Descriptive Statistics

This chapter introduces descriptives statistics. Descriptive statistics are mathematical concepts that describe–as the name implies–the characteristics of a sample. In this chapter only the basic descriptives will be covered.

Review: Measurement Scales

Understanding descriptive statistics requires paying attention to the type of data that are being described. As a review, recall that we distinguish among four types of measurement scales: nominal, ordinal, interval, and ratio.

Nominal scales simply place numeric labels on otherwise non-quantitative concepts, such as giving males a value of "1" and females "2". Many category schemes are essentially nominal variables, such as religion, nationality, and ethnic group. In some research studies, we count how many people fall into the levels of a category system, such as how many of our subjects come from various nations.

Ordinal scales allow us to rank items in a comparative manner, such as rank in a class or highest to lowest income in a community. Mathematical operations cannot be performed on ordinal data.

Interval scales allow us to assign values to levels of a phenomena such that the numeric values are on an equal interval scale, such as a 7-point attitude scale. Because the values are on an equal interval scale, we can do some mathematical operations such as adding and averaging on such data.

Ratio scales are interval scales with a true zero point, such as weight or number of number of observed behaviors. Having a true zero point allows us to form ratios with such data.

Central Tendency

We often need to find the middle of a distribution. The middle point has various meanings depending on the measurement scale on which the data were collected.

The Mode

On a nominal scale, the "middle" is the most frequent category, termed the **mode**. In a sample of Introductory Psychology subject pool research participants, we might look at the distribution of participants on the nominal scale "nation of origin." For example, if the subject pool had 100 students:

US citizens of all ethnic groups:	75
French:	10
Indian:	5
East Asian (all nations combined):	5
Arab (all nations combined):	3
Caribbean (all nations combined):	2

The most common category of this sample is US citizens, so the mode is "US citizens."

The mode can be used with any measurement scale. On a 5-point Likert opinion scale, we might find this distribution of responses: (values are counts of how many people out of a sample of 100 chose each answer)

Strongly agree:	10
Agree:	15
Neutral:	20
Disagree:	35
Strongly disagree:	20

The modal response is "disagree."

The Median

The median requires ordinal, interval or ratio scale data. Such measures can be ranked, and the middle case (i.e., person) in the ranking can be identified. For example, it is common to rank all Americans according to their incomes and determine the median income. "Median cost of single family house" is also commonly reported. Here are the yearly salaries of the Psychology department faculty, ranked:

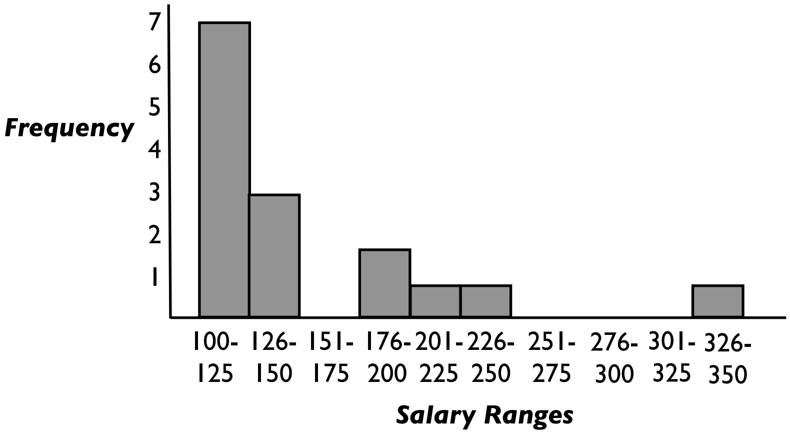
Dean	$350,000
Undergrad Chair	$250,000
Director of Clinical Training	$225,000
Chair of I/O Program	$185,000
Chair of ABA Program	$185,000
Faculty member #1	$150,000
Faculty member #2	$140,000
Faculty member #3	$130,000
Faculty member #4	$120,000
Faculty member #5	$110,000
Faculty member #6	$105,000
Faculty member #7	$104,000
Faculty member #8	$103,000
Faculty member #9	$103,000
Faculty member #10	$101,000

First, note that all faculty earn six-figure incomes, because they are brilliant and work so hard.[1] The middle of this distribution is Faculty member #3, earning a cool $130K. So "the median psychology faculty income" is $130,000.

[1] If you believe these salaries are accurate, you are aren't reading the *Chronicle of Higher Eduction* as much as you should.

The Mean

The mean is the arithmetic average of the distribution. Means can only be calculated for interval and ratio scale data because ordinal data cannot be added. The average of the psych department faculty salaries is $157,400. The mode is $100-125K. Note that the mean is not the same as the median and both are far from the mode. This happened because the shape of the distribution is not balanced or symmetrical: a few faculty make some real money while a large number are just getting by. A histogram illustrates how odd the distribution is:



The distribution is **skewed**. Skewing takes two directions: toward the high end and toward the low end of the distribution. In this case, we would say that the distribution is **positively skewed.** The positive skew in this distribution is extreme because one faculty member, the Dean, is making so much money.

When distributions are skewed, the median provides a better sense of the middle. The salary distribution illustrated here is similar to the distribution of income in the United States, where median income is around $40,000 but the range is from zero to tens of millions. The median tells us about "the average guy" better than the mean.

Variability

Just as important as central tendency, we must be able to quantify the variability or dispersion of a distribution. Variability refers to the extent to which values on the distribution differ from its middle. Variability is most important and most frequently used in distributions involving interval and ratio scale data, but some measures can be used for ordinal data.

Range

The range is the simplest and least interesting measure of variability. As one might suspect, it is the maximum spread of the data. Technically, it is the highest value minus the lowest value, plus 1. The range in faculty salaries is 350,000 - 101,000 + 1 = 249,001. The "+1" part of the equation is often forgotten, and for numbers this large, it is trivial.

Semi-Interquartile Range

When values in a distribution are ordered by value (regardless of whether they are ordinal or interval measures), they can be divided into quarters: top quarter, bottom quarter, etc. (The two middle quarters are divided by the median, right?) The difference between the top of the high middle quarter and the bottom of the low middle quarter, divided by 2, is the **semi-interquartile range**. Calculating the SIR is complicated because the precise values at the top of the high middle quarter and at the bottom of the low middle quarter must be interpolated from the particular values that appear in the data.

	High Quarter	$450,000
		$350,000
		$250,000
		$225,000
Interquartile Range	High Middle Quarter	$185,000
		$185,000
		$150,000
		$140,000
	Low Middle Quarter	$130,000
		$120,000
		$110,000
		$105,000
	Low Quarter	$104,000
		$103,000
		$103,000
		$101,000

Average Deviation

The average deviation (AD) is the mean of the absolute values of the differences between the data and their mean. In other words, calculate all the differences from the mean, discard the signs (positive or negative) and calculate the average. The AD of the salary data is 54,400. The AD is intuitively appealing but otherwise useless because it has mathematical properties of little value to statistical analyses that require measures of variability.

Standard Deviation

The standard deviation (SD) is similar to the average deviation except that the differences between the values and their mean are squared prior to averaging (then the square root of the average is taken to bring the SD down to units similar to the original data).

$$\mathrm{SD} = \sqrt{\frac{\sum(X_i - M)^2}{N}}$$

where M is the mean

The SD squared is the **variance**, another useful and common measure of variability.

The SD of the salary sample is 71,321. Note that the SD is so much larger than the AD. This difference is due to the squaring that takes place in the numerator of the formula: squaring gives higher weight to values far from the mean, and because of the distribution's positive skew, one value is way out and contributes to a large SD. It is common in data analysis to refer to these deviant values as "outliers" and to find justifications to drop them from the dataset.