Chapter 1 - Looking at Data: Distributions

Introduction

In science, we run various experiments and collect data - qualitative or quantitative observations of the objects we want to study. Statistics is the science of learning from this data. The goal of statistics, particularly in psychology, is to be able to make predictions about a population based on a sample.

1.1: Data

In order to be able to do statistics, you must begin with a set of data. There are many components that make up a set of data:

- *Cases* are the objects that your data describes
- A *variable* is a particular trait of a case
  - A *qualitative* or *categorical* variable is a variable that does not have a numeric value. They place cases into groups of categories
  - A *quantitative* variable is a variable that has a numeric value
- A *label* is a special variable used to tell the different cases apart
- A *value* is something that a variable holds
- A *distribution* of a variable tells us what values the variable takes, and how often these values occur for that particular variable

For example: If you think about a list of psychology students in Groningen, each student is a case; characteristics about them, such as age, sex, year of study, etc. are variables. Each student has a student number - this is a label. If a student, for example, were 24 years old, 24 would be the value associated with the variable "age."

Key Characteristics of a Dataset

Data sets often come with background information that helps us understand the data better. The following traits are important to note:

- Who/What are we describing? What are the cases and how many cases are included in the dataset?
- What are the variables, how are they defined, what are the units of measurement for each quantitative variable, and what are the definitions of these variables?
- Why are we looking at this data? What is the greater purpose? Are the recorded variables appropriate for answering the research question?

In considering your variables, it is important to ask yourself whether the variables measure what you actually want to measure. If you find that the variables in your data set do not align with the goal of your research, it is also possible to create a new variable by adjusting another variable. For example, if your recorded variables are time and distance, but your research actually wants to investigate velocity, you could simple make a new variable, velocity, by dividing distance by time.

1.2: Displaying Distributions with Graphs

Exploratory data analysis is the examination of data to describe their main features. You can either begin by looking at each variable individually and then seeing how they all relate to each other, or by examining graphs and adding appropriate numerical summaries to them.

For Categorical (Qualitative) Variables: Bar Graphs and Pie Charts

For *categorical variables*, bar graphs and pie charts are used because of the nature of the distribution of these *qualitative* variables. The distribution of categorical variables lists each category and gives the *count* or *percentage* of cases that fall in each category. These distributions can then be turned into bar graphs or pie charts.

- A *bar graph* lists each category (in any order) along the X-axis and the count along the Y-axis. Bars for each category are then drawn according to the count of each category. While the categories can be listed in any order, you should consider presenting your data in an order that makes sense to you and fits the purpose of your research
- A *pie chart* shows the percentages of the total count that each category takes up. Here, it is important to include every category, so that the total of the percentages is always 100%. In some cases, when specific categories have very low counts, it is acceptable to include an "Other" category on the pie chart.

For Quantitative Variables: Stemplots and Histograms

For qualitative variables, stemplots (or stem-and-leaf plots) and histograms are used.

- In a stemplot, each observation is separated into a stem and leaf. The stem is everything except the last digit in the value and the leaf is simply the last digit. In a vertical column, the stems of the dataset are written from least to most, and each leaf is written in the row of its corresponding stem, in ascending order. This allows us to have an overview of our dataset.
    - If you would like to compare two distributions for the same variable, for example the IQ scores of boys vs. girls, you can construct a *back-to-back stemplot,* with common stems and leaves on either side of the stem.

- If your dataset is extremely large, it might be useful to make a *split stemplot*, by splitting each stem into two: one for leaves from 0-4 and one for leaves from 5-9.

- A Histogram separates the range of values into classes and shows the count or percentage of each class, similar to a bar graph. In a histogram, any number of classes can be used; however it is important to use classes of equal width.
  - In a histogram, we react to the area (size) of the bars in the graph. By using bars of the same width, we ensure that all of the classes are fairly represented.
  - Tip: Play around with the classes and find the right amount and ranges to make an aesthetically representative graph. Too many classes may result in a "skyscraper" effect, while too few may lead to an overly flat graph.

Bar Graphs vs. Histograms

\While bar graphs and histograms share many characteristics, there are several notable differences:

Bar Graphs	**Histograms**
• are used for qualitative or categorical variables • compare the counts of different items • do not need to have a measurement scale on the X-axis • have spaces between the bars	• are used for quantitative variables • show the distribution of counts of a variable • use a continuous scale along the X-axis • do not have spaces between bars

Examining Distributions

By plotting your data, you can make statistical graphs to help you understand your data. In examining your graph, there are several features you should pay attention to. The *tails* of a graph refer to their extreme values of distribution. The higher values make up the *right tail* or *high tail,* and the lower values make up the *left tail* or *lower tail.*

Look at the shape of your graph and try to see an *overall pattern:*

- Center: the midpoint of the data
- Spread: the range that the data covers
- Shape:

- Modes: peaks in the data. Distributions that have one main peak are called *unimodal*
- Symmetry: is the distribution on one side of the midpoint a mirror image of the other side?
- The distribution is *skewed tot he right* if the right tail is much longer than the left tail, and *skewed to the left* if the opposite is true.

Individual data points that fall outside the overall pattern are called *outliers*. These are identified by using your best judgment and it is important to search for explanations behind these outliers. Remember to look beyond just the extreme data points. In some cases, outliers are useful in pointing out mistakes that were made during the experiment, for example: errors in recording, malfunctions in equipment, or other unusual circumstances.

Time Plots

It is always a good idea to collect data collected over time in chronological order. This is to avoid misunderstandings, as statistical displays that ignore time as a variable (histograms and stem plots) do not clearly show a systematic change over time. A *time plot* shows data plotted against the time they occur. Time is always plotted on the horizontal (x) axis and the variable measured over time is plotted on the vertical (y) axis.

1.3: Describing Distributions with Numbers

While graphs are a good way to get an overview of your data, numerical descriptions are much more specific. It is important to remember that these numbers, like graphs, are tools to help us understand and interpret the data; the numbers, themselves, are not answers.

Measuring the Centre

The numerical description of any dataset begins with a description of the middle. There are two common ways to describe the midpoint of a distribution:

Mean

The *mean* is the average value of all of your data points. To find the mean $\bar{x}$, for a set of observations, you simply sum all of their values and divide by the total number of observations. Thus, for a data set, $x_1 + x_2 + x_3 + \ldots + x_n$, the mean can be found using the following equation:

- $\bar{x} = x_1 + x_2 + x_3 + \ldots + x_n / n$.
- Another way of expressing the mean is: $\bar{x} = 1/n \Sigma x_i$ In this formula, Σ denotes the function "sum", or "add everything together." The bar over the *x* signifies the mean of all the x-values. Said aloud, it is pronounced, "x-bar."

The main disadvantage of the mean is that it is very sensitive to extreme values in the data set - outliers and skewed distributions will undercut the integrity and accuracy of using the mean as the midpoint of your data. Because the mean cannot help but be influenced by these extreme values, it is not a *resistant* or *robust* measure. Robust measures are not easily influenced by a few data points.

The Median

The median, *M,* is the literal mid point of a distribution. Half of the observations in a dataset fall above, and half fall below the median. To find the median:

1. Order all observed values in ascending of descending order
2. If the number of observations is uneven, the median is the observation in the exact centre of the list. The median can be found by counting (n+1)/2 observations up from the bottom of the ordered list.
3. If the number of observations is even, the median is the mean of the two centre observations

Mean vs. Median

If a distribution is completely symmetrical, then the median and mean are the same thing. In a distribution that deviates to the left or the right, the average is located in the tail more than the median. This is because the mean is much more affected by extreme scores. The tails of a distribution consist of extreme scores.

Range (Variability)

The simplest numerical description of a distribution should consist of a measure of the mid-point (such as the average and the median), but also a measurement of the spread of a distribution. We can describe the spread of a distribution by calculating various percentiles. The median splits the distribution exactly in half, and that is why we say that the median is the fiftieth percentile. However, there are also upper and lower quartiles on either side of the median. Each quartile is about a quarter of the data. Quartiles can be calculated as follows.

- First put all scores in ascending order. Then, calculate the median of the data set.
- The first quartile (Q1) is the median of the lower half of the distribution\

- The third quartile (Q3) is the median of the higher half of the distribution

The p^{th} percentile of a distribution is the value by which p percent of the scores is the same or below it

The Five-Number Summary and Boxplots

In order to make a description of the mid-point and the spread of a distribution, it is useful to (1) the lowest score (minimum), (2), Q1, (3) M (the median), (4) Q3, and (5), the highest score (maximum).

These five values are clearly visible in box plots:

- The outer two edges of the box in a box plot stand for Q1 and Q3.
- The median is shown by the line in the middle of the box.
- Two lines (upwards and downwards from the box) show the maximum and minimum values.

Interquartile Range (IQR)

An overview of largest and the smallest value says very little about the variation within the data. The distance between the first and the third quartile is a more robust measure of spread. This distance is referred to as the interquartile range, and is calculated as follow

IQR: Q3 - Q1

The IQR is often used as a rule of thumb to determine outliers. Often, a score is an outlier named as this 1.5 x IQR above the third quartile or 1.5 x IQR falls below the first quartile.

Differing Distributions

Quartiles and the IQR are not affected by changes in the tail of a distribution; they are quite robust. However, no single numerical value of dispersion (such as the IQR) is very useful to describe the spread of skewed distributions (left or right). It is often possible to detect skewedness using the five-number summary. A deviation to the left or right can be seen by looking at how far the first quartile and the lowest score are from the median (left tail) and by looking at how far the third quartile of the highest score is (right tail).

Variance and Standard Deviation

Standard deviation measures the spread of the distribution to be by looking at how far the observations are from the mean.

- The variance (s^2) of a data set is the average of the standard deviations, squared. The formula is: $s^2=(x_1-\bar{x})^2+(x_2-\bar{x})^2+ \ldots+(x_n-\bar{x})^2/ n-1$. Another correct formula is: $s^2=1/n-1\ \Sigma(x_i-\bar{x})^2$. In this context, n-1 stands for the degrees of freedom.
- For the standard deviation (s), find the square root of the variance.

The variance and standard deviation measure the distance between the observations and the mean. Since some observations fall above and some observations below the mean, squaring all the values will make all of the variances (and consequently, standard deviations) positive. Therefore, s^2 and s will be large if observations are widely spread about the mean, and small if the observations are relatively close to the mean.

The standard deviation is particularly useful in normal distributions. The standard deviation is preferred over the variance because finding the square root of the variance ensures that spread is measured according to the original scale of the variable.

Some Important Points About the Standard Deviation:

- Standard deviation s measure of the dispersion from the mean, and should only be used if the mean (and not the median) is chosen as a measure of midpoint.
- The standard deviation is zero when there is no spread is present in a distribution. This only happens if all values are the same. If this is not so, which standard deviation is greater than zero. The more there is spread, the greater will be *s*.
- The standard deviation, like the mean, is not robust. The presence of outliers can make *s* very large. The standard deviation is even more sensitive to extreme scores than the mean.
- Distributions with a strong deviation (left or right) have large standard deviations. In this case, it is not very practical to calculate the standard deviation. The five-number summary is often more suitable than the average and the standard deviation when an abnormal distribution needs to be described or when a distribution has extreme outliers. The use of the mean and the standard deviation is just more convenient when there are few outliers are present and if the distribution is symmetrical.

Changing the Unit of Measurement

The same variable may be often measured in different units of measurement. For example, temperature can be measure in Fahrenheit and Celsius. Fortunately, it is easy to convert measurement units by using linear transformations. Linear transformations do not change the shape of a distribution; for example, if temperature data that is measured in Fahrenheit shows a right skew, converting the values to Celsius will not eliminate that right skew.

A linear transformation changes the original variable x into a new variable (x_{new}) on the basis of the following formula:

- $x_{new} = a+bx$.
- The addition of the constant a will change all the values of x in the same degree. For example, an adjustment will change the zero point of *a* variable. Multiply by the positive constant *b* changes the size of the measuring unit.
- To see the effect of linear transformation on measures of dispersion and size of the center point, it is important to multiply each observation with the positive number *b*. This ensures that the median, mean, standard deviation, and IQR are multiplied by b.
- Adding the same number *a* (*a* can also be negative) to each observation, adds *a* to the mean, median, quartiles and percentiles. Measures of spread, however, are not affected.

1.4: Density Curves and Normal Distributions

Density Curves

Because the manual creation of histograms is time consuming and impractical, scientists often use computer programs to create histograms. The advantage of using computer programs is that they can also make an appropriate curve on the basis of the histogram. These are called *density curves*. Density curves "flow" with the peaks of a histogram.

- A density curve is always made above the horizontal axis.
- The total area within the curve is always equal to 1.
- A density curve describes the general pattern of distribution.

As with distributions, density curves can have different shapes. A special variant is the *normal distribution*, in which both halves of the curve are symmetrical. Outliers are not described by a density curve.

The mode of a distribution describes the peak point of the curve. It therefore comes to the place where the curve is the highest. Because areas under the curve stand for proportions of the observations, the median is the point that is located exactly in the middle.

The quartiles can be estimated by dividing the curve into approximately four equal parts. The IQR is than the distance between the first and the third quartile. There are mathematical ways to calculate the areas under the curve. Through this arithmetic, ways we can precisely calculate the median and quartiles.

The average of a density curve is the point at which the curve would balance if it were made of solid material. In a symmetrical curve, the median and the mean are the same.

In other distributions, that is not the case: For example, in a curve with a deviation to the right, the mean will be to the right of the median. With an abnormal distribution, it is difficult to determine the balance point with the naked eye. There are ways to calculate the arithmetic mean and the standard deviation of a density curve:

- The *median* of a density curve is the point that divides the area under the curve in half
- The *mean* of a density curve is the balance point at which the curve would balance if it would be made of solid material.
- The median and the mean are the same for a symmetric density curve. The average of a different distribution lies more in the direction of the long tail, while the median lies more in the direction of the peak.

Normal Distributions

Normal distributions are an important subset of density curves. They are unimodal, symmetrical, and bell-shaped. The mean and standard deviation determine the shape of a normal distribution:

The mean of a curve indicated with the letter μ. Changing μ (while the standard deviation is unchanged) will ensure that the position of the curve moves on the horizontal axis, while the distribution remains the same.

The standard deviation is represented with the symbol σ. The standard deviation is the measure of dispersion associated with a normal distribution. A curve with a larger standard deviation is wider and lower.

Why are normal distributions important in statistics?

- Normal distributions are good descriptions of real data. Many real-life examples of data re normally distributed, including distributions of height, weight and IQ.
- Normal distributions are good approximations of the outcomes of probability calculations, for example in the case of tossing a coin.
- Normal distributions are useful because many *statistical inference* procedures are based on normal distributions

The 65-95-99.7 Rule

While there are many types of normal distributions, but they have some common characteristics. The most important characteristics are:

- Approximately 68% of the scores fall within one standard deviation (σ) of the mean (μ).
- Approximately 95% of the scores fall within two standard deviations of the mean.
- Approximately 99.7% of the scores fall within three standard deviations of the mean.

This is known as the 68-95-99.7 Rule. The normal distribution with mean μ and standard deviation σ is written as N (μ, σ).

Standardised Values

If, for example, someone has scored sixty points on a test, you do not know whether this is a high or low score in comparison to all the other scores. It is therefore important to standardize the value.

If x is a score from a distribution with mean μ and standard deviation σ, then the standardized value of x is:

- $z = (x-\mu) / \sigma$.

A standardized value is often referred to as a *z-score*. A z-score tells us how many standard deviations away from the mean a particular observation is, and in which direction. The standardized values of a distribution have a mean of 0 and a standard deviation of 1. Together, the standardized normal distribution has the N (0,1) distribution.

Cumulative Proportions

The calculation of the proportions in a precise manner within the normal distribution can be done by means of z-tables or software.

- Z-tables and software often calculate a cumulative proportion: this is the proportion of observations in a distribution that is exactly equal to, or is there below a certain value.

The Z-table can be used to determine proportions under the curve. To do this it must first be standardized scores. Suppose you wanted to know how many students had a score above or below 820 on a particular test. Assuming you have a mean score of 1026 and a standard deviation of 209:

- The corresponding z-score would be: 820-1026 / 209 = -0.99.
- Using the z-table, look up the proportion that belongs to -0.99. You will find the p-value to be 0.1611. This area refers to the area to the left of -0.99. The area to the right of -0.99 is therefore 1-0.1611 = 0.8389.
- This means that 16% of the test-takers scored below 820 and below, while 84% of the test-takers scored above 820.

Normal Quantile Plot

Stem-and-leaf plots and histograms are often used to see if a distribution is normally distributed. However, the normal quantile plot is the best graphical way to discover normality. It is uncommon to make a normal quantile plot by hand, however in order to understand how software would make one, we would follow these steps:

- First, all scores must be put in ascending order. The percentile that each value occupies is then recorded
- Then, z-values associated with these values must then be found. These are also referred to as z-normal scores.
- Finally, each data point is to be graphically connected with the corresponding normal score. If the distribution is (almost) normally distributed, then the data points will lie on an approximately straight line. Systematic deviations from the straight line indicate a non-normal distribution. Outliers are data points that are far from the general pattern of the plot.