Classical Item Analysis

PSY 600K
Dr. Alyssa Gibbons
March 27, 2018

Agenda

- Properties of items:
  - Item difficulty
  - Item discrimination
    - Semistandardized covariances
    - Point-biserial correlation
    - Biserial correlation
  - Item discrimination, reliability, & validity
    - Choosing items
  - Inter-item correlations
    - Tetrachoric correlations

Classical Item Analysis

- "Legacy" material – many of these techniques are not in common use anymore.
- BUT you need to know about them! Why?
  - Some are still used (and still useful).
  - Others were used in the development of scales we use regularly today.
  - Many are precursors to IRT and other elements of modern test theory.
- Useful for **selecting** items to make up a test.
  - We can use different indices in different ways depending on our purpose.

Item Difficulty

- We've talked about this before… the **mean** of an item is an index of its **difficulty.**
  - Binary items: probability of passing the item.
  - Continuous items: the item mean.
- Note that **higher** values = **easier** items.
  - But "item easiness" sounds funny.
- We can still think about difficulty for items with no "right" answer.
  - Extreme-ness of the item.
  - How "difficult" would it be for an average person to strongly endorse the item?

Choosing Items by Difficulty

- Often, you want a **range** of item difficulty.
  - So that you can measure across the range of the construct.
- Some recommend an average difficulty of .50.
  - The key word here is **average** – you should have a range around this average.
  - Why don't you want all your items to have a difficulty of .50?
- At times, you may want to focus your scale on a specific portion of the construct spectrum and choose item difficulties accordingly.
  - When would you do this? Can you give an example?

Item Discrimination

- Has **absolutely nothing** to do with race, gender, bias, etc., etc.
- Discrimination in the old-fashioned sense – sensitive to distinctions between people.
  - Think "a discriminating palate."
- In other words, a **highly discriminating** item **gives us a lot of information about the person's standing on the construct.**
- There are lots of ways we can calculate item discrimination – all are based on this one underlying idea.

Item Discrimination Parameters

- 1. *d* Index.
  - Simple procedure:
    - Rank-order your sample by total test score and divide it into thirds.
    - Find the average score for the top 1/3 and the average score for the bottom 1/3.
    - Find the difference between the two.
  - Advantages: computationally easy (except for the sorting part).
  - Disadvantages: loses lots of information! Tedious to calculate.
  - Yet some older tests were developed this way and some older test users still expect to see this information.

Item Discrimination Parameters

- 2. Item-total covariance.
  - Covariance of each item score with the total test score.
- 3. Semi-standardized covariance.
  - Covariance of item score with the standardized total test score.
  - But why would you use either of these when you can just use the…

Item Discrimination Parameters

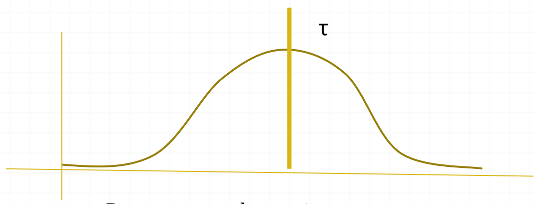
- 4. Item-total correlation!
  - Standardized with respect to both the factor and the item.
  - When the items are binary, we often call this the **point-biserial** correlation.
  - But it is **absolutely identical** to the ordinary Pearson correlation.
    - A special case – we can use a simpler formula to calculate it – but it's the same thing.

Biserial Correlations

- The point-biserial correlation is just an ordinary Pearson correlation… but the **biserial correlation** is completely different.
  - Still used with binary items, but…
  - Different theoretical model for what the binary item means.
- The biserial correlation assumes that although the *item* only has 2 possible scores, the underlying *variable* or *response tendency* is continuous.
  - Ex: "I often feel hopeless." YES NO
  - Are people either depressed or not?
  - Or is there a range of depression such that people who are above a certain threshold will answer yes?

Biserial Correlations

- So instead of a truly binary item *X*, we really have a continuous response tendency *X**:



Response tendency →

- Our item "cuts" the response tendency at some threshold τ.
  - If $X^* > \tau$, the response is 0.
  - If $X^* > \tau$, the response is 1.

Biserial Correlations

- The biserial correlation is the correlation between *X** and the total test score.
  - In other words, it "un-dichotomizes" the item response, restoring the "lost" variance.
  - Not as shady as it sounds – if we assume *X** is normally distributed, we can use basic calculus to do the conversion.
  - You would not do this by hand – there are converter programs for this!.
- There is a mathematical relationship between the point-biserial and the biserial correlations.
  - The point-biserial is never more than .798 x biserial.
  - This is true only when the item difficulty = .50; at other item difficulties, the point-biserial underestimates the biserial by a larger amount.
- Does this suggest why you might bother with the biserial?

Item Discrimination Parameters

- So far, we've talked about item discrimination parameters in relation to the *total test score*.
- But this includes the item we care about, so it's inflating our correlation somewhat.
  - More of a problem with shorter tests.
- So we could calculate **all** of these indices using the **remainder score** instead.

So What?

- Does it matter which item discrimination parameter we use?
- Implicit in all of these measures is an assumption that items are *homogeneous* and that the scale as a whole is reasonably reliable.
  - So assess these things first!
- Item-test correlations **converge on the standardized factor loadings.**
  - Convergence is better when we have more items.

Selecting Items

Selecting Items

- We want to remove **poor** items.
  - Loadings < .30.
  - Negative correlations with other items.
  - Inappropriate difficulty.
- Sometimes, we want to keep the **best** items.
  - How do we choose? Lots of factors to consider.
- Common strategies:
  - Select items to maximize alpha.
  - Select items to maximize prediction of an outcome.
  - **You can't do both of these at once!**
    - And both have substantial limitations.

Item-Total Correlations and Total Test Variance

- We know already that the variance of the total test score (σ^2_Y) is the sum of all the variances and covariances.
- The sum of any item's variance + covariances with other items (its row or column in the var/cover matrix) is the item-total covariance.
- This means that the total test variance can be written as the sum of the item-total covariances.
  - So having items with high discrimination increases your total test variance! (and variance is a good thing).

Implications

- We can extend this to rewrite our formula for alpha in terms of item-total covariances.
  - And maximize alpha by choosing the items with the highest item-total covariances.
  - "Alpha if item deleted…"
  - BUT…
- This only works if your items fit a single-factor model!
  - Otherwise, you will select items that maximize reliability for **one** factor… and not select the others.
  - Potentially losing key pieces of your construct along the way.

Maximizing Prediction

- If we standardize both our predictor Y and our criterion V, the correlation between the Y and V can be written as the ratio of the sums of the semi-standardized covariances of all of our items with Y and V:
  - $\rho_{YV} = \frac{\sigma_{x_j z_v}}{\sigma_{x_j z_Y}}$
- We maximize ρ_{YV} by choosing items that have large semistandardized covariances with V compared to their semistandardized covariances with Y.

Reliability/ (predictive) Validity Tradeoff

- To maximize alpha, we choose items that have large semistandardized covariances with the total test score.
- To maximize prediction, we choose items that have large semistandardized covariances with the criterion.
- Do you see the contradiction?
- Recommendation: let your purpose drive your process.
  - For prediction only, choose based on the criterion and don't worry about measuring a construct.
  - To measure a construct, choose homogeneous/reliable items first and assess predictive validity later.

Item Information

- We can calculate a statistic called **item information** from the results of our factor analysis:
  - $I(x_i) = \frac{\lambda^2}{\psi^2}$
  - Square the loading; divide by the uniqueness.
    - Don't square the uniqueness again!
- The sum of these values across all items = $1/\sigma^2_E$, or the reciprocal of the error variance.
  - So information = the amount the item contributes to reducing error!
- If you want to select the best items to measure a construct, **information** is a sensible and justifiable way to do it.

Interitem Relationships

Interitem Correlations

- As we've seen, useful in and important for factor analysis!
- Problematic for binary items:
  - Pearson correlation between a binary item & a continuous item will **underestimate** the real relationship (thus the biserial correlation).
- Correlation between 2 binary items is even messier!
- Old approach: *phi over maximum phi*
  - *Phi* = Pearson correlation between 2 items.
  - This **must** be < 1 if the items have different difficulty.
  - But we can estimate the maximum possible correlation if we know the difficulty parameters; thus, phi-over-maximum-phi.
  - Ultimately, not very successful – no longer commonly used.

Tetrachoric Correlations

- Extending the logic of the biserial correlation to 2 binary items.
  - Now we have 2 continuous response tendencies X^*_j and X^*_k, both dichotomized by binary items.
  - We assume that X^*_j and X^*_k are standardized (mean 0, variance 1) and have a bivariate normal distribution.
- ~~Very difficult~~ Impossible to compute by hand!

Tetrachoric Correlations

- Why do we use it?
  - Estimates the “true” correlation between the latent response tendencies – not attenuated by dichotomizing responses.
  - Tetrachoric correlations will be bigger than product-moment (phi) correlations.
  - Tetrachoric correlations *can* equal 1 or -1, even if the items are not exactly equal in difficulty.
    - To equal 1, the probability of getting 1 item right and the other wrong must be 0 (in one direction or the other; not necessarily both).
    - To equal -1, the probability of getting both right or both wrong should equal 0 (again, in one direction or the other).
  - This is the basis for *item factor analysis*, one way to estimate parameters in IRT.

Questions?

For next time: Validity: Content & Response Processes

Read: R & M 8.1 – 8.3