SOC613 Applied Analytical Statistics for Social Scientists

LINGNAN UNIVERSITY · POSTGRADUATE · TERM 1, 2026-27

Full guide + ask Sia about SOC613

SIDE 1/2 OPEN-BOOK EXAM SHEET · decisions, assumptions, formulas, interpretation METHODS FIRST claim → measure → data → summary → inference

01 · Exam map COURSE OUTLINE

10 | Replication 45% | Final 45%

ASSESSMENT	WEIGHT	FORMAT
Class & tutorial participation	10%	Preparation, discussion, applied work
Quantitative Paper Replication Project	45%	Group report, 1,500–2,000 words; reproducible code
Final exam	**45%**	Individual · open-book · in class · two hours

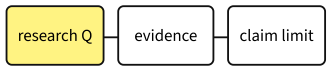
"The final individual exam (open-book, in-class format) assesses students' ability to apply multivariate statistical techniques and principles of causal inference to analyze social data." *SOC613 course outline v2, Term 1 2026-27*

PUBLISHED EMPHASIS	VISIBLE WORKING
Application	Choose a technique because it fits the question and measurement.
Causal reasoning	Separate association, timing and alternative explanations.
Analytical rigour	Check assumptions, denominators, valid N and transformations.
Clear interpretation	Return from output to social meaning; match verb to design.

Captured materials do not publish sections or question types. Prepare to choose, check, calculate and interpret, rather than memorising an unverified exam structure.

02 · Start with the claim C1

Question sets the conclusion limit



describe obs/assign relate/cause

verb ≤ design

Match the conclusion verb to the comparison that produced it.

QUESTION	EVIDENCE JOB	SAFE ANSWER
Descriptive	Measure what exists	Level, distribution, frequency
Explanatory	Describe co-movement	Association + direction/magnitude
Predictive	Anticipate unused outcomes	Performance on outcomes not used to construct model
Causal	Ask what changes under intervention	Effect only with credible counterfactual comparison

FIRST FOUR LINES

```
Research question · unit of analysis
· dependent variable (outcome) ·
independent variable (explanatory
factor).
```

03 · Causal-claim audit C1

ALL THREE REQUIRED

```
1 Association observed
2 Proposed cause precedes outcome
3 Plausible alternative explanations
addressed
```

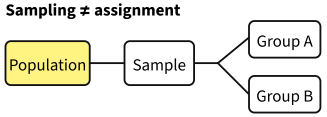
AUDIT	ASK	FAILURE MEANS
Association	Do X and Y vary together?	No causal effect supported
Temporal order	Was X measured/assigned before Y?	Reverse ordering remains possible
Alternatives	Could C cause both X and Y?	Confounding can generate the pattern

Confounder = a variable related to both proposed cause and outcome. Draw C → X and C → Y. Statistical control can address measured confounders; it cannot convert an unmeasured assumption into observed fact.

Observational conclusion stem: "X is associated with Y in the studied sample, in [direction/magnitude], subject to [adjustment or remaining confounding]." Avoid causal verbs unless design and assumptions support the counterfactual comparison.

04 · Sampling ≠ assignment C1

Sampling ≠ assignment



sampling → reach assign → causal

One mechanism does not imply the other; name what actually occurred.

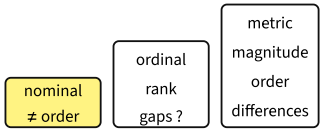
MECHANISM	PRIMARY SUPPORT	THREATS
Random sampling	Representation of target population	Coverage, nonresponse, attrition
Random assignment	Causal contrast within study; balances pre-existing traits on average	Noncompliance, attrition
Neither	Careful sample-specific association	Selection and confounding

A representative survey can describe a city while programme uptake remains self-selected. A randomised study can identify a treatment contrast in its sample without representing a wider population.

Internal validity concerns the studied contrast; ***external validity*** concerns reach beyond studied units. Attrition, nonresponse and noncompliance can weaken the original mechanism, so report the process achieved—not merely its label.

05 · Measurement ladder C2

Level limits valid operations



Numbered category codes do not move a variable up the ladder.

LEVEL	WHAT VALUES MEAN	DEFENSIBLE SUMMARIES
Nominal	Different categories; no rank	Counts, percentages, mode
Ordinal	Difference + meaningful order; gaps unspecified	Counts, percentages, mode, median, IQR
Interval-ratio	Ordered magnitude; meaningful differences	Mode, median/IQR, mean/SD where suitable
Binary 0/1	Two categories; one category designated 1	Mean = proportion coded 1

BINARY INTERPRETATION

```
x̄ = n(coded 1) / n(valid). Name the
category coded 1; convert to per
cent. It is not an average quantity.
```

06 · From concept to variable C2

LAYER	QUESTION	RECORD
Concept	Which abstract idea?	Meaning and scope
Indicator	What observable evidence represents it?	Question/observation
Operational definition	Exactly how observed and scored?	Responses, recall period, rule, direction, missing handling
Variable	How stored across cases?	Name, labels, codes, level

Reliability asks whether measurement is consistent across repetitions/items. ***Validity*** asks whether it captures the intended concept. A stable measure can be reliably wrong.

COMPOSITE/RECODING AUDIT

- State item set, reverse-coded items, weighting, missing-response rule and what higher values mean.
- Collapse categories only for a research-relevant, substantively defensible boundary.
- Preserve the original field; name the transformed variable; show how frequencies changed.
- Before comparing datasets, reconcile wording, recall period, response categories, scoring and direction.

Strong answers name permission and prohibition: ordinal order permits a median, but unspecified gaps do not justify equal-step arithmetic.

07 · Reproducible pipeline C3

Visible steps make work checkable



keep raw data unchanged

Unchanged source + ordered code + checks = traceable output.

STAGE	REQUIRED ACTION	DIAGNOSTIC
Import	Use project-relative path; keep received file unchanged	Dimensions/types match documentation
Inspect	Read codebook; identify row unit, labels, valid/nonresponse codes	Raw frequencies expose rare/impossible values
Transform	Create descriptively named variables; document boundary/direction/unit	Trace known cases through rule
Verify	Cross-tab, frequency or summary immediately	Check missingness, scale direction, valid N
Analyse	State analytic sample for every result	Added variable may change complete-case N
Re-run	Restart; run ordered entry script from source	No hidden objects/manual steps/machine path

A row is one observation and a column one variable; this is substantive. Person-level data cannot silently support a household conclusion. Repeated person-years need an identifier and time index.

08 · Missing + recoding C3

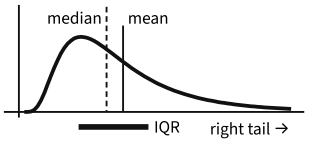
SAFE ORDER

```
codebook → convert documented
missing codes → define new variable
from valid values → verify original
× new → report valid N.
```

TRAP	WHY IT BITES	REPAIR
Threshold before missing conversion	Large nonresponse codes satisfy inequalities	Handle missing codes first
Overwrite source variable	Rule cannot be checked/reversed	Create a new named field
Silent complete-case deletion	Models use different observations	Report N per result
Manual correction	No audit trail	Encode decision in script
Comment repeats syntax	Reason remains unknown	Explain substantive purpose

09 · Describe before modelling C4

Shape guides centre + spread



In a right-skewed distribution the upper tail can pull mean above median.

Profile every variable in this order: measurement level → valid N → centre → spread → shape → unusual cases. Statistics are also diagnostics: polished models still calculate with miscoded missing values and implausible records.

MEASURE	QUESTION ANSWERED	USE/CAUTION
Mode	Most frequent value/category	Available at every level
Median	Half below/above	Requires order; resistant to extremes
Mean	Arithmetic centre using every value	Interval-ratio; sensitive; retain unit
Range	Full observed span	Tail-sensitive
IQR	Spread of middle half	Pairs with median for skew/extremes
SD	Spread around mean in original unit	Use where mean is meaningful; inspect shape

SOURCE-GROUNDED IDENTITIES

```
IQR = Q₃ - Q₁
binary mean = proportion coded 1
```

10 · Graph + outlier rules C4

DATA/QUESTION	GRAPH	READ
Categorical frequency	Bar chart	Category counts/percentages
Quantitative shape	Histogram	Skew, clusters, gaps, peaks
Centre/spread/extremes	Box plot	Median, IQR, unusual cases

An extreme value is a request for investigation, not an automatic deletion rule. Check coding, units, source record and substantive plausibility. Retain genuine cases; if exclusion is justified, state the rule and show how results change.

Mean ≠ median alone does not prove skew. Inspect the graph. Similar centre measures do not prove symmetry. If graph and summaries disagree, investigate coding, missing values, multiple groups or extreme observations.

Pair centre with spread deliberately: median + IQR for resistance; mean + SD when arithmetic magnitude is meaningful and sensitivity is acknowledged.

11 · Cross-tab first C5

Cross-tabulation jointly classifies each case by two categorical variables. Interior cells are joint frequencies; edges are row/column marginals. Identify the independent variable from the question *before* calculating percentages.

CONDITIONAL PERCENTAGE

```
100 × cell count / total within that
independent-variable category
```

IF INDEPENDENT VARIABLE IS...	PERCENTAGE WITHIN...	COMPARE...
Columns	Each column (must total 100%, rounding allowed)	Across rows
Rows	Each row (must total 100%, rounding allowed)	Down columns

INTERPRETIVE LAYER	QUESTION
Existence	Do conditional distributions differ?
Strength	How large is the pattern?
Direction	Only when both variables carry meaningful order

Wrong percentage direction answers a different question. State the denominator so direction is visible.

12 · χ^2 method C5

EXPECTED COUNT UNDER INDEPENDENCE

$E_{ij} = (\text{row total}_i \times \text{column total}_j) / N$

PEARSON STATISTIC

```
cell contribution = (O - E)² / E
χ² = sum of all cell contributions
df = (r - 1)(c - 1)
```

STEP	DO	INTERPRET
1	Use categorical **counts**; independent observations	Check table matches research question
2	Compute expected cells from marginals	Expected table must reproduce observed marginals
3	Check expected-cell adequacy	If too small, qualify; consider justified merger/alternative
4	Sum contributions; find df; read p	p = surprise of ≥ this discrepancy under independence, not P(null true)
5	Return to conditional percentages/cells	Explain where pattern lies
6	Add Cramer's V	0–1 strength; nondirectional; separate magnitude from significance

Lambda asks about proportional reduction in prediction error. It may miss a patterned table when modal predictions do not change.

Cramer's V separates magnitude from sample-size-driven significance. It is bounded between zero and one but remains nondirectional; the conditional percentages provide the substantive pattern.

SOC613 Worked drills, traps & fast recall

APPLIED ANALYTICAL STATISTICS FOR SOCIAL SCIENTISTS · LINGNAN UNIVERSITY

Full guide + ask Sia about SOC613

SIDE 2/2 WORKED STEPS · method chooser · error diagnoses · reporting stems SHOW YOUR LOGIC Use counts to test · percentages to explain

13 · Worked: limit a causal claim C1

Prompt. A survey finds that frequent study-room users report higher grades. Classify the evidence and explain one reason it does not identify a causal effect.

Each decision constrains the next



weak input → weak conclusion

Research question and design constrain every later inference.

STEP	WORKING
1	Classify the observed relationship as an **association**.
2	Name plausible confounder: prior motivation.
3	Explain path: motivation may increase room use *and* grades.
4	Match verb to survey design; do not claim an effect.

Answer. The survey provides explanatory evidence of association. Prior motivation could produce both greater study-room use and higher grades. Therefore reported room use is associated with grades in the surveyed sample, subject to unmeasured confounding.

COUNTERFACTUAL CHECK

```
What comparison shows what would have happened to the same units without the proposed cause?
```

14 · Causal-language traps C1 FAQ

WEAK ANSWER	WHY WRONG	UPGRADE
"Temporal order proves causation."	An earlier common cause may drive both.	Add association + alternatives audit.
"The study was random."	Mechanism unspecified.	Name sampling or assignment and inference supported.
"Control proves the cause."	Only measured confounders addressed.	State assumptions and residual confounding.
"Significant means causal."	Test evidence does not supply design.	Describe association; limit by design.

FOUR QUESTION TYPES

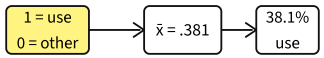
```
descriptive: what exists · explanatory: how variables move together · predictive: how well information anticipates unused outcomes · causal: what changes under intervention.
```

Prediction may work without identifying a cause. Causation needs a credible counterfactual comparison; observational explanation remains association unless stronger design evidence exists.

15 · Worked: binary mean C2

Prompt. Service use is coded 1 for use and 0 otherwise. Mean = 0.381. Interpret.

0/1 mean = share coded 1



name 1 · valid N · category

The arithmetic is only interpretable after identifying which category equals one.

STEP	WORKING
1	Identify 1 = service use.
2	0.381 × 100 = 38.1%.
3	Limit denominator to observations with valid data.
4	Interpret category membership, not amount of service.

Answer. Among observations with valid data, 38.1% are coded as service users. This is a proportion created by 0/1 coding, not an average quantity of service received.

16 · Measurement quick test C2 FAQ

CLUE	LEVEL/ISSUE	WHAT FOLLOWS
Names only	Nominal	Classify; do not rank/average codes
Order meaningful, gaps unknown	Ordinal	Median/IQR possible; equal-step arithmetic unjustified
Differences meaningful	Interval-ratio	Mean/SD possible after shape check
Stable repeated values	Reliability	Does not prove construct validity
Same label, different question/period	Comparability threat	Reconstruct operational definitions
Collapsed categories	New substantive boundary	Defend, preserve original, verify counts

OPERATIONAL-DEFINITION ANSWER

```
indicator + question/observation + response categories + recall period + scoring + missing rule + scale direction.
```

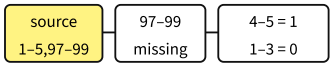
COMPOSITE SCORE MINI-AUDIT

- Which items enter, and do they cover the intended concept?
- Which items are reversed; what does a higher total mean?
- How are missing item responses handled and are items weighted?
- Would two datasets with the same label produce comparable scores?

17 · Worked: repair recode order C3

Prompt. Valid satisfaction = 1–5; 97–99 = nonresponse. An analyst defines high satisfaction as values ≥4 before handling missing codes.

Missing first, then threshold



new field → cross-tab → valid N

Thresholding first would falsely classify 97–99 as highly satisfied.

STEP	REPAIR
1 Diagnose	97–99 satisfy the faulty ≥4 condition.
2 Preserve	Keep the received field unchanged.
3 Missing first	Convert documented 97–99 to missing.
4 Recode	New indicator: valid 4–5 = 1; valid 1–3 = 0.
5 Verify	Cross-tab source × new; compare valid N with codebook.

Answer. The original rule misclassifies every nonresponse code as highly satisfied. Missingness must be resolved before the inequality, and the mapping must be auditable in a new variable.

18 · Pipeline diagnostics C3 FAQ

SYMPTOM	CHECK
Unexpected category	Codebook, raw frequency, transformation boundary
Result changes after predictor added	Specification and complete-case sample N
Collaborator cannot run code	Absolute paths, packages, file versions, hidden objects, order
Direction reversed	Reverse-coded items and "higher means…" statement
Manual fix needed	Encode the rule; rerun from unchanged source

REPRODUCIBILITY TEST

```
restart session → unchanged received files → documented entry script → cleaned data + tables + reported results, without manual intervention.
```

DATA-STRUCTURE CHECK

State what a row represents before analysis. If rows are persons, a household conclusion needs an explicit aggregation; if rows repeat person-years, keep both an identifier and time index. Then compare imported dimensions and types with documentation.

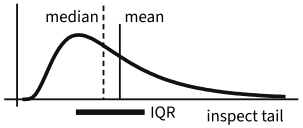
After each transformation, trace source values and inspect a frequency, cross-tabulation or summary. Check impossible categories, direction, missingness and valid N before modelling.

EXAM *A useful comment explains why a transformation/exclusion/ordering decision exists, not what obvious syntax says.*

19 · Worked: skewed care hours C4

Prompt. Weekly care hours have a long upper tail with plausible high values. Choose centre/spread and handle extremes.

Long upper tail: use robust pair



Lead with median + IQR; show shape and investigate, do not auto-delete.

STEP	WORKING
1	Identify right-skewed shape from the long upper tail.
2	Primary centre = median.
3	Primary spread = IQR.
4	Check high records against units, coding and related fields.
5	Retain plausible observations; interpret all values in hours.

Answer. Lead with median and IQR because they describe the typical case and middle half without allowing the upper tail to dominate. Use a histogram or box plot; retain plausible high values and explain their influence. Mean and SD may be supplementary if their sensitivity is stated.

20 · Distribution answer builder C4 FAQ

SENTENCE	REQUIRED CONTENT
1 Measure	Variable, unit, level, valid N
2 Centre/spread	Values + reason pair suits level/shape
3 Shape	Symmetry/skew, peaks, gaps or clusters from graph
4 Extremes	Which records checked; plausible/error; action
5 Limit	Sensitivity or exclusion effect where relevant

Never delete because a box plot marks a point. The plot flags investigation; documentation and substantive plausibility decide treatment.

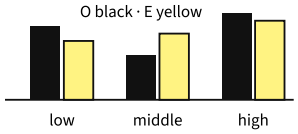
GRAPH	BEST RETRIEVAL CUE	DO NOT INFER ALONE
Bar chart	Category frequencies	Arithmetic distance between bars
Histogram	Quantitative shape, peaks, gaps, tails	Cause of the shape
Box plot	Median, IQR, possible extremes	Automatic error/deletion

If an exclusion is substantively justified, state the rule and compare the result with and without affected cases. This keeps sensitivity visible instead of hiding a consequential cleaning choice.

21 · Worked: expected count C5

Prompt. Row total = 180; column total = 240; valid N = 600. Find expected frequency under independence.

χ^2 adds observed–expected gaps



Larger observed–expected gaps contribute more after division by E.

CALCULATION

```
E = (180 × 240) / 600 = 43,200 / 600 = 72 cases
```

STEP	WORKING
1	Write row total × column total / N.
2	Substitute 180, 240 and 600.
3	Calculate 72; retain count units.
4	Interpret as the cell count implied by independence with observed marginals fixed.

Answer. 72 is neither forecast nor observed count. Pearson χ^2 compares O with E; it does not compare percentages inside the formula.

EXAM *Counts calculate; conditional percentages explain.*

22 · Read cells after omnibus C5

Largest cells locate evidence

small	largest	moderate
small	moderate	largest

contribution ≠ causal importance

Largest contributions locate evidence; O versus E supplies direction.

Squaring prevents discrepancies cancelling; division by E makes gaps comparable. A large contribution locates evidence against independence, but direction requires O versus E and conditional percentages.

CELL-TO-TABLE LOGIC

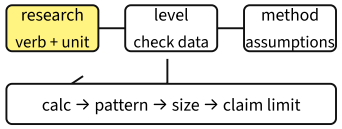
```
For every cell: find E → compare O with E → compute (O - E)²/E → sum to χ². Then use conditional percentages—not contribution size—to describe the social pattern.
```

AFTER χ^2	PURPOSE
Largest contributions	Locate where departure from independence concentrates
O − E sign	Show over-/under-representation relative to independence
Conditional percentages	Describe the substantive comparison
Cramer's V	Summarise bounded, nondirectional strength

23 · Method chooser C1–C5

PROMPT CLUE	CHOOSE	OUTPUT/CHECK
"What exists/how common?"	Descriptive summary	Level + valid N + suitable centre/frequency + spread/shape
"Do two categorical variables relate?"	Cross-tab + χ^2	Conditional % + assumptions + χ^2(df), p + Cramer's V
"Is X a cause?"	Causal audit	Association + temporal order + confounding + design
Data changed or inconsistent	Reproducibility audit	Codebook + source preservation + verification + analytic N

Prompt → defensible conclusion



Do not begin with software output; make the decision chain explicit.

24 · χ^2 report + trap bank C5 FAQ

FULL REPORT STEM

```
Among N valid [units], [conditional percentage contrast]. The variables [were/were not] statistically independent, χ²(df) = …, p = …. Cramer's V = … indicates [strength]. Cells … depart most from independence. This is association, not causation, because ….
```

TRAP	CORRECTION
Use % in E or χ^2	Use counts; percentages interpret pattern
Include marginals in r/c	Dimensions count non-marginal categories only
p = probability H_0 true	p concerns data at least this discrepant under H_0
Significance = large/important	Report conditional pattern + Cramer's V
Cramer's V gives direction	It is strength only, bounded 0–1
Merge sparse cells mechanically	Only if new category remains substantively meaningful
Largest cell "causes" result	Omnibus statistic aggregates all contributions

FINAL 30-SECOND AUDIT

1. Did I name the unit, variables and measurement levels?
2. Is denominator visible and valid N reported?
3. Did I state assumptions before interpretation?
4. Did I separate statistical evidence, magnitude and causation?
5. Does my final verb match the design?