Section 1.10 – Measures of Position, Outliers, and Boxplots

This section introduces measures of relative standing, which numbers are showing the location of data values relative to the other values within a data set. They can be used to compare values from different data sets, or to compare values within the same data set. The most important concept is the z score. We will also discuss percentiles and quartiles, as well as a new statistical graph called the boxplot.

Definition

A z score (or standardized value) is the number of standard deviations that a given value x is above or below the mean. The z score is calculated by using one of the following:

Sample	***Population***
$z = \dfrac{x - \overline{x}}{s}$	$z = \dfrac{x - \mu}{\sigma}$

Example

Compare those two data values by finding z score.

Men heights

70.8	66.2	71.7	68.7	67.6	69.2	66.5	67.2	68.3	65.6	63.0	68.3	73.1	67.6
68.0	71.0	61.3	76.2	66.3	69.7	65.4	70.0	62.9	68.5	68.3	69.4	69.2	68.0
71.9	66.1	72.4	73.0	68.0	68.7	70.3	63.7	71.1	65.6	68.3	66.3		

Men weights

169.1	144.2	179.3	175.8	152.6	166.8	135.0	201.5	175.2	139.0	156.3	186.6	191.1
151.3	209.4	237.1	176.7	220.6	166.1	137.4	164.2	162.4	151.8	144.1	204.6	193.8
172.9	161.9	174.8	169.8	213.3	198.0	173.3	214.5	137.1	119.5	189.1	164.7	170.1
151.0												

Solution

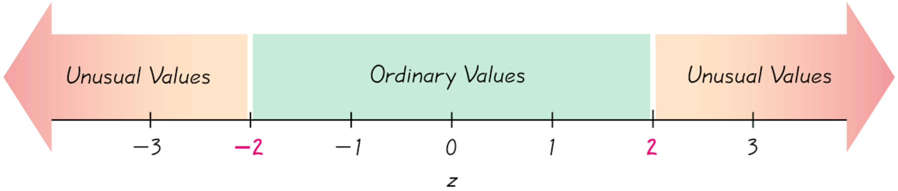
The heights: $\overline{x} = 68.34$ in. and $s = 3.02$ in.

$$z = \frac{x - \overline{x}}{s} = \frac{76.2 - 68.34}{3.02} = 2.60$$

The weights: $\overline{x} = 172.55$ lb. and $s = 26.33$ lb.

$$z = \frac{x - \overline{x}}{z} = \frac{237.1 - 172.55}{26.33} = 2.45$$

Z Scores, Unusual Values, and Outliers



Whenever a value is less than the mean, its corresponding *z* score is negative

Ordinary values: $-2 \le z$ score ≤ 2

Unusual Values: z score < -2 or z score > 2

Percentiles

Definition

Percentiles are measures of location. There are 99 percentiles denoted $P_1, P_2, \ldots, P_{99}$, which divide a set of data into 100 groups with about 1% of the values in each group.

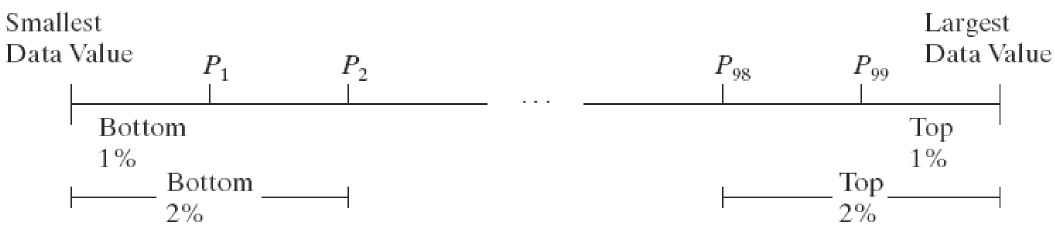
The *k*th percentile, denoted, P_k, of a set of data is a value such that *k* percent of the observations are less than or equal to the value.

Finding the Percentile of a Data Value

The process of finding the percentile that corresponds to a particular data value *x* is given by the following:

$$percentile\ of\ value\ x = \frac{number\ of\ values\ less\ than\ x}{total\ number\ of\ values} \cdot 100$$

(*Round to the nearest whole number*)



Example

The table below lists the 35 sorted budget amounts (in millions of dollars) from the simple random sample of movies. Find the percentile for the value of $29 million

4.5	5	6.5	7	20	20	29	30	35	40	40	41
50	52	60	65	68	68	70	70	70	72	74	75
80	100	113	116	120	125	132	150	160	200	225	

Solution

From the table, there are 6 budget amounts less than 29, so

$$percentile\ of\ 29 = \frac{6}{35} \cdot 100 \approx 17$$

✓ The budget amount of \$29 million is the 17^{th} percentile. This can be interpreted loosely as: The budget amount of \$29 million separates the lowest 17% of the budget amounts from the highest 83%.

Notation

$$L = \frac{k}{100} \cdot n$$

n total number of values in the data set

k percentile being used

L locator that gives the position of a value

P_k *k*th percentile

Example

The table below lists the 35 sorted budget amounts (in millions of dollars) from the simple random sample of movies. Find the value of the 90^{th} percentile, P_{90}

4.5	5	6.5	7	20	20	29	30	35	40	40	41
50	52	60	65	68	68	70	70	70	72	74	75
80	100	113	116	120	125	132	150	160	200	225	

Solution

$k = 90$ and $n = 35$ because there are 35 data values.

$$\begin{aligned} L &= \frac{k}{100} \cdot n \\ &= \frac{90}{100} \cdot 35 \\ &\approx 32 \end{aligned}$$

The 32nd value is 150 that is, $P_{90} = \$150$ million.

So, about 90% of the movies have budgets below \$150 million and about 10% of the movies have budgets above \$150 million.

Example

The list of setting speed limits are recorded (in mi/h.) and listed below

68	68	72	73	65	74	73	72	68	65	65	73	66	71
68	74	66	71	65	73	59	75	70	56	66	75	68	75
62	72	60	73	61	75	58	74	60	73	58	74		

That section has a posted speed limit of 65 mi/h. Traffic engineers often establish speed limits by using the "85th percentile rule" whereby the speed limit is set so that 85% of drivers are at or below the speed limit.

a) Find the 85th percentile of the listed speeds.

b) Given that speed limits are usually rounded to a multiple of 5, what speed limit is suggested by these data? Explain your choice.

c) Does the existing speed limit conform to the 85th percentile rule?

Solution

Sorting the data

56	58	58	59	60	60	61	62	65	65	65	65	66	66	66	68	68	68	68	68
70	71	71	72	72	72	73	73	73	73	73	73	74	74	74	74	75	75	75	75

a) $n = 40$ because there are 40 sample. To find the 85th percentile, then $k = 85$.

$$L = \frac{k}{100} \cdot n = \frac{85}{100} \cdot 40 = 34$$

That indicated that the 85th percentile is 34th speeds, the 34th speed is 74 mi/h.

b) A speed of 75 mi/h is the multiple of 5 closest to the 85th percentile, but it is probably safer to round down, so that a speed of 70 mi/h is the closest multiple of 5 below the 85th percentile.

c) The existing speed limit of 65 mi/h is below the speed limit determined by the 85th percentile rule, so the existing speed limit does not conform to the 85th percentile rule.

Example

Find P_{75} for the set of test scores below

51	54	64	68	72	74	76	82	89	94	99

Solution

Data already sorted.

$$k = \frac{L}{100} \cdot n = \frac{75}{100} \cdot 11 = 8.25$$

1. we find the 8th number in the list = 82

2. $.25\left(9^{th}\#-8^{th}\#\right) = .25(89-82) = 1.75$

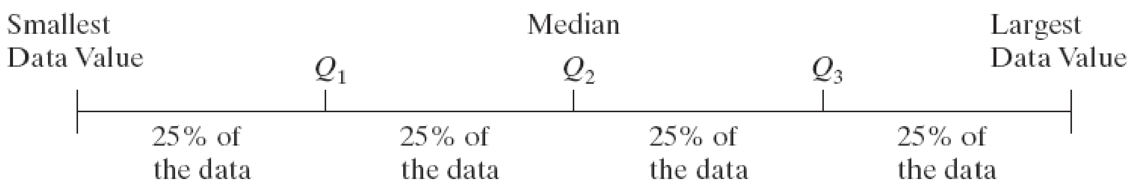
3. $P_{75} = 82 + 1.75 \underline{= 83.75}$

Quartiles

Definition

Quartile are measures of location, denoted Q_1, Q_2, and Q_3, which divide a set of data into four groups with about 25% of the values in each group.

- Q_1 (First Quartile) separates the bottom 25% of sorted values from the top 75%.
- Q_2 (Second Quartile) same as the median; separates the bottom 50% of sorted values from the top 50%.
- Q_3 (Third Quartile) separates the bottom 75% of sorted values from the top 25%.



Example

Find the value of the first quartile Q_1.

4.5	5	6.5	7	20	20	29	30	35	40	40	41
50	52	60	65	68	68	70	70	70	72	74	75
80	100	113	116	120	125	132	150	160	200	225	

Solution

Finding Q_1 is the same as finding P_{25}

$L = \frac{k}{100} \cdot n = \frac{25}{100} \cdot 35 = 8.75 \approx 9$.

Therefore, the first quartile is given by $Q_1 = \$35$ million.

- ✓ Interquartile range (or ***IRQ***) $= Q_3 - Q_1$
- ✓ Semi-interquartile range $= \dfrac{Q_3 - Q_1}{2}$
- ✓ Midquartile $= \dfrac{Q_3 + Q_1}{2}$
- ✓ 10–90 percentile range $= P_{90} - P_{10}$

Definition

For a set of data, the *5-**number summary*** consists of the minimum value; the first quartile Q_1; the median (or second quartile Q_2); the third quartile, Q_3; and the maximum value.

A ***boxplot*** (or ***box-and-whisker-diagram***) is a graph of a data set that consists of a line extending from the minimum value to the maximum value, and a box with lines drawn at the first quartile, Q_1; the median; and the third quartile, Q_3.

Example

Use the movie budget amount to find the 5–number summary.

4.5	5	6.5	7	20	20	29	30	35	40	40	41
50	52	60	65	68	68	70	70	70	72	74	75
80	100	113	116	120	125	132	150	160	200	225	

Solution

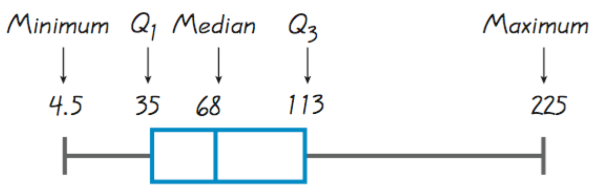
From the table:

The minimum is \$4.5 million and the maximum is \$225 million.

$L = \frac{25}{100} \cdot 35 \approx 9 \quad \Rightarrow \quad P_{25} = 35 \rightarrow Q_1 = \$35$ million

$L = \frac{50}{100} \cdot 35 \approx 18 \quad \Rightarrow \quad P_{50} = 68 \rightarrow Q_2 = \$68$ million.

$L = \frac{75}{100} \cdot 35 \approx 27 \quad \Rightarrow \quad P_{75} = 113 \rightarrow Q_3 = \$113$ million.



Procedure for Constructing a Boxplot

1. Find the 5-number summary consisting of the minimum value, Q_1, the median, Q_3, and the maximum value .
2. Construct a scale with values that include the minimum and maximum data values.
3. Construct a box (rectangle) extending from Q_1 to Q_3, and draw a line in the box at the median value.
4. Draw lines extending outward from the box to the minimum and maximum data values.

Outliers and Modified Boxplots

Definition

An ***outlier*** is a value that lies very far away from the vast majority of the other values in a data set.

For purposes of constructing *modified boxplots*, we can consider outliers to be data values meeting specific criteria.

In modified boxplots, a data value is an outlier if it is . . .

above Q_3 by an amount greater than $1.5 \times$ IQR

or

below Q_1 by an amount greater than $1.5 \times$ IQR

Example

The pulse rates of females listed below

76	72	88	60	72	68	80	64	68	68	80	76	68	72	96
72	68	72	64	80	64	80	76	76	76	80	104	88	60	76
72	72	88	80	60	72	88	88	124	64					

Use the data to construct a modified boxplot.

Solution

60	60	60	64	64	64	64	68	68	68	68	68	72	72	72	72
72	72	72	72	76	76	76	76	76	76	80	80	80	80	80	80
88	88	88	88	88	96	104	124								

From the table: The minimum is 60 and the maximum is 124.

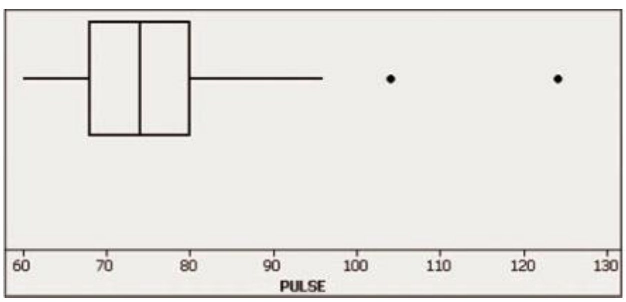
$$L = \frac{25}{100} \cdot 40 \approx 10 \quad \Rightarrow \quad P_{25} = 68 \rightarrow Q_1 = 68$$

$$L = \frac{75}{100} \cdot 40 \approx 30 \quad \Rightarrow \quad P_{75} = 80 \rightarrow Q_3 = 80$$

Interquartile range (or ***IRQ***) $= Q_3 - Q_1 = 80 - 68 = 12$

For pulse rates above the third quartile of 80 by an amount that is greater than $1.5 \times IQR = 1.5 \times 12 = 18$ so high outliers are greater than 98.

The pulse rates of 104 and 124 satisfy this condition, so those two values are outliers.



Store the list values in L1.
Press [2nd] (Y=) [STAT PLOT] [1]
Press [ENTER} to turn on the stat plot.
Scroll down to Type, then right 4 times first pictures in 2nd row.
Be sure that Xlist is L1
Allow Freq: 1
Press [GRAPH], Press [ZOOM] and select 9, and press [ENTER]

Exercises *Section* 1.10 – Measures of Position, Outliers, and Boxplots

1. When Reese Witherspoon won an Oscar as Best Actress for the movie *Walk the Line*, her age was converted to a z-score of -0.61 when included among the ages of all other Oscar-winning Best Actress at the time of this writing. Was her age above the mean or below the mean? How many standard deviations away from the mean is her age?

2. Hoffman was 38 years of age when he won a Best Actor Oscar for his role in Capote. The Oscar-winning Best Actors have a mean age of 43.8 years and a standard deviation of 8.9 years.

a) What is the difference between Hoffman's age and the mean age?
b) How many standard deviations is that (the difference found in part (a))?
c) Convert Hoffman's age to a z-score.
d) If we consider "usual" ages to be those that convert to z-scores between -2 and 2, is Hoffman's age usual or unusual?

3. Eruptions of the Old Faithful geyser have duration times with a mean of 245.0 sec and a standard deviation of 36.4 sec. One eruption had a duration time of 110 sec.

a) What is the difference between a duration time of 110 sec and the mea?
b) How many standard deviations is that (the difference found in part (a))?
c) Convert duration time of 110 sec to a z-score.
d) If we consider "usual" ages to be those that convert to z-scores between -2 and 2, is a duration time of 110 sec usual or unusual?

4. Human body temperatures have a mean of 98.20°F and a standard deviation of 0.62°F. Convert each given temperature to a z-score and determine whether it is usual and unusual.

a) 101.00°F *b*) 96.90°F *c*) 96.98°F

5. Scores on SAT test have a mean of 1518 and a standard deviation of 325. Scores on the ACT test have a mean of 21.1 and standard deviation of 4.8. Which is relatively better: a score of 1840 on the SAT test or a score of 26.0 on the ACT test? Why?

6. Scores on SAT test have a mean of 1518 and a standard deviation of 325. Scores on the ACT test have a mean of 21.1 and standard deviation of 4.8. Which is relatively better: a score of 1190 on the SAT test or a score of 16.0 on the ACT test? Why?

7. Use the given sorted values, which are the numbers of points scored in the Super Bowl for a recent period of 24 years. Find the percentile corresponding to the given number of points

36 37 37 39 39 41 43 44 44 47 50 53 54 55 56 56 57 59 61 61 65 69 69 75

a) 47 *b*) 65 *c*) 54 *d*) 41

8. For the given data, find the indicated percentile or quartile

36 37 37 39 39 41 43 44 44 47 50 53 54 55 56 56 57 59 61 61 65 69 69 75

a) P_{20} *c)* P_{50} *e)* P_{25} *g)* Q_1

b) P_{80} *d)* P_{75} *f)* P_{95} *h)* Q_3

9. The number of hours of television watched per day by a sample of 28 people

2 4 1 5 7 2 5 4 4 2 3 6 4 3 5 2 0 3 5 9 4 5 2 1 3 6 7 2

a) Find the data set's first, second, and third quartiles.
b) Draw a box-and-whisker plot that represents the data set.
c) About 75% of the people watched no more than now many hours of television per day?
d) What percent of the people watched more than 4 hours of television per day?
e) If you randomly selected one person from the sample, what is the likelihood that the person watched less than 2 hours of television per day? Write your answer as a percent.

10. The hourly earnings (in dollars) of a sample of 25 railroad equipment manufacturers

15.6 18.75 14.6 15.8 14.35 13.9 17.5 17.55 13. 14.2 19.05 15.35 15.2
19.45 15.95 16.5 16.3 15.25 15.05 19.1 15.2 16.22 17.75 18.4 15.25

a) Find the data set's first, second, and third quartiles.
b) Draw a box-and-whisker plot that represents the data set.
c) About 75% of the manufacturers made less than \$15.80 per hour?
d) What percent of the manufacturers made more than \$15.80 per hour?
e) If you randomly selected one manufacturer from the sample, what is the likelihood that the manufacturer made less than \$15.80 per hour? Write your answer as a percent.

11. A certain brand of automobile tire has a mean life span of 35,000 miles, with a standard deviation of 2250 miles. (Assume the life spans of the tires have a bell-shaped distribution)

a) The life spans of three randomly selected tires are 34,000 miles, 37,000 miles, and 30,000 miles. Find the z-score that corresponds to each life span. According to the z-scores, would the life spans of any of these tires be considered unusual?
b) The life spans of three randomly selected tires are 30,500 miles, 37,250 miles, and 35,000 miles. Using the Empirical Rule, find the percentile that corresponds to each life span.

12. The life spans of species of fruit fly have a bell shaped distribution, with mean of 33 days and a standard deviation of 4 days.

a) The life spans of three randomly selected fruit flies are 34 days, 30 days, and 42 days. Find the z-score that corresponds to each life span and determine if any of these life spans are unusual.
b) The life spans of three randomly selected fruit flies are 29 days, 41 days, and 25 days. Using the Empirical Rule, find the percentile that corresponds to each life span.

13. Find the Q_1 and Q_3 for the given data: 49 52 52 52 74 67 55 55

14. Find the Q_1 and Q_3 for the given weights (in pounds) of 30 newborn babies listed below:

5.5	5.7	5.8	6.0	6.1	6.1	6.3	6.4	6.5	6.6
6.7	6.7	6.7	6.9	7.0	7.0	7.0	7.1	7.2	7.2
7.4	7.5	7.7	7.7	7.8	8.0	8.1	8.1	8.3	8.7

15. Find the percentile for the data value:

113 125 117 111 119 121 111 109 116 113 117 127 109 113 115 110

Data value: 119

16. The test scores of 40 students are listed below:

30 35 43 44 47 48 54 55 56 57 59 62 63 65 66 68 69 69 71 72
72 73 74 76 77 77 78 79 80 81 81 82 83 85 89 92 93 94 97 98

Find P_{56}