Probability and Statistics

Topic 4 - Measure of Position and Outliers

Aamir Alaud Din, PhD

October 10, 2023

TABLE OF CONTENTS

TABLE OF CONTENTS

RECAP

- The meaning of the measure of central tendency and measure of dispersion for grouped and ungrouped data are the same.
- To determine classes, there is no exact rule and different classes with varying class widths must be tested.
- The formulas for sample and population standard deviation for grouped and ungrouped data are symbolically different but exactly logically exactly same as for the ungrouped data.
- The estimated are approximations and not the deterministic values, after all, statistics is all about the logics.
- To estimate the weighted mean, the weights must be determined carefully after conduction of several logical meetings.

TABLE OF CONTENTS

OBJECTIVES

After learning this topic and studying, you should be able to:

1. Determine and interpret z-scores
2. Compute and interpret percentiles
3. Determine and interpret quartiles
4. Determine and interpret interquartile range
5. Check a set of data for outliers

TABLE OF CONTENTS

THE WHY SECTION

- In Topics 1, 2, and 3 we studied the measure of central tendency and dispersion which are single numbers and tell us the numberical summary of the data in terms of central point and dispersion in the data.
- What if we have to determine which data points divide the data set into four equal parts (quartiles)?
- What if we have to divide the data set into 100 equal parts and determine which numbers divide the data set into 100 equal parts (percentiles)?
- What if we have to decide the standing of a data point in the data set?
- Are the extreme values in the data set part of data?
- Our aim is to answer these questions which is the focus of this topic.

TABLE OF CONTENTS

Z-SCORES

- At the end of the 2014 season, the Los Angeles Angels led the American League with 773 runs scored, while the Colorado Rockies led the National League with 755 runs scored.
- It appears that the Angels are the better run-producing team.
- However, this comparison is unfair because the two teams play in different leagues.
- The Angels play in the American League, where the designated hitter bats for the pitcher, whereas the Rockies play in the National League, where the pitcher must bat (pitchers are typically poor hitters).
- To compare the two teams' scoring of runs, we need to determine their relative standings in their respective leagues.
- We can do this using a *z-score*.

Z-SCORES

z-Scores

The z-score represents the distance that a data value is from the mean in terms of the number of standard deviations. We find it by subtracting the mean from the data value and dividing this result by the standard deviation. There is both a population z-score and a sample z-score:

Population z-Score: $z = \frac{x-\mu}{\sigma}$

Sample z-Score: $z = \frac{x-\bar{x}}{s}$

The z-score is unitless. It has mean 0 and standard deviation 1.

- If a data value is larger than the mean, the z-score is positive.
- If a data value is smaller than the mean, the z-score is negative.

Z-SCORES

- If the data value equals the mean, the z-score is zero.
- A z-score measures the number of standard deviations an observation is above or below the mean.
- For example, a z-score of 1.24 means the data value is 1.24 standard deviations above the mean.
- A z-score of -2.31 means the data value is 2.31 standard deviations below the mean.
- Check if the computation of z-scores is available in python's statistics module (mandatory).
- If not available, write your own function to compute the sample and population z-scores (optional).

EXAMPLE 1

Determine whether the Los Angeles Angels or the Colorado Rockies had a relatively better run-producing season. The Angels scored 773 runs and play in the American League, where the mean number of runs scored was $\mu = 677.4$ and the standard deviation was $\sigma = 51.7$ runs. The Rockies scored 755 runs and play in the National League, where the mean number of runs scored was $\mu = 640.0$ and the standard deviation was $\sigma = 55.9$ runs.

- In Example 1, the team with the higher z-score was said to have a relatively better season in producing runs.
- With negative z-scores, we need to be careful when deciding the better outcome.
- For example, suppose Bob and Mary run a marathon.

Z-SCORES

- If Bob finished the marathon in 213 minutes, where the mean finishing time among all men was 242 minutes with a standard deviation of 57 minutes, and Mary finished the marathon in 241 minutes, where the mean finishing time among all women was 273 minutes with a standard deviation of 52 minutes, who did better in the race?
- Since Bob's z-score is $z_{Bob} = \frac{213-242}{57} = -0.51$ and Mary's z-score is $z_{Mary} = \frac{241-273}{52} = -0.62$, Mary did better.
- Even though Bob's z-score is larger, Mary did better because she is more standard deviations below the mean.
- Why the function for computation of z-score is not available in python's statistics module?
- What should you do now?

TABLE OF CONTENTS

PERCENTILES

- Recall that the median divides the lower 50% of a set of data from the upper 50%.
- The median is a special case of a general concept called the percentile.

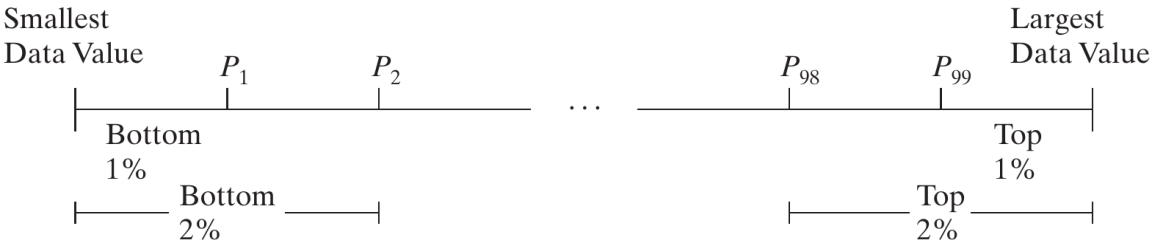
Percentile

The kth percentile, denoted P_k, of a set of data is a value such that k percent of the observations are less than or equal to the value.

- So percentiles divide a set of data that is written in ascending order into 100 parts; thus 99 percentiles can be determined.
- For example, P_1 divides the bottom 1% of the observations from the top 99%, P_2 divides the bottom 2% of the observations from the top 98%, and so on.

PERCENTILES

- Figure below displays the 99 possible percentiles.



- Percentiles are used to give the relative standing of an observation.
- Many standardized exams, such as the SAT college entrance exam, use percentiles to let students know how they scored on the exam in relation to all other students who took the exam.

PERCENTILES

EXAMPLE 2

Jennifer just received the results of her SAT exam. Her SAT Mathematics score of 600 is in the 74th percentile. What does this mean?

SOLUTION

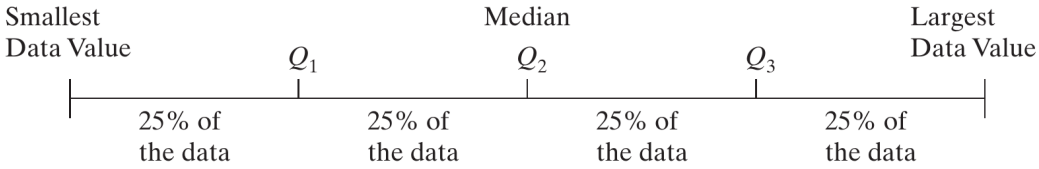
- A percentile rank of 74% means that 74% of SAT Mathematics scores are less than or equal to 600 and 26% of the scores are greater.
- So 26% of the students who took the exam scored better than Jennifer.

TABLE OF CONTENTS

QUARTILES

- The most common percentiles are quartiles. Quartiles divide data sets into fourths, or four equal parts.
    1. The first quartile, denoted Q_1, divides the bottom 25% of the data from the top 75%. Therefore, the first quartile is equivalent to the 25th percentile.
    2. The second quartile, Q_2, divides the bottom 50% of the data from the top 50%; it is equivalent to the 50th percentile or the median.
    3. The third quartile, Q_3, divides the bottom 75% of the data from the top 25%; it is equivalent to the 75th percentile.
- Find the functions to find quartiles in statistics module of python and if not available, find the package which contains the functions.
- Figure below shows the concept of quartiles.

QUARTILES



Finding Quartiles

1. Arrange the data in ascending order.
2. Determine the median, M, or second quartile, Q_2.
3. Divide the data set into halves: the observations below (to the left of) M and the observations above M. The first quartile, Q_1, is the median of the bottom half of the data and the third quartile, Q_3, is the median of the top half of the data.

EXAMPLE 3

The Highway Loss Data Institute routinely collects data on collision coverage claims. Collision coverage insures against physical damage to an insured individual's vehicle. The data in Table 16 represent a random sample of 18 collision coverage claims based on data obtained from the Highway Loss Data Institute. Find and interpret the first, second, and third quartiles for collision coverage claims.

Table 16

$6751	$9908	$3461	$2336	$21,147	$2332
$189	$1185	$370	$1414	$4668	$1953
$10,034	$735	$802	$618	$180	$1657

TABLE OF CONTENTS

INTERQUARTILE RANGE

- So far we have discussed three measures of dispersion: range, standard deviation, and variance, all of which are not resistant.
- Quartiles, however, are resistant.
- For this reason, quartiles are used to define a fourth measure of dispersion.

Interquartile Range

The interquartile range, IQR, is the range of the middle 50% of the observations in a data set. That is, the IQR is the difference between the third and first quartiles and is found using the formula

$$IQR = Q_3 - Q_1$$

INTERQUARTILE RANGE

- The interpretation of the interquartile range is similar to that of the range and standard deviation.
- That is, the more spread a set of data has, the higher the interquartile range will be.

EXAMPLE 5

Determine and interpret the interquartile range of the collision claim data from Example 3.

- Let's compare the measures of central tendency and dispersion discussed thus far for the collision claim data.
- The mean collision claim is \$3874.4 and the median is \$1805.

INTERQUARTILE RANGE

- The median is more representative of the "center" because the data are skewed to the right (only 5 of the 18 observations are greater than the mean).
- The range is \$21,147 - \$180 = \$20,967.
- The standard deviation is \$5301.6 and the interquartile range is \$3933.
- The values of the range and standard deviation are affected by the extreme claim of \$21,147.
- In fact, if this claim had been \$120,000 (let's say the claim was for a totaled Mercedes S-class AMG), then the range and standard deviation would increase to \$119,820 and \$27,782.5, respectively.
- The interquartile range would not be affected.

INTERQUARTILE RANGE

- Therefore, when the distribution of data is highly skewed or contains extreme observations, it is best to use the interquartile range as the measure of dispersion because it is resistant.

Summary: Which Measures to Report

Shape of Distribution	Measure of Central Tendency	Measure of Dispersion
Symmetric	Mean	Standard deviation
Skewed left or skewed right	Median	Interquartile range

- For the remainder of this text, the direction describe the distribution will mean to describe its shape (skewed left, skewed right, symmetric), its center (mean or median), and its spread (standard deviation or interquartile range).

TABLE OF CONTENTS

OUTLIERS

- When performing any type of data analysis, we should always check for extreme observations in the data set.
- Extreme observations are referred to as outliers.
- Outliers can occur by chance, because of error in the measurement of a variable, during data entry, or from errors in sampling.
- For example, in the 2000 presidential election, a precinct in New Mexico accidentally recorded 610 absentee ballots for Al Gore as 110.
- Workers in the Gore camp discovered the data-entry error through an analysis of vote totals.
- Outliers do not always occur because of error.

OUTLIERS

- Sometimes extreme observations are common within a population.
- For example, suppose we wanted to estimate the mean price of a European car.
- We might take a random sample of size 5 from the population of all European automobiles.
- If our sample included a Ferrari F430 Spider (approximately $175,000), it probably would be an outlier, because this car costs much more than the typical European automobile.
- The value of this car would be considered unusual because it is not a typical value from the data set.
- Steps to check outliers in the data are given below.

Checking for Outliers by Using Quartiles

1. Determine the first and third quartiles of the data.
2. Compute the interquartile range.
3. Determine the fences. Fences serve as cutoff points for determining outliers.

$$\text{Lower fence} = Q_1 - 1.5(IQR), \qquad \text{Upper fence} = Q_3 + 1.5(IQR)$$

4. If a data value is less than the lower fence or greater than the upper fence, it is considered an outlier.

EXAMPLE 6

Check the collision coverage claims data in Table 16 for outliers.

TABLE OF CONTENTS

SUMMARY

- The z-score of a data point is the distance of the data point from the mean in terms of number of standard deviations.
- The population z-score is given by the formula

$$z = \frac{x - \mu}{\sigma}$$

- The sample z-score is given by the formula

$$z = \frac{x - \bar{x}}{s}$$

- The z-score can be used to compare the data points from two different data sets provided the mean and standard deviations for both data sets are known.
- The percentile divides the data set into 100 equal parts.

SUMMARY

- The kth percentile of a data set means k percent of the observations are less than or equal to the value.
- Quartiles divide the data set into four equal parts.
- First quartile Q_1 divides the bottom 25% data from the top 75% data.
- Second quartile Q_2 divides the bottom 50% data from the top 50% data which is in fact the median of the data set.
- Third quartile Q_3 divides the bottom 75% data from the top 25% data.
- Interquartile range is the range of the middle 50% of the observations in a data set and is given by the formula

$$\text{IQR} = Q_3 - Q_1$$

SUMMARY

- Outliers are the extreme observations in the data set.
- Outliers may or may not be a part of the data set.
- If an extreme observation is a typo mistake, for example, it must be removed from the data set.
- If an extreme observation, no matter how much extreme, is an observation from the sample, it must be kept in the data set and included in computations.
- The observations below the lower fence (shown below) and above the upper fence (also shown below) must be checked for outliers.

$$\text{Lower Fence} = Q_1 - 1.5(\text{IQR}) \qquad \text{and} \qquad \text{Upper Fence} = Q_3 + 1.5(\text{IQR})$$


Thank You!