DATA CODING AND SCREENING

Jessica True
Mike Cendejas
Krystal Appiah
Amy Guy
Rachel Pacas

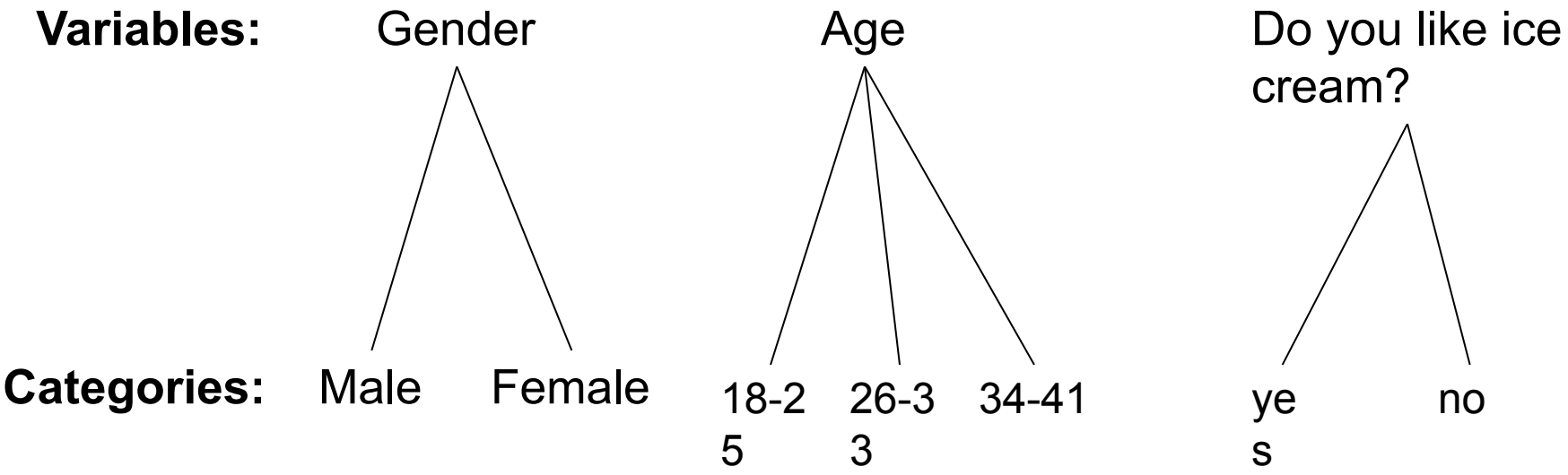
WHAT IS DATA CODING?

- "A systematic way in which to condense extensive data sets into smaller analyzable units through the creation of categories and concepts derived from the data."[1]

- "The process by which verbal data are converted into variables and categories of variables using numbers, so that the data can be entered into computers for analysis."[2]

1. Lockyer, Sharon. "Coding Qualitative Data." In The Sage Encyclopedia of Social Science Research Methods, Edited by Michael S. Lewis-Beck, Alan Bryman, and Timothy Futing Liao, v. 1, 137-138. Thousand Oaks, Calif.: Sage, 2004.
2. Bourque, Linda B. "Coding." In The Sage Encyclopedia of Social Science Research Methods, Edited by Michael S. Lewis-Beck, Alan Bryman, and Timothy Futing Liao, v. 1, 132-136. Thousand Oaks, Calif.: Sage, 2004.

Categories and Variables



WHEN TO CODE

- When testing a hypothesis (deductive), categories and codes can be developed before data is collected.
- When generating a theory (inductive), categories and codes are generated after examining the collected data.
  - Content analysis
  - How will the data be used?

Adopted from Bourque (2004) and Lockyer (2004).

LEVELS OF CODING
(FOR QUALITATIVE DATA)

- Open
  - Break down, compare, and categorize data
- Axial
  - Make connections between categories after open coding
- Selective
  - Select the core category, relate it to other categories and confirm and explain those relationships

Strauss, A. and J. Corbin. *Basics of qualitative research: Grounded theory procedures and techniques.* Newbury Park, CA: Sage, 1990 as cited in Lockyer, S., 2004.

WHY DO DATA CODING?

- It lets you make sense of and analyze your data.
- For qualitative studies, it can help you generate a general theory.
- The type of statistical analysis you can use depends on the type of data you collect, how you collect it, *and* how it's coded.
- "Coding facilitates the organization, retrieval, and interpretation of data and leads to conclusions on the basis of that interpretation."[1]

1. Lockyer, Sharon. "Coding Qualitative Data." In The Sage Encyclopedia of Social Science Research Methods, Edited by Michael S. Lewis-Beck, Alan Bryman, and Timothy Futing Liao, v. 1, 137-138. Thousand Oaks, Calif.: Sage, 2004

DATA SCREENING

- Used to identify miscoded, missing, or messy data
- Find possible outliers, non-normal distributions, other anomalies in the data
- Can improve performance of statistical methods
- Screening should be done with particular analysis methods in mind

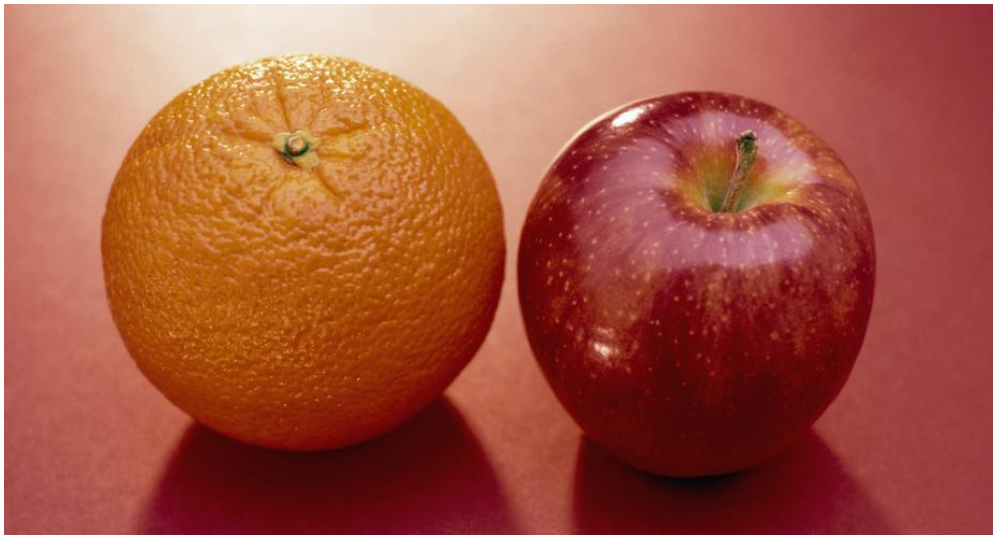
From *Data Screening: Essential Techniques for Data Review and Preparation* by Leslie R. Odom and Robin K. Henson. A paper presented at the annual meeting of the Southwest Educational Research Association, Feb. 15, 2002, Austin, Texas.

DETERMINING CODES (BOURQUE, 2004)

- For surveys or questionnaires, codes are finalized as the questionnaire is completed
- For interviews, focus groups, observations, etc. , codes are developed inductively after data collection and during data analysis

IMPORTANCE OF CODEBOOK (SHENTON, 2004)

- Allows study to be repeated and validated.
- Makes methods transparent by recording analytical thinking used to devise codes.
- Allows comparison with other studies.



DETERMINING CODES, CONT.

- **Exhaustive** – a unique code number has been created for each category ex. if religions are the category, also include agnostic and atheist
- **Mutually Exclusive** – information being coded can only be assigned to one category
- **Residual other** – allows for the participant to provide information that was not anticipated, i.e. “Other”

_____

DETERMINING CODES, CONT.

- **Missing Data** - includes conditions such as “refused,” “not applicable,” “missing,” “don’t know”
- **Heaping** – is the condition when too much data falls into same category, ex. college undergraduates in 18-21 range (variable becomes useless because it has no variance)

CREATING CODE FRAME PRIOR TO DATA COLLECTION (BOURQUE, 2004; EPSTEIN & MARTIN, 2005)

- Use this when know number of variables and range of probable data in advance of data collection, e.g. when using a survey or questionnaire
- Use more variables rather than fewer
- Do a pre-test of questions to help limit “other” responses

TABLE OF CODE VALUES (EPSTEIN & MARTIN, 2005)

Table I Possible Dispositions in Cases Decided by the U.S. Courts of Appeals

Value	*Value label*
0	Stay, petition, or motion granted
1	Affirmed; or affirmed and petition denied
2	Reversed (including reversed & vacated)
3	Reversed and remanded (or just remanded)
4	Vacated and remanded (also set aside & remanded; modified and remanded)
5	Affirmed in part and reversed in part (or modified or affirmed and modified)
6	Affirmed in part, reversed in part, and remanded; affirmed in part, vacated in part, and remanded
7	Vacated
8	Petition denied or appeal dismissed
9	Certification to another court

Source: U.S. Court of Appeals Data Base, available at: http://www.polisci.msu.edu/pljp/databases.html

TRANSCRIPT **(SHENTON, 2004)**

- Appropriate for open-ended answers as in focus groups, observation, individual interviews, etc.
- Strengthens “audit trail” since reviewers can see actual data
- Use identifiers that anonymize participant but still reveal information to researcher

  ex. Y10/B-3/II/83 or “Mary”

THREE PARTS TO TRANSCRIPT (SHENTON, 2004)

1. Background information, ex. time, date, organizations involved, participants.
2. Verbatim transcription (if possible, participants should verify for accuracy)
3. Observations made by researcher after session, ex. diagram showing seating, intonation of speakers, description of room

POSTCODING **(SHENTON, 2004)**

1. Post-meeting observations
2. Post-transcript review
   a. Compilation of insightful quotations
   b. Preliminary theme tracking
   c. Identification of links to previous work
3. Create categories and definitions of codes

DATA DICTIONARY (SHENTON, 2004)

Table 1
Extract showing codes and categories within specimen data dictionary

IS/LIB	Library used
IS/LIB/S	School library used
IS/LIB/S/NEED	School library used to meet particular need
IS/LIB/S/NEED/SCH	School library used for school assignment
IS/LIB/S/NEED/LEI	School library used for leisure information needs
IS/LIB/S/NEED/FIC	School library used for fiction
IS/LIB/S/FIND	Particular strategy used for finding information in school library
IS/LIB/S/FIND/CAT	Catalogue used to find materials
IS/LIB/S/FIND/SHELF	Shelf signs used to find materials
IS/LIB/S/FIND/SUB-IN	Subject index used to find materials
IS/LIB/S/EOU	School library considered easy to use
IS/LIB/S/FRU	Frustration caused when using school library
IS/LIB/S/UN	School library unable to meet need
IS/LIB/S/PROB	Problems in relation to school library
IS/LIB/S/PC	Poor concept of school library
IS/LIB/S/NO	School library not used
IS/LIB/P	Public library used

REFERENCES

Bourque, Linda B. "Coding." In *The Sage Encyclopedia of Social Science Research Methods.* Eds. Michael S. Lewis-Beck, Alan Bryman, and Timothy Futing Liao, v. 1, 132-136. Thousand Oaks, Calif.: Sage, 2004.

Lee, Epstein and Andrew Martin. "Coding Variables." In *The Encyclopedia of Social Measurement.* Ed. Kimberly Kempf-Leonard, v.1, 321-327. New York: Elsevier Academic Press, 2005.

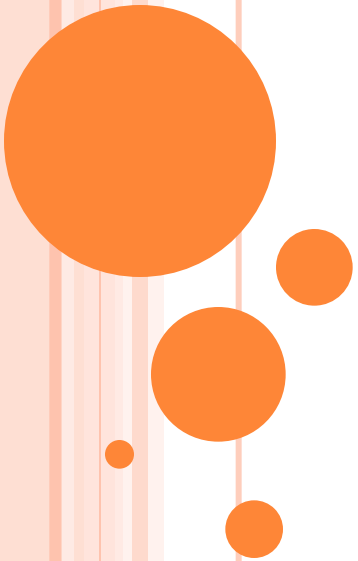
Shenton, Andrew K. “The analysis of qualitative data in LIS research projects: A possible approach.” *Education for Information* 22 (2004): 143-162.

Levels of Measurement

Scale/Levels	Basic Operations	Permissible Statistics
Nominal	Determination of Equality	Number of cases Mode
Ordinal	Determination of greater or less (rank)	Median Percentiles
Interval	Determination of equality of intervals	Mean Standard Deviation
Ratio	Determination of equality of ratios	Coefficient of variation

Coding Mixed Methods:

Advantages and Disadvantages



Position 1 v. Position 2

- "When compared to quantitative research, qualitative research is perceived as being less rigorous, primarily because it may not include statistics and all the mumbo jumbo that goes with extensive statistical analysis. Qualitative and quantitative research methods in librarianship and information science are not simply different ways of doing the same thing."

- Source: Riggs, D.E. (1998). Let us stop apologizing for qualitative research. *College & Research Libraries*, 59(5).
- Retrieved from: http://www.ala.org/ala/acrl/acrlpubs/crljournal/backissues1998b/september98/ALA_print_layout_1_179518_179518.cfm

Move Toward P1 and P2 Cooperation

- **Cooperation** – last 25 years –

- Limitations of only using one method:

  - Quantitative – lack of thick description
  - Qualitative – lacks visual presentation of numbers

Source: Grbich, Carol. "Incorporating Data from Multiple Sources." In *Qualitative Data Analysis*. (Thousand Oaks, Calif.: Sage Publications, 2007): 195-204.

Advantages of Mixed Methods:

- Improves validity of findings
- More in-depth data
- Increases your capacity to cross-check one data set against another
- Provides detail of individual experiences behind the statistics
- More focused questionnaire
- Further in-depth interviews can be used to tease out problems and seek solutions

Disadvantages of Mixed Methods

- Inequality in data sets
- "Data sets must be properly designed, collected, and analyzed"
- "Numerical data set treated less theoretically, mere proving of hypothesis"
- Presenting both data sets can overwhelm the reader
- Synthesized findings might be "dumbed-down" to make results more readable

- Source: Grbich, Carol. "Incorporating Data from Multiple Sources." In *Qualitative Data Analysis*. (Thousand Oaks, Calif.: Sage Publications, 2007): 195-204.

Key Point in Coding Mixed Methods Data

- "The issue to be most concerned about in mixed methods is ensuring that your qualitative data have not been poorly designed, badly collected, and shallowly analyzed."

- Source: Grbich, Carol. "Incorporating Data from Multiple Sources." In *Qualitative Data Analysis*. (Thousand Oaks, Calif.: Sage Publications, 2007): 195-204.

Examining a Mixed Methods Research Study

- Makani, S. & Wooshue, K. (2006). Information seeking behaviors of business students and the development of academic digital libraries. *Evidence Based Library and Information Practice*, 1(4), 30-45.

Study Details

- **Population**: Purposive population, 10 undergraduates (2 groups) / 5 graduate students
  - Undergraduate business students at Dalhousie University in Canada
- **Objectives**: To explore the information-seeking behaviors of business students at Dalhousie University in Canada to determine if these behaviors should direct the design and development of digital academic libraries.

Methods

- Data: Used both **qualitative and qualitative data collected** through a **survey, in-depth semi-structured interviews, observation, and document analysis**.
- **Qualitative case study data** was coded using **QSR N6 qualitative data analysis software**.

Study Observations

- Followed 3 groups of business students working on group project assignments. The assignments involved formulating a topic, searching for information and writing and submitting a group project report.

Coding Methods

- Used pre-selected codes from literature review:
  - Time
  - Efficiency of use
  - Cost
  - Actors
  - Objects (research sources)

Coding: Ordinal Measures

- Opinion Survey

  What sources do you use to get started on your research?

	Graduate	Undergraduate
Consult your textbook/class notes	16.7%	15.6%
Browse the Business section of the library stacks	0%	0%
Talk to a librarian	0%	0%
Talk to your classmates	0%	3.9%
Search Google (or another similar search engine)	37.5%	63.6%
Go to the library's Business subject page	4.2%	1.3%
Search the library's online catalogue (Novanet)	12.5%	5.2%
Search a Business database	20.8%	9.1%
Other, Please Specify	8.3%	1.3%

Examples of Ratio-Interval Coding and Level of Measurement

- The age of the survey participants (survey and group study) ranged from 18 – 45 years.
- Most of the undergraduates were between 18 and 25 years of age (95%)

- While 56% of graduate students fell within the same age range.

Study Conclusions

- This study reveals that in order to create an effective business digital library, an understanding of how the targeted users do their work, how they use information, and how they create knowledge is essential factors in creating a digital library for business students.

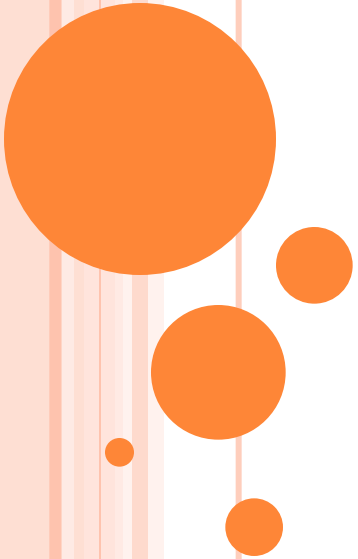
Study Weaknesses: Use of Mixed Methods Data

- No discussion of how the survey was delivered electronically
- Survey questions were not included in the published article
- Created for a long results section

Study Advantages: Use of Mixed Methods Data

- Numeric data helped create a clearer picture of the participants
- Numeric data from the survey questions nicely compliments the excerpts from the semi-structured interviews

OUTLIERS IN DATA ANALYSIS



WHAT IS AN OUTLIER?

- Miller (1981): '... An outlier is a single observation or single mean which does not conform with the rest of the data... .'
- **Barnett & Lewis (1984): '... An outlier in a set of data is an observation which appears to be inconsistent with the remainder of that set of data....'**

WHY ARE OUTLIERS IMPORTANT IN DATA ANALYSIS?

- Outliers can influence the analysis of a set of data
  - Objective analysis should be done in order to determine the cause of an outlier appearing in a data set

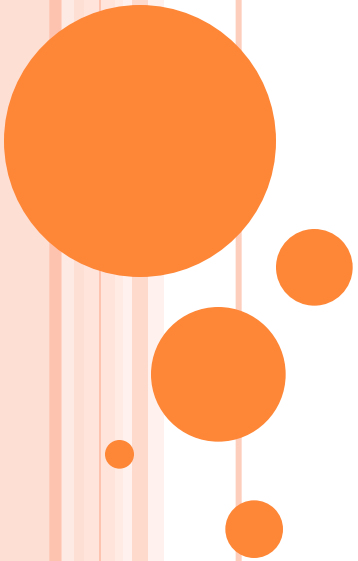
ISSUES CONCERNING OUTLIERS

- Rejection of Outliers
  - "From the earliest efforts to harness and employ the information implicit in collected data there has been concern for "unrepresentative", "rogue", "spurious", "maverick", or "outlying" observations in a data set. What should we do about the "outliers" in a sample: Should we automatically reject them, as alien contaminants, thus restoring the integrity of the data set or take no notice of them unless we have overt practical evidence that they are unrepresentative?"

What do we do with outliers?

- There are four basic ways in which outliers can be handled:
  - The outlier can be *accommodated* into the data set through sophisticated statistical refinements
  - An outlier can be *incorporated* by replacing it with another model
  - The outlier can be used *identify* another important feature of the population being analyzed, which can lead to new experimentation
  - If other options are of no alternative, the outlier will be *rejected* and regarded as a “contaminant” of the data set

A CLASSIC EXAMPLE ON THE USE OF OUTLIERS



Hadlum vs. Hadlum (1949)

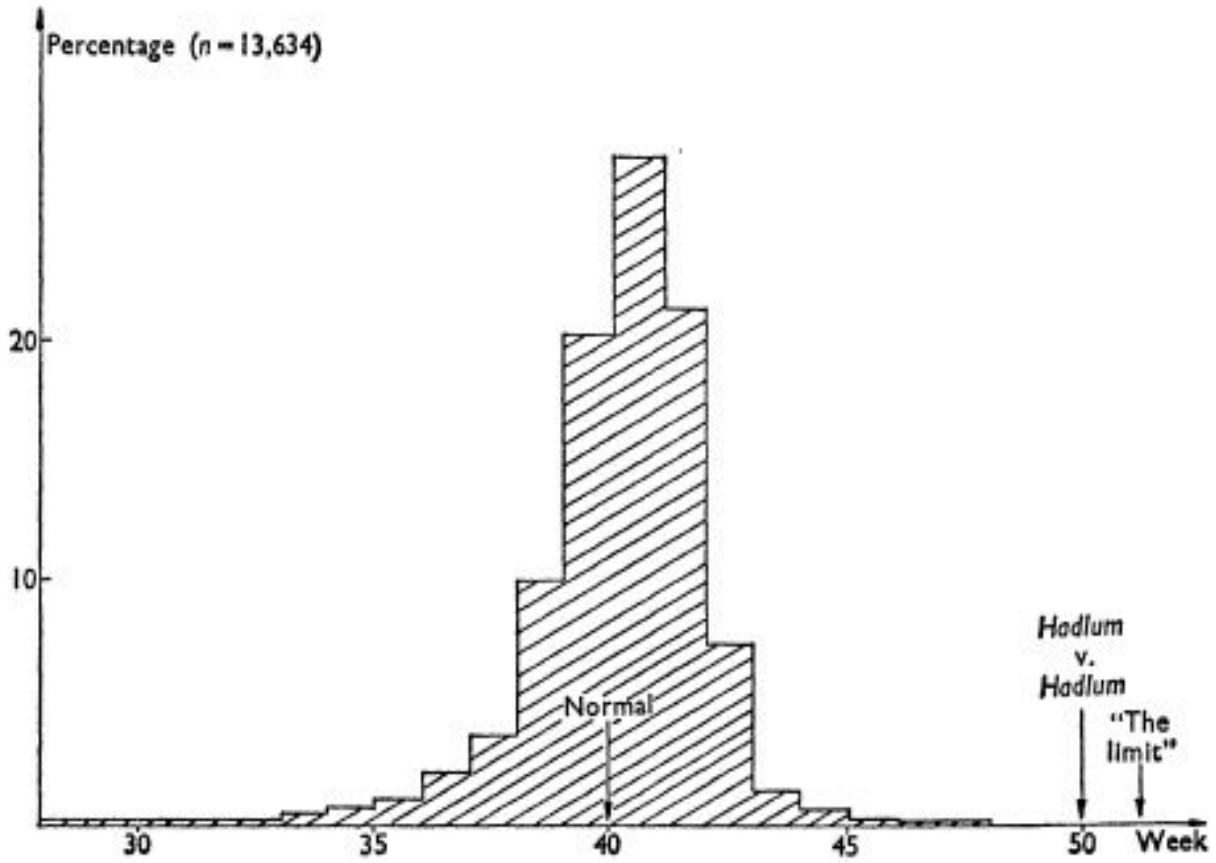


FIG. 1. Distribution of human gestation periods.

Sources

Barnett, Vic. 1978. The study of outliers: purpose and models. *Applied Statistics* 27: 242-250.

Munoz-Garcia, J., J.L. Moreno-Rebollo, and A. Pascual-Acosta. 1990. Outliers: a formal approach. *International Statistical Review* 58: 215-226.