0837
1	interval 0.616–0.784
1	tie claim to sampling process; selection bias remains possible

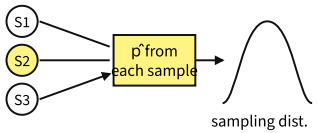
Answer. The estimate is 0.70 and the approximate 95% interval is 0.616 to 0.784. It quantifies repeated-sample uncertainty under the stated random-sampling model; it does not correct an unrepresentative sample.

SIA → *Name the population and sampling method before interpreting the endpoints.*

12b · Arithmetic checks

SAME EXAMPLE

- 0.70 lies midway between the endpoints.
- Both endpoints lie between 0 and 1.
- 0.70(0.30)=0.21 and dividing by 120 makes the variance small.
- Rounding at the end preserves 0.616 and 0.784.



Source-derived sampling map: the interval's SE describes how p̂ varies over repeated samples from the same design.

12c · Final priorities

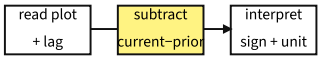
EXAM WINDOW

13 · Worked drill 2: ordered change

W1 WORKED METHOD

Prompt. Use the one-period change relationship while separating level, change and time structure in a plotted pattern.

STEP	ANSWER PATH
1	Define y_t, y_{t-1}, time unit and lag.
2	Write $\Delta y_t = y_t - y_{t-1}$; substitute with units and sign visible.
3	Check by rearranging $y_t = y_{t-1} + \Delta y_t$ and compare with the plot.
4	Interpret positive/negative movement in the original unit.
5	State that the difference does not identify cause or independence.



check: $y_{t-1} + \Delta y_t = y_t$

movement ≠ cause

The chapter's worked method: define, transform, independently check, interpret, then enforce the stated limit.

Answer stem: "From the preceding time point to t, the measured quantity changed by Δy_t in the original unit. The ordered plot also shows [trend/cycle/seasonal structure/unusual event]. This describes change, not its cause."

14 · Worked drill 3: classification

W2–3 WORKED METHOD

Prompt. Compute or compare a classification rate with an explicit denominator.

define evaluated cases + reference category

overall error	subgroup error
wrong / evaluated	wrong-g / eval-g

compare both; minority failure can hide

Source-derived classification drill: overall accuracy is not enough when subgroup denominators differ.

COMMON SETUP ERROR	REPAIR
numerator not a subset of denominator	redefine eligibility before calculation
subgroup count ÷ overall n	use the subgroup's eligible denominator
only overall accuracy reported	inspect class-specific misclassification
sample p̂ described as population p	justify the sampling link first

If the prompt asks "what fraction or percentage" → begin with p̂=x/n and name the eligible denominator in words.

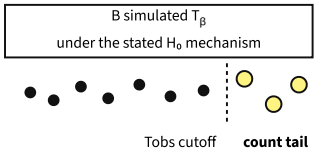
SIA → *State the denominator in words before writing it in symbols.*

15 · Worked drill 4: null simulation

W4 WORKED METHOD

Prompt. Use the adjusted simulation p-value to compare an observed statistic with a null reference distribution.

STEP	ANSWER PATH
1	State the null mechanism and statistic T.
2	Generate B null statistics T_β using that mechanism.
3	Define "at least as extreme" in the relevant tail(s).
4	Compute (1 + extreme count)/(B + 1).
5	Describe compatibility with H_0 and qualify the model/design.



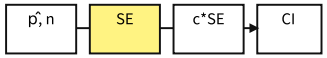
The finite-run adjustment includes the observed statistic and prevents an estimated zero.

Answer stem: "Under the stated null mechanism, the adjusted proportion of simulated statistics at least as extreme as Tobs is $\hat{p}_{sim}$. This measures compatibility with that benchmark; it does not give P(H_0 is true)."

Trap: changing the tail after seeing Tobs, or simulating a process that does not match the study design.

16 · Worked drill 5: SE → interval

W5–6 WORKED METHODS



check centre + symmetry

plausible bounds; bias remains

Source-derived inference path: calculate sampling error, form the margin, check endpoints, then state the claim boundary.

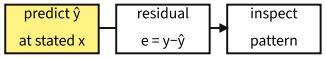
IF YOU SEE…	DO…
larger n	expect smaller SE at $1/\sqrt{n}$, all else fixed
wider confidence target	expect a larger critical multiplier and margin
narrow interval from biased sample	call it precise but not automatically valid
fixed realised interval	avoid assigning posterior probability to θ

Answer path: define the statistic's repeated-sample distribution → calculate SE → choose the stated critical multiplier → form $\hat{\theta} \pm c^*SE$ → interpret for the named parameter and sampling process.

17 · Worked drill 6: fitted line

W8 WORKED METHOD

Prompt. Fit and interpret a linear relationship while checking residual structure and extrapolation.



slope: y-units per x-unit

stay in x range; no causal claim

The chapter's worked route keeps fitted value, observed value and residual as separate objects.

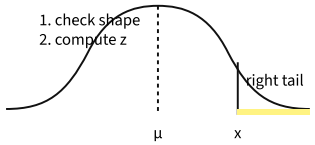
STEP	SELF-CHECK
Define	x explanatory; y outcome; units visible
Predict	$\hat{y} = b_0 + b_1 x$
Depart	e = y − ŷ; sign matches plot
Diagnose	look for structured residuals
Conclude	conditional association, supported x range only

Classic sign trap: residual is observed minus fitted. A point above the fitted line has e>0; a point below has e<0.

18 · Worked drill 7: z & a tail

W9 WORKED METHOD

Prompt. Standardise a value and interpret its tail location under a defensible normal model.

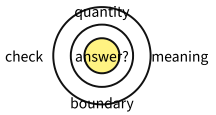


Source-derived normal drill: model fit comes before standardisation; tail direction comes after the sign of z.

ANSWER MOVE	WORDS
standardise	$z = (x-\mu)/\sigma$
sign	above/below the reference mean
magnitude	|z| standard deviations from μ
tail	match "at least", "above", "below" or two-sided wording
limit	z does not prove normality

18b · One-minute answer audit

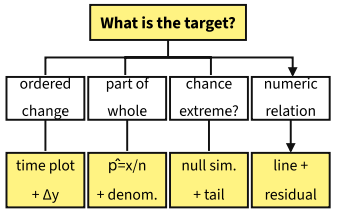
W1–12 LIMITS



Source-derived final audit: calculation sits inside check, meaning and boundary.

19 · Method chooser

ALL WORKED METHODS



Need uncertainty or benchmark?

SE/CI or standardised test statistic

New decision tree, using only the source methods: classify the target before choosing the calculation.

IF YOU SEE X	DO Y
"from the previous time"	$\Delta y_t = y_t - y_{t-1}$; keep order and lag
"under chance / no association"	state null generator, statistic and tail
"how variable over samples"	sampling distribution and SE
"plausible range"	point estimate ± critical multiplier × SE
"predict y from x"	fitted line; inspect residuals and x range

20 · Data judgement cards

W7

low overall rate → check subgroup	systematic absence → claim at risk
unclear eligibility → redefine n	privacy / consent → changes use

Source-derived data-quality comparison: amount, pattern, eligibility and ethical use are separate checks.

Unclear source → trace provenance. **Concentrated absence →** report subgroup rates and narrow the population claim. **Consent, privacy, representation or fairness concern →** treat it as an analytic constraint.

20b · Rapid recall table

ALL FORMULAS

RELATIONSHIP	VERBAL CHECKSUM
$\Delta y_t = y_t - y_{t-1}$	current minus preceding
$\hat{p} = x/n$	category count over eligible sample
$(1+\#extreme)/(B+1)$	finite null run includes observed
$SE(\hat{p}) = \sqrt{[p(1-p)/n]}$	sampling spread falls at $\sqrt{n}$ rate
$\hat{\theta} \pm c^*SE$	centre ± margin
$n_{miss}/n_{eligible}$	missing among expected values
$\hat{y} = b_0 + b_1 x$; $e = y - \hat{y}$	prediction; observed minus fitted
$z = (x-\mu)/\sigma$	signed SD distance from mean
$z = (\hat{\theta} - \theta_0)/SE$	null departure in SE units

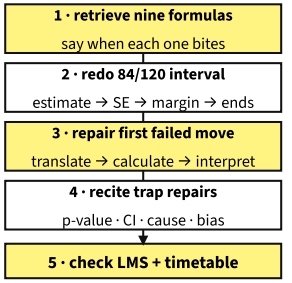
21 · Classic traps: say the repair

W1–12 FAQS

TRAP	REPAIR SENTENCE
Difference proves cause	It describes movement; causal mechanism is not identified.
Successes divided by any total	The numerator must be contained in the eligible denominator.
Accuracy is enough	Inspect class-specific error, especially minority-class performance.
p-value = P(H_0 true)	It is extremeness under the stated null process.
More simulations fix the null	Larger B stabilises estimation; model misspecification remains.
Bigger n removes all error	It reduces random error, not selection, dependence or measurement error.
95% probability θ is inside	Confidence is long-run coverage of the procedure.
Low missing rate is harmless	Check whether absence is systematic or concentrated.
Fitted line proves cause	It describes conditional association; design controls causal scope.
Predict beyond observed x	Label it extrapolation; direct support ends at the observed range.
z proves normality	It standardises under a chosen reference; check shape separately.
Significance = importance	Report practical magnitude and design limits separately.

22 · Exam-morning plan

COURSE STUDY STRATEGY



New exam-morning flow: retrieve, calculate, diagnose, repair, then verify current operational instructions.

ERROR LOG	DIAGNOSTIC QUESTION	REPAIR
Translation	Did I identify target, variables, unit, denominator and reference?	Rewrite the setup in words before symbols.
Calculation	Did sign, substitution, operation or rounding first diverge?	Use an inverse or boundary check.
Interpretation	Did the conclusion outrun design, model or population?	Attach the specific boundary, not generic caution.