Introduction to Statistics for Psychologists
Claremont McKenna College
Professor Cook

Measurement, Frequency, & Probability Distributions

1. Measurement is the *systematic assignment* of numbers to an event, or object, based on the properties inherent in that event, or object. We begin with a discussion of measurement and its characteristics.

   (a) Types of variables

      i. The **independent variable (IV)** (also called a *factor*) is what the experimenter manipulates in order to influence, or change, some aspect of behavior (e.g., manipulating mood by showing cartoons or wars scenes).

      ii. The **dependent variable (DV)** (also called a *measure* or *outcome variable*) is used to describe the behavioral outcome (e.g., smiling) of an experimental situation that usually changes as a function of the independent variable.

         - The dependent variable is also called a **unit of analysis** because statistical analyses are conducted on the DV measured in experiments.

         - In terms of measurement, X_i represents the DV or the outcome variable

      iii. **Extraneous** variables, or **confounds**, are uncontrolled influences that affect the dependent variable. Extraneous influences can increase in the likelihood of making decision errors when conducting research.

      iv. Qualitative, quantitative discrete, quantitative continuous variables

(b) Ideally, the dependent variable would be a precise measure of behavior. However, the object of measurement, the measurement environment, and the measurement device often prevent precise measurements. Thus, each data point, X , is a composite score reflecting the influence of the IV, DV, controlled variables, and uncontrolled variables. **Measurement error** takes on a few forms, some are good and some are bad.

i. The *unit of analysis* for each measured object $\left(X_i\right)$ is composed of **true ability** $\left(\tau\right)$ and some degree of **error** $\left(\varepsilon\right)$; $X = \tau + \varepsilon$

ii. In many statistical models, **systematic** influences are reflected by variables that are under the analyst's control (e.g., true ability determined by an IV), whereas **nonsystematic** influences are considered outside the analyst's control (e.g., random error).

iii. However, error may be **systematic** $\left(\varepsilon_S\right)$ (e.g., room temperature, time of day, or some variable that you did not measure) or it may be **random** $\left(\varepsilon_R\right)$ or unexplainable (e.g., accidentally responding correctly or incorrectly).

iv. Error may also be **intrinsic** to the unit of analysis (e.g., stress, headache, accidentally circling the incorrect answer on the SAT, etc.) or **extrinsic** (e.g., room temperature, time of day, faulty measurement device, etc.).

v. A key component of statistics is making inferences about unknown data based on known data. In order to increase inferential accuracy, we should aim to reduce error so that we can make inferences based on scores that approximate true ability.

vi. All statistics are essentially a comparison of systematic variance (good stuff) to unsystematic variance (bad stuff) (e.g., regression, *t*-tests, ANOVAs, etc.)

(c) Whether or not in error, measurements can be scaled hierarchically according to their characteristics. We use **scales of measurements** (Stevens, 1946; 1951) for categorizing variables according to their quantitative properties. Measurement scales take one of 3 (or 4) different forms that ultimately assign values to variables.

i. The **nominal scale** is used to describe objects that belong to qualitatively different groups or categories; numeric assignment is arbitrary; values have no quantitative meaning (e.g., major, political affiliation, etc.).

ii. The **ordinal scale** is used to describe the rank order of objects based on some property; numeric values reflect order only (e.g., preference ranking for candy bars, Moh's scale, floors in a building, etc.).

iii. Equal-interval data

- The **interval scale** is used to describe quantitative differences of objects based on a measurable property; numeric values reflect rank and the interval between values is equal. No true zero exists (e.g., Fahrenheit, Celsius, SAT scores, etc.).

- The **ratio scale** is used to describe quantitative differences of objects based on a measurable property; numeric values reflect rank, the interval between values is equal, and a true zero-point exists. Numeric values reflect magnitude information about the property (e.g., Kelvin, height, weight, etc.).

iv. N.B. On exams, you are expected to know these four "scales of measurement"; no excuses. Master these four scales of measurement in order to survive this class.

2. A frequency distribution arranges event classes systematically and shows the number of events or objects (i.e., frequency) comprising each class.

   (a) Visual representations of data

      i. Tables/tabular forms

      ii. **Bar graphs** and **pie charts** for *qual*itative variables

      iii. **Histograms** (frequency distributions) for *quant*itative variables

      iv. **Frequency polygons** for *quant*itative variables

3. A close kin to the frequency distribution is the probability distribution. The machinery of inferential statistics depends critically on the concept of probability distributions and the random variables that these distributions depict.

   (a) Random variables

   (b) **Probability** distributions simply represent a distribution of relative frequencies of events (e.g., $90/100, 2/36,$ etc.) for discrete or continuous variables

Homework #1

Due: See Syllabus

Unit of analysis , Types of variables (independent, dependent, extraneous, qualitative, quantitative discrete, quantitative continuous), Types of measurement error (systematic, random, intrinsic, extrinsic), Scales of measurement (nominal, ordinal, interval, ratio and examples of them) Frequency distributions (ungrouped, grouped), Visual representations of data (tables, bar graphs, pie charts, histograms, frequency polygons), Shapes of frequency distributions (bimodal, rectangular, positively skewed, negatively skewed, Normal, symmetrical, rectangular, platykurtic, and leptokurtic distributions)

After reading this Topic, you should be able to:

1. identify the two major branches (or types) of statistical methods
2. describe the differences between the two major branches (or types) of statistical methods
3. provide an example of descriptive and inferential statistics
4. explain the difference between a variable and a constant
5. explain the difference between an independent variable and a dependent variable
6. explain the different levels or scales of measurement
7. identify variables that are nominal, ordinal, interval, or ratio in nature
8. identify variables that are discrete versus continuous
9. identify variables that are qualitative versus quantitative
10. provide examples of discrete and continuous variables
11. provide examples of qualitative and quantitative variables
12. explain why we use frequency tables
13. create intervals for a grouped frequency table
14. ~~explain the suggested number of intervals to include in a grouped frequency distribution~~
15. explain what information in an ungrouped frequency distribution allows for it to be preferred (sometimes) over a grouped frequency distribution
16. explain what information in a grouped frequency distribution allows for it to be preferred (sometimes) over an ungrouped frequency distribution
17. locate and label the X and Y axes on a graph
18. locate and label the abscissa and ordinate on a graph
19. explain when you would use a particular type of graph rather than another type (hint: types of variables)
20. know how to create, label, and interpret a histogram
21. know how to create, label, and interpret a bar graph
22. know how to create, label, and interpret a pie chart
23. know how to create, label, and interpret a frequency polygon
24. know how to create, label, and interpret a line graph
25. ~~recognize and correct misleading graphs based on your knowledge about how to construct graphs appropriately~~
26. define and identify an experimental unit vs. unit of analysis
27. explain the difference between intrinsic and extrinsic error
28. explain the concept of measurement error
29. explain the difference between systematic and random error
30. explain the concept of extraneous variables

Notes

Introduction to Statistics for Psychologists
Claremont McKenna College
Professor Cook

Descriptive Statistics: Central Tendency & Variability

1. Although both frequency and probability distributions are quite useful, more succinct methods of summarizing the outcomes of simple experiments are needed. In general, distributions are depicted well by two classes of measurement: **central tendency** & **dispersion** (i.e., variability/deviance).

   (a) Central tendency for frequency distributions

      i. The **mode** describes the most frequently occurring score in a distribution of scores. Often symbolized as *Mo.*

      ii. The **median** describes the score about which 50% of the scores fall above and below. Symbolized as *Mdn*. The middle score or the midpoint between two middle scores. How is the median affected by extreme scores?

      iii. The **mean** describes the average of all the scores in your distribution. Symbolized as $\overline{X}$ (or less commonly, M). The formula is $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{N} = \frac{\sum X_i}{N}$, where X represents the measurement of some property (e.g., a participant's reaction time). How is the mean affected by extreme scores?

      iv. Physical analogies of the mean (as a segue into dispersion): $\sum_{i=1}^{n}\left(X_i - \overline{X}\right) = 0$

(b) Central tendency measures help describe scores that best represent a sample or a population; scores most central to the entire distribution. However, all scores help define the shape of a distribution. The **shape of a distribution** of scores will change depending on whether most scores are clustered near its center, in the negative direction, or in the positive direction. As such, skewed distributions will affect central-tendency measures as well as the preference for using.

i. **Symmetrical** distributions

ii. **Negatively-skewed** distributions $\left(Sk < 0\right)$

iii. **Positively-skewed** distributions $\left(Sk > 0\right)$

2. Measures of central tendency depict one aspect of a distribution (*viz.*, location). That value, however, is relatively uninformative about the distribution of scores unless it is supplemented with a measure of **dispersion** (i.e., how variable or "spread out" the scores are in a distribution). There are several ways to numerically represent dispersion, including:

(a) **Range**: *noninclusive* $\left(R_{ni} = X_{hi} - X_{lo}\right)$

(b) **Semi-interquartile range (SIR)** $\left(Q = \frac{Q_3 - Q_1}{2}\right)$

- Because the semi-interquartile range depends only on the 50% of the scores nearest the mean, the SIR is not affected much by outliers, or extreme scores.

(c) **Variance** is a measure of the **average squared deviations** about the mean.

Unbiased Sample: $SD^2 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{N-1}$ *or* $\frac{SS}{N-1}$;
SS = Sum of Squared deviations OR Sum of Squares

Population: $\sigma^2 = \frac{\sum_{i=1}^{n}\left(X_i - \mu\right)^2}{N}$

Although the *variance* is not as typically used as a descriptive statistic (whereas the *standard deviation* is), the variance is nevertheless used extensively in various statistical analyses. You must know it.

Because individual scores are deviations from the mean, sometime you might see the formula written in shorthand as $\frac{\sum d_i^2}{N-1}$ where d stands for *deviations.*

(d) **Standard deviation** is a measure of average variation; the square root of the average squared deviation (or variance).

Unbiased Sample: $SD = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{N-1}}$ *or* $\sqrt{\frac{SS}{N-1}}$

Population: $\sigma = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i - \mu\right)^2}{N}}$

3. Although the median *could* be used as the measure of central tendency, in practice the median has historically been reserved for skewed distributions.

(a) Medians and Average Deviations (AD) - $\left(A.D. = \frac{\sum_{i=1}^{n} | x_i - Mdn |}{N} \right)$

(b) The mean and the variance are only two summary measures that characterize a distribution; other *moments* describe additional aspects of a distribution. We can understand the shape of distributions by considering the four moments of distributions

i. First moment, ***mean***, measure of center

ii. Second moment, ***variance***, measure of dispersion

iii. Third moment, ***skewness,*** represents a measure of a distribution's symmetry. Asymmetrical distributions are skewed either positively or negatively:

iv. Fourth moment, ***kurtosis***, represents a measure of a distribution's central peak and the fullness of its tails. **Leptokurtic** distributions have a positive peakedness, whereas **platykurtic** distributions have a negative peakedness.

Key terms, Concepts, Important Formula, and Study Tips

Measures of central tendency (mean, median, mode), Measures of dispersion/variability (variance, standard deviation), Samples statistics versus population parameters, Shapes of distributions (mesokurtic, platykurtic, leptokurtic, skewed positively, skewed negatively)

$$\left(R_{ni} = X_{hi} - X_{lo}\right) \qquad \sum\left(x_i - \overline{X}\right) = 0 \qquad \frac{\sum d_i^2}{N-1} \qquad \frac{\sum\left(X_i - \overline{X}\right)^2}{N-1} \qquad \frac{SS}{N-1}$$

$$\frac{\sum\left(X-\mu\right)^2}{N} \qquad \sqrt{\frac{\sum\left(X-\overline{X}\right)^2}{N-1}} \qquad \sqrt{\frac{SS}{N}} \qquad \sqrt{\frac{\sum\left(X-\mu\right)^2}{N}}$$

After reading this Topic, you should be able to:

1. explain the concept of central tendency
2. identify and calculate the three measures of central tendency
3. mathematically explain why the mean is analogous to a fulcrum
4. determine what measure of central tendency is best for describing different samples of data
5. locate the three measures of central tendency on symmetrical distribution
6. locate the three measures of central tendency on skewed distributions
7. explain why/when you should prefer one measure of central-tendency over another
8. explain the concept of dispersion; identify the 3 most commonly used measures of dispersion
9. explain the concept of variance
10. explain the concept of sums of squares
11. explain what the variance tells you about a distribution of scores and how that information is different from what a mean tells you about that same distribution
12. calculate the population variance when given sums of squares (definitional formula)
13. calculate the population standard deviation when given a set of scores from a population (definitional formula)
14. explain the relationship between the variance and the standard deviation
15. calculate the sample variance when given a sample of scores (definitional formula)
16. calculate the sample variance if you are given the sums of squares from a sample of scores
17. calculate the sample standard deviation if given a set of scores (definitional formula)
18. calculate the sample standard deviation if given the sums of squares from a sample of scores
19. explain the meaningfulness of the mean for nominal, ordinal, interval, and ratio data
20. identify the formulae for calculations of a sample's mean, variance, and standard deviation
21. identify the shapes of distributions
22. provide an example of rectangular, symmetrical, positively-skewed, and negatively-skewed distributions
23. identify a data set as either positively or negative skewed
24. identify a distribution as bimodal or multimodal
25. provide examples of symmetrical and skewed distributions
26. create and label a bar graph if given some sample means
27. understand that a "floor effect" describes a situation when most scores in a distribution are near the lower end (floor) of the distribution with fewer scores at the top end (ceiling); this results in positive skew
28. understand that a "ceiling effect" describes a situation when most scores in a distribution are near the higher end (ceiling) of the distribution with fewer scores at the bottom end (floor); this results in negative skew

Homework #2

Due: See Syllabus

See Homework Packet

Notes