Shape and spread of data

Statistics shape center and spread of data. What is spread of data. Identifying the center spread and shape of a data set. What graph displays the center the spread and the shape of the data.

Â Â Â AboutStatisticsNumeroTeoriaJava Data StructuresPrecalculusCalculusA population parameter is a feature or measure obtained using all data values of a population. A sample statistics is a feature or measure obtained using the data values of a sample. The parameters and statistics we are dealing with in the first place try to quantify the "centre" (i.e. location) and the "dissemination" (i.e. variability) of a set of data. Note that there are different center measurements and different diffusion measures that can be used -- you have to be careful to use appropriate measures given the form of data distribution, the presence of extreme values, and the nature and level of data involved. When we consider different central and diffusion measures, remember that we really want to know the center and diffusion of the population in question (i.e., a parameter) -- but normally we only have sample data. As such, we calculate sample statistics to estimate these population parameters. The form of a distribution We can characterize the form of a set of data by observing its istogram. First, if data values seem to accumulate in a single "mound", we say that distribution is unimodal. If there are two "mountains", we say that the distribution is bimodal. If there are more than two "mountains", we say that the distribution is multimodal. Secondly, we focus on whether the distribution is symmetric, or whether it has a longer 'tail' on either side. In case there is a longer "tail" we say that the distribution is tilted in the direction of the longest queue. In case the longest queue is associated with larger data values, we say that the distribution is distorted right or (possibly distorted). In case the longer tail is associated with smaller (or more negative) values, we say that the distribution is oblique on the left or (negatively oblique). If the distribution is symmetrical, it will often be necessary to check whether it is approximately bell, or if it has a different shape. In the case of a distribution in which each rectangle has approximately the same height, we say that we have a uniform distribution. The chart below provides some examples of the above distribution forms. Center Measures For range or ratio level data, a center measure is the average. The population average is μ, while the sample average is $\overline{x}$. Both values are calculated very similarly. Assuming that the population has size N, a sample has size n, and x covers all available values in the population or sample, depending on the case, we find these means by calculating $$\mu = \frac{\sum x}{N}\quad \textrm{ and } \quad \overline{x} = \frac{\sumThe median, indicated by Q_2 (or med) is the average value of a data set when written in order. In the case of an even number of data values (and therefore no exact mean), it is the average of the two mean data values. It is not affected by Presence of extreme values in the data set. Unlike the average, sometimes it can also suggest a central value for the ordinal data. Â «: You can list the ordinal data Â« in orderÂ »and find the central value when there is an odd number of total values. However, when there is a number number of values, there is a complication: we cannot mediate two ordinary values as we can do with the report values or interval to find an «intermediate value.â €» As an example , suppose the data of one involve ranks of poker cards: $ A, 7,7,10, J, Q, Q, K, K, K $. The two average ranks are a fante (J) and a queen (Q). What would their average be? Given the difficulty of answering this question, some texts suggest that for a list of equal length data, you should instead simply choose the lower of the two average values such as median. The mode is the most frequent data value in the population or in the sample. There may be more mode, even if in the event that there are no repeated data values, let's say that there is no mode. The ways can also be used for nominal data. Midrange is only the average of higher and lower data values. Although easily understandable, it is strongly influenced by extreme values in the data set, and it does not reliably find the center of a distribution. Dissemination measures In addition to knowing where the center is located for a Date Distribution, we often want to know how it is "diffused" distribution - this gives us a measure of the variability of the values taken from that distribution. The graph below shows the general form of three symmetrical aimodent distributions with identical center measures, but very different quantities of â â â â â â â â â € "as there were multiple center measures, there are multiple spread measures - each with some advantages in Certain situations and disadvantages in others: the interval is technically the difference between the maximum and minimum distribution values, even if it is often reported simply by listing the minimum values and maximum visas. It strongly influenced by the extreme values in the distribution. Another measure of the spread is given by the absolute media deviation, which is the average distance from the average. Remember that the distance between two $ X $ and $ Y $ values is given by the absolute value of their difference $ | x â € "Y | $, then the distance between a $ X $ value and the average population $ MU $ would be $ | x â € œ MU | $. To find the average of this distance, add the population and divide for the number of things in the population, $ N $: $$ Mad = Frac {SUM | X â € "MU |} {n} $$ Although Simple to be expressed, the absolute media deviation creates some problems for us along the line (not horring different from how the introduction of an absolute value in a function - like those who have studied the SA calculation - can cause problems for how much concerns differentiability). Furthermore, the statistics of the sample is a distorted estimate of the mean absolute deviation of the population. This means that its average value does not agree with the Mad. When the average is the most appropriate measure of the center, then the most appropriate measure of the spread is the standard deviation. This measure is obtained by taking the square root of the variance - which is essentially the average square distance between the values of the population (or the sample values) and the average. Using the square distance between these values and the average are the difficulties introduced by the absolute value in the absolute media deviation, even if it exaggerates the contribution to the dissemination of the population given by values far from the average. In the whole, however, for our purposes, the advantages deriving from the use of standard variance and deviation to measure variability and dispersion on the absolute media deviation far exceed the disadvantages. Keeping our minds, the variance of the population, $ SIGMA ^ $ 2, and the standard deviation of the population, $ SIGMA $, are given by $$ SIGMA ^ 2 = FRAC {SUM (X- MU) ^ 2} {n} quads {e} sigma = sqrt {frac {sm (x- mu) ^ 2} {n}} $$ when you are dealing with a sample , A slight modification must be made to the denominators of these formulas so that © $ 2 $ is an impartial estimate of the corresponding population parameter $ SIGMA ^ $ 2 (see Bessel correction), as shown below. $$ s ^ 2 = frac {sm (x- overline {x}) ^ 2} {n-1} quads {e} quad s = sqrt {frac {sm (x- overline {x}) ^ 2} {n-1} $$ When median is the most appropriate measure of the center, then the interquartile interval (or iQR) is the most appropriate measure of diffusion. When the data is ordered, the IQR is simply the range of central data half. If the data has quartiles $ Q_1, Q_2, Q_3, Q_4 $ (noting that $ Q_2 $ is the median and $ Q_4 $ is the maximum value), then $$ IQR = Q_3 â € "Q_1 $$ unlike € ™ interval itself, the IQR is not easily influenced by the presence of extreme data values. Determination of a significant note sdewness, the presence of sdewness (or abnormal values) can affect the point where the center's measurements are positioned with respect to the other, as the underlying chart suggests. As you can see, when there is a significant deviation, the media and the median end up in different places. Turning this point, if the media and the median are quite distant, we can determine if an observed deviation is significant. To this end, the Pearson Skewness Index, I, is defined as $$ i = frac {3 (overline {x} â € q_2)} {s} $$ as regards the fact that The media and the median are sufficiently distant (relative to the dissemination of distribution), let's say that if $ | I | Ge $ 1, then the data set is significantly distorted. Identify outliers An outlier is a data value significantly distant from the main body of a data set. Remember that in calculating the IQR we measure the middle range of a data set, from $ Q_1 $ to $ Q_3 $. AND' That if a data value goes away too much from this interval, we should call it an outlier. Of course, we expect the values are farther from the center (here, $ Q_2 $) when the spread (here, the IQR) is great, and closer to the center when the IL It's small. With this in mind, let's say that any value outside the following range is an outlier. $$[Q_1 – 1.5 \times IQR, Q_3 + 1.5 \times IQR]$$ You might wonder where the 1.5 in the above range comes from -- Paul Velleman, a statistician at Cornell University, was a student of John Tukey, who invented this test for outliers. He was wondering the same thing. When he asked Tukey, "Why 1.5?" Tukey said, "Because 1 is too small and 2 is too big." Cost to Treat Tuberculosis in India Step 1: Design the study. Tuberculosis (TB) is the deadliest bacterial disease in the world. In 2009, nine million new cases of tuberculosis were diagnosed, leading to almost 2 million deaths worldwide. Currently, the main vaccine used to prevent tuberculosis is Bacille Calmette Guerin (BCG). Unfortunately, BCG is only moderately effective in preventing tuberculosis. Historically, India has had a high number of cases of tuberculosis. The Indian government wants to reduce the prevalence of this disease. In this activity, we compare the average costs of treating a person who gets TB with the costs of preventing a case of TB in India. The health care records of TB patients in India were reviewed to estimate the cost of treating TB patients. The following figures are representative of the total costs (in US dollars) incurred by the company in treating 10 randomly selected TB patients in India. These costs include medical treatment, time lost from work, and in some cases lost utility due to death. Step 3: Describe the data. The following figures are representative of the total costs (in US dollars) incurred by the company in treating 10 randomly selected TB patients in India. 15,100 19,000 4,800 6,500 14,900 600 23,500 11,500 12,900 32,200 To help us visualize this data, we will create a graph called a histogram. To make a histogram, we're going to divide the line number from 0 to 35,000 into seven equal parts. We will then count the number of data points in each of these ranges: At least 0 and less than 5,000 2 At least 5,000 and less than 10,000 1 At least 10,000 and less than 15,000 3 At least 15,000 and less than 20,000 2 At least 20,000 and less than 25,000 1 At least 25,000 and less than 30,000 0 At least 30,000 and less than 35,000 1 For each of these ranges, we draw a bar on the histogram. The width of the bars is determined by the width of the range (5000 in this example). The height of the bars is equal to the number of observations falling in each range. As we look at the histogram shown below, we see bars ranging from $0 to $35,000. We also see higher bars in the middle between $10,000 to $20,000 show that these values are more commonly occurring than the other values. If we calculated the average of the values In our histogram, we would compute the number \[\frac{15,100 + 19,000 + 4,800 + 6,500 + 14,900 + 600 + 23,500 23,500 11.500 + 12.900 + 32.200} {10} = 14.100] Shows that the center of the histogram (or media) is at $ 14,100. This is a histogram created in Excel: you can watch this short video of how to create a histogram in Excel, or follow these steps: Start by typing the data in a cell column in Excel: each data point in your cell, as Showing below. Then highlight the data. Go to insert the ribbon in Excel and select the histogram icon from the "Change" section of the tape. Then select the first option of PRESENTED ISTOGRAMS. You will notice that the histogram does not look a lot to the histogram over again. The number and size of the containers in a histogram can dramatically change the apparent form of the distribution. It is worth experiencing with a different number of containers so that the true form of data distribution is revealed. To change the number of containers go ahead to the next steps. Make sure the graph is selected, so you can select the â € œformatâ € tape. In the upper left corner of the ribbon, select â € œHorizontal category Axisâ € in the drop-down box. Then click "Selecting the format" An option menu will open on the right side of the screen. You can adjust the width of the trash or the number of containers. In this case, even if we cannot make it match, in order to make the histogram look as far as possible like the one shown above, we will choose a bin width of 5000. Excel shows the initial and final values for each basket. For example, the first basket ranges from 600 to 5,600. Let's see that there are two data points contained in that basket, represented by the height of the trash bar. The next bin contains only one of our data points, includes any value greater than 5,600 up through, and understood, 10,600. And so on. This histogram doesn't look exactly like the histogram imagined above, but it's close. To make it exactly you can use the options â € œVoverflow binâ € and â € œunderflow binâ €; But this is beyond what we will discuss here. After summarizing the data of our sample of the populations both numerically and graphically, we can use this information to make inference on the total population. Step 4: Make the inferences. In the past, the total average cost for the company for the treatment of a case of tuberculosis in India was known to be $ 13.800. As shown in our step 3 calculations, the 10 selected patients randomly showed an average cost higher than the historical value at $ 14.100. This could make us believe that the average actual average cost for the company is also $ 14,100. However, in statistical calculations of depth (which will be taught you how to do this semester later) show that there is a probability of 46% that our champion had an average of $ 14,100 only by random chance. This is not too difficult to since we only had a sample size of 10 people, and $ 14,100 is only $ 300 above $ 13,800, so it turns out to be quite probable (46% possibilities) that because of random possibility our sample had an average that was a bit higher of real value from the population. # # Conclude that the average total cost to the company is still essentially the same as it was in the past. Step 5: Take action. After making inferences, you act. The motivation for conducting a study like this is usually to see if there is inflation in the costs. Answer the following question: given our conclusion in Step 4 (that the results of our random sample averaging $14,100 had a 46% chance of being caused by a random probability) Do you think the government of India needs to take any special action to stop the rising cost of treating tuberculosis? Show / hide Solution Responses may vary. â¨"However, we could not say that the real average cost has really changed from $13,800. So, there's not enough evidence of inflation. There is no need for the government of India to act. One advantage of using a histogram is that it allows you to view the distribution of the data. A histogram illustrates the overall shape of the data distribution. The height of the bars shows how many observations fall into that range. Answer the following question: which bin of the TB cost histogram contained the most data points? Show / Hide Solution The basket ranged from $10,000 to $15,000 contained 3 observations ($11,500, $12,900 and $14,900), which was most of any of the containers in the histogram. This can be seen visually in the histogram by looking at the height of each bar and the starting and stopping points of the bar along the x-axis of the graph. We will describe the shape of the distribution of a data set using the following basic categories: symmetrical, bell-shaped, right-hand distorted and left-hand distorted. In addition, we can label the shape of a distribution as uniform, nonimodal, bimodal or multimodal. A distribution is symmetrical if both the left side and the right side of the distribution appear to be roughly a mirror image of each other. A special symmetrical distribution is a bell distribution. When the data follows a bell distribution, the histogram looks like a bell. Bell distributions play an important role in statistics and will play a role in most future lessons. A distribution is distorted if a histogram of the distribution shows a long tail to the right. This can happen if there are some very large anomalous values on the right side of the distribution. A distribution is left to the left if a histogram shows it has a long tail to the left. If a distribution has only one peak, it is said to be unimodal. The three distributions shown above are all non-simulated distributions. Some people might argue that there are several peaks in the GPA data, so it should not be considered uncontested. Even if there are jagged bumps in the histogram, it is important the overall shape in the data. When interpreting a histogram, it may be helpful to blur the eyes and imagine the overall shape after smoothing the bumps. If the general trend indicates that there is more than one shot, then we don't do it we do the distribution to be unimodal. Usually we will only work with unimodal datasets in this course. Some distributions do not have a distinct peak, others have more than one peak. When there is no distinct peak, and the histogram shows a relatively flat shape, we could say that the data follows an even distribution. If there are two distinct peaks, a distribution is called bimodal. If there are more than two peaks, we refer to the distribution as multimodal. multimodal.