Reliability and Validity

Downloaded form Slide Share

Lecture objectives

- To review the definitions of reliability and validity
- To review methods of evaluating reliability and validity in survey research
- EBM prospective

Reliability

Definition

- The degree of **stability** exhibited when a measurement is **repeated** under identical conditions

- Lack of reliability may arise from divergences between **observers** or **instruments** of measurement or **instability** of the attribute being measured

(from Last. Dictionary of Epidemiology)

Assessment of reliability

- Reliability is assessed in 3 forms
  1. Test-retest reliability
  2. Alternate-form reliability
  3. Internal consistency reliability

Test-retest reliability

- Most common form in surveys
- Same respondents complete a survey at **two** different points in **time**
- Usually quantified with a **correlation coefficient (*r* value)**
- *r* values are considered good if $r \geq 0.70$

Test-retest reliability (2)

- If data are recorded by an observer, you can have the **same observer** make **two** separate measurements
- The comparison between the two measurements is **intra*a*observer** reliability
- What does a difference mean?

Test-retest reliability (3)

- You can test-retest **specific questions** or the **entire** survey instrument
- Variables likely to change over a short period of time, such as energy, happiness, anxiety
- Test-retest over very short periods of time

Test-retest reliability (4)

- Potential problem with test-retest is the **practice effect**
  - Individuals become familiar with the items
- What effect does this have on your reliability estimates?
  - It inflates the reliability estimate

Alternate-form reliability

- Use differently worded forms to measure the same attribute
- Questions or responses are reworded
- Or their order is changed
- To produce two items that are similar but not identical

Alternate-form reliability (2)

- Two items address:
  - The same aspect of behavior
  - Same vocabulary
  - Same level of difficulty
  - Items should differ in wording only
- It is common to simply change the order of the response alternatives
  - This reduces practice effect

Example: Assessment of depression

Circle one item

Version A:

During the past 4 weeks, I have felt downhearted:

Every day	1
Some days	2
Never	3

Version B:

During the past 4 weeks, I have felt downhearted:

Never	1
Some days	2
Every day	3

Alternate-form reliability (3)

- You could also **change the wording of the *response*** alternatives without changing the meaning

Example: Assessment of urinary function

Version A:

During the past week, how often did you usually empty your bladder?

1 to 2 times per day

3 to 4 times per day

5 to 8 times per day

12 times per day

More than 12 times per day

Example: Assessment of urinary function

Version B:

During the past week, how often did you usually empty your bladder?

- Every 12 to 24 hours
- Every 6 to 8 hours
- Every 3 to 5 hours
- Every 2 hours
- More than every 2 hours

Alternate-form reliability (4)

- You could also change the actual wording of the **question**
  - The two items must be equivalent
  - Items with different degrees of difficulty do not measure the same attribute
  - What might they measure?
    - Reading comprehension or cognitive function

Example: Assessment of loneliness

Version A:

How often in the past month have you felt alone in the world?

Every day

Some days

Occasionally

Never

Version B:

During the past 4 weeks, how often have you felt a sense of loneliness?

All of the time

Sometimes

From time to time

Never

Example of nonequivalent item rewording

Version A:

When your boss blames you for something you did not do, how often do you stick up for yourself?

All the time

Some of the time

None of the time

Version B:

When presented with difficult professional situations where a superior censures you for an act for which you are not responsible, how frequently do you respond in an assertive way?

All of the time

Some of the time

None of the time

Alternate-form reliability (5)

- You can measure alternate-form reliability at the **same timepoint** or **separate timepoints**
- If large enough sample:
  - You can split it in half and administer one item to each half
  - Then compare the two halves
  - This is called a split-halves method
  - Can split into thirds and administer three forms of the item

Internal consistency reliability

- Applied to **groups of items** that are thought to measure **different aspects** of the **same concept**
- Cronbach's coefficient alpha
  - Measures internal consistency reliability
  - It is a reflection of how well the different items complement each
  - Interpret like a correlation coefficient (≥0.70 is good)

Example: Assessment of physical function

	Limited a lot	Limited a little	Not limited
Vigorous activities, such as running, lifting heavy objects, participating in strenuous sports	1	2	3
Moderate activities, such as moving a table, pushing a vacuum cleaner, bowling, or playing golf	1	2	3
Lifting or carrying groceries	1	2	3
Climbing several flights of stairs	1	2	3
Bending, kneeling, or stooping	1	2	3
Walking more than a mile	1	2	3
Walking several blocks	1	2	3
Walking one block	1	2	3
Bathing or dressing yourself	1	2	3

Calculation of Cronbach's coefficient alpha

Example: Assessment of emotional health

During the past month:	Yes	No
Have you been a very nervous person?	1	0
Have you felt downhearted and blue?	1	0
Have you felt so down in the dumps that nothing could cheer you up?	1	0

Results

Patient	Item 1	Item 2	Item 3	Summed scale score
1	0	1	1	2
2	1	1	1	3
3	0	0	0	0
4	1	1	1	3
5	1	1	0	2
Percentage positive	3/5=.6	4/5=.8	3/5=.6	

Calculations

Mean score=2

Sample variance= $\frac{(2-2)^2+(3-2)^2+(0-2)^2+(3-2)^2+(2-2)^2}{(5-1)} = 1.5$

$$CC\ alpha = \left[1 - \frac{\sum (\%pos)_i (\%neg)_i}{Var}\right]\left[\frac{k}{k-1}\right]$$

$$= \left[1 - \frac{(.6)(.4)+(.8)(.2)+(.6)(.4)}{1.5}\right]\left[\frac{3}{2}\right] = 0.86$$

Conclude that this scale has good reliability

Internal consistency reliability (2)

- If internal consistency is
  - low: You can add more
  - items

    Re-examine existing items for clarity

Interobserver reliability

- How well ***two*** evaluators agree in their assessment of a variable
- Use **correlation coefficient** to compare data between observers
- May be used as **property of the test** or as an **outcome variable**

Validity

Definition

How well a survey measures what it sets out to measure

Assessment of validity

- Validity is measured in four forms
  - Face validity
  - Content validity
  - Criterion validity
  - Construct validity

Face validity

- Cursory review of survey items by untrained judges
  - Ex. Showing the survey to **untrained individuals** to see whether they think the items look okay
  - Very casual, soft
  - Many don't really consider this as a measure of validity at all

Content validity

- **Subjective** measure of how appropriate the items seem to a set of reviewers who have **some knowledge** of the subject matter
  - Usually consists of an organized review of the survey's contents
  - Still very qualitative

Criterion validity

- Measure of how well **one instrument** stacks up **against another instrument** or predictor
  - Concurrent: assess your instrument against a "***gold standard***"
  - Predictive: assess the ability of your instrument to forecast future events, behavior, attitudes, or outcomes
- Assess with correlation coefficient

Construct validity

- Most **valuable** and most **difficult** measure of validity
- Basically, it is a measure of how meaningful the scale or instrument is ***when it is in practical use***

Construct validity (2)

- ***Convergent***: Implies that **several different methods** for obtaining the **same information** about a given trait or concept produce similar results
  - Evaluation is analogous to alternate-form reliability ***except*** that it is more theoretical and requires a great deal of work-usually ***by multiple investigators with different approaches***

Construct validity (3)

- ***Divergent***: The ability of a measure to estimate the underlying truth in a given area-must be shown not to correlate too closely with similar but ***distinct concepts or traits***

EBM Prospective

Introduction

- Three Steps in Using Medical Literature Articles :
  - Are the results of the study valid?
  - What are the results?
  - How can I apply these results to patient care?

Introduction

- Four types of
  - papers: Therapy
  - Diagnostic Intervention
  - Prognosis
  - Systematic review

Therapy

- Study design: RCT
- Were Patients Randomized?
- Was Randomization Concealed?
- Were Patients Analyzed in the Groups to Which They Were Randomized?
  - Intention to treat analysis

Therapy

- Were Patients in
  - The Treatment
  - And Control Groups
  - Similar With Respect to Known Prognostic Factors?
- Were Patients Aware of Group Allocation?

Therapy

- Were Clinicians Aware of Group Allocation?
- Were Outcome Assessors Aware of Group Allocation?
- Was Follow-up Complete?
- Was Follow-up Long Enough?

Diagnostic Intervention

- Study Design: Cross-sectional
- Was there an independent, blind comparison with a reference standard?
- Spectrum of patients
- Did the results of the test being evaluated influence the decision to perform the reference standard?
- Were the methods description permit replication?

Prognosis

- Study design: Cohort
- Was a
  - Defined,
  - representative sample of patient
  - assembled at a common point in the course of their disease?
- Inception Cohort; early
- Late stage prognosis
- Patient equal in all prognostic factors
  - Stratified analysis?
- Follow up complete and long enough
- Valid and reliable data collection

Thank You