3.3 Measure of Position

Objectives

- Determine an interpret z-scores
- Determine an interpret percentile
- Determine an interpret quartile
- Check a set of data for outliers

Z-Score (or standard score)

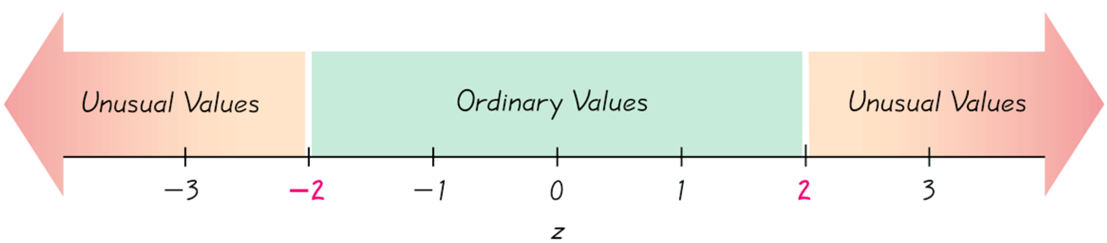
The number of standard deviations that a given value x is above or below the mean. The following formula allows a raw score (x) from a data set to be converted to its equivalent standard value (z) in a new data set with a mean of zero and a standard deviation of one.

$$\boxed{z = \frac{data - mean}{standard\ deviation}}$$

For Population: $= \frac{x-\mu}{\sigma}$, while for Sample: $z = \frac{x-\bar{x}}{s}$

A z-score can be positive or negative:

- positive z-score – raw score (data item) greater than the mean
- negative z-score – raw score less than the mean



Ordinary values: z-score between –2 and 2 standard deviation.

Unusual Values: z-score < -2 standard deviation or z-score > 2 standard deviation.

Example (3.23):

Who is relatively taller, a 67-inch man or a 62 inch women? The average height for men is 69.9 inches with a standard deviation of 3 inches and the average height for women is 64.6 inches with a standard deviation of 2.8 inches. Find:

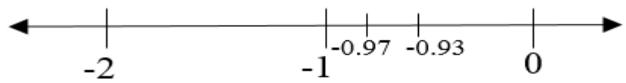
a. z-score for the 67 inch men?
b. z-score for the 62 inch women?
c. Who is taller?

Solution:

a. $z_{men} = \frac{x-\mu}{\sigma}$ >>>> $z = \frac{67-69.9}{3}$ >>>> z= -0.9666 >>>>> z= -0.97

b. $z_{women} = \frac{x-\mu}{\sigma}$ >>>> $z = \frac{62-64.6}{2.8}$ >>>> z= -0.928 >>>>> z= -0.93

67 inch men has a z-score -0.97, and 62 inch women has a z-score of -0.93



c. 62 inch tall women is relatively taller for her gender.

Example (3.24)

If the annual rainfalls in a certain city are 1, 22, 26, 33, and 123. cm over a 5-year period. Determine the z-score for each raw score, is there any UN-USUAL Data?

Measures of Position

There are several ways of measuring the relative position of a specific member of a set.

1. **Percentiles:** Divides an ordered list into **100 parts**. Just as there are quartiles separating data into four parts, there are 99 percentiles denoted P_1=1%, P_2=2% . . . P_{99}=99%.

A person with percentile rank of 20, means that he /she scored the same as or better than 20 percent of the group. If the data item is at P_{65}, the 65th percentile, then 65% of the data items are below this item and (100-65=35) 35% of the data items are above.

Don't count (x)

$$\text{Percentile of data item (x)} = \frac{Count\ of\ items\ below\ (x)}{\text{Total number of items}} * 100$$

Example (3.25):

Find the percentile of (x=8) from the **ordered list** of data: x= 5, 5, 7, 8, 10, 11, 11, 13, 17 and 19

Solution:

$\text{Percentile of (8)} = \frac{3\ items\ before\ 8}{\text{10 item as total \#}} * 100$ >>>> $\text{Percentile of (8)} = 30\%$

$\text{Percentile of (11)} = \frac{5}{10} * 100 =$

$\text{Percentile of (17)} = \frac{8}{10} * 100 =$

Example (3.26): (Homework)

The following data is the minimum monthly flow in (m^3/s) for 10 years of a certain river.
36, 4, 21, 21, 23, 11, 10, 10, 12 and 17. Answer the following questions:

- What percentage of the data is less than 11?
- What percentage of the data is less than 23?

Location of percentiles:

$$\boxed{\text{Location} = \frac{\text{Percent} * \ Count\ of\ items}{100}}$$

This is only gives you the location of the percentile for an ordered list, not the actual data value.
Case 1. When the location is not a whole number, **BUMP up** (not round) to the next whole number.
Case 2. When the location is whole number you must find the **average** of the item in that location and **next** location.

Example (3.27):

A teacher gives a 20-point test to 10 students. Find the value corresponding to the 25th and 60th percentile.
18, 15, 12, 6, 8, 2, 3, 5, 20, 10

Solution:
Sort in ascending order as: 2, 3, 5, 6, 8, 10, 12, 15, 18, and 20

a) For 25th percentile

$$\text{Location} = \frac{25 * 10}{100} = 2.5$$

The value 5 corresponds to the 25th percentile.
(a student who had 5, did better than 25th percent of all student).

In part (a) the location value was not a whole number (2.5), Hence, a decimal bump up (here is round up) to 3 and take the 3rd item.
And the corresponding value to the 25th percentile is the 3rd from the lowest which is 5.

b) For 60th percentile

$\text{Location} = \frac{60*10}{100} = 6$ ….. Taking average of 6th and 7th value.

2, 3, 5, 6, 8, 10, 12, 15, 18, 20

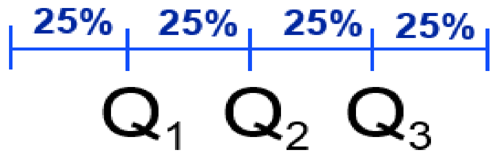
7th value

6th value

In Part (b) the L value is a whole number (6), use the value halfway between *L* and *L*+1.

Average = (10+12)/2 = 11
Hence, a score of 11 correspond to the 60th Percentile.

2. <u>Quartiles:</u> Quartiles split a set of ordered data into **four equal parts** by Q_1, Q_2, and Q_3 as shown:



- Q_1 is the First Quartile defined as: (**25% of the observations are smaller than Q_1** and 75% of the observations are larger).
- Q_2 is the Second Quartile defined as: (**50% of the observations are smaller than Q_2** and 50% of the observations are larger. Same as the Median. It is also the 50th percentile.
- Q_3 is the Third Quartile defined as: (**75% of the observations are smaller than Q_3** and 25% of the observations are larger).

The lower quartile is the median of the lower half of the data and the upper quartile is the median of the upper half. The median divides the data in the data in half. The upper and lower quartiles divide each half into two parts.

Example (3.28):

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

Solution:

Sort in ascending order as: 5, 6, 12, 13, 15, 18, 22, and 50

$Q_1 = \frac{25*8}{100} = 2$ taking 2nd & 3rd value in the data set above: like 6 & 12 >>>>> $Q_1 = \frac{6*12}{2} = 9$

$Q_2 = \frac{50*8}{100} = 4$ taking 4th & 5th value in the data set above: like 13 & 15 >>>>> $Q_2 = \frac{13*15}{2} = 14$

$Q_3 = \frac{75*8}{100} = 6$ taking 6th & 7th value in the data set above: like 18 & 22 >>>>> $Q_3 = \frac{18*22}{2} = 20$

Example (3.29):

Find Q_1, Q_2, and Q_3 for the following data set. 77, 78, 86, 88, 93 and 95.

Solution:

$Q_1 = P_{25} = 25^{th}$ percentile.

$Q_1 = \frac{25*6}{100} = 1.5$ this is the location and the decimal is bump up to next entry which will 2 >>>> Q_1=78.

$Q_2 = \frac{50*6}{100} =$

$Q_3 = \frac{75*6}{100} =$

Exploratory Data Analysis

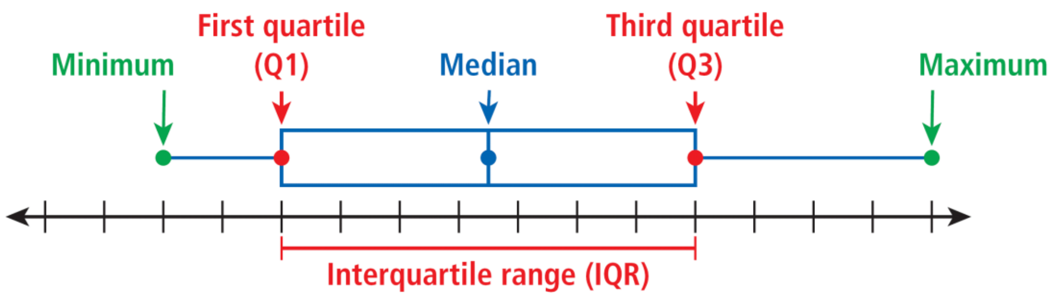
Exploratory Data Analysis is the process of using statistical tools (such as graphs, measures of centre, and measures of variation) to investigate data sets in order to understand their important characteristics.

Five Number Summary:

Made up of minimum, Q_1, Q_2, Q_3 and maximum.
Minimum is Q_0 and maximum is Q_4

Boxplots (Whisker Diagram)

A box-and-whisker plot shows the spread of a data set. It displays 5 key points: the minimum and maximum values, the median, and the first and third quartiles.



A box-plot is a graph of the five-number summary.

- A central box spans the quartiles.
- A line in the box marks the median.
- Lines extend from the box out to the smallest and largest observations.
- Useful for side-by-side comparison of several distributions.

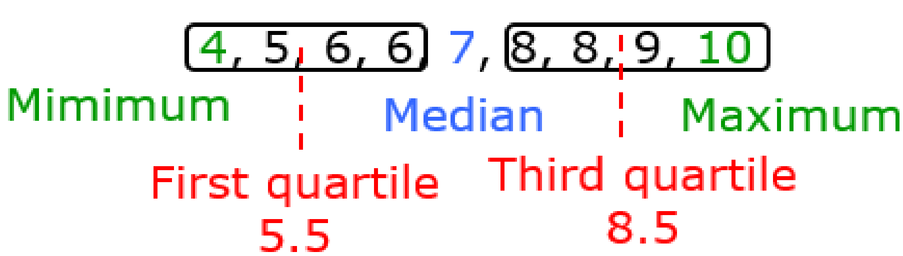
Example (3.30):

Make a box-and-whisker plot of the data below. Find the interquartile range.
{6, 8, 7, 5, 10, 6, 9, 8, 4}

Solution:

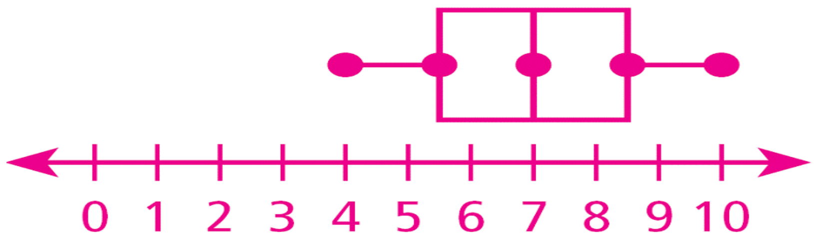
Step 1. Order the data from least to greatest.
4, 5, 6, 6, 7, 8, 8, 9, 10

Step 2. Find the minimum, maximum, median, and quart



Step 3. Draw a box-and-whisker plot.

Draw a number line, and plot a point above each of the five values. Then draw a box from the first quartile to the third quartile with a line segment through the median. Draw whiskers from the box to the minimum and maximum.



IRQ = Q_3-Q_1 = 8.5 – 5.5 = 3

The interquartile range is 3, the length of the box in the diagram.

Outlier:

An outlier is a value that is located very far away from almost all the other values.

Important Principles

- An outlier can have a dramatic effect on the mean
- An outlier have a dramatic effect on the standard deviation
- An outlier can have a dramatic effect on the scale of the histogram so that the true nature of the distribution is totally obscured

Method of detecting an outlier

To determine whether a data value can be considered as an outlier:

- Step 1: Compute Q_1 and Q_3.
- Step 2: Find the IQR = $Q_3 - Q_1$.
- Step 3: Compute (1.5) (IQR).
- Step 4: Compute Q_1 – (1.5) (IQR) and Q_3 + (1.5) (IQR).
- Step 5: Compare the data value (say X) with Q_1 – (1.5) (IQR) and Q_3 + (1.5) (IQR).
- If $X < Q_1$ – (1.5) (IQR) or if $X > Q_3$ + (1.5) (IQR), then X is considered an outlier.

Example (3.31): homework

Given the data set 5, 6, 12, 13, 15, 18, 22, 50, can the value of 50 be considered as an outlier?