Measurement, methods, statistics and history

Supported GRE teaching and original transfer practice

Measurement, methods, statistics and history

- Write operational definitions and judge **reliability** 信度 and **validity** 效度.
- Separate random sampling from random assignment.
- Describe data with averages and spreads.

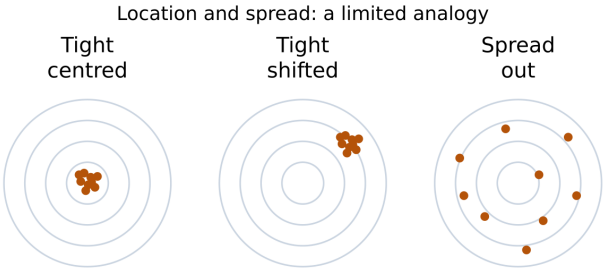
Measurement, methods, statistics and history (2)

- Interpret p-values, effect sizes and **statistical power** 统计检验力.
- Apply research ethics.

Measurement

- An operational definition says exactly how a construct is measured: "anxiety" as a score on a named scale, or as heart rate during a speech.
- Reliability is consistency; validity is evidence that the scores mean what is claimed.
- Naming a questionnaire "memory test" does not make it one.

Figure: Measurement

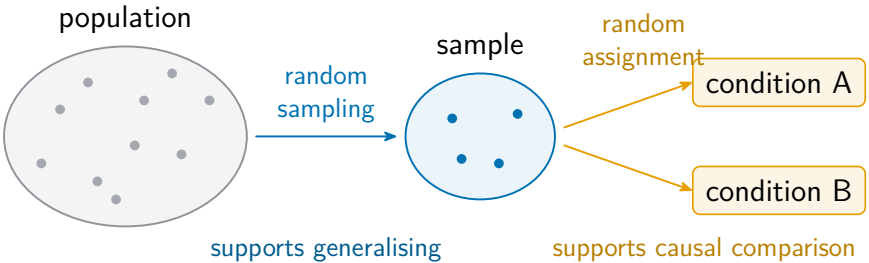


- Tight grouping is like reliability, being on target like validity: a test can be reliable but not valid
- Match each part of the diagram to the text.

Sampling and assignment

- Random sampling decides who enters the study. It supports generalising to the population.
- Random assignment decides which condition each participant gets. It supports causal conclusions.
- Volunteers randomly assigned to conditions give a fair comparison, but only for people like the volunteers.

Figure: Sampling and assignment



- Two different random steps answer two different questions.
- Match each part of the diagram to the text.

Describing data

- Mean, median and mode describe the centre. With a long right tail (such as income), the mean is pulled above the median.
- Standard deviation describes spread. In a normal distribution about 68% of scores lie within 1 SD of the mean.
- A correlation runs from -1 to $+1$; it describes a linear association, not a cause.

exercises.lol exercises.com.cn

Inference

- A p-value is the probability of data at least this extreme if the null hypothesis were true. It is not the probability that the null hypothesis, or a theory, is true.
- An effect size, such as Cohen's d (the difference in means divided by the SD), describes how large an effect is. Significance does not.
- Power is the probability of detecting a real effect. Small samples have low power, and their significant results tend to overestimate effects.

Ethics and history

- Ethics requires informed consent, the right to withdraw, minimal risk, confidentiality and debriefing; review boards approve studies first. Deception must be justified and explained afterwards.
- History questions match schools to their questions and methods: structuralism (elements of experience), functionalism (purposes), behaviourism (observable behaviour), Gestalt (organised wholes), cognitive psychology (information processing).

Worked example: a reliable test

- It has test–retest reliability.
- It lacks validity evidence for the memory interpretation.
- Validity needs more: correlations with other memory measures, and weak correlations with reading once memory is controlled.

Worked example: misreading p

- Wrong: p is the probability of such data if the null were true, not the probability that the theory is true.
- Wrong: p does not measure size. Report an effect size with a confidence interval.
- With a small sample and many analyses tried, the result may not replicate.

Common errors

- Treating random assignment as random sampling.
- Treating reliability as validity.
- Reading p as the probability that a hypothesis is true.

Common errors (2)

- Calling a significant result large or important without an effect size.
- Using the mean to describe a strongly skewed distribution.

Key terms

- **reliability**: The consistency of a measure.
- **validity**: Evidence that a measure supports its intended interpretation.
- **statistical power**: The probability that a study detects an effect that really exists.

Check yourself

- Most students are aged 18–20, but a few are over 50. Which is larger, the mean age or the median?
- Which random step supports causal conclusions?
- A huge study finds a significant but tiny effect. What should be reported as well as p?