3.4: Measures of Position and Outliers

The *z*-score:

The *z*-score for a data point represents the number of standard deviations that lie between the data point and the mean. The *z*-score is sometimes known as the standardized value, and it allows us to compare data points from different distributions.

You can think of the *z*-score as representing the "distance from the mean," with distance measured in standard deviations. A positive *z*-score indicates the data point lies above the mean; a negative *z*-score indicates the data point lies below the mean.

Therefore, for bell-shaped distributions,

from empirical rule

- about 68% of the data points have *z*-scores between –1 and 1;
- about 95.7% of the data points have *z*-scores between –2 and 2;
- about 99% of the data points have *z*-scores between –3 and 3.

99.7%

For all distributions,

Chebyshev's inequality

- at least 75% of the data points have *z*-scores between –2 and 2;
- at least 88.9% of the data points have *z*-scores between –3 and 3.

The *z*-score:

The *z*-score for a value *x* is

$$z = \frac{x-\mu}{\sigma} \text{ (for a population), or}$$

$$z = \frac{x-\bar{x}}{s} \text{ (for a sample), where}$$

μ and σ are the population mean and standard deviation, or
$\bar{x}$ and s are the sample mean and standard deviation.

Note: The *z*-score is unitless. All distributions of *z*-scores have mean 0 and standard deviation 1.

Example 1: Suppose a data set has mean 52 and standard deviation 8. Find the *z*-scores for the scores 44, 64, 38, and 52.

$$Z_{44} = \frac{x-\bar{x}}{s} = \frac{44-52}{8} = -1$$

(1 std. dev. below the mean)

Note that $Z_{60} = 1$ (1 st. dev. above the mean)

$$Z_{64} = \frac{64-52}{8} = \frac{12}{8} = 1.5$$ (so this is 1.5 std devs above the mean)

$$Z_{52} = \frac{52-52}{8} = 0$$ (52 is 0 std devs from the mean)

$$z_{38} = \frac{38-52}{8} = \frac{-14}{8} = -1\tfrac{3}{4} = -1.75$$

what value has a z-score of -2?

$$\bar{x} - 2s = 52 - 2(8) = 52-16 = 36$$

Example 2: In 2014, the mean of the ACT mathematics test was 20.9 and the standard deviation was 5.3. In the same year, the mean of the SAT mathematics test was 513 and the standard deviation was 120. Suppose Tamara, a high school student, received a score of 24 on the ACT mathematics test, and 600 on the SAT mathematics test. On which test did she perform better?

(ACT data from the National Center for Education Statistics, https://nces.ed.gov/programs/digest/d14/tables/dt14_226.50.asp?current=yes; SAT data from the College Board, https://www.collegeboard.org/program-results/2014/sat)

Tamara's z-score on the ACT math:

$$z_{ACT} = \frac{x-\mu}{\sigma} = \frac{24-20.9}{5.3} \approx 0.5849$$

Her z-score on SAT math

$$z_{SAT} = \frac{x-\mu}{\sigma} = \frac{600-513}{120} = 0.725$$

(She did better on the SAT)

Percentiles:

The <u>*k*th percentile</u>, denoted P_k, of a data set is the value such that *k* % of the data points are less than or equal to that value. The <u>percentile rank</u> of a score is the percent of scores equal to or below that score.

For example, a value is known as the 85th percentile if 85% of the data points are less than or equal to that score.

Example 3: Here are the 50 randomly generated scores from Example 8 in Section 3.2. Estimate the 70th percentile, 80th percentile and the 90th percentile.

[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
						[illegible]

See next page for a scanned copy of these numbers, along with the solution to this problem.

Example 2: In 2014, the mean of the ACT mathematics test was 20.9 and the standard deviation was 5.3. In the same year, the mean of the SAT mathematics test was 513 and the standard deviation was 120. Suppose Tamara, a high school student, received a score of 24 on the ACT mathematics test, and 600 on the SAT mathematics test. On which test did she perform better?

(ACT data from the National Center for Education Statistics, https://nces.ed.gov/programs/digest/d14/tables/dt14_226.50.asp?current=yes; SAT data from the College Board, https://www.collegeboard.org/program-results/2014/sat)

(See next page for Example 9 - Finding quartiles for this data set)

Percentiles:

The kth percentile, denoted P_k, of a data set is the value such that k % of the data points are less than or equal to that value. The percentile rank of a score is the percent of scores equal to or below that score.

For example, a value is known as the 85th percentile if 85% of the data points are less than or equal to that score.

70th percentile: 63.15417, 80th percentile: 63.79043, 90th percentile: 67.85567

Example 3: Here are the 50 randomly generated scores from Example 8 in Section 3.2. Estimate the 70th percentile, 80th percentile and the 90th percentile.

37.48295	53.07996	54.94143	57.29676	60.95421	63.16013	66.48368
44.16628	53.20456	55.31494	57.37955	61.43636	63.3329	67.79641
47.40146	54.25092	55.90412	58.99277	61.91373	63.39574	67.85567
50.54246	54.41687	56.48669	59.10063	62.14886	63.61741	68.12883
51.77209	54.42467	56.64306	59.74812	62.52829	63.79043	68.23415
52.06366	54.87849	56.84053	60.00459	62.58302	63.93691	70.72309
53.05055	54.91449	57.00922	60.59386	63.15417	66.44211	73.3014
						87.41814

Bottom 80%

90th percentile

80th percentile

Bottom 90%

70th percentile

Bottom 70%

70th percentile ⇒ 0.70 (50) = 35 scores. So separate these from the top 15 (30%) of scores.

80th percentile ⇒ 0.80(50) = 40 scores. Separate these from the top 10 (20%) of scores

90th percentile ⇒ 0.9 (50) = 45 scores. Separate these from the top 5 (top 10%).

Example 2: In 2014, the mean of the ACT mathematics test was 20.9 and the standard deviation was 5.3. In the same year, the mean of the SAT mathematics test was 513 and the standard deviation was 120. Suppose Tamara, a high school student, received a score of 24 on the ACT mathematics test, and 600 on the SAT mathematics test. On which test did she perform better?

(ACT data from the National Center for Education Statistics, https://nces.ed.gov/programs/digest/d14/tables/dt14_226.50.asp?current=yes; SAT data from the College Board, https://www.collegeboard.org/program-results/2014/sat)

Ex 9: are there any outliers in this data set?

$Q_3 - Q_1 = 63.396 - 54.878 = 8.518$

$1.5(IQR) = 1.5(8.518) = 12.777$

Lower fence: $Q_1 - 1.5(IQR) = 54.878 - 12.777 = 42.101$

Upper fence: $Q_3 + 1.5(IQR) = 63.39574 + 12.777 = 76.173$

So, there are 2 outliers, data points that are more extreme than those fences. 37.483 and 87.418 are outliers.

Percentiles:

The kth percentile, denoted P_k, of a data set is the value such that k % of the data points are less than or equal to that value. The percentile rank of a score is the percent of scores equal to or below that score.

For example, a value is known as the 85th percentile if 85% of the data points are less than or equal to that score.

Example 3: Here are the 50 randomly generated scores from Example 8 in Section 3.2. Estimate the 70th percentile, 80th percentile and the 90th percentile.

Q3

37.48295	53.07996	54.94143	57.29676	60.95421	63.16013	66.48368
44.16628	53.20456	55.31494	57.37955	61.43636	63.3329	67.79641
47.40146	54.25092	55.90412	58.99277	61.91373	63.39574	67.85567
50.54246	54.41687	56.48669	59.10063	62.14886	63.61741	68.12883
51.77209	54.42467	56.64306	59.74812	62.52829	63.79043	68.23415
52.06366	54.87849	56.84053	60.00459	62.58302	63.93691	70.72309
53.05055	54.91449	57.00922	60.59386	63.15417	66.44211	73.3014
						87.41814

Q1

Q2 = M: Between 25th and 26th data point: Bottom half has 25 scores. So the Q1 is the 13th score. Similarly, Q3 is the 13th score in the top half of scores

Quartiles:

Quartiles are values that divide a data set into fourths. The 25th percentile, 50th percentile, and 75th percentile are often referred to as the first quartiles, second quartile, and third quartile.

The second quartile, Q_2 , is the median M of the data set.
The first quartile, Q_1, is the median of the <u>bottom</u> half of the data set (the values less than M).
The third quartile, Q_3, is the median of the <u>top</u> half of the data set (the values greater than M).

<u>Example 4:</u> Calculate the quartiles for the data set $A = \{17,1,9,3,4,10,12,11,5,9,12,8,13,2,7\}$.

Put them in order: 1, 2, 3, 4, 5, 7, 8, 9, 9, 10, 11, 12, 12, 13, 17

Median = $m = Q_2 = 9$

Q_1 is the median of $\{1,2,3,4,5,7,8\}$, so $Q1 = 4$

Q_3 is the median of $\{9, 10, 11, 12, 12, 13, 17\}$, so $Q_3 = 12$

↑ Q3

<u>Example 5:</u> Calculate the quartiles for $B = \{2,3,3,4,5,5,6,7,9,9,9,10,11,12,12,13\}$.

$M = Q2 = 8$ $\quad \frac{7+9}{2} = \frac{16}{2} = 8$

Q_1 is median of $\{2,3,3,4,5,5,6,7\}$, so $Q_1 = \frac{4+5}{2} = 4.5$

Q_3 is median of $\{9,9,9,10,11,12,12,13\}$, so $Q_3 = \frac{10+11}{2} = 10.5$

<u>Example 6:</u> Calculate the quartiles for $C = \{1,2,3,8,11,15,16,19,27,29,31,34,40,51,52,52,53\}$.

$Q_2 = m = 27$

$Q_1 = \frac{8+11}{2} = 9.5$

$Q_3 = \frac{40+51}{2} = 45.5$

<u>Example 7:</u> Calculate the quartiles for $D = \{1,1,3,5,10,10,15,15,19,20,22,24,24,30,31,32,32,38\}$.

$Q_2 = m = \frac{19+20}{2} = 19.5$

There are 9 values below the median, so the 5th will be Q_1

$Q_1 = 10$

There are 9 values above the median, so the 5th of these will be Q_3. $Q_3 = 30$

Definition: The *interquartile range*, denoted *IQR*, is the difference between the first and third quartiles.

$$IQR = Q_3 - Q_1$$

The *IQR* is the range of the middle 50% of the data set. The interquartile range is a measure of dispersion (how spread out the data are); the standard deviation, variance, and range of the data set are also measures of dispersion. The IQR is resistant to extreme values (outliers); the range and standard deviation are not resistant to extreme values.

An *outlier* is an extreme value (extremely low or extremely high, relative to other values in the data set).

One common definition for an outlier: A data point is considered an outlier if it lies beyond these *fences*:

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

So, a data point *x* is an outlier if $x < Q_1 - 1.5(IQR)$ or if $x > Q_3 + 1.5(IQR)$.

Example 8: Using the definition above, find any outliers in these data sets.

$1.5(IQR) = 1.5(9) = 13.5$

a. $A = \{2, 5, 7, 10, 12, 14, 30\}$

$M = Q_2 = 10$
$Q_1 = 5$, $Q_3 = 14$

$IQR = 14 - 5 = 9$

Lower fence: $Q_1 - 1.5(IQR) = 5 - 1.5(9)$
$= 5 - 13.5 = -8.5$

upper fence: $Q_3 + 1.5(IQR) = 14 + 13.5 = 27.5$
So 30 is an outlier.

b. $B = \{2, 14, 16, 19, 23, 24, 30\}$

$M = Q_2 = 19$
$Q_1 = 14$
$Q_3 = 24$

$IQR = 24 - 14 = 10$
$1.5(IQR) = 1.5(10) = 15$
Lower fence: $Q_1 - 1.5(IQR) = 14 - 15 = -1$
Upper fence $= Q_3 + 1.5(IQR) = 24 + 15 = 39$

No outliers

Example 9: Does the randomly generated data set in Example 3 contain any outliers?

See previous page

Some researchers and statisticians consider a data point to be an extreme outlier if it lies beyond the two outer fences $Q_1 - 3(IQR)$ and $Q_3 + 3(IQR)$. Does the Example 3 data set contain extreme outliers? $3(IQR) = 3(8.518) = 25.554$

$Q_1 - 3(IQR) = 54.878 - 25.554 = 29.324$
$Q_3 + 3(IQR) = 63.396 + 25.554 = 88.95.$

So No, this data set has no extreme outliers.