Psychology 253: Statistical Theory, Models and Methodology

Instructors: Steph, Dan, Ewart

Topics include: `lm()`, `glm()`, `lmer()`; reliability, factor analysis, penalised regression, SEM

Texts: *Howell; articles, Wiki*

Package: *R (**3.1.3**)*

Work: Group work encouraged on **HW**, but write up your **own solutions.** Use of solutions from previous Psy 253 classes **not** allowed. Take-home **Midterm** and **Final** must be your **own work**.

Hope: To discuss 'modern' methods, 'modern' applications of off-the-shelf methods, odd questions from my 'consulting'. At times, a workshop atmosphere?

HW-1, Sec 1 on Fri

- HW-1, due 4/7: on Cohen's *kappa*, an index of the **reliability** of coding *qualitative* data; and Pearson's *r*, another index for *quantitative* data.
- Our 1st topic is **Qualitative Research**.
- Use the R packages, 'irr' and 'psych', (or SPSS: Analyze > Scale …).
- This week's Section on Friday 11-12 in 420-358 will look at HW-1, and include, if necessary, a review of R.

Outline of Module

- Examples of qualitative research: values, thoughts
- Goals: validity, reliability (= consistency, agreement, accuracy)
- Data formats: ‘scores’ vs ‘counts’
- Cohen’s κ: reliability = agreement corrected for chance
- Packages, R functions, & scripts: ‘skappa0.r’, ‘skapreliab2.r’ (see HW-1)

Outline of Module (cont'd)

- More **examples** of 'reliability' as bridge from qualitative to quantitative research
- *Reliability* as *distance*
- A **problem** with κ when a category is rare

Qualitative Research *(Wiki)*

- Origins in Anthropology as *Ethnomethodology* (‘methodology of the people’, c. 1960’s) or *Ethnography*. E.g., P’s in a new culture identify flora as ‘food’, ‘medicine’ or ‘other’. What is ‘true’? Who is ‘expert’?
- Seeks to ans: **How** and **Why**? - rather than What, When, How Much?
- Focus on (spoken) language, meanings; and effect of context
- Most familiar to psychologists as focus group testing, open-ended items, structured interviews, pilot testing; thought protocols

Example 1: *Values*

- Given an interest in ***values***: What are the antecedents & consequences of a person's *value priorities (*or *ranking)* - within a **culture**?
- An analysis of *values* into, e.g., **content** (or **domains,** or **dimensions**); **extent** or **comprehensiveness**; **structure**, e.g., the *conflict* between 'independence' & 'conformity', or the *compatibility* between 'equality' and 'helpfulness'.
- We can study the **cultural** determination of **meaning** of the various dimensions of value.
- Or, cultural **moderators** (e.g., ecology, history, technology, social stratification & politics) of **relations** among values, antecedents & consequences.

- **Example 2**: From *T. Thomas' study of "ethnic group stereotypes".* List the positive (or negative) attributes that most people would associate with certain groups, e.g., black immigrants, immigrants, and whites.
- The adjectives given by a participant are categorized into a set of useful, **reliable** categories or "traits."
- 'Positive' **adjectives:** hard-working, rich, intelligent, business-savvy, ambitious, clean, stable family, …

Example 3: Measures of *Thoughts*

- **Extent** or **number** of thoughts (arising, e.g., from reading an essay)
- **Valence** or **favorability**
- **Integrative Complexity**, e.g., number of aspects or dimensions of the issue; presence of arguments both pro and con, or of counter-arguments; use of graded vs dichotomous opinions
- In both examples (*values* and *thoughts*), we need to process the qualitative raw data into 'aspects', etc. before we can extract **counts** and other **quantities** from the data for use in interesting **statistical analyses**. Is this extraction **reliable**? [Karmarkar & Tormala (*J Consumer Res*, Apr 2010)]

Example 4. *Attenuation* in the measurement of brain *adaptation*

- *fMRI-Adaptation and Category Selectivity in Human Ventral Temporal Cortex*, by **Kevin S. Weiner**, Rory Sayres, Joakim Vinberg, and Kalanit Grill-Spector. In *J Neurophysiol* 103: 3349–3365, 2010.
- "When stimuli are repeated, cortical responses in high-level visual cortex generally decrease. When the responses are measured with fMRI, this **reduction** in activation is labeled *fMRI-Adaptation (fMRI-A)."*
- Important to determine when this reduction is due to neural noise (i.e., to unreliability of measurement) or is meaningful (i.e., adaptation).

Goals of *Qualitative* Data Analysis

- Carefully discern and document (i.e., '**code**') themes in the data in a **consistent** and **reliable** way.
- Establish **content validity** - Does an index measure what a researcher thinks it measures? This is a strength of qualitative research.
- But **quantitative** analysis of these codes is the capstone analytical step for many analyses of qualitative data.

- **Question**: How to define ‘reliability’ when X is categorical, e.g., ‘aggression’ of Type 1, 2 or 3; or Yes/No?
- **Ans**. As X for Rater 1, X_1, and X for Rater 2, X_2, vary across ‘objects’, we would regard the rating procedure as reliable if the responses tend to **agree** with each other. That is, **‘reliability’ = ‘level of agreement’**.
- **Question**: What **categories** best capture the **ideas and meanings** contained in a typical ‘object’ (e.g., narrative). Are these categories used **reliably** by trained, disinterested raters?

Objects/Stimuli coded into *Categories* by *Raters*

- Many ‘Objects’/‘stimuli’ (e.g., narratives, adjectives, colored objects, snippets of triadic play, flora, thoughts) are coded into categories, ‘A’, ‘B’, ‘C’, … (e.g., themes, traits, emotions, types of play by triads of pre-schoolers).
- Coder = Rater = ‘Method’
- ‘Rating’ also called ‘Score’

Raters label objects as ‘A’, ‘B’, ‘C’

	Rater			
Object	*R1*	*R2*	*R3*	*R4*
1	A	C	B	A
2	B	A	A	A
3	B	B	A	B
4	A	A	C	C
5		…		

Reliability = level of agreement

Two types of data: (a) scores

- Rows = many ‘objects’ to be coded
- Columns = few ‘raters’ who code objects
- The (*i, j*)’th cell contains the code or rating or score given to Object *i* by Rater *j*.
- Data arrayed in ‘Object’ by ‘Rater’ matrix of **scores,** as in the previous slide.

Two types of data: (b) counts

- Rows = many ‘objects’ to be coded, as before
- Columns = few ‘categories’ into which ‘many’ raters code the objects**; objects may lie in more than 1 category.**
- The (*i, j*)’th cell contains the number or count of raters who assign Object *i* to Category *j*.
- Data arrayed in ‘Object’ by ‘Category’ matrix of **counts.**

Objects assigned to 1 or more of ‘A’, ‘B’, ‘C’

	Category		
Object	*A*	*B*	*C*
1	5	7	1
2	3	0	0
3	3	2	2
4	2	4	1
5		…	

‘Reliability’ based on correlations & ANOVA calculations involving these *counts*.

Outline of Module

- Examples of qualitative research: values, thoughts
- Goals: validity, reliability (= consistency, agreement, accuracy)
- Data formats: ‘scores’ vs ‘counts’
- Cohen’s κ: reliability = agreement corrected for chance
- Packages and R functions: ‘skappa0.r’, ‘skapreliab2.r’ (see HW-1)

- **Example 1 (Values)**: From *Markus, Ryff, Curhan & Palmersheim's study of "Well-Being".* Let us focus on **coding the narrative answer** to just 1 question, "What does it mean to you to have a good life?"
- **Ans**: "A good life is having the things you need. Being happy and content. Having your health and the things you need. I mean, I always want more than what I have. I guess that's human nature. But I am talking about having a roof over your head, a job, some kind of security."

- When we search initially for ‘reliable’ categories, what is our goal?
- **Ans**. To assure that **the variable to be used in subsequent substantive analyses** is ‘valid’ and has ‘high’ reliability.
- E.g., which scheme, {‘below normal’, ‘normal’, ‘above normal’} or {‘unacceptable’, ‘good’, ‘excellent’} should we use to categorise a set of objects? Is a consensus on ‘good’ easier to reach than one on ‘normal?
- **Ans**. Maybe, calculate the reliability of each scheme, and use the more reliable scheme.

- Train 2 or 3 raters by giving them a set of mutually exclusive and exhaustive categories. Check **all and only** those categories that are present in each narrative.
- **Categories (Rater 1)**

#	**RelIns**	**Health**	**Family**	**$$**	**Self**	**Other**
1	0	1	1	0	1	0
2	1	1	1	0	0	0
3	1	..	..			

- More columns for Rater 2, Rater 3, etc.
- Select a category, crosstabulate Rater 1's responses (row) and Rater 2's responses (column) to see how much they agree on this category.

Cat=1	**R2**		
	0	**1**	R_i
R1= 0	20	4	24
R2= 1	8	11	19
C_j	28	15	43 = N

- The two raters agreed 20 times in not seeing category 1, and agreed 11 times in seeing category 1; the overall level of agreement is P_a = (20 + 11)/43 = 31/43 = 0.721.
- **Cohen's *kappa*** is degree of agreement, corrected for chance agreement, P_c.
- $K = (P_a - P_c)/(1 - P_c)$ = 0.422,

where $$P_c = \sum_{i=1}^{k} \left(\frac{R_i}{N} \right) \left(\frac{C_i}{N} \right) \quad .$$

Reliability = Agreement *corrected for chance*

- A rating of a patient, e.g., ‘depressed’ or ‘not depressed’, is ‘reliable’ if similarly trained raters show ‘high’ agreement. But some agreement is expected by chance alone. How should we **correct** observed agreement for chance?

- The observations in the diagonal cells (using the familiar notation for 2-way tables) gives % agreement:

- $P_a = \sum_{i=1}^{k} O_{ii} / N$.

- Next, what is the **expected** contingency table, $\{E_{ij}\}$, if the ratings of the 2 raters were independent. **Ans**. $E_{ij} = R_i C_j / N$.
- What is the % agreement, P_c, in this 'expected' table? **Ans**. $P_c = (R_1 C_1 + R_2 C_2)/N$.

$$E_{ii} = R_i C_i / N, \quad and \quad P_c = \sum_{i=1}^{k} E_{ii} / N.$$

$$\kappa = \frac{P_A - P_C}{1 - P_C}.$$

- To get the standard error of κ, let

$$U = \sum_{i=1}^{k} \frac{R_i^2 C_i + R_i C_i^2}{N^3}.$$

- Then $$S.E.(\kappa) = \left(\frac{P_C + P_C^2 - U}{N(1 - P_C)^2} \right)^{1/2}.$$
- Use s.e. to compare κ across groups.

Interpretation of *kappa* values

- J. Cohen suggested the following convention:
  - .2-.4 is ‘fair’;
  - .4-.6 is ‘moderate’;
  - .6-.8 is ‘substantial’.

Outline of Module

- Examples of qualitative research: values, thoughts
- Goals: validity, reliability (= consistency, agreement, accuracy)
- Data formats: ‘scores’ vs ‘counts’
- Cohen’s κ: reliability = agreement corrected for chance
- Packages and R functions: ‘skappa0.r’ (see HW-1)

R package, `psych`

- **`install.packages('psych')`**
- Contains a function, **`cohen.kappa()`**, that computes reliability when there are **2 raters**. (See HW-1 for details and difficulties.)
- **`> library(psych)`**
- **`> ?cohen.kappa  #for details`**
- **Relevant questions**
  - **`Are data raw or in form of agreement matrix?`**
  - **`2 raters or more than 2 raters?`**
  - **`Are the data 'scores' or 'counts'?`**

‘kappadata1.csv’

Object	psy1	psy2	psy3	psy4
1	1	1	2	1
2	2	1	1	1
3	2	2	1	2
4	1	1	1	1
5		…		

‘skappa1.r’

```
library(psych)
sink("rkappa2.r")
d0 = read.csv("kappadata1.csv")
ctab12 = with(d0, table(psych1, psych2))
print(ctab12)
kap12 = cohen.kappa(ctab12)   #for c1,c2
print(kap12)

library(irr)          #input raw data
kap12a = kappam.light(d0[, c(1,2)])
kap12b = kappa2(d0[, c(1,2)])
kap1234 = kappam.light(d0[, c(1,2,3,4)])
      #for c1-c4
```

Outline of Module (cont'd)

- More examples of 'reliability' as bridge from qualitative to quantitative research
- A problem with κ when a category is rare

‘Reliability’ as bridge from Qualitative to Quantitative research

- The index or coding scheme, the reliability of which is being assessed, should be closely related to the DV’s in the planned statistical analyses
- **Example 1**: Markus et al. want to compare Prob(‘Wealth’) in narrative between ‘hi school only’ and ‘some college’ people. Thus, we need to calculate the reliability with which a rater reports the **presence or absence of ‘Wealth’** in a narrative.

- **Example 2**: T. Thomas Tormala similarly examined the dependence of Prob(**positive** trait in stereotype) on ethnicity of *participant* and ethnicity of *target group* being rated. Thus, we need the reliability with which the adjectives produced by Ss are assigned to a '**positive** trait'.
- **Example 5**: McLoyd et al. (Describe study) did 2 analyses:
  - Compare P(*triadic* play) in *N* mins of pretend play for young and old kids, and
  - Analyze the minute-by-minute social dynamics of play, e.g., P(*triadic* at *t* | *solitary* at *t-1*)

1st reliability analysis in McLoyd et al.:
Play period divided into, e.g., 5-min ('long') segments

- Reliability in estimating, e.g., P(triadic play) in a segment.
  - 'Objects' are, e.g., 20 5-min (i.e., 'long') segments of play from different triads
  - 2 Raters each estimate *s* = P(solitary), *d* = P(dyadic) and *t* = P(triadic) for each segment (*s*+*d*+*t* = 1); these are **quantitative** responses
  - Find *r* = corr(t_1, t_2) across objects; **use *r* to estimate reliability** (details later in course)

2nd reliability analysis in McLoyd et al:
Play period divided into 30 1-min (i.e., 'short') periods

- Reliability in coding play in each period as solitary, dyadic, or triadic
  - 'Objects' are, e.g., 100 1-min periods of play from different triads
  - 2 Raters code each period (these are **categorical** ratings)
  - Find the agreement between raters across objects; use agreement to estimate Cohen's *kappa* index of reliability

Difference between 2 analyses

- Consider 2 raters' codes for 6 periods:
  - Rater 1: SSDDTT
  - Rater 2: DDTTSS
- Agreement on P(triadic), P(dyadic) for the **sample** is perfect, but agreement on the coding of **individual** periods is 0!
- Agreement on coding individual periods is more stringent, and is **necessary** for sequential analysis

Reliability and *Distance*

	Rater	
Object	*R1*	*R2*
1	A	C
2	B	A
3	B	B
4	A	A
5 …		

The **Hamming distance** between *r1 = c(a,b,b,a,…)* and *r2 = c(c,a,b,a,…),* of equal length, is the number of slots at which the 2 strings differ; i.e., the number of **disagreements** between R1 and R2.

```
library(stringdist)
```

```
h0 = sum(stringdist(r1, r2, method = 'h'))
pa1 = 1 - h0/length(r1)  # % agreement
```

Reliability and *Distance*

- The simplest strings are those consisting of 0's and 1's, and 'reliability as agreement' is easy to implement.
- For more complex strings, e.g., meaningful text, speech, of variable length, some disagreements, e.g., transpositions, are more serious than others. Various string distance metrics have been proposed to deal with such complexity, and reliability can be defined through these generalised distance metrics.
- **End of Lecture**