Warsaw Summer School 2023, OSU Study Abroad Program

Frequency Distribution

Central Tendency

Why Do We Test Hypotheses?

- **Hypothesis testing is a foundation of science.**

- **In statistical inference, hypotheses generally take one of the two forms: substantive and null.**
- **A *substantive hypothesis* represents an actual expectation. E.g.: higher education increases the likelihood of upward mobility.)**
- **To decide whether a substantive hypothesis is supported by the evidence it is necessary to test a related hypothesis called the *null hypothesis*. (E.g.: education has no effect on upward mobility.)**

A Framework for Statistical Work

Units of observation/analysis (cases)

Variables: data characterizing units of observation

Levels of Measurement

The level of measurement of a variable refers to the type of information that the numbers assigned to units of observation contain.

Four levels of measurement:

- **nominal (categorical; discrete)**
- **ordinal (rank-order)**
- **interval (distance)**
- **ratio (zero-reference)**

Specifying Levels of Measurement

- **R distinguishes between the scale level (that is numerical: interval and ratio) from ordinal and nominal levels.**

Recoding

Recoding into “metric variables”

- Any nominal variable can be recoded into a set of 0,1 variables, called also dummies.
- Ordinal variables can be recoded into interval variables, if

  (a) ranks are interpretable as having a property of equal distances between them (e.g. Likert scale);

  (b) we can assign some know values to the ranks on the basis of this variable (e.g. years of schooling);

  (c) we can derive the values from the distribution properties (e.g. mid-points of the cumulative distribution);

  (d) we can assign some values to the basis of another (correlated) variable.

Mid-points

- L150, 1-50 = 25
- L2 30, 51-80 = 65
- L3 20, 81-100 = 90

Frequency distribution

- **A frequency distribution is the simplest way of representing sociological observations. It contains at least two columns: the left-hand one contains the values that a variable may take, and the right-hand one contains the number of times each value occurs. Additional right-hand columns show the <u>percentage distribution</u> in two forms: unadjusted and adjusted for missing data:**

Value	Frequency	Unadjusted %	Adjusted %
1	2 300	56.1	60.5
2	1 500	36.6	39.5
3 (missing)	300	7.3	-----
Total	4 100	100.0	100.0

Counting

Proportions and Percentages (all type of variables)

- **A *proportion* is a special ratio by which a subset of frequencies in a distribution is divided by the total number of cases.**

  Proportion = f(i) / N

  Proportion = Part/Whole

- **A *percentage* is a proportion multiplied by 100:**

  Percentage = (f(i) / N) * 100

%

- **1 Strongly agree 20**
- **2 Agree somewhat 30**
- **3 Not sure 20**
- **4 Disagree somewhat 15**
- **5 Strongly disagree 15**

Counting

Cumulative Distributions

Cumulative percentage (c%) **is the percentage of cases having any given score or a score that is lower. To calculate the cumulative percentage, we use the formula:**

$c\% = (cf/N) * 100$ where cf = cumulative frequency

Cumulative frequencies (cf) **are defined as the total number of cases having any given value or a value that is lower. The cumulative frequency *cf* for any value is obtained by adding the frequency for that value to the total frequency for all scores below.**

Charts

- **PIE CHART: Graph drawn as a circle where the category (value) is paired with a segment that represents the frequency of that category (value).**
- **BAR GRAPH: Graph of the data where the category (value) is paired with a bar that represents the frequency of that category (value). Used with qualitative data.**
- **HISTOGRAM: Graph of the data where the bars for scores or intervals are connected. Used also with quantitative data.**
- **FREQUENCY POLYGON: Graph in which a smooth line connects the top of the bars in a histogram.**

Central Tendency and Dispersion

Two basic characteristics to describe with descriptive statistics:

- the middle of the distribution: Central Tendency Measures, CT

- how spread out the distribution is: Measures of Variability (Dispersion), VR

Central Tendency, CT, is a way to describe the overall trend of a variable in one number. It says something about a typical response in the data; refers to the middle of the distribution.

Central Tendency

Three measures of CT, depending on level of measurement of the variables:

Mode

Median

Mean

Mode = the most common response; the value with the largest frequency. The only way we can measure CT for nominal variables; however it is also use for ordinal and interval variables.

Central Tendency

- **<u>Median</u> = the middle point of the distribution; <u>value</u> where 50% above and 50% below.**

  The middle point of the distribution: variable's value for the case (N+1) / 2. However, for the continues distribution or intervals, the formula is:

Median = L + [(N/2 – cf_b) / f] * i

L = lower limit of the interval, N = total number of cases,

f = frequency within the critical interval, i – interval size

cf_b = cumulative frequency below the lower limit of the critical interval

Median

Meadian value?

Eg. I 1,3,6,9,14,15,18,25,30,31,32

Eg. II

1-2	**10**
3-4	**20**
5-6	**40**
7-8	**20**
9-10	**10**

$$5 + [(50 - 30) / 40] * 1 = 5.5$$

Central Tendency

Median = L + [(N/2 – cf_b) / f] * I

L = lower limit of the interval, N = total number of cases,

f = frequency within the critical interval, i – interval size

cf_b = cumulative frequency below the lower limit of the critical interval

5 + [(50 – 30) / 40] * 1 = 5.5

Median: for both ordinal and interval variables; not for nominal.

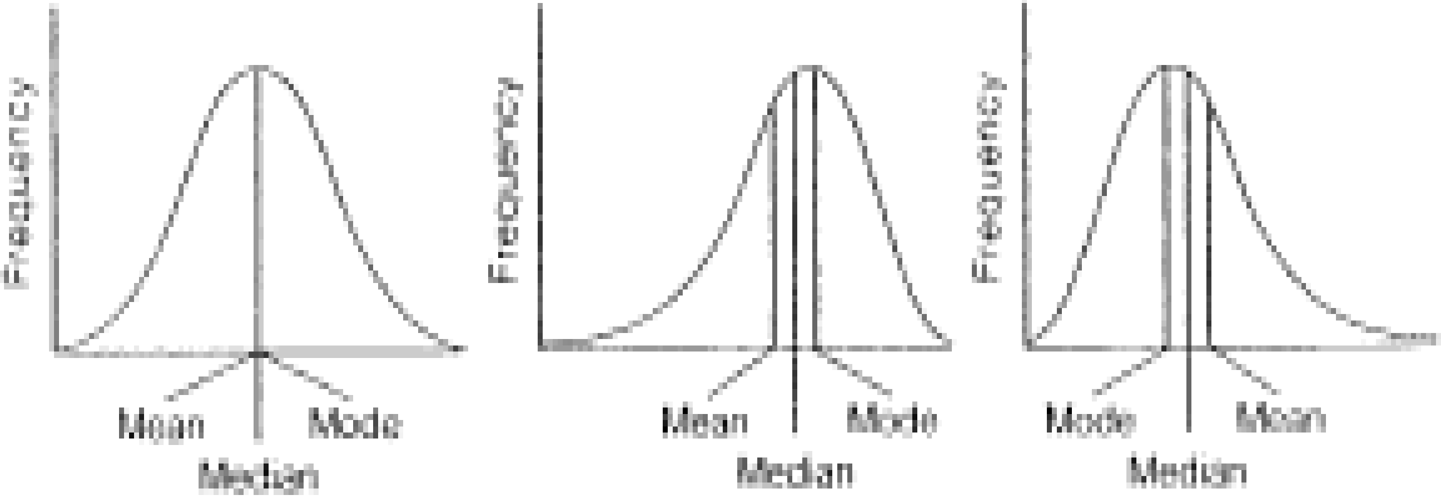
Central Tendency

Mean = the average.

- Symbol: bar over the letter used for the particular variable. (If X symbolizes age, the x with bar is the mean age)
- Mean = (ΣX_i) / N, where X_i means "the value for each case," and Σ means "add all of these up."

Mean: only for metric data!

Figure 4.6 **Types of Frequency Distributions**



a. Symmetrical distribution

b. Negatively skewed distribution

c. Positively skewed distribution

Summary of CT

Scale	Mode	Median	Mean
Nominal	Yes	No	No
Ordinal	Yes	Yes	No
Interval	Yes	Yes	Yes