3.3: Measures of Position

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 1: Here are the 50 randomly generated scores from Example 4 in Section 3.3. 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

90th percentile

80th percentile

70th percentile

divides the bottom 70% from the top 30%

70th percentile: 0.70 (50) = 35

70th percentile is 63.15417

80th percentile 0.8(50) = 40

80th percentile is 63.79043

90th percentile: 0.9 (50) = 45

90th percentile is 67.85567

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.

Our book is not using either of these

Method 1 (Tukey's Method): Used in our book:
The second quartile, Q_2 , is the median M of the data set.
The first quartile, Q_1, is the median of the *bottom half of the data set.
The third quartile, Q_3, is the median of the *top half of the data set.
* If the data set has an odd number of data points, the median is included in both halves.

Method 2 (NOT Used in our book):
The second quartile, Q_2 , is the median M of the data set.
The first quartile, Q_1, is the median of the bottom half of the data set (the values less than M).
The third quartile, Q_3, is the median of the top half of the data set (the values greater than M).

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

Reorder: 1, 2, 3, 4, 5, 7, 8, 9, 9, 10, 11, 12, 12, 13, 17

Q1   Q2 = median = 9   Q3

Q_1: $L_1 = 0.25n = 0.25(15) = 3.75 \Rightarrow$ round up to 4 $\Rightarrow Q_1$ is the 4th data point

Q_3: $L_3 = 0.75n = 0.75(15) = 11.25 \Rightarrow$ round up to 12 $\Rightarrow Q_3$ is the 12th data point

Q1: 4
Q2 = M: 9
Q3 = 12

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

Q_1 Q_3

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

n = 16

Q1: $L_1 = 0.25(16) = 4 \Rightarrow$ Q1 is the average of the 4th and 5th data points
$Q_1 = \frac{4+5}{2} = 4.5$

Q3: $L_3 = 0.75(16) = 12 \Rightarrow$ Q3 is the average of the 12th and 13th data points: Q3 = $\frac{10+11}{2} = 10.5$

Q1 = 4.5
Q2 = m = 8
Q3 = 10.5

Example 4: Calculate the quartiles for $C = \{1, 2, 3, 8, 11, 15, 16, 19, 27, 29, 31, 34, 40, 51, 52, 52, 53\}$.

To find Q1 and Q3:

$L_1 = 0.25n$

$L_3 = 0.75n$

If L_1 or L_3 is a whole number, the quartile (Q_2 or Q_3) is the average of the number in position L_1 and position L_1+1 (or L_3 and L_3+1)

If L_1 or L_3 is not a whole number, round it up to nearest whole number above. the quartile is the data point in that position

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

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 (or a potential outlier) if it lies beyond these *fences*:

Lower fence (lower limit) = $Q_1 - 1.5(IQR)$
Upper fence (upper limit) = $Q_3 + 1.5(IQR)$

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

Example 6: Using the definition above, find any outliers in these data sets.

Q1 ↓ Q1=M ↓ ↙ Q3

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

n = 7
Q1: 0.25(7) = 1.75 ⇒ Q1 is the 2nd data point
Q3: 0.75(7) = 5.25 ⇒ Q3 is the 6th data point
IQR = Q3 − Q1 = 14 − 5 = 9

5-number Summary:

min	Q1	Q_2 = median	Q3	Max
2	5	10	14	30

1.5(IQR) = 1.5(9) = 13.5
Upper Fence: Q3 + 1.5(IQR) = 14 + 13.5 = 27.5
Lower fence: Q1 − 1.5(IQR) = 5 − 13.5 = −8.5
Outliers will be points that are below −8.5 or above 27.5

$\boxed{\text{30 is the only outlier}}$

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

Example 7: Does the randomly generated data set in Example 1 contain any outliers?

Some researchers and statisticians consider a data point to be an extreme outlier if it lies beyond the two <u>outer fences</u> $Q_1 - 3(IQR)$ and $Q_3 + 3(IQR)$. Does the Example 1 data set contain extreme outliers? (don't worry about "extreme outliers")

The five-number summary:

We can get a fairly useful and descriptive picture of any data set from just 5 numbers: the minimum (smallest value), first quartile, second quartile (median), third quartile, and maximum (largest value).

<u>Five-number summary</u>:

Minimum Q_1 Q_2 (= median) Q_3 Maximum

Boxplots:

A boxplot, or box-and-whisker plot, visually depicts these five numbers.

<u>How to make a boxplot</u>:

1. Determine the minimum, quartiles, and maximum of the data set.
2. Set up a horizontal scale, and draw a box that has Q_1 and Q_3 for endpoints, and a vertical line at Q_2 (the median). The length of the box is $IQR = Q_3 - Q_1$.
3. Calculate the upper and lower fences, and temporarily mark them on the graph:

   Lower fence = $Q_1 - 1.5(IQR)$
   Upper fence = $Q_3 + 1.5(IQR)$

4. Draw a line from Q_1 to the smallest data point that is larger than the lower fence. Draw a line from Q_3 to the largest data point that is smaller than the upper fence.
5. Use an asterisk to mark any data values that lie outside the fences.

5-number summary:

min	Q1	Q2=m	Q3	max
3	5	6.5	7.5	20

Example 8: Construct a box plot for the data set.

3, 4, 4, 5, | 5, 5, 6, 6, | 7, 7, 7, 7 | 8, 8, 9, ~~9~~ 20

$Q1 = \frac{5+5}{2} = 5$ $Q2 = m = 6.5$ $Q3 = \frac{7+8}{2} = 7.5$

$n = 16$

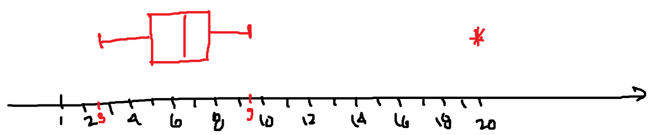
$L_1 = 0.25n = 0.25(16) = 4$

$L_3 = 0.75n = 0.25(16) = 12$

$IQR = Q3 - Q_1 = 7.5 - 5 = 2.5$

$Q1 - 1.5(IQR) = 5 - 1.5(2.5)$
$= 1.25$ Lower fence

$Q3 + 1.5(IQR) = 7.5 + 1.5(2.5)$
$= 11.25$ Upper fence



Most extreme non-outlier on left side is 3

Most extreme non-outlier on right side is 9

(20 is an outlier... it is beyond upper fence

Example 9: Construct a box plot for the data set.

21, 1, 5, 3, 7, 14, 12, 10, 5, 9, 12, 4, 6, 13, 2, 8

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.

For all distributions (from Chebyshev's Rule),

- 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.

<u>The *z*-score (or standardized score)</u>:

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.

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

<u>Example 10:</u> Suppose a data set has mean 52 and standard deviation 8. Find the *z*-scores for the scores 44, 64, 38, and 52.

Example 11: 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)