Reliability & Errors

Dewayne E Perry
ENS 623
Perry@ece.utexas.edu

Errors in Measurement

- All measurements subject to fluctuations
  - Affects reliability and validity
- Reliability : constancy or stability
- Validity : appropriateness or meaningfulness
- Reliability coefficient : degree that what is measure is free from measurement fluctuation
- Observer agreement coefficient : objectivity and repeatability of rating procedures
- Random vs systematic errors
  - Random: cancel out on average over repeated measurements
  - Systematic: do not cancel out
- Systematic errors are known as Biases
  - Main concern of internal validity
  - Can compensate for known biases
    - Eg, in astronomy, known biases of observations

Reliability Criteria

- Principle criteria of test reliability
  - Test-retest reliability
  - Reliability of test components
    - ie internal consistency
- Stability (Test-Retest)
  - Temporal stability from one session to the next
  - Problem: distinguishing between real change and the effect of memory
    - Too short an interval between: memory effect possible
    - Too long an interval: real changes may interfere
    - May use changes to test sensitivity of tests

Reliability (of Test Components)

- Internal consistency reliability
- Depends on the average of Intercorrelations among all the single test items
- Coefficients of internal consistency increase as the number of test items goes up (if the new items are positively correlated with the old)
- The more items, the more internally consistent the test; if other relevant factors remain the same
  - Not always the same for different length tests
  - Boredom & fatigue can result in attenuation

Spearman Brown Formula

- $$R = \frac{n\bar{r}}{1+(n-1)\bar{r}}$$
  - R is the reliability coefficient
  - n is the factor by which the test is lengthened
  - $\bar{r}$ is the mean correlation among all items
- Suppose mean correlation is .50, determine reliability of test for twice, thrice:
  - 2(.50)/[1+(2-1).50] = .667 – increase *R* by a third
  - 3(.50)/[1+(3-1).50] = .75 – increase *R* by half
- Other Tests
  - *Kuder-Richardson formula 20 (K-R 20)*
    - Used to measure internal consistency when items of the test are scored 1 if marked correctly, 0 otherwise
  - Cronbach's *alpha coefficient*
    - Employ the use of analysis of variance procedures for estimating reliability of test components

Acceptable Reliability

- Need to evaluate whether low validity is due to low reliability
  - If so can it be improved by adding items
- What is the acceptable range of reliability?
  - Depends on situation and nature of variable being measured
  - For clinical testing *R* = .85 is considered as indicative of dependable psychological tests
  - In experimental research, accept much lower *R*
- Problem:
  - Reliability test reflects both individual differences and measurement fluctuations
  - If everyone alike, the only differences are in error variations
  - Hence, lower reliability where fewer differences
    - Eg, IQ at highly selective where students are more similar than at a public university

Acceptable Reliability

- Reliabilities of major psychological tests
  - MMPI – MN Multiphasic Personality Inventory
  - WAIS – Winchester Adult Intelligence Scale
  - Rorschach inkblot test
- MMPI and Rorschach most widely used, WAIS used as control
- Internal consistency – all three acceptable
  - WAIS *R* = .87, 12 studies with 1759 subjects
  - MMPI *R* = .84, 33 studies with 3414 subjects
  - Rorschach *R* = .86, 4 studies with 154 subjects

Acceptable Reliability

- Stability – respectable scores
  - Fewer studies available
  - WAIS as .82 – 4 studies with total N = 93
  - MMPI as .74 – 5 studies with total N = 171
  - Rorschach as .85 – 2 with total N = 125
    - WAIS/Rorschach difference not significant;
    - MMPI/Rorschach and WAIS/MMPI difference is highly significant
- Internal consistency usually higher than stability
- Problem of inter-rater reliability
  - Use test reliability measures to assess their aggregate internal consistency
  - Arises in SWE in classifying faults, root causes, evaluating designs, reviewing papers, evaluating developers, etc

Effective Reliability of Judges

- Problem: correlation of .60 between the ratings of two judges tells us only the reliability of either single judge in this situation
- For aggregate or effective reliability, use approach as in “how many test items”
  - Use *Spearman-Brown* where
    - *n* is the number of judges and
    - $\bar{r}$ is the mean correlation among them
  - Aggregate reliability of
    - 2 judges: 2(.60)/[1+(2-1).60] = .75
    - 3 judges: 3(.60)/[1+(3-1).60] = .82
  - The more judges, the higher the reliability
  - Table 3.3 very useful for planning/analysis

% Agreement & Reliability

- Many se percent agreement as an index of reliability
  - A agreements and D disagreements
    - %: [A/(A+D] x 100
    - Net: [(A-D)/(A+D)] x 100
- Misleading – fails to differentiate between accuracy and variability
- Better - use the product moment correlation phi
  - can be computed from the chi-square

ANOVA & Reliability

- Sometimes need more than 2-3 judges
- Excellent approach based on analysis of variance
  - Tedious to do average of large number of correlations of previous approach
  - Assess how well judges are able to discriminate among sampling units (*MS* persons) minus the judge's disagreements (*MS* residuals) controlling for rating bias or main effect, divided by a standardizing quantity

$$R_{est} = \frac{MS_{persons} - MS_{residuals}}{MS_{persons}}$$

$$\bar{r}_{est} = \frac{MS_{persons} - MS_{residuals}}{MS_{persons} + (n-1)(MS_{residuals})}$$

Replication & Reliability

- Reliability in research implies generalizability as indicated by replicability (repeatability) of the results
  - ★ Across time (test-retest reliability)
  - ★ Across different measurements, observers, or manipulations (reliability of components)
  - ★ Note that may not be possible to repeat and authenticate every observation with perfect precision

Replication Factors

- Same experiment can never be repeated
  - At very least everyone is older
- 3 important factors affect the utility of a replication as an indicator of reliability:
  - When the replication is conducted
    - Earlier better than later; 2nd doubles our info
  - How the replication is conducted
    - The more imprecise, the more generalizability
  - By whom is the replication conducted
    - Independence is critical – rule out pre-correlations
    - Selection and training considerations
    - Correlated observers a critical problem in all fields

Statistical Analysis

- Rationale
  - Essential aspect of the rhetoric of justification in behavioral sciences evaluation, defense and confirmation of claims of truth
  - Traditional ways to shore up facts and inductive inferences
  - Imposes a sense of order and lawfulness
- 4 problems in the methodological spirit of statistical data analysis
  - Dichotomous decisions on significance
  - Low power
  - Significance as defining results
  - Over emphasis on single studies

Statistical Analysis

- Over reliance on dichotomous on significance testing decisions
  - Anti-null if p is not greater than .05
  - Pro-null if p is greater than .05
  - .05 α considered to be axiomatic: on the one side joy; on the other side ruin
  - Comes from the fact we ought to avoid Type I errors
  - A convenient and stringent enough fail safe standard
  - Not axiomatic: strength of evidence is continuous on the magnitude of p
- Tendency to do many research studies in situations of low power
  - Often ignore the extent to which the sample size is stacking the deck against themselves
  - May be considered to be to complicated
  - Seminal work of Cohen on Power in the 60s – has resurfaced as an important issue

Statistical Analysis

- Defining results in terms of significance alone
  - Need to consider effect size estimation procedures
  - Both when p is significant as well as when not significant
  - Guides our judgment about sample size
  - Significant p values should not be interpreted as reflecting large effects or the practical importance of the results
- Over emphasis on single studies at the expense of accumulating results
  - Accumulating results critical for increasing weight of evidence
  - Evaluate impact on things other than p value – use multiple criteria
  - Make more use of meta-analysis
  - Accumulate data via meta-analysis, not just results
  - Often need to compute effect size and significance where it does not exist

Methodological Problems

- 4 problems on methodological substance
  - Omnibus tests
  - Need for contrasts
  - Misinterpretation of interaction effects
  - Hidden nesting
- Omnibus tests
  - In SWE, to much reliance on shotgun metrics
  - Need to ask focused questions
  - Focused test more relevant
  - Omnibus tests
    - Of dubious practical or theoretical significance
    - Effect size estimates are of doubtful utility
- Need for contrasts
  - Specific predictions are analyzed by comparing them to the data
  - Temporal progression levels are emphasized in in contrast approach
  - Increased statistical power results from contrasts
    - Avoid Type II error

Methodological Problems

- Misinterpretation of interaction effects
  - Mathematical meaning of interaction effects is unambiguous
  - But only a tiny fraction of results interpreted correctly
  - May be due to lack of correspondence between the meaning of "interaction" in the analysis of variance model and its meaning in other discourse
- Hidden nesting
  - Concealed non-independence of observations
    - results from sampling without regard to sources of similarity in the persons sampled
  - Significance and effect size estimation become problematic
  - Samples too similar
    - Usual assumptions underlying analysis do not hold
  - Degrees of freedom fall somewhere between the number of people and the number of groups of people in the study

Re-Emphasis

- There will almost always be two kinds of information we want to have for each of our research questions:
  - The size of the effect and
  - Its statistical significance
- Magnitude of significance test = size of effect x size of study
  - Significance will increase for any given size of study
  - For any given size of effect and for any give size of study, there will be a corresponding test of significance
- Much of the analysis we will look at is about how to determine these three elements in a study

Errors Revisited

- One reality
  - H0 (Null Hypothesis) is True
  - H1 (Alternative Hypothesis) is False
  - There is no relationship, no difference, theory is wrong
- We accept H0, reject H1
  - Match reality
  - Confidence level: $1-\alpha$ (eg, .95)
    - The odds of saying there is no relationship or difference when in fact there is none
    - The odds of correctly not confirming our theory
    - Ie, 95 time out of 100 when there is no effect, we will say there is none.
- Type I Error: we reject H0, accept H1
  - Contradict reality – say there is a relationship when there is none
  - Significance level: α (eg, .05)
    - The odds of saying there is a relationship or difference when there is none
    - The odds of confirming our theory incorrectly
    - 5 times out of 100, when there is no effect, we will say there is
    - We should keep this small when we can't afford/risk wrongly concluding our treatment works

Errors Revisited

- The other reality
  - H0 (Null) is False
  - H1 (Alternative) is True
  - There *is* a relationship, *is* a difference, and our theory *is supported*
- Type II Error: we *accept* H0, *reject* H1
  - Contradict reality – say there is no relationship when there is one
  - β (eg, .20)
    - The odds of saying there is no relationship or difference when in fact there is one
    - The odds of not confirming out theory when it is true
    - 20 times out 100, when there is an effect, we will say there isn't
- We *accept* H1, *reject* H0
  - Match reality
  - Power: $1-\beta$ (eg, .80)
    - The odds of saying there is a relationship or difference when there is one
    - The odds of confirming our theory correctly
    - 80 times out 100 when there is an effect we will say there is
    - We generally want this to be as large as possible

Decreasing Errors

- Decrease Type I Error by setting a more stringent α
  - Eg, .01 instead of .05
  - Decreasing Type I increases the likelihood of Type II Error
- Decrease Type II Error by setting less stringent α
  - Eg, .10 instead of .05
- Seek a balance between the two
  - As Type I goes up, Type II goes down and vice versa

Purpose of Power Analysis

- Planning of research
  - Determine size of sample needed
  - To reach a given α level
  - For any particular size of effect expected
- Evaluation of research completed
  - Determine if failure to detect an effect at a given α is primarily due to too small a sample
- Level of Power determined by
  - Statistic used to determine the level of significance
  - Level of α selected, size of the sample, size of the effect
- Increasing Power an be achieved by
  - Raising the level of significance required,
  - Reducing the standard deviation,
  - Increasing the magnitude of the effect by using strong treatments, and
  - Increasing the size of the sample

Example

- X compares OO programming against standard programming randomly assigning 40 programmers to use OO and 40 as the control group
  - The OO treatment programs have significantly fewer bugs
  - Using *t* test (comparing means), $t(78) = 2.21$, $p < .05$
- Y is skeptical and replicates X's work
  - Assigns 10 programmers to each
  - Results: $t(18) - 1.06$, $p > .30$
  - Y claims X results unrepeatable
- Misleading conclusions
  - Y's results in the same direction as X's
  - Y's effect size same as X's ($1/2\sigma = 2t / \sqrt{df}$)
  - Y's sample size too small: X's power = .6, Y's power = .2

Effect Size (ES)

- Effect Size: *standardized measure of the change in the dependent variable as a result of the independent variable*
- Standardization of effect size is done in the simplest case by dividing the change in the dependent measure by the standard deviation of the control group
- If ES=1, the experimental and control results differ by 1 standard deviation
- Effect Sizes are usually less than 1
- Cohen 1988 argues
  - Small effect size = 0.2
  - Medium effect size = 0.5
  - Large effect size = 0.8
- Enables us to compare the effects in different studies of the same phenomena
- Enables us to combine results from different studies in meta-analyses

Example

- Comparison:
  - ★ Treatment: 8 designers, design method X
  - ★ Control: 8 designers, std design method Y
- Results in terms of errors:
  - ★ Treatment: 5 6 9 4 8 3 7 6
  - ★ Control: 10 11 10 9 9 8 9 14
- Means:
  - ★ Treatment: 6
  - ★ Control: 10
- Standard deviations
  - ★ Calculate sum of squared deviations from the mean via shortcut formula:

$$\sum x^2 - \left(\sum x\right)^2 \backslash N$$

Example

- Treatment:
  - Squares: 25, 36, 81, 16, 64, 9, 49, 36
  - Sum = 48, sum of squares = 316
  - 316 – 2304/8 = 316 – 288 = 28
  - Std dev is $\sigma = \sqrt{(28/7)} = \sqrt{4} = 2$
- Control:
  - Squares: 100, 121, 100, 81, 81, 64, 81, 196
  - Sum = 80, sum of squares = 824
  - 824 – 6400/8 = 824 – 800 = 24
  - Std dev is $\sigma = \sqrt{(24/7)} = \sqrt{3.53} = 1.85$
- Effect size d = mean 1 – mean 2 / σ
  - (6 – 10) / 1.85 = 2.16
  - A very large effect (Cohen: 0.8 is a large effect)

Power Tables

- Cohen 1969, 1977, 1988
  - Comprehensive, elegant and useful discussion of power analysis in behavioral research
  - Defines small, medium and large effects for 7 statistics from *t* to *F*
  - Tables provide sample sizes vs power and significance

Neglect of Power

- Behavioral researcher faces a high risk of committing Type II errors
  - For medium effect sizes and α = .05 the odds are better than 50:50 that the null hypothesis would not be rejected when its false
  - Since Cohen's work, situation has gotten worse apparently
  - Continue to work at low power
  - Continue to rate Type I errors as more significant than Type II errors

Neglect of Power

- Assessing relationship of Type I vs Type II errors
  - Use ratio β/α
    - Remember β is the likelihood we will make a Type II error, α the likelihood of making a Type I error
  - Eg, α = .05 and power = .40,
    - β/α = .6/.05 = 12, ie Type I errors are considered to be 12 times more serious than Type II
  - What would we need to do if we wanted α = .05 and power = .95, β/α = .05/.05 = 1
    - ie, consider I & II equally serious