Measures of Position

Lecture objectives

- Identify the position of a data value in a data set, using various measures of position, such as percentiles, deciles, and quartiles.

In addition to measures of central tendency and measures of variation, there are measures of position or location. These measures include standard scores, percentiles, deciles, and quartiles. They are used to locate the relative position of a data value in the data set.

2 Standard Scores

A standard score or z score tells how many standard deviations a data value is above or below the mean for a specific distribution of values. If a standard score is zero, then the data value is the same as the mean.

A z score or standard score for a value is obtained by subtracting the mean from the value and dividing the result by the standard deviation.
$z = \frac{vallue - mean}{standard\, deviation}$

for the sample	for the Population
$z = \frac{x-\bar{x}}{s}$	$z = \frac{X-\mu}{\sigma}$

The z score represents the number of standard deviations that a data value falls above or below the mean.

Example 1. A student scored 65 on a calculus test that had a mean of 50 and a standard deviation of 10; she scored 30 on a history test with a mean of 25 and a standard deviation of 5. Compare her relative positions on the two tests.

Solution

For calculus the z score is
$z = \frac{x-\bar{x}}{s} = \frac{65-50}{10}$ =1.5
For history the z score is
$z = \frac{x-\bar{x}}{s} = \frac{30-25}{5}$ =1

Since the z score for calculus is larger, her relative position in the calculus class is higher than her relative position in the history class.

Note

If the z score is positive, the score is above the mean. If the z score is 0, the score is the same as the mean. And if the z score is negative, the score is below the mean.

Example 2. Find the z score for each test, and state which is higher.
Test A X=38, $\bar{x}$=40, s=5
Test B X=94, $\bar{x}$=100, s=10

Solution

For test A the z score is
$z = \frac{x-\bar{x}}{s} = \frac{38-40}{5}$ $=-0.4$
For test B the z score is
$z = \frac{x-\bar{x}}{s} = \frac{94-100}{10}$ $=-0.6$

Comment on the tests' scores?

3 Percentiles

Percentiles divide the data set into 100 equal groups.
They are symbolized by
$P_1, P_2, P_3, ..., P_{99}$
The percentile corresponding to a given value X is computed by using the following formula
$Percentile = \frac{(number\ of\ values\ below X)+0.5}{Total\ number\ of\ values} * 100$

Example 3. A teacher gives a 20-point test to 10 students. The scores are shown here. Find the percentile rank of a score of 12 in 18, 15, 12, 6, 8, 2, 3, 5, 20, 10

Solution

Arrange the data in order from lowest to highest.
2, 3, 5, 6, 8, 10, 12, 15, 18, 20
Then substitute into the formula.
$Percentile = \frac{(number\ of\ values\ below X)+0.5}{Total\ number\ of\ values} *100 = \frac{6+0.5}{10} *100 = 65th\ percentile$
Thus, a student whose score was 12 did better than 65% of the class.

Example 4. using the previous example data find the value corresponding to the 25th and 60th percentile.

Solution

to find the value of the 25 the percentile

First	Arrange the data in order from lowest to highest.
Second	Compute. $c = \frac{n*P}{100}$ Where C is the order of the percentile n is the total number of values P is the percentile
Third	$c = \frac{10*25}{100} = 2.5.$

If c is not a whole number, round it up to the next whole number; in this case, c=3.Start at the lowest value and count over to the third value, which is 5. Hence, the value 5 corresponds to the 25th percentile.

for the 60th percentile use the same rule
$c = \frac{10*60}{100} = 6$.
If c is a whole number, use the value halfway between the c and c+1 values when counting up from the lowest value—in this case, the 6th and 7th values are 10 and 12 respectively .
find the half way value by adding the two values and dividing by 2.
$\frac{10+12}{2} = 11$
11 corresponds to the 60th percentile. Anyone scoring 11 would have done better than 60% of the class.

4 Quartiles and Deciles

Quartiles divide the distribution into four groups, separated by Q_1, Q_2, Q_3.
Q_1 is the same as the 25th percentile; Q_2 is the same as the 50th percentile, or the median; Q_3 corresponds to the 75th percentile.
finding the quartile using these steps.

First	Arrange the data in order from lowest to highest.
Second	Find the median of the data values. This is the value for Q_2.
Third	Find the median of the data values that fall below Q_2. This is the value for Q_1.
Fourth	Find the median of the data values that fall above Q_2. This is the value for Q_3.

In addition to dividing the data set into four groups, quartiles can be used as a rough measurement of variability.

4.1 Interquartile Range

The **interquartile range (IQR)** is defined as the difference between Q_1 and Q_3 and is the range of the middle 50% of the data.
The interquartile range is used to identify outliers, and it is also used as a **measure of variability** in exploratory data analysis.

Example 5. Find Q_1, Q_2, Q_3 and interquartile range for the data set 15, 13, 6, 5, 12, 50, 22, 18.

Solution

Arrange the data in order.
5, 6, 12, 13, 15, 18, 22, 50
Find the median (Q_2).
5, 6, 12, 13, 15, 18, 22, 50
$MD = \frac{13+15}{2} = 14$
Find the median of the data values less than Q_2=14.
5, 6, 12, 13
$Q_1 = \frac{6+12}{2} = 9$
Find the median of the data values greater than Q_2=14.
15, 18, 22, 50
$Q_3 = \frac{18+22}{2} = 20$
$Q_1 = 9, Q_2 = 14, and\, Q_3 = 20$.
interquartile range= $Q_3 - Q_1 = 20 - 9 = 11$

Deciles divide the distribution into 10 groups. They are denoted by $D_1, D_2, etc.$ Deciles can be found by using the formulas given for percentiles.

5 Outliers

An outlier is an extremely high or an extremely low data value when compared with the rest of the data values.
finding the outlier using these steps.

First	Arrange the data in order and find Q1 and Q3.
Second	Find the interquartile range: $IQR = Q_3 - Q_1$.
Third	Multiply the IQR by 1.5
Fourth	Subtract the value obtained in previous step from Q_1 and add the value to Q_3.
Fifth	Check the data set for any data value that is smaller than $Q_1 - 1.5 * (IQR)$ or larger than $Q_3 + 1.5 * (IQR)$.

Example 6. Check the following data set for outliers. 5, 6, 12, 13, 15, 18, 22, 39

Solution

find Q1 and Q3.
$Q_1 = 9\, Q_3 = 20$
Find the interquartile range (IQR)
$IQR = Q_3 - Q_1 = 20 - 9 = 11$
Multiply the IQR by 1.5
$IQR * 1.5 = 11 * 1.5 = 16.5$
Subtract from Q_1, Q_1-16.5=9-16.5=-7.5 and add the value to Q_3, $Q_3 + 16.5 = 20 + 16.5 = 36.5$
Check the data set for any data values that fall outside the interval from -7.5 to 36.5. The value 39 is outside this interval ,so it can be considered an outlier.

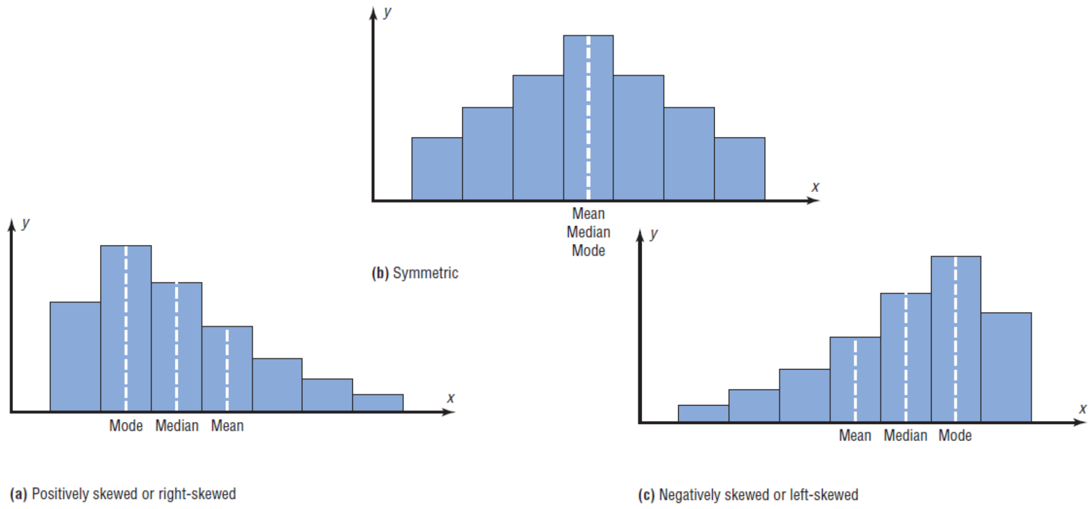
6 Distribution Shapes

Frequency distributions can assume many shapes. The three most important shapes are

1. positively skewed.
2. symmetric
3. negatively skewed

In *a positively skewed or right-skewed distribution*, the majority of the data values fall to the left of the mean and cluster at the lower end of the distribution; the "tail" is to the right. Also, the mean is to the right of the median, and the mode is to the left of the median.

In *a symmetric distribution*, the data values are evenly distributed on both sides of the mean in addition, when the distribution is unimodal, the mean, median, and mode are the same and are at the center of the distribution and we call the distribution is normal distribution.



In *a negatively skewed or left-skewed* the majority of the data values fall to the right of the mean and cluster at the upper end of the distribution, with the tail to the left.Also, the mean is to the left of the median, and the mode is to the right of the median

6.1 Measures of Skewness

In Fig above you noticed that the mean, median and mode are not equal in a skewed distribution. The **Karl Pearson's measure of skewness** is based upon the divergence of mean from mode and median in a skewed distribution.

$$sk = \frac{3(Mean - Median)}{S}$$

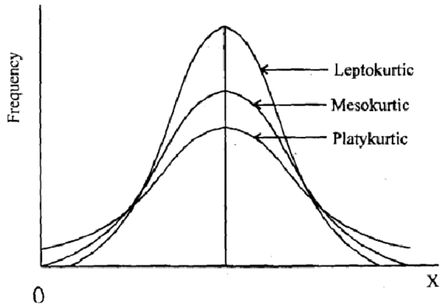
or

$$sk = \frac{Mean - Mode}{S}$$

If Karl Pearson's measure of skewness is positive, the data are positively skewed or skewed right, meaning that the right tail of the distribution is longer than the left. If Karl Pearson's measure of skewness is negative, the data are negatively skewed or skewed left, meaning that the left tail is longer. And if skewness $= 0$, the data are perfectly symmetrical and normal.

6.2 Measure of Kurtosis

Kurtosis is another measure of the shape of a distribution. Whereas skewness measures the lack of symmetric of the frequency curve of a distribution, kurtosis is a measure of the relative peakedness of its frequency curve. Various frequency curves ca nbe divided into three categories depending upon the shape of their



peak.
A normal distribution has kurtosis exactly 3 is called *mesokurtic*. A distribution with kurtosis less than 3 is called *platykurtic*. Compared to a normal distribution, its central peak is lower and broader, and its tails are shorter and thinner. And a distribution with kurtosis greater than 3 is called *leptokurtic*. Compared to a normal distribution, its central peak is higher and sharper, and its tails are longer and fatter.
Kurtosis measure by:

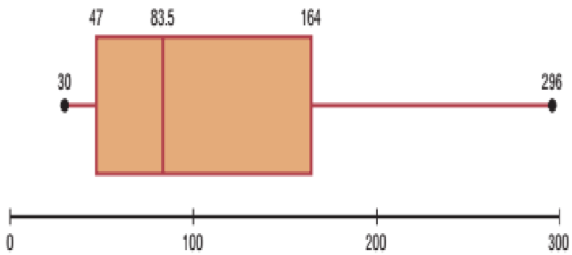
$$ku = \frac{Q_3 - Q_1}{2(P_{0.90} - P_{0.10})}$$

7 Boxplot

A boxplot is a graph of a data set obtained by drawing a horizontal line from the minimum data value to Q_1, drawing a horizontal line from Q_3 to the maximum data value, and drawing a box whose vertical sides pass through Q_1 and Q_3 with a vertical line inside the box passing through the median or Q_2.
These plots involve five specific values:

1. The lowest value of the data set (i.e., minimum)
2. Q_1
3. The median
4. Q_3
5. $The highest value of the data set (i.e., maximum)$

Example 7. The number of meteorites found in 10 states of the United States is 89, 47, 164, 296, 30, 215, 138, 78, 48, 39. Construct a boxplot for the data.



solution

First	Find the five-number summary for the data values. Arrange the data in order: 30, 39, 47, 48, 78, 89, 138, 164, 215, 296 Find the median $\frac{78+89}{2} = 83.5$ Find $Q_1 = 47$ Find $Q_3 = 164$
Second	Draw a horizontal axis with a scale such that it includes the maximum and minimum data values.
Third	Draw a box whose vertical sides go through Q_1 and Q_3, and draw a vertical line though the median.
Fourth	Draw a line from the minimum data value to the left side of the box and a line from the maximum data value to the right side of the box.

The distribution is somewhat positively skewed.

There are many of information obtained from a boxplot, if the median is near the center of the box and the lines are about the same length, the distribution is approximately symmetric, and if the median falls to the left or right of the center of the box, the distribution is positively or negatively skewed. And if the lines are about the same length, the distribution is approximately symmetric.