

Week 4: February 13, 2008

Design of Quantitative Survey Instruments / Introduction to SPSS and Descriptive Statistics

"There are three kinds of lies: lies, damned lies, and statistics."

Benjamin Disraeli (19th Century British Statesman)

"Grown ups love figures...only from these figures do they think they have learned anything..."

The Little Prince

Concepts You Should Know:

- Descriptive statistics
- Frequency (of a variable)
- Frequency distribution
- Absolute frequency distribution
- Cumulative frequency distribution
- Cumulative percentage frequency distribution
- Bar graph
- Histogram
- Frequency polygon
- Measures of central tendency
- Mode
- Median
- Mean (average)
- Outlier
- Measures of variability
- N =
- Range
- Percentile
- Quartiles
- Variance
- Standard deviation

I. Design of Data Collection Instruments

A. Review: asking the right questions:

1. Avoid leading questions
2. Avoid double barreled questions
3. Match the language with age and cognitive levels of target population

4. Match the question with the intended cultural meaning of the target population
5. Build in ways to minimize error in response due to memory limitations
6. Remember—the attributes for “check one” questions should be *exhaustive and mutually exclusive*
7. Consider the order of questions in the instrument/survey:
 - a) Least sensitive (such as basic demographic info) to most sensitive
 - b) Whenever possible, questions should flow like a conversation
 - c) Prepare a few prompts for each question (could be in the instructions to the interviewer) to encourage response

B. A few important technical formatting principles:

1. See Rubin & Babbie, Chapter 9 for good examples of formatting
2. Number and identify sub-questions (1, 1a, 2, 2a, 2b etc.) to help respondents and research staff.
3. Allow sufficient space for responses and to maintain a “clean” look of the form. Keep general appearance in mind. Use large and clear font-types. Use colored paper covers (actual project version only). Leave space for data notes.
4. A vertical format is recommended with consistent patterns for codes, alignment, and use of response labels.
5. Include instructions and directions for question direction (e.g., “if yes go to...”). Assume that you will not be the only interviewer and that instructions are needed to maintain consistency of data collection.
6. Keep the set of response categories for a particular question all on same page
7. Make check boxes or lines that are clearly visible to an interviewer, a respondent, or someone doing data entry
8. Pilot test the instrument with a classmate or friend to make sure the questions make sense, are easily answerable, and flow in the intended order

C. Use of Existing Instruments—Aligning Your Variables

- In selecting the questions to ask about demographic characteristics and key variables, researchers should start with items comparable to those found in major databases or published works. For example, in health surveys, variables are often measured according to the National Center for Health Statistics – National Health Interview Survey or related federally sponsored health surveys. (See Rubin & Babbie for examples.)
- Give examples about how the following variables can be assessed:

- Household composition
- Sex
- Race and Ethnicity
- Employment Status
- Socioeconomic Status
- Age
- Marital Status
- Education
- Occupation
- Income

D. Examples of SES Demographic Questionnaires—What the U.S. Census Asks and Why

II. **Descriptive Statistics**—are the most basic statistics that summarize (or “describe”) characteristics of variables. *With any statistic (including a descriptive statistic) we’re trying to estimate a population parameter from a sample.*

A. Counting

1. **Frequency (of a variable)** – a count of observations falling in a value category (or in an attribute) of a variable, e.g. “the sample consists of 40 men and 60 women”
 - a) **Absolute frequency distribution**—simple counts
 - b) **Absolute percentage**—percentage of total N
 - c) **Cumulative frequency distribution**—running summed count
 - d) **Cumulative percentage frequency distribution**—cumulative percentage of the frequency distribution.

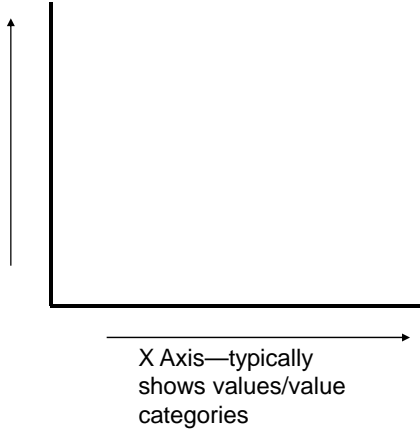
<i>Ethnicity</i>	Absolute Frequency	Absolute Percentage	Cumulative Frequency	Cumulative Percentage
<i>Caucasian</i>	20	40%	20	40%
<i>African American</i>	15	30%	35	70%
<i>Hispanic</i>	13	26%	48	96%
<i>Other</i>	2	4%	50	100%
<i>Total (N)</i>	50	100%		

2. Frequencies can be displayed graphically, with **frequency distributions** – tables or graphs the present the number (frequency) with which different values (attributes) of a variable occur

a) First, an introduction to the Y axis, X axis

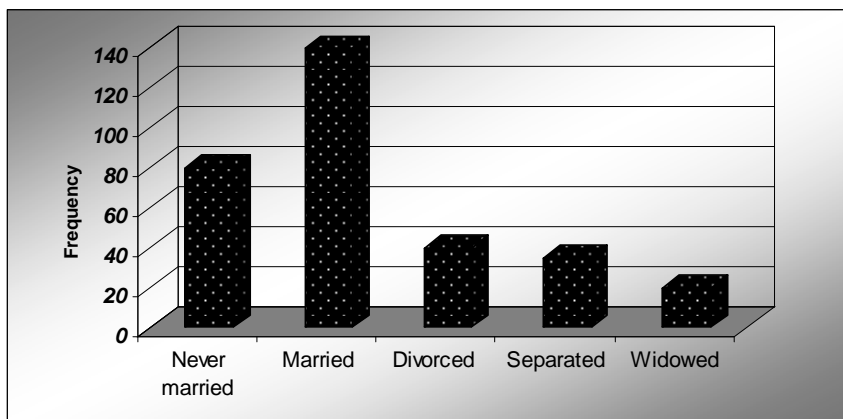
Y axis, X axis

Y Axis—typically shows frequencies



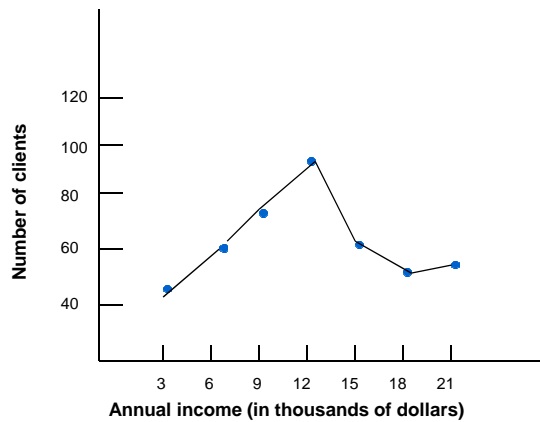
b) **Bar graph** – represents nominal-level data (usually) with height (or length) of bars representing counts

Bar Graph—Nominal Level Measure



- c) **Frequency polygon** – graphical representation using dots and connecting lines to display the shape of the distribution of a ratio or interval level data

Frequency Polygon—Interval or Ratio Level of Measures

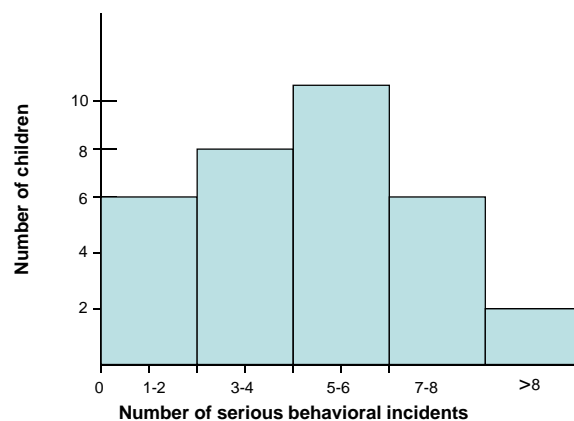


ScWk 242 Week 4

7

- d) **Histogram** – like a bar graph, except the bars touch. Used to display frequency distributions of ratio, interval or ordinal data

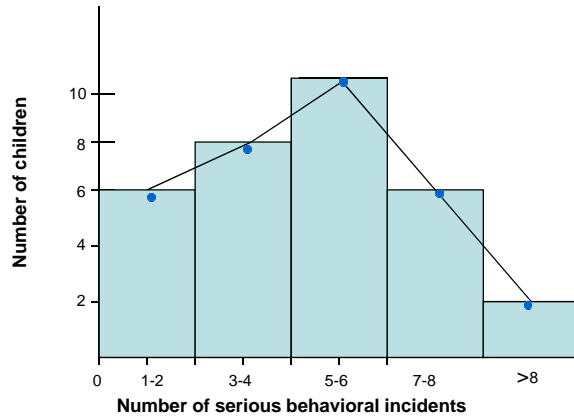
Histogram—Ordinal, Interval or Ratio Level of Measures



ScWk 242 Week 4

8

Plus Frequency Polygon



ScWk 242 Week 4

9

- Percentile**—indicates the percentage of cases that fall below a certain value. For example, saying that “Johnny’s test score is in the 75th percentile” means that 75% of students have test scores below Johnny’s.

The percentile is derived from the cumulative percentage, e.g.

Sixth Grade Achievement Score Distribution

Achievement Score Categories	Number of children	Percentage of total	Cumulative Percentage
10-25	2,000	3.39%	3.39%
26-50	10,000	16.95%	20.34%
51-75	35,000	59.32%	76.27%
76-100	12,000	20.34%	79.66%
Total	59,000	100%	

Here, we can say that 12,000 children score near the 80th percentile, and 47,000 children (35,000 + 12,000) score in the 76th percentile.

- Quartiles** -- Dividing a group of data into sets of four (or quarters), indicating for an individual the percentage of people whose scores are smaller (a.k.a. “interquartile range” – 25th percentile, 50th percentile, and 75th percentile)

B. What’s “typical”? Measures of central tendency

Measures of central tendency – we want to be able to describe what is “typical.” In statistics, we describe typical through measures of central tendency (mode, median, mean). Why?

- They summarize data; one number explains a lot
- They provide a common reference point for comparing two groups of data

- Mode** – the value in a distribution of values within a data set that occurs most

frequently (also see book examples)

0 0 0 1 1 2 2 3 3 3 3 3 4 4 5 N = Mode =

0 0 0 0 0 1 2 2 3 4 5 5 6 7 7 7 7 7 8 9 11 14 N = Mode =

Example of use: What age group is the most frequently seen at our agency?

- 2. **Median** – if data can be ranked into a list or array (interval or ratio variables) the median is that value which divides that list or array of values into two equal halves (also see book examples)

1 2 4 5 6 9 9 N = Median =

1 3 4 4 4 5 5 5 6 7 N = Median =

The median is often used for data that are highly influenced by outliers, such as average income, or average home price.

***True or False: *The Median is the same thing as the 50th percentile*

- 3. **Mean** – called the arithmetic mean or average, is the sum of the values (interval or ratio variables) divided by the total number of values (also see book examples). This is the most frequently used measure of central tendency for continuous variables.

The mean is sensitive to **outliers** – very extreme values in a frequency distribution

Subject ID	Data1 N=10	Data2 N=10
1	6	6
2	6	6
3	6	6
4	6	6
5	6	6
6	8	8
7	8	8
8	8	8
9	8	8
10	8	28
Total	70	90
Mean	7	9

Mean of Data1 = Sum of observations divided by N = 70/10 = 7

Mean of Data2 (with outlier) = Sum of observations divided by N = 90/10 = 9

***Compare to the median.

What's the median of Data1? _____

What's the median of Data2? _____

C. How much variation? Measures of variability (or dispersion)

1. **Range** – (the largest value minus the smallest value + 1)
2. **Variance** – the average value of the squared deviations from the mean
3. **Standard deviation--**
 - a) “Square root of the variance”
 - b) The standard deviation says something about the size of the average *deviation from the mean*
 - c) A “standardized” way to summarize variability
 - d) A way to compare the variability of different variables

D. Example of how to calculate the variance and standard deviation:

Subject ID	Data1	Mean1	Data1 - mean	Squared deviations from mean
1	6	7	-1	1
2	6	7	-1	1
3	6	7	-1	1
4	6	7	-1	1
5	6	7	-1	1
6	8	7	1	1
7	8	7	1	1
8	8	7	1	1
9	8	7	1	1
10	8	7	1	1
Sums	70		0	10

$$\begin{aligned}\text{Variance} &= \frac{\text{Sum of squared deviations from the mean}}{N-1} \\ &= 10 / 9 \\ &= 1.11\end{aligned}$$

Standard Deviation

$$= \sqrt{\text{Variance}}$$

$$= \sqrt{1.11}$$

$$= 1.05$$

Data with outlier:

Subject ID	Data2	Mean2	Data2 - mean	Squared deviations from mean
1	6	9	-3	9
2	6	9	-3	9
3	6	9	-3	9
4	6	9	-3	9
5	6	9	-3	9
6	8	9	-1	1
7	8	9	-1	1
8	8	9	-1	1
9	8	9	-1	1
10	28	9	19	361
Sums	90		0	410

$$\text{Variance} = \frac{\text{Sum of squared deviations from the mean}}{N-1}$$

$$= 410 / 9$$

$$= 45.56$$

Standard Deviation

$$= \sqrt{\text{Variance}}$$

$$= \sqrt{45.56}$$

$$= 6.75$$

***How does the outlier affect variability?

E. Why is the standard deviation so important? Why do we care about variability?

The standard deviation is one of the most important indicators of how well the sample reflects the population from which it was drawn. A sample that has low variability will better represent the population than a sample that has high variability. (Although, if you sampled incorrectly

from the population, e.g. didn't represent ethnicity in the right proportions, your sample will never represent ethnicity in the population regardless of the amount of statistical variability of the observations.)

Look at this SPSS output—a comparison of the average scores from the two exams from last semester (both 240 sections). For both tests, the highest possible grade was 15. What can you infer from the difference in the means and standard deviations about how students performed in the 2nd test vs. the 1st one?

Comparison of Means, ScWk 240 Exams, Fall 2008

	N	Range	Mean	Std. Deviation	Variance
Test_1	50	4.00	14.2549	1.29373	1.674
Test_2	50	4.00	14.6800	.84370	.712

What difference does sample size make?

What if we randomly sampled five students' scores—what happens to the standard deviation?

	N	Range	Mean	Std. Deviation	Variance
Test_1	5	4.00	13.6000	1.94936	3.800
Test_2	5	2.00	14.6000	.89443	.800

What does this say about the importance of sample size?

***Discussion question: why not just use the variance when analyzing variability?