DESCRIPTIVE STATISTICS

MAT 152 – Statistical Methods I

Lecture 2

Instructor: Dustin Roten

Fall 2020

Measures of the Location of the Data

Suppose a student claims that she scored in the 90th **percentile** on the SAT. What does this mean?

Answer: 90% of the other test scores were at or below her score. 10% of the scores were higher.

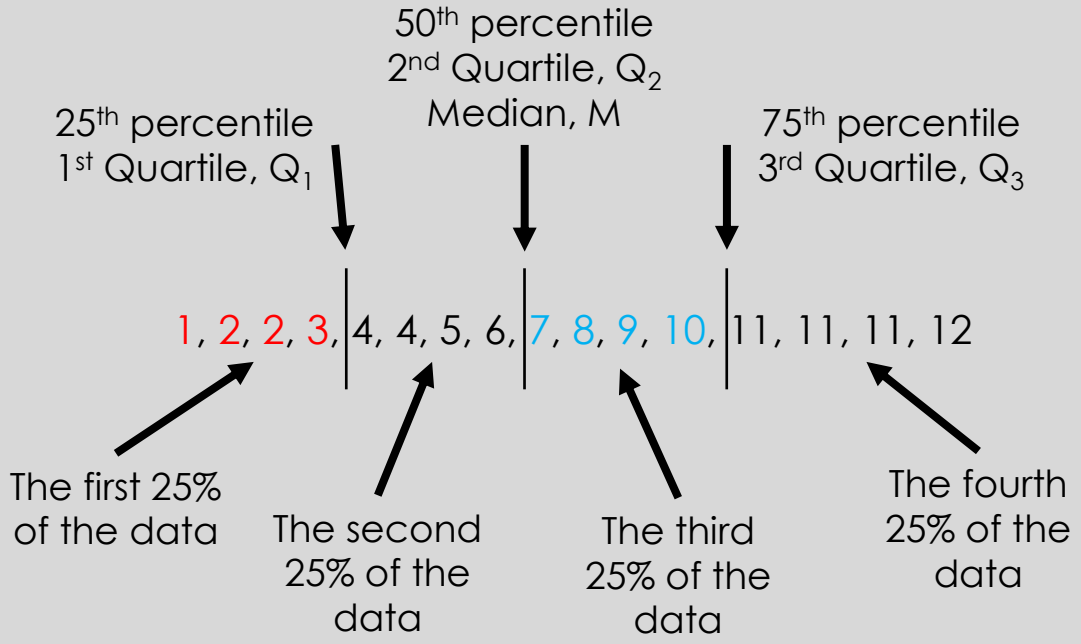
Suppose the exam was really challenging. It is possible to score a grade of 70% but still score in a high percentile (85th). This would indicate that 85% of test takers scored a 70% or less on the exam.

Percentiles indicate <u>where</u> values are with respect to the entire dataset.

Measures of the Location of the Data

Consider the following (**ordered**) data:

1, 2, 2, 3, 4, 4, 5, 6, 7, 8, 9, 10, 11, 11, 11, 12.



Percentiles and quartiles aren't always present in the data. Here, they are **in between** values.

Quartiles divide data into quarters
Percentiles divide data into hundredths

Percentiles are useful for comparing data.

How are Q_1, Q_3, and M calculated?

Median (M) – The median is the **center** of the data. It doesn't have to be a value that exists in the dataset.

1, 2, 2, 3, 4, 4, 5, 6, 7, 8, 9, 10, 11, 11, 11, 12

Count in from the outsides of the dataset until the middle number (or numbers) are found. Here, the middle values are 6 and 7. Thus, these two values must be added then divided by 2.

$$M = \frac{6+7}{2} = \frac{13}{2} = 6.5$$

Half the values are smaller than 6.5 and half are larger.

How are Q_1, Q_3, and M calculated?

$$M = 6.5$$

1, 2, 2, 3, 4, 4, 5, 6,	7, 8, 9, 10, 11, 11, 11, 12
50% of the data	50% of the data

The first quartile, Q_1, is the middle value of the **lower half** of the data. Here, 3 and 4 are the middle values.

$$Q_1 = \frac{3+4}{2} = \frac{7}{2} = 3.5$$

25% of the data are smaller than 3.5 and 75% are larger than 3.5.

How are Q_1, Q_3, and M calculated?

The third quartile, Q_3, is the middle value of the **upper half** of the data. Here, 10 and 11 are the middle values.

$$Q_3 = \frac{10 + 11}{2} = \frac{21}{2} = 10.5$$

75% of the data are smaller than 10.5 and 25% are larger than 10.5.

$$M = 6.5$$

1, 2, 2, 3, | 4, 4, 5, 6, | 7, 8, 9, 10, | 11, 11, 11, 12

$$Q_1 = 3.5 \qquad Q_3 = 10.5$$

Example

Sometimes, values need not be averaged. Consider the following (unordered) data:

5, 8, 3, 4, 9, 4, 2, 6, 7, 6, 2

Step 1: order the data

2, 2, 3, 4, 4, 5, 6, 6, 7, 8, 9

Step 2: Find the min and max of the data

$$\min = 2, \max = 9$$

Step 3: Find the Median (M)

~~2~~, ~~2~~, ~~3~~, ~~4~~, ~~4~~, 5, ~~6~~, ~~6~~, ~~7~~, ~~8~~, ~~9~~

Step 4: Find the first and third quartiles (Q_1 and Q_3)

2, 2, 3, 4, 4, 5, 6, 6, 7, 8, 9

$$\min = 2; \mathrm{Q}_1 = 3; \mathrm{M} = 5; \mathrm{Q}_3 = 7; \max = 9$$

Interquartile Range

The interquartile range is a number that indicates the **spread** of the middle 50% of the data.

$$IQR = Q_3 - Q_1$$

The IQR can be used to determine outliers. Outliers may be errors or abnormalities.

Any values above $Q_3 + (1.5)(IQR)$ are considered outliers

Any values below $Q_1 - (1.5)(IQR)$ are considered outliers

1, 2, 2, 3, 4, 4, 5, 6, 7, 8, 9, 10, 11, 11, 11, 12. $[IQR = 10.5 - 3.5 = 7.0]$

Any values above $10.5 + (1.5)(7.0) = 21$ are considered outliers

Any values above $3.5 - (1.5)(7.0) = -7$ are considered outliers

Finding the value corresponding to the k^{th} percentile

Suppose the 30th percentile needs to be found. What value would this correspond to?

Finding the value corresponding to the k^{th} percentile.

$$i = \frac{k}{100}(n+1)$$

$i = \text{index of the value}$
$k = k^{th} \text{ percentile}$
$n = \text{total number of data}$

1, 2, 2, 3, 4, 4, 5, 6, 7, 8, 9, 10, 11, 11, 11, 12

$$i = \frac{30}{100}(16+1) = 0.3 \cdot 17 = 5.1$$

The 30th percentile lies between the 5th and 6th values.

These values are averaged to determine the 30th percentile, which is 4. In this data, the 5th and 6th values are both 4.

Example (From textbook)

EXAMPLE 2.17

Listed are 29 ages for Academy Award winning best actors *in order from smallest to largest.*
18; 21; 22; 25; 26; 27; 29; 30; 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

a. Find the 70^{th} percentile.
b. Find the 83^{rd} percentile.

a.
- $k = 70$
- i = the index
- $n = 29$

$i = \frac{k}{100}(n + 1) = (\frac{70}{100})(29 + 1) = 21$. Twenty-one is an integer, and the data value in the 21^{st} position in the ordered data set is 64. The 70^{th} percentile is 64 years.

b.
- $k = 83^{rd}$ percentile
- i = the index
- $n = 29$

$i = \frac{k}{100}(n + 1) =)\frac{83}{100})(29 + 1) = 24.9$, which is NOT an integer. Round it down to 24 and up to 25. The age in the 24^{th} position is 71 and the age in the 25^{th} position is 72. Average 71 and 72. The 83^{rd} percentile is 71.5 years.

Finding the k^{th} percentile corresponding to a specific value

Now, suppose the percentile of a value is needed.

The percentile associated with a specific value can be found by:

$$\frac{x + 0.5y}{n}(100)$$

x = the number of values **below** the value in question
y = the number of data values equal to the data value in question
n = total number of data

What is the percentile associated with the value "11"?

1, 2, 2, 3, 4, 4, 5, 6, 7, 8, 9, 10, 11, 11, 11, 12

$$\frac{12 + 0.5 \cdot 3}{16}(100) = 84.375$$

Round to the nearest integer: 84th percentile.

Example (From textbook)

EXAMPLE 2.19

On a timed math test, the first quartile for time it took to finish the exam was 35 minutes. Interpret the first quartile in the context of this situation.

Solution 2.19

- Twenty-five percent of students finished the exam in 35 minutes or less.
- Seventy-five percent of students finished the exam in 35 minutes or more.
- A low percentile could be considered good, as finishing more quickly on a timed exam is desirable. (If you take too long, you might not be able to finish.)

EXAMPLE 2.20

On a 20 question math test, the 70^{th} percentile for number of correct answers was 16. Interpret the 70^{th} percentile in the context of this situation.

Solution 2.20

- 70% of students answered 16 or fewer questions correctly.
- 30% of students answered 16 or more questions correctly.
- A higher percentile could be considered good, as answer more questions correctly is desirable.

Moving Quartiles to Boxplots

Boxplots provide a "picture" of the concentration of the data.

They demonstrate the overall **spread** of the data and show how far extreme values are from most of the data.

A boxplot is constructed with 5 values from a dataset:

1. The minimum value
2. The first quartile
3. The median
4. The third quartile
5. The maximum value

The smallest and largest data values mark the endpoints of the plot. The first and third quartiles mark the edges of the "box".

Roughly 50% of the data fall inside the box.

IT IS VERY IMPORTANT THAT BOXPLOTS ARE ADDED TO SCALED NUMBER LINES!!!

Example (From textbook)

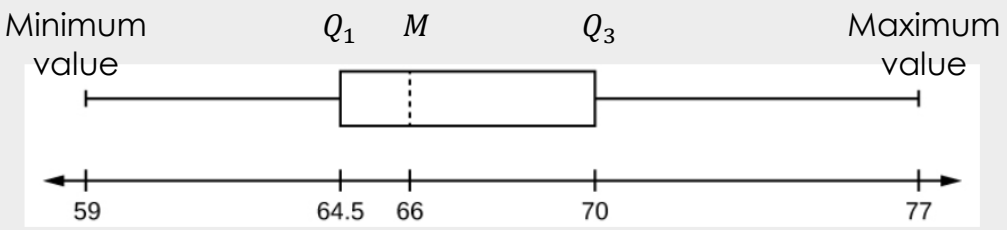
EXAMPLE 2.23

The following data are the heights of 40 students in a statistics class.

59; 60; 61; 62; 62; 63; 63; 64; 64; 64; 65; 65; 65; 65; 65; 65; 65; 65; 65; 66; 66; 67; 67; 68; 68; 69; 70; 70; 70; 70; 70; 71; 71; 72; 72; 73; 74; 74; 75; 77

Construct a box plot with the following properties; the calculator intructions for the minimum and maximum values as well as the quartiles follow the example.

- Minimum value = 59
- Maximum value = 77
- Q1: First quartile = 64.5
- Q2: Second quartile or median= 66
- Q3: Third quartile = 70

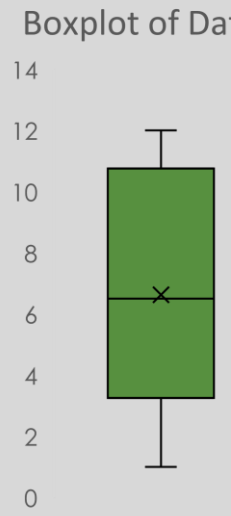


Example

What does the boxplot look like for the original data?

1, 2, 2, 3, 4, 4, 5, 6, 7, 8, 9, 10, 11, 11, 11, 12

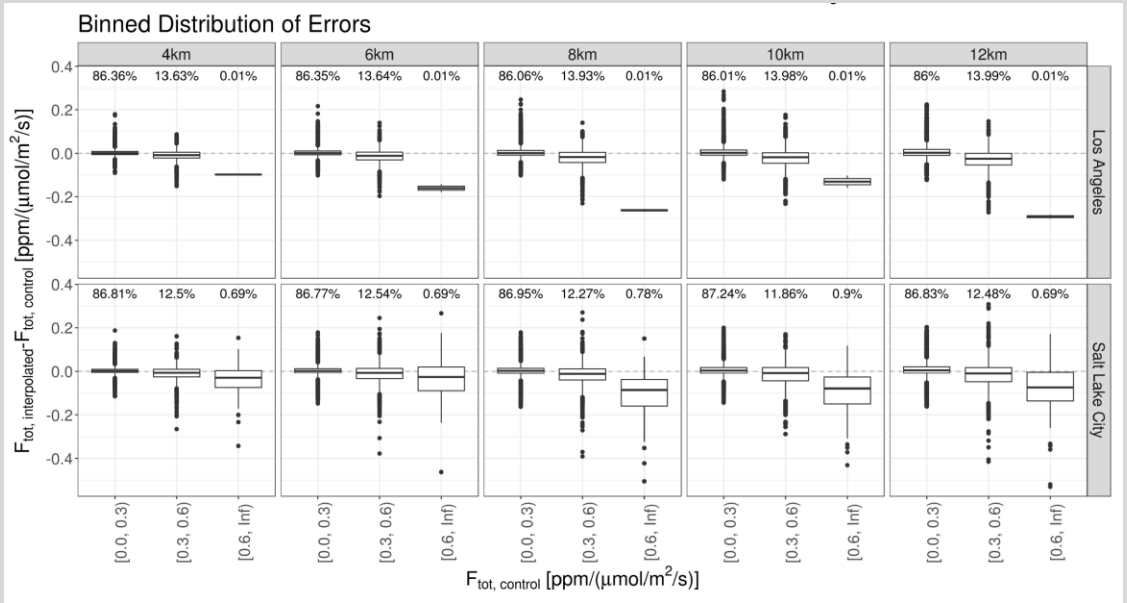
Boxplots can be vertical or horizontal



Some "whiskers" only extend to the outlier thresholds instead of the max/min values.

In the graph below, the whiskers extend to the $Q_3 + 1.5(\text{IQR})$ and $Q_1 - 1.5(\text{IQR})$ thresholds.

Outliers are shown as black points



A Quick Review

- **Quartiles** divide ordered data into quarters
- **Percentiles** divide ordered data into hundredths
- Median (M) – The median is the **center** of the data
- The first quartile, Q_1, is the middle value of the **lower half** of the data
- The third quartile, Q_3, is the middle value of the **upper half** of the data
- The **interquartile range** is a number that indicates the **spread** of the middle 50% of the data: $IQR = Q_3 - Q_1$
- The percentile associated with a specific value can be found by $\frac{x+0.5y}{n}(100)$
- Finding the value corresponding to the k^{th} percentile by $i = \frac{k}{100}(n+1)$
- Boxplots added to scaled number lines provide a "picture" of the spread of the data