Section H

<u>Measures of Position</u>

Another way to summarize a data set it to determine where data values lie within the data set. These measures of position are z-scores, percentiles, and quartiles.

<u>z-scores</u> indicate the location of a value with respect to the mean of a data set. The z-score of a value expresses how many standard deviations above or below the value is from the mean. In addition, z-scores provide a way to compare data sets which have different means and standard deviations.

A **z-score** is calculated using the following formulas depending on whether you are using a population or a sample:

Population | sample

$$z = \frac{x-\mu}{\sigma} \quad \text{or} \quad z = \frac{x-\bar{x}}{s}$$

If the z-score is positive (+) the value is above the mean.

If the z- score is negative (−) the value is below the mean.

The mean for z-scores is equal to zero and the standard deviation equal to 1.

The mean for a distribution would have a z-score of 0.

Solving the above equations for *x*: $x = \mu + z\sigma$ or $x = \bar{x} + zs$

Examples:

1) Suppose you have a data set in which the population mean = μ = 45 and the population standard deviation = σ = 4

   a) Find the *z*-score for 52.

   b) Find the *z*-score for 32.

   c) Find the *x*-value that corresponds to a *z*-score of −2.25.

   d) Find the *x*-value that corresponds to a *z*-score of 1.87.

2) Suppose you have a data set in which the sample mean = $\bar{x}$ = 75 and the sample standard deviation = s = 10.6

a) Find the *z*-score for 57.

b) Find the *z*-score for 95.

c) Find the *x*-value that corresponds to a *z*-score of 2.65.

d) Find the *x*-value that corresponds to a *z*-score of –2.38.

3) Before applying to colleges. Jimmy took both the SATs and the ACTs. He scored a 1350 on the SATs and a 28 on the ACTs. The mean and standard deviation for the SATs are 1083 and 194, respectively. The mean and standard deviation for the ACTs are 21 and 5.4, respectively.

On which exam did he do relatively better on? Why?

4) During the year, Jennifer ran in a full marathon as well as a half marathon. She ran the full marathon in 262 minutes and the half marathon in 129 minutes. The mean and standard deviation for the full marathon are 287 minutes and 45 minutes, respectively. The mean and standard deviation for the half marathon are 143 minutes and 18 minutes, respectively. In which marathon did she do relatively better in? Why?

Percentiles indicate what percent of the values in the data set are below a particular data value.

The k^th percentile, denoted P_k, of a data set is a value such that k percent of the observations are less than or equal to the value. Note: The median would be at the 50^th percentile, i.e. median = P_{50}.

Example: Suppose 75 is at the 68^th percentile ($P_{68} = 75$) this means that 68% of the data values are less than 75.

Suppose 62 is at the 42^nd percentile ($P_{42} = 62$) this means that 42% of the data values are less than 62.

1) In a particular county, records indicated the assessed value of each of the 150,000 houses there. The following percentiles were obtained.

P_{15} = \$120,000 P_{50} = \$175,000 P_{65} = \$210,000 P_{90} = \$255,000

a) What percent of houses were assessed below $120,000?________________

b) What percent of houses were assessed below $255,000?________________

c) What percent of houses were assessed between $120,000 and $210,000?_______________

d) What percent of houses were assessed above $210,000?___________________

e) What percent of houses were assessed above $120,000?_________________

f) How many houses were assessed above $120,000?__________________

g) What is the median value of the assessed houses?____________________

A special type of percentile are the quartiles.

Quartiles are the 25^th, 50^th and 75^th percentiles. Denoted Q_1, median, and Q_3, respectively. Quartiles divide the data set into quarters or in other words four parts. Quartiles are used to determine the shape of a distribution and are used to determine if the data set has what are called outliers or extreme values; data values that differ significantly from the other observations in the data set.

Example: Given the data set: 45 47 50 53 56 59 62 65 67 74 76 Find the quartiles.

A way to describe a data set using quartiles is called the **five-number summary**.

The **five-number summary** consists of the minimum, Q_1, median, Q_3, maximum written in this order.

[min, Q_1, median, Q_3, max]

For the data set above find the five-number summary. ______________________________

Interquartile Range **(IQR)** is the difference between the third and first quartiles of a data set. Note: The IQR is actually a measure of spread since it is the range of the middle 50% of the observations.

$IQR = Q_3 - Q_1$

For the data set above, **IQR = _______________**

Using the quartiles and IQR it can be determined if the data set contains outliers or not. An **outlier** is a value that is considerably larger or smaller than most of the values in a data set.

Boundaries (fences) serve as cutoff points for determining outliers:

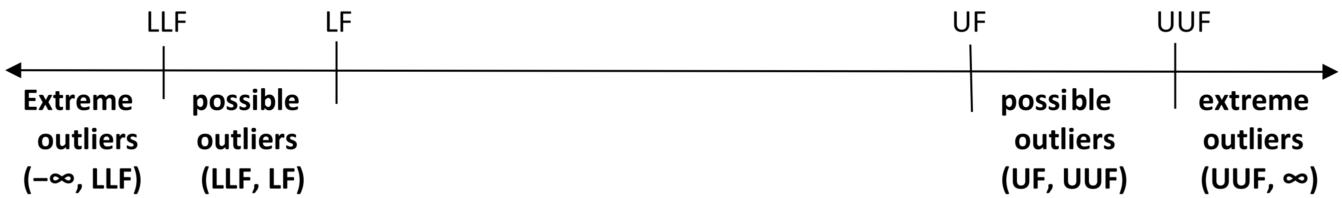
Possible outlier boundaries: Lower Fence = LF = $Q_1 - 1.5(IQR)$

Upper Fence = UF = $Q_3 + 1.5(IQR)$

Extreme outlier boundaries : Lower Lower Fence = LLF = $Q_1 - 3(IQR)$

Upper Upper Fence = UUF = $Q_3 + 3(IQR)$

Therefore any values that are between the lower lower fence and the lower fence or between the upper fence and the upper upper fence are considered possible outliers. Any values less than the lower lower fence and the upper upper fence are considered extreme outliers.



Example 1a: Given the following data set: 14 34 38 43 45 47 53 54 55 56 58 85

Find the five-number summary, the IQR and determine if there are any outliers.

five – number summary ________________________

IQR = _______________ LF = _____________ UF = ________________

Outliers (if any)_____________

Modified Boxplots are a graphical display of quantitative data. Boxplots are created using the five-number summary and outliers, if any. Boxplots are useful for comparing two or more data sets. You can also use a boxplot to identify the approximate shape of the distribution of a data set especially for large data sets; histogram and stem-and-leaf plots are better graphical displays for small data sets.

To draw a box-plot,

Step 1) Draw lines at Q_1, the median, and Q_3 and draw a box using the lines.

Step 2) Put an asterisk at the outliers, if any.

Step 3) Draw the lower whisker out to either the minimum value, if there are no outliers, or to the smallest value that is not an outlier

Step 4) Draw the upper whisker out to either the maximum value, if there are no outliers, or to the largest value that is not an outlier

Example 1b: Draw a Modified Boxplot using the data in example 1a:

Example 2: Given the following data set:

67 68 78 79 80 81 83 85 86 87 88 89 90 91 92 93 95

Find the five-number summary, the IQR, determine if there are any outliers and draw a modified box plot.

five-number summary ____________________

IQR = ___________ LF = ____________ UF = _____________

Outliers (if any) ___________

Modified Box-plot:

Example 3: On a baseball team, the ages of each of the players are as follows:

19 24 24 25 25 25 26 26 26 26 27 27 27 28 28 31

Find the five-number summary, the IQR, determine if there are any outliers and draw a modified box plot.

five-number summary________________________

IQR = _____________ LF = ______________ UF = _____________

Outliers (if any) _______________

Modified box-plot:

4) The U.S. National Center for Health Statistics compiles data on the length of stay by patients in short-term hospitals and publishes its findings in *Vital and Health Statistics.* A random sample of 21 patients yielded the following data on length of stay, in days.

1	1	3	3	4	4	5	6	6	7	7
9	9	10	12	12	13	15	18	23	55	

Find the five-number summary, the IQR, determine if there are any outliers and draw a modified box plot.

five-number summary________________________

IQR = _____________ LF = ______________ UF = _____________

Outliers (if any)_______________

Modified box-plot: