**DATA**

**Data Coding**

***TIP! Keep a list of codes/symbols in a document & embedded in data set (SPSS variable tab, or separate excel sheet)***

**Checkboxes/True-False** are typically entered as 0 (=no) and 1 (=yes).

**Categories/ratings** are entered as numbers associated with each level e.g., gender: male=0, female=1, other=2).

**Open-ended** are entered as text “strings” into excel and SPSS. Enter bullets with \* and symbols with words (e.g., “at”)

**Missing data** is either left blank or coded with 999 (or another unused number). Designate in SPSS or other programs so treated like missing data when encountered in analysis (see software resources for procedure).

**Entering, Formatting and Storing Data**

During entry note which values are Nominal, Ordinal or Continuous (see process handout)

**LAYOUT: Each Row** = 1 Case (e.g., participant, site). **Each column** = 1 variable with all its levels (e.g., gender: male…)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Participant/site** | **Gender** | **Importance\_Weather** | **Ski** | **Run** | **Bike** | **Summer\_activities** |
| 1 | 0 | 5 | 0 | 1 | 0 | During the summer, I am always… |
| 2 | 2 | 3 | 1 | 1 | 1 | I like all seasons, but I do like to go… |
| 3 | 1 | 4 | 0 | 1 | 1 | Every summer I try to get jobs tha… |

**Data Screening & Cleaning** (Chapter 4 orTabachnick & Fidell, 2007)

**Missing data** (look through raw data, compute frequencies and look for totals < n)

 Identify the common approach to missing data in your field and study type, such as listwise deletion (deleting the case/person/source but reduces sample size), substitution with average or predicted response etc.

**Outlier analysis** (compute min, max values or check *frequency table* listing counts of all response values)

Outliers, which are extreme high or low responses to an item, can occur due to very unique respondents, data entry or collection errors. When found, double check data entry (include values assigned online survey answers). Also look for unexpected ratings such as incongruent score and comment (e.g, low 1 but “best ever!”). If not an error, then you need to decide if you will delete to meet statistical analysis data needs, or be inclusive of the outlier or change tests.

**Sufficient variation (Restricted range, ceiling effects or floor effects):**

**Restricted range** is when there are insufficient response options leading to distorted/squished dataset.

**Ceiling effect** is when an item’s responses are skewed towards the upper end and respondents may have needed further high options (e.g., how many miles do you walk each day? 0, 1, 2 or more)

**Floor effect** is restriction at the lower end with responses skewed (e.g., adults age? Under 60, 60-69…)

To reduce the likelihood: Check if enough variation in sample, enough Likert-scale numbers with wide but not extreme end points, enough categories e.g., age is <18, 18, 19, 20, >20 vs. <15, 16…25…). You can always collapse/combine categories so err on side of too many categories (e.g., 0 – 110 possible for age)

Check: Create and examine frequency table. Also run Skewness or Kurtosis statistics

To manage: As restricted range, ceiling and floor effect data is distorted and skewed, the data may need to be altered through transformation (e.g., square root or Logarithm transformation), categories combined or other methods. Consult a resources such as Tabachnick & Fidell, 2007, Chapter 4 “Cleaning up your act.”

**Validity** – Determine if what you think you are measuring is actually what you are measuring by checking expert validity, predictive validity, face validity, or other measures of validity. The focus is on the strength of the relationship between the observed/measured variable and the intended latent/immeasurable concept. (See a statistic textbook e.g., McMillan & Schumacher p 179-183)

**Reliability** – There are several approaches to seeing if your approach to measuring would get the same response every time (e.g., thermometer show same value when temperature is consistent). Common scale reliability tests include test-retest, split-half etc. (see McMillan & Schumacher p183-189)

**Resources**

Tabachnick, B. G., & Fidell, L. S. (2007). *Using Multivariate statistics (5th ed.).* Boston, MA: Pearson Education.

Healey, J. F. (2009). *Statistics: A tool for social research* (8th ed.). Belmont, CA: Wadsworth Cengage Learning. – *Covers categorical inferential statistics, and the theory & math underlying statistics.*