Stat 201: Statistics I
Chapter 3

date

Chapter 3
Statistics for Describing, Exploring, and Combining Data

Section 3.1
Measures of Center

Measures of center

In order to understand a data set, values are calculated which summarize the distribution of the data or describe various properties of the data. These are, unsurprisingly, called **descriptive** or **summary** statistics.

Measures of center

In order to understand a data set, values are calculated which summarize the distribution of the data or describe various properties of the data. These are, unsurprisingly, called **descriptive** or **summary** statistics.

Perhaps the most important of these, **measures of center** are a way of representing the value of the middle of the data.

There are four measures of center discussed in this section:

- mean
- median
- mode
- midrange

Mean

The **mean** (the arithmetic mean) is the measure of center calculated by adding the values of the data set and dividing by the size of the data set. Also known as the average.

- Only makes sense with quantitative data
- Sensitive to outliers (extreme or unusual values) and skewed data distributions.

Mean

The **mean** (the arithmetic mean) is the measure of center calculated by adding the values of the data set and dividing by the size of the data set. Also known as the average.

- Only makes sense with quantitative data
- Sensitive to outliers (extreme or unusual values) and skewed data distributions.

To calculate

Let X be sample of size n of quantitative data with values $x_1, \ldots, x_n$. Then,

$$\bar{x} = \frac{\sum x_i}{n}$$

$\bar{x}$ is the mean of the sample.

- $\sum$ means add all the x_i's, where i is between 1 and n

Mean, example

Example

Suppose we find a sample of 8 students and ask for their ages.

- The sample is $X = \{22, 32, 46, 50, 33, 38, 20, 24\}$

Mean, example

Example

Suppose we find a sample of 8 students and ask for their ages.

- The sample is $X = \{22, 32, 46, 50, 33, 38, 20, 24\}$
- The sample size is $n = 8$

Mean, example

Example

Suppose we find a sample of 8 students and ask for their ages.

- The sample is $X = \{22, 32, 46, 50, 33, 38, 20, 24\}$
- The sample size is $n = 8$
- The sum of the data is

$$\sum x_i = 22 + 32 + 46 + 50 + 33 + 38 + 20 + 24 = 265$$

Mean, example

Example

Suppose we find a sample of 8 students and ask for their ages.

- The sample is $X = \{22, 32, 46, 50, 33, 38, 20, 24\}$
- The sample size is $n = 8$
- The sum of the data is

$$\sum x_i = 22 + 32 + 46 + 50 + 33 + 38 + 20 + 24 = 265$$

- The mean is

$$\bar{x} = \frac{\sum x_i}{n} = \frac{265}{8} = 33.125 \text{ years}$$

Median

The **median** is the value that is greater than or equal to at least 50% of the data and less than or equal to at least 50% of the data.

- Can be used with quantitative and ordinal data
- Not sensitive to extreme values (**resistant** measure of center)

Median

The **median** is the value that is greater than or equal to at least 50% of the data and less than or equal to at least 50% of the data.

- Can be used with quantitative and ordinal data
- Not sensitive to extreme values (**resistant** measure of center)

To calculate

Arrange the data in order, from lowest to highest.

- If n is odd, the middle value is the median.
- If n is even, the median is the mean of the two middle values.

Median, example

Example

Returning to the ages of 8 students.

- The sample is $X = \{22, 32, 46, 50, 33, 38, 20, 24\}$

Median, example

Example

Returning to the ages of 8 students.

- The sample is $X = \{22, 32, 46, 50, 33, 38, 20, 24\}$
- Arranged in order, the sample looks like

$$20 \quad 22 \quad 24 \quad 32 \quad 33 \quad 38 \quad 46 \quad 50$$

Median, example

Example

Returning to the ages of 8 students.

- The sample is $X = \{22, 32, 46, 50, 33, 38, 20, 24\}$
- Arranged in order, the sample looks like

$$20 \quad 22 \quad 24 \quad 32 \quad 33 \quad 38 \quad 46 \quad 50$$

- Since n is even, find the the mean of the two middle values.

$$20 \quad 22 \quad 24 \quad \underbrace{32 \quad 33}_{(32+33)/2=32.5} \quad 38 \quad 46 \quad 50$$

Median, example

Example

Returning to the ages of 8 students.

- The sample is $X = \{22, 32, 46, 50, 33, 38, 20, 24\}$
- Arranged in order, the sample looks like

$$20 \quad 22 \quad 24 \quad 32 \quad 33 \quad 38 \quad 46 \quad 50$$

- Since n is even, find the the mean of the two middle values.

$$20 \quad 22 \quad 24 \quad \underbrace{32 \quad 33}_{(32+33)/2=32.5} \quad 38 \quad 46 \quad 50$$

- The median is $\tilde{x} = 32.5$ years.

Mean vs. Median

Suppose in our age data set, we replaced the 50 with a 85.

- Mean goes from 33.125 to 37.5
- Median remains unchanged at 32.5

Mean vs. Median

Suppose in our age data set, we replaced the 50 with a 85.

- Mean goes from 33.125 to 37.5
- Median remains unchanged at 32.5

This is why median is called a **resistant** statistic.

- Median is used when we don't want a few extreme values to distort a more reasonable middle, such as house prices or incomes.

Mean vs. Median, cont.

Suppose instead of calculating a grade point average (mean), we calculated a grade point median. Consider a student who got A's in 3 classes and D's in 2.

- The median grade point is 4, an A.
- The GPA for such a student would be 2.8.

Mean vs. Median, cont.

Suppose instead of calculating a grade point average (mean), we calculated a grade point median. Consider a student who got A's in 3 classes and D's in 2.

- The median grade point is 4, an A.
- The GPA for such a student would be 2.8.

The median does not consider all values of a data set. The mean does.

- Mean is used when all values are important or when we expect to have roughly symmetric data.

Mode

The **mode** is the data value with highest frequency.

- Can be used with any kind of data.
- A data set might have more than one mode, or there might not be any mode.

Mode, example

Examples

- The age data, $\{22, 32, 46, 50, 33, 38, 20, 24\}$, has no mode.

Mode, example

Examples

- The age data, $\{22, 32, 46, 50, 33, 38, 20, 24\}$, has no mode.
- From a TACO survey, favorite kind of taco had these responses:

  {Beef, Beef, Fish, Shrimp, Beef, Pork, Chicken, Beef, Chicken, Beef}

  The mode is "Beef" with a frequency of five.

Mode, example

Examples

- The age data, $\{22, 32, 46, 50, 33, 38, 20, 24\}$, has no mode.
- From a TACO survey, favorite kind of taco had these responses:

  {Beef, Beef, Fish, Shrimp, Beef, Pork, Chicken, Beef, Chicken, Beef}

  The mode is "Beef" with a frequency of five.
- Suppose a sample from a class got the following grades on a quiz:

  {A, C, B, A, A, B, C, B}

  The modes are A and B, with frequencies of three each.

Midrange

The **midrange** is the value half way between the minimum and maximum values. Calculate by finding the mean of the minimum and maximum.

- Only makes sense with quantitative data.
- *Very* sensitive the extreme values.
- Easy to calculate, but rarely used.

Midrange

The **midrange** is the value half way between the minimum and maximum values. Calculate by finding the mean of the minimum and maximum.

- Only makes sense with quantitative data.
- *Very* sensitive the extreme values.
- Easy to calculate, but rarely used.

Example

The age data is $X = \{22, 32, 46, 50, 33, 38, 20, 24\}$.

- The minimum age is 20 and the maximum age is 50.
- The midrange is

$$\frac{\min(X) + \max(X)}{2} = \frac{20 + 50}{2} = 35$$

Measures of center in StatCrunch

- Stat $\rightarrow$ Summary Stats $\rightarrow$ Columns
- Select the column (or columns) which contains the data
- Under "Statistics" select statistics to calculate (ctrl-click to select multiple statistics)
- For measures of center select "Mean:, "Median", "Min", "Max" and "Mode"
- Click "Compute!"
- Mean, median and mode will be displayed
- Midrange must be calculated by hand using min and max

Note

Mode will display "No mode", "Multiple modes" or the mode if exactly one exists. If there are multiple modes, you must identify them yourself.

Group work

- For each question, complete part (a).

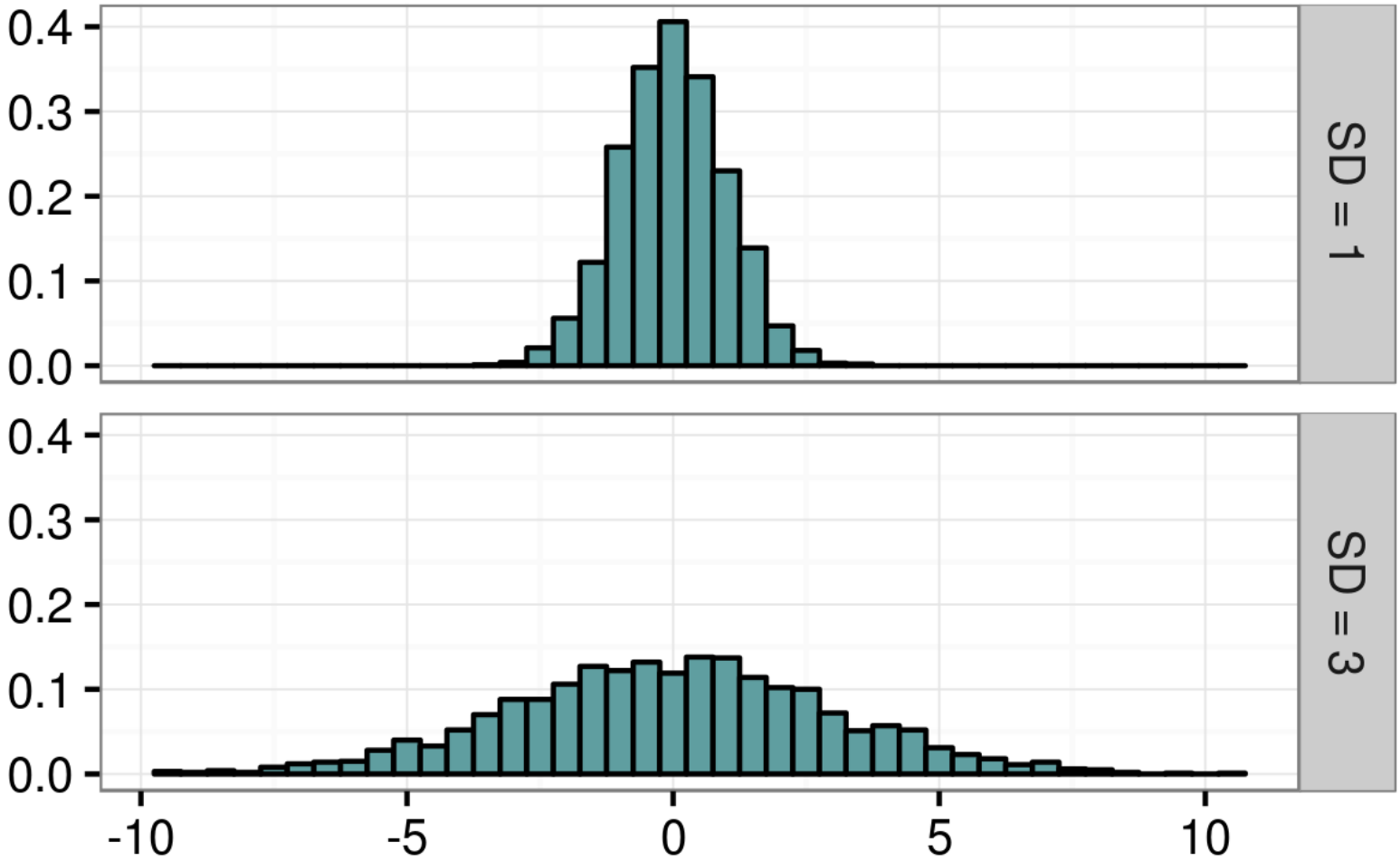
Section 3.2
Measures of Variation

Other measures

Other measures are necessary to adequately describe data.

Other measures

Other measures are necessary to adequately describe data.



Measures of variation

Another important class of descriptive statistics are **measures of variation** which describe how much the data is spread out.

There are four measures of variation discussed in this section:

- Range
- Variance
- Standard deviation

Range

The **range** is the difference between the maximum and minimum values.

- Like the midrange, very sensitive to extreme values.

Range

The **range** is the difference between the maximum and minimum values.

- Like the midrange, very sensitive to extreme values.

Example

The age data is $X = \{22, 32, 46, 50, 33, 38, 20, 24\}$.

- The minimum age is 20 and the maximum age is 50.

Range

The **range** is the difference between the maximum and minimum values.

- Like the midrange, very sensitive to extreme values.

Example

The age data is $X = \{22, 32, 46, 50, 33, 38, 20, 24\}$.

- The minimum age is 20 and the maximum age is 50.
- The range is

$$\max(X) - \min(X) = 50 - 20 = 30 \text{ years}$$

Variance and standard deviation

The **variance** is the mean of the squared difference of the data from the mean. The **standard deviation** is the square root of the variance.

Variance and standard deviation

The **variance** is the mean of the squared difference of the data from the mean. The **standard deviation** is the square root of the variance.

- More simply, the standard deviation is the average distance of the data from the data mean (the center).

Variance and standard deviation

The **variance** is the mean of the squared difference of the data from the mean. The **standard deviation** is the square root of the variance.

- More simply, the standard deviation is the average distance of the data from the data mean (the center).
- Always non-negative. A zero standard deviation means all the data are the same value.

Variance and standard deviation

The **variance** is the mean of the squared difference of the data from the mean. The **standard deviation** is the square root of the variance.

- More simply, the standard deviation is the average distance of the data from the data mean (the center).
- Always non-negative. A zero standard deviation means all the data are the same value.
- Sensitive to extreme values.

Variance and standard deviation

The **variance** is the mean of the squared difference of the data from the mean. The **standard deviation** is the square root of the variance.

- More simply, the standard deviation is the average distance of the data from the data mean (the center).
- Always non-negative. A zero standard deviation means all the data are the same value.
- Sensitive to extreme values.
- The units of standard deviation are the same as the data. Variance units are the data units squared.

Variance and standard deviation, calculation

To calculate

Let X be sample of size n of quantitative data with values $x_1, \ldots, x_n$ and sample mean $\bar{x}$. Then,

$$\mathrm{Var}\,(X) = s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \qquad \text{and} \qquad \mathrm{SD}\,(X) = s = \sqrt{s^2}$$

Variance and standard deviation, calculation

To calculate

Let X be sample of size n of quantitative data with values $x_1, \ldots, x_n$ and sample mean $\bar{x}$. Then,

$$\mathrm{Var}\,(X) = s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \qquad \text{and} \qquad \mathrm{SD}\,(X) = s = \sqrt{s^2}$$

- Note: Never calculate this by hand. Use technology.

Variance and standard deviation, example

Example

The age data is $X = \{22, 32, 46, 50, 33, 38, 20, 24\}$. The sample size is $n = 8$ and the sample mean is $\bar{x} = 33.125$

Variance and standard deviation, example

Example

The age data is $X = \{22, 32, 46, 50, 33, 38, 20, 24\}$. The sample size is $n = 8$ and the sample mean is $\bar{x} = 33.125$

- The variance is

$$\begin{aligned} s^2 &= \frac{\sum (x_i - \bar{x})^2}{n-1} \\ &= \frac{(22 - 33.125)^2 + \cdots + (24 - 33.125)^2}{7} \\ &= 122.125 \text{ years}^2 \end{aligned}$$

Variance and standard deviation, example

Example

The age data is $X = \{22, 32, 46, 50, 33, 38, 20, 24\}$. The sample size is $n = 8$ and the sample mean is $\bar{x} = 33.125$

- The variance is

$$\begin{aligned} s^2 &= \frac{\sum (x_i - \bar{x})^2}{n-1} \\ &= \frac{(22 - 33.125)^2 + \cdots + (24 - 33.125)^2}{7} \\ &= 122.125 \text{ years}^2 \end{aligned}$$

- The standard deviation is

$$s = \sqrt{s^2} = \sqrt{122.125} = 11.05 \text{ years}$$

Notation

Recall, values that describe the properties of populations are called **parameters** and values that describe samples are called **statistics**. Notationally, in math formulas or when abbreviating, Greek letters are used to refer to parameters and Latin letters are used to refer to statistics.

Notation

Recall, values that describe the properties of populations are called **parameters** and values that describe samples are called **statistics**. Notationally, in math formulas or when abbreviating, Greek letters are used to refer to parameters and Latin letters are used to refer to statistics.

Property	Parameter	Statistic
Mean	μ (mu)	$\bar{x}$
Variance	σ^2 (sigma-squared)	s^2
Standard deviation	σ (sigma)	s

Measures of variation in StatCrunch

- Stat $\rightarrow$ Summary Stats $\rightarrow$ Columns
- Select the column (or columns) which contains the data
- Under "Statistics" select statistics to calculate (ctrl-click to select multiple statistics)
- For measures of variation select "Variance", "Std. Dev." and "Range"
- Click "Compute!"
- The statistics will be displayed

Group work

- For each question, complete part (b).

Section 3.3
Measures of Relative Standing and Boxplots

Measures of relative standing

Measures of relative standing describe the location of a given data value with a data distribution or a data set.

Two measures of relative standing are discussed in this section:

- Z-scores
- Percentiles

Z-scores

A **z-score** describes the relative position of a data value within a data distribution.

Z-scores

A **z-score** describes the relative position of a data value within a data distribution.

- Another way to put it is a z-score is the number of standard deviations that a particular value is above or below the mean.

Z-scores

A **z-score** describes the relative position of a data value within a data distribution.

- Another way to put it is a z-score is the number of standard deviations that a particular value is above or below the mean.
- Z-scores are standardized and unit-less, so they can be used to compare values from different populations.

Z-scores

A **z-score** describes the relative position of a data value within a data distribution.

- Another way to put it is a z-score is the number of standard deviations that a particular value is above or below the mean.
- Z-scores are standardized and unit-less, so they can be used to compare values from different populations.
- A positive z-score means the value is greater than the mean and a negative z-score means that it is below the mean.

Z-scores

A **z-score** describes the relative position of a data value within a data distribution.

- Another way to put it is a z-score is the number of standard deviations that a particular value is above or below the mean.
- Z-scores are standardized and unit-less, so they can be used to compare values from different populations.
- A positive z-score means the value is greater than the mean and a negative z-score means that it is below the mean.
- Z-scores can be calculated for samples or populations, if the population mean and standard deviation are known.

Z-scores, calculations

To calculate

- For a sample X with sample mean $\bar{x}$ and standard deviation s, the z-score for a value x is

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

Z-scores, calculations

To calculate

- For a sample X with sample mean $\bar{x}$ and standard deviation s, the z-score for a value x is
$$z = \frac{x - \bar{x}}{s}$$

- For a population with population mean μ and standard deviation σ, the z-score for value x is
$$z = \frac{x - \mu}{\sigma}$$

Z-scores, example

Example

The age data has mean of $\bar{x} = 33.125$ and SD of $s = 11.05$.

Z-scores, example

Example

The age data has mean of $\bar{x} = 33.125$ and SD of $s = 11.05$.

- Suppose a new student joins the class. His age is 61. He has an age z-score of

$$z = \frac{x - \bar{x}}{s} = \frac{61 - 33.125}{11.05} = \frac{27.875}{11.05} = 2.52$$

Z-scores, example

Example

The age data has mean of $\bar{x} = 33.125$ and SD of $s = 11.05$.

- Suppose a new student joins the class. His age is 61. He has an age z-score of

$$z = \frac{x - \bar{x}}{s} = \frac{61 - 33.125}{11.05} = \frac{27.875}{11.05} = 2.52$$

His age two and a half standard deviations above the class mean.

Z-scores, example

Example

The age data has mean of $\bar{x} = 33.125$ and SD of $s = 11.05$.

- Suppose a new student joins the class. His age is 61. He has an age z-score of

$$z = \frac{x - \bar{x}}{s} = \frac{61 - 33.125}{11.05} = \frac{27.875}{11.05} = 2.52$$

  His age two and a half standard deviations above the class mean.

- Another student joins the class. Her age is 27. She has an age z-score of

$$z = \frac{x - \bar{x}}{s} = \frac{27 - 33.125}{11.05} = \frac{-6.125}{11.05} = -0.554$$

Z-scores, example

Example

The age data has mean of $\bar{x} = 33.125$ and SD of $s = 11.05$.

- Suppose a new student joins the class. His age is 61. He has an age z-score of

$$z = \frac{x - \bar{x}}{s} = \frac{61 - 33.125}{11.05} = \frac{27.875}{11.05} = 2.52$$

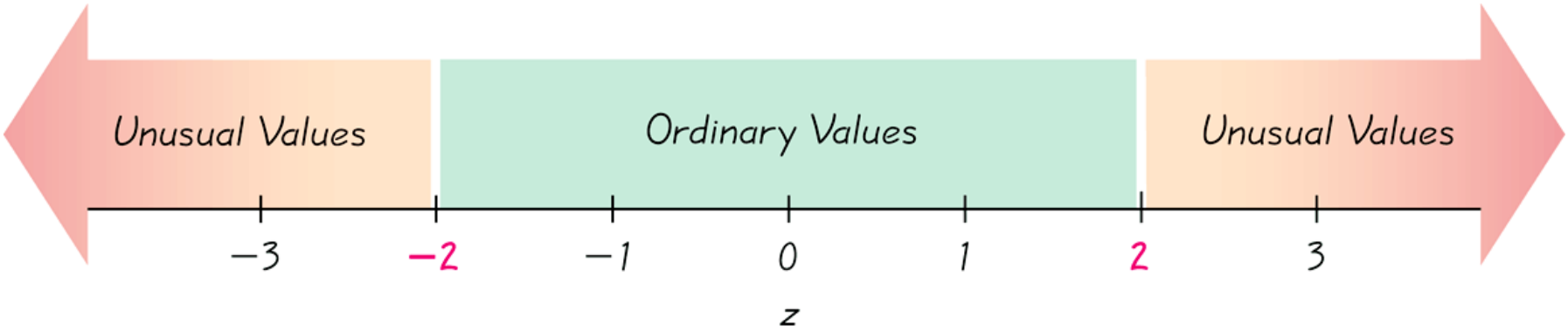
  His age two and a half standard deviations above the class mean.

- Another student joins the class. Her age is 27. She has an age z-score of

$$z = \frac{x - \bar{x}}{s} = \frac{27 - 33.125}{11.05} = \frac{-6.125}{11.05} = -0.554$$

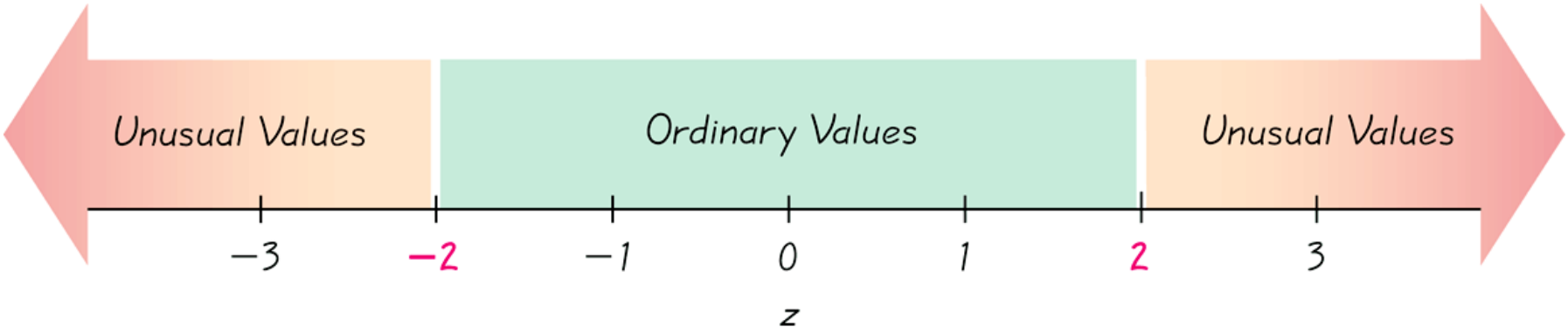
  Her age is about a half standard deviation below the class mean.

Unusual values



A value is called **unusual** if it has a z-score z such that $z < -2$ or $z > 2$. A value is **ordinary** if z is between -2 and 2.

Unusual values



A value is called **unusual** if it has a z-score z such that $z < -2$ or $z > 2$. A value is **ordinary** if z is between -2 and 2.

Example

Consider the new students to the class:

- The 61 year old ($z = 2.52$) has an unusual age for the class.
- The 27 year old ($z = -0.554$) has an ordinary age for the class.

Percentiles

Percentiles measure relative position within a data set as order rank. In other words, the value at the pth percentile (written as P_p) in a data set is greater than $p\%$ of the data.

Percentiles

Percentiles measure relative position within a data set as order rank. In other words, the value at the pth percentile (written as P_p) in a data set is greater than $p\%$ of the data.

To calculate

- To find the percentile of a value x in a data set,

$$\%\text{ile} = \frac{\text{number of values} < x}{n} \times 100\%$$

Percentiles

Percentiles measure relative position within a data set as order rank. In other words, the value at the pth percentile (written as P_p) in a data set is greater than $p\%$ of the data.

To calculate

- To find the percentile of a value x in a data set,

$$\%\text{ile} = \frac{\text{number of values} < x}{n} \times 100\%$$

- To find the value of P_p (the pth percentile), calculate the rank,

$$r = \frac{p}{100} \times n$$

If r is a whole number, P_p is the mean of the rth and $(r+1)$th values. If not, round up. Then, P_p is the rth value in an ordered list.

Percentile, example

Example

The age data, in order is,

$$20 \quad 22 \quad 24 \quad 32 \quad 33 \quad 38 \quad 46 \quad 50$$

Percentile, example

Example

The age data, in order is,

$$20 \quad 22 \quad 24 \quad 32 \quad 33 \quad 38 \quad 46 \quad 50$$

- The percentile of the value 38 is

$$\frac{\text{number of values} < x}{n} \times 100\% = \frac{5}{8} \times 100\% = 62.5\,\% \Rightarrow P_{63}$$

Percentile, example

Example

The age data, in order is,

$$20 \quad 22 \quad 24 \quad 32 \quad 33 \quad 38 \quad 46 \quad 50$$

- The percentile of the value 38 is

$$\frac{\text{number of values} < x}{n} \times 100\% = \frac{5}{8} \times 100\% = 62.5\,\% \Rightarrow P_{63}$$

- To find the 30th percentile, P_{30}, calculate rank

$$r = \frac{p}{100} \times n = \frac{30}{100} \times 8 = 2.4$$

Round up r to 3. P_{30} is the 3rd value, 24.

Percentiles in StatCrunch

To find percentile of value x:

- If data is not ordered, order it using Data $\rightarrow$ Sort
- The row number of the value **before** x is number of values $< x$
- The row number of the last value is n
- Use the formula to calculate percentile (round down)

To find value for percentile p:

- Stat $\rightarrow$ Summary Stats $\rightarrow$ Columns
- Select the column (or columns) which contains the data
- Under "Percentiles..." enter p (can enter multiple p's separated by commas)
- Click "Compute!"
- The value(s) will be displayed as "pth Per."

Quartiles

The **quartiles** are values that divide the data set into 4 parts, or quarters.

$$Q_1 = P_{25} \qquad Q_2 = P_{50} \qquad Q_3 = P_{75}$$

Quartiles

The **quartiles** are values that divide the data set into 4 parts, or quarters.

$$Q_1 = P_{25} \qquad Q_2 = P_{50} \qquad Q_3 = P_{75}$$

- Note: The median is equivalent to Q_2 and P_{50}.

5 number summary

The **5 number summary** summarize the distribution of a data set.

The 5 numbers are:

- Minimum
- Q_1
- Median (or Q_2)
- Q_3
- Maximum

5 number summary, example

Example

The age data, in order is,

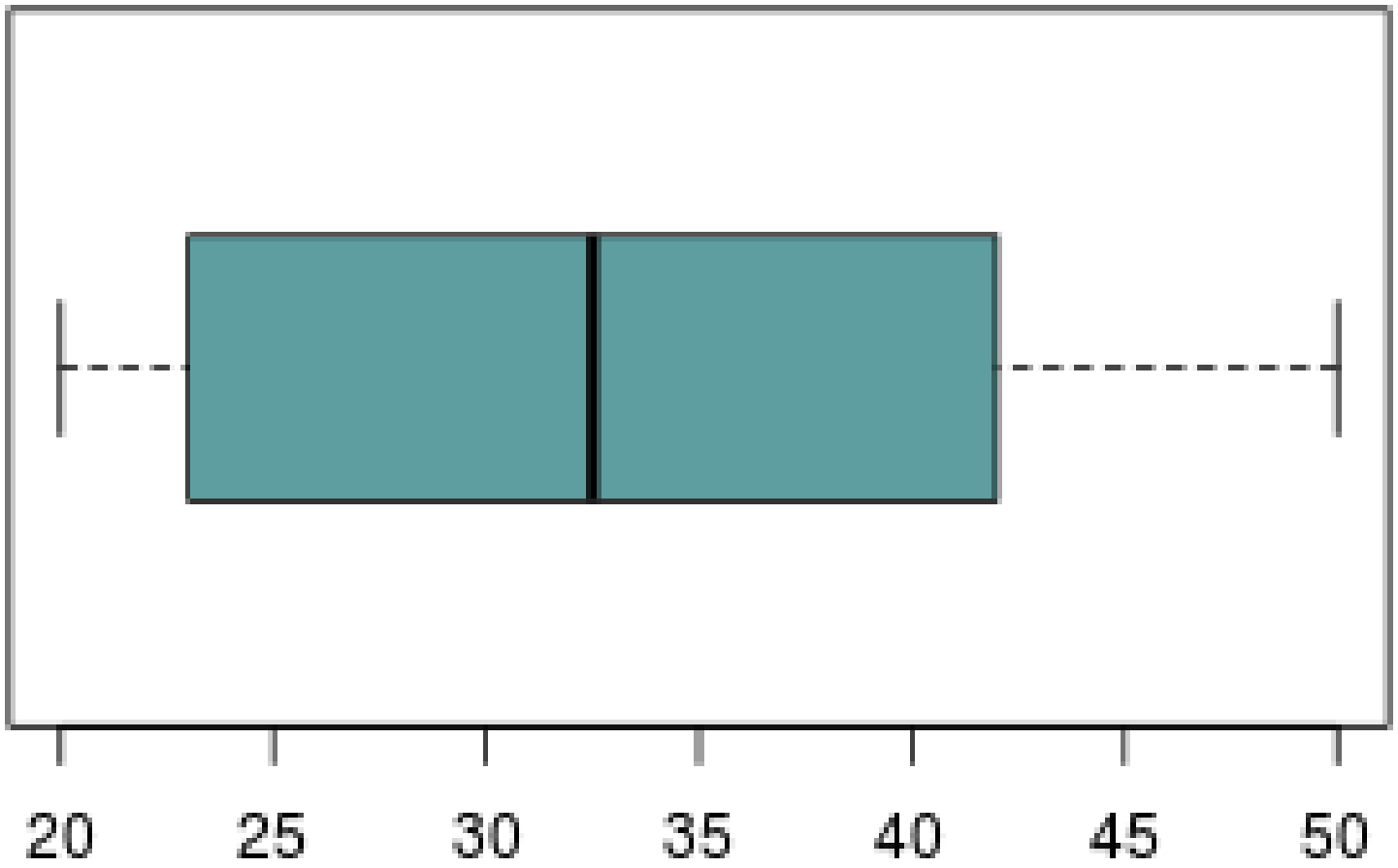
$$20 \quad 22 \quad 24 \quad 32 \quad 33 \quad 38 \quad 46 \quad 50$$

The 5 number summary is

$$\underbrace{20}_{\min} \quad \underbrace{23}_{Q_1} \quad \underbrace{32.5}_{\text{med}} \quad \underbrace{42}_{Q_3} \quad \underbrace{50}_{\max}$$

Boxplots

A **boxplot** is a graph depicting the 5 number summary.



Five number summaries in StatCrunch

- Stat $\rightarrow$ Summary Stats $\rightarrow$ Columns
- Select the column (or columns) which contains the data
- Under "Statistics" select "Min", "Q1", "Median", "Q3" and "Max"
- Or, under "Percentile" enter "1, 25, 50, 75, 99" (might not give correct values for the min and max in large data sets)
- Click "Compute!"
- The values will be displayed, but maybe not in the right order

Boxplots in StatCrunch

- Graph $\rightarrow$ Boxplot
- Select the column (or columns) which contains the data
- (Optional) Under "Other options", click "Draw boxes horizontally"
- Click "Compute!"
- Hold pointer over plot to get IQR (inter-quartile range) and five number summary

Group work

- For each question, complete part (c).