MAKING SENSE OF DATA

Essentials series

DESCRIPTIVE STATISTICS

Prerequisites
Descriptive statistics
Charts and graphs
The normal distribution
Surveys and sampling
Correlation and regression

Descriptive statistics are numbers that summarise some characteristic about a data set. They provide easy-to-understand information that helps answer questions. They enable researchers to get a rough idea about what's happening for more targeted analyses. Descriptive statistics make a point clearly and concisely.

Types of data

Data come in a wide range of formats. For example, the Census might ask questions about gender, ethnicity, or education, while other questions might be about age, occupation, or the distance travelled to work each day. Different types of questions result in different types of data to be collected and analysed. The type of data you have determines the type of descriptive statistics that can be found and interpreted.

There are two main types of data:

- **Categorical** – or qualitative data, and
- **Numerical** – or quantitative data

Categorical (or nominal) data record qualities or characteristics about the individual, such as eye colour, gender, political party, or opinion on some issue using categories such as agree, disagree, or no opinion. *Numerical data* record measurements or counts regarding each individual, which may include weight, age, height, or time to take an exam; counts may include number of pets, or the number of red lights you hit on your way to work. The scales used with numerical data include *interval* (e.g. temperature), or *ratio*.

The important difference between the two is that with categorical data, any numbers involved do not have real numerical meaning (e.g., using 1 for male and 2 for female), while all numerical data represents actual numbers for which mathematical operations make sense.

A third type called *ordinal data* falls between the two, where data appear in categories, but the categories have a meaningful order, such as ratings from 1 to 5. Ordinal data can be analysed like categorical data, and the basic numerical data techniques also apply when categories are represented by numbers that have meaning.

Counts and percents

Categorical data place individuals into groups. For example, male/female, own your home/don't own, or Labour/Conservative/Liberal. Categorical data often come from survey data such as the Census.

Categorical data are typically summarised by reporting either the number of individuals falling into each category, or the percentage of individuals falling into each category. For example, pollsters may report the percentage of Labour, Conservative, Liberal voters who took part in a survey. To calculate the percentage of individuals in a certain category, find the number of individuals in that category, divide by the total number of people in the study, and then multiply by 100%. For example, if a survey of 2,000 teenagers included 1,200 females and 800 males, the resulting percentages would be (1,200 ÷ 2,000) * 100% = 60% female and (800 ÷ 2,000) * 100% = 40% male.

You can further break down categorical data by creating crosstabs. *Crosstabs* (also called *two-way tables*) are tables with rows and columns. They summarize the information from two categorical variables at once, such as gender and political party, so you can see (or easily calculate) the percentage of individuals in each combination of categories. For example, if you had data about the gender and political party of your respondents, you would be able to look at the percentage of Conservative females, Labour males, and so on. In this example, the total number of possible combinations in your table would be the total number of gender categories times the total number of party affiliation categories.

If you're given the number of individuals in each category, you can always calculate your own percents. But if you're only given percentages without the total number in the group, you can never retrieve the original number of individuals in each group.

Measures of center

The most common way to summarize a numerical data set is to describe where the center is. One way of thinking about what the center of a data set means is to ask, "What's a typical value?" Or, "Where is the middle of the data?" The center of a data set can be measured in different ways, and the method chosen can greatly influence the conclusions people make about the data. The two most common measures of center: the mean (or average) and the median.

The **mean** (or average) of a data set is simply the average of all the numbers.

$$\bar{x} = \sum x_i / n$$

Where $\bar{x}$ = the mean
$\sum x_i$ = the sum of values
n = the count of values

For example, the mean of the five values: 4, 36, 45, 50, 75 is

$$\bar{x} = \frac{4 + 36 + 45 + 50 + 75}{5}$$

$$= \frac{210}{5} = 42$$

The **median** of a data set is the place that divides the data in half, once the data are *ordered* from smallest to largest. It is denoted by M or $\tilde{x}$. For example, consider the sequence 1, 2, 2, 6, 13 14. The place that divides this sequence in half is the average of the 3rd and 4th value, calculated as

$$M = (2 + 6)/2 = 4$$

Note that if the sequence has an odd number of values, the median will be one of the numbers in the data set itself.

A third less often used measure is called the **mode**, which is defined as the most frequently occurring value.

Measures of variability

Variability is what the field of statistics is all about. Results vary from individual to individual, from group to group, from city to city, from moment to moment. Variation always exists in a data set, regardless of which characteristic you're measuring, because not every individual will have the same exact value for every characteristic you measure. Without a measure of variability you can't compare two data sets effectively.

By far the most commonly used measure of variability is the standard deviation. The *standard deviation* of a data set, denoted by s, represents the typical distance from any point in the data set to the center. It's roughly the average distance from the center, and in this case, the center is the average. Most often, you don't hear a standard deviation given just by itself; if it's reported it's usually in the fine print, in parentheses, like "($s = 2.68$)."

The formula for the standard deviation of a data set is

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

Where
x = a value
$\bar{x}$ = the mean
$\sum(\ )^2$ = the sum of $(x - \bar{x})$-squared
$\sqrt{\ }$ = the square root
n = the count of values

Suppose you have four numbers: 1, 3, 5, and 7. The mean is $16 \div 4 = 4$. Subtracting the mean from each number, you get $(1 - 4) = -3$, $(3 - 4) = -1$, $(5 - 4) = +1$, and $(7 - 4) = +3$. Squaring the results you get 9, 1, 1, and 9, which sum to 20. Divide 20 by $4 - 1 = 3$ to get 6.67. The standard deviation is the square root of 6.67, giving $s = 2.58$.

Here are some properties that can help when interpreting a standard deviation:

- ✓ The standard deviation can never be a negative number.
- ✓ The smallest possible value for the standard deviation is 0 (when every number in the data set is exactly the same).
- ✓ Standard deviation is affected by outliers, as it's based on distance from the mean, which is affected by outliers.
- ✓ The standard deviation has the same units as the original data, while variance is in square units.

Percentiles

The most common way to report relative standing of a number within a data set is by using *percentiles*. A percentile is the percentage of individuals in the data set who are below where your particular number is located. If your exam score is at the 90th percentile, for example, that means 90% of the people taking the exam with you scored lower than you did.

To calculate percentiles, sort the data values so that x_1 is the smallest value, and x_n is the largest, with $n =$ total number of values.

x_i is the p_ith percentile of the data set where:

$$p_i = 100.\frac{i - 0.5}{n}$$

For example:

(original data)

5	1	9	3	14	9	7

(sorted data)

x_i	1	3	5	7	9	9	14
i	1	2	3	4	5	6	7
p_i	(calculate, using this equation, as shown below…)						

$p_1 = 100(1 - 0.5) / 7 = 7.1$
$p_2 = 100(2 - 0.5) / 7 = 21.4$
$p_3 = 100(3 - 0.5) / 7 = 35.7$
$p_4 = 100(4 - 0.5) / 7 = 50$
etc…

(filling in the final row, we get)

x_i	1	3	5	7	9	9	14
i	1	2	3	4	5	6	7
p_i	7.1	21.4	35.7	50.0	64.3	78.6	92.7

A percentile is *not* a percent; a percentile is a number that is a certain percentage of the way through the data set, when the data set is ordered.