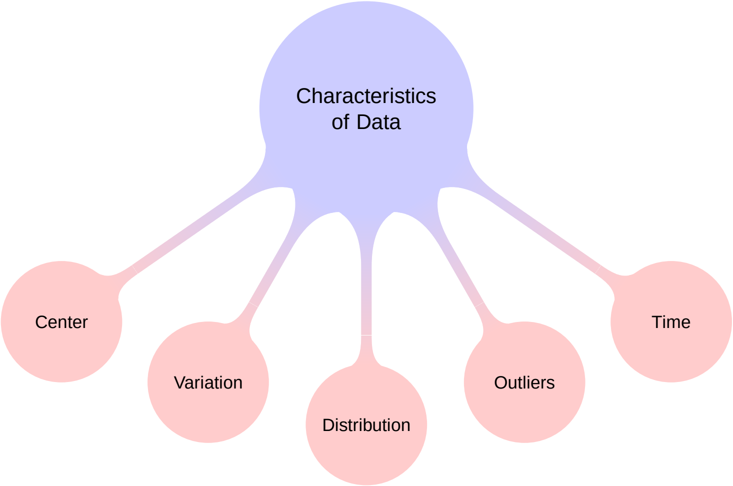
Chapter 3

Professor Tim Busken

Mathematics Department

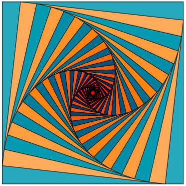
February 18, 2014

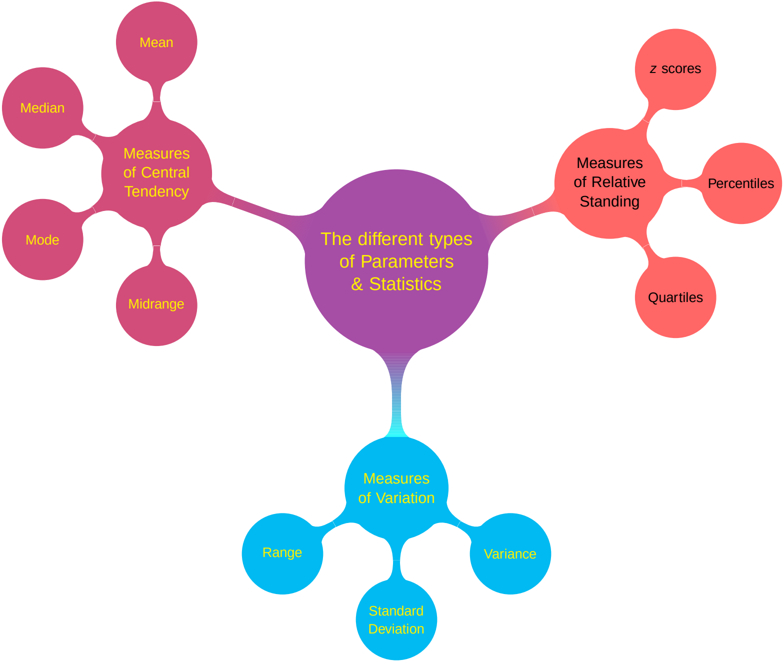
Table of Contents


Characteristics of Data
Center
Variation
Distribution
Outliers
Time

Characteristics of Data [1]

1. **Center:** A representative or average value that indicates where the middle of the data set is located.
2. **Variation:** A measure of the amount that the data values vary.
3. **Distribution:** The nature or shape of the spread of data over the range of values (such as bell-shaped, uniform, or skewed).
4. **Outliers:** Sample values that lie very far away from the vast majority of other sample values.
5. **Time:** Changing characteristics of the data over time.




The different types of Parameters & Statistics
Measures of Central Tendency
Mean
Median
Mode
Midrange
Measures of Relative Standing
z scores
Percentiles
Quartiles
Measures of Variation
Range
Standard Deviation
Variance

Notation

$\sum$ denotes the sum of a set of values.

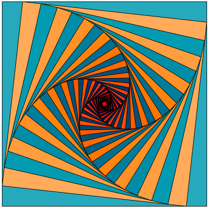
x is the variable usually used to represent the individual data values.

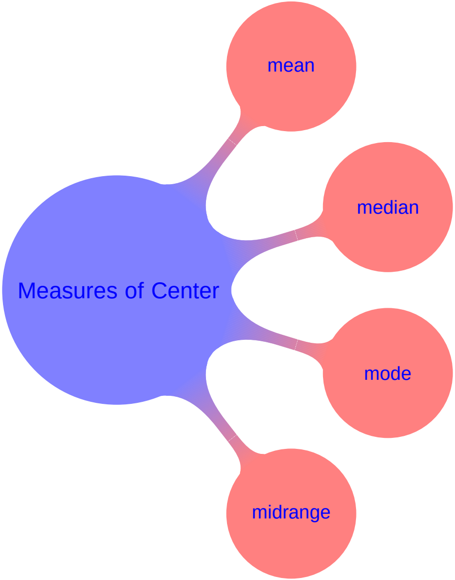
n represents the number of data values in a sample.

N represents the number of data values in a population.

$\bar{x}$ the symbol that represents the sample mean.

μ the symbol that represents the population mean

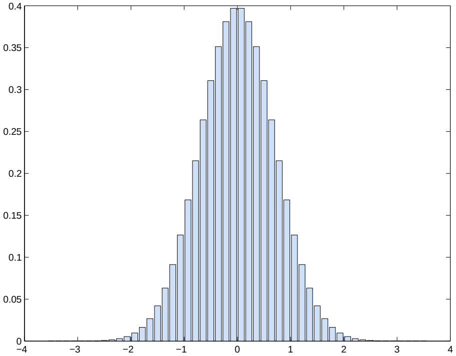




These are Statistics and Parameters!

Definition

A **Measure of Center** is a value at the center or middle of a data set.[1]



Definition

The **mean** (average) is the value obtained by adding all of the data values and dividing the total by the number of values.

Definition

The **median** is the middle value when the original data values are arranged in order of increasing (or decreasing) magnitude

Definition

The **mode** is the value that occurs with the greatest frequency.

Definition

The **midrange** is the value midway between the maximum and minimum values in the original data set. [1]

$$\text{midrange} = \frac{\text{max. value} + \text{min. value}}{2}$$

classroom worksheet

KEY.

Mode

A data set can have one mode, more than one mode, or no mode.

Definition

Whenever two data values occur with the same greatest frequency, we say the data is **bimodal**.

Definition

Whenever more than two data values occur with the same greatest frequency, we say the data is **multimodal**.

Definition

Whenever no data value is repeated, there is **no mode**.

Median

Example: What is the median of the following data set?

21 85 15 43 75 12

Median

Example: What is the median of the following data set?

21 85 15 43 75 12

We begin answering the question by sorting the data in a ascending fashion:

12 15 21 43 75 85

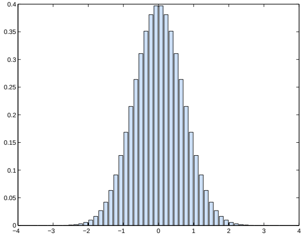
Since the number of data entries is even, there is no single data entry representing the median. Instead, we take the median to be the midpoint between the two middle numbers then divide by 2.

$$\text{median} = \frac{21 + 43}{2} = 32$$

Finding the mean from a Distribution

Suppose you are presented with a frequency distribution table related to a particular data set, but not with the actual data set. It is possible to compute a good approximation of the average, $\bar{x}$, with the following formula:

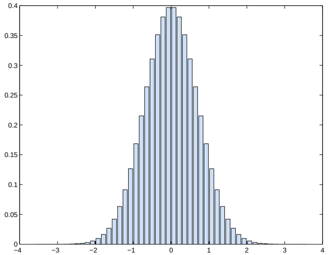
$$\bar{x} = \frac{\sum(f \cdot x)}{\sum f}$$



Definition

When data values are assigned different weights, *w*, then we can compute a **weighted mean**, given by the formula

$$\bar{x} = \frac{\sum (w \cdot x)}{\sum w}$$



Measures of Center: Advantages and Disadvantages

	Advantages	Disadvantages
Mean	Is relatively reliable, means of samples drawn from the same population dont vary as much as other measures of center. Takes every data value into account	Is sensitive to every data value, one extreme value can affect it dramatically; is not a resistant measure of center.
Median	is not affected by an extreme value - is a resistant measure of the center	Doesn't always reflect the true center
Mode	is fairly easy to find	Doesn't always reflect the true center often a data set has no mode
Midrange	very easy to compute reinforces that there are several ways to define the center	Sensitive to extremes because it uses only the maximum and minimum values, so rarely used

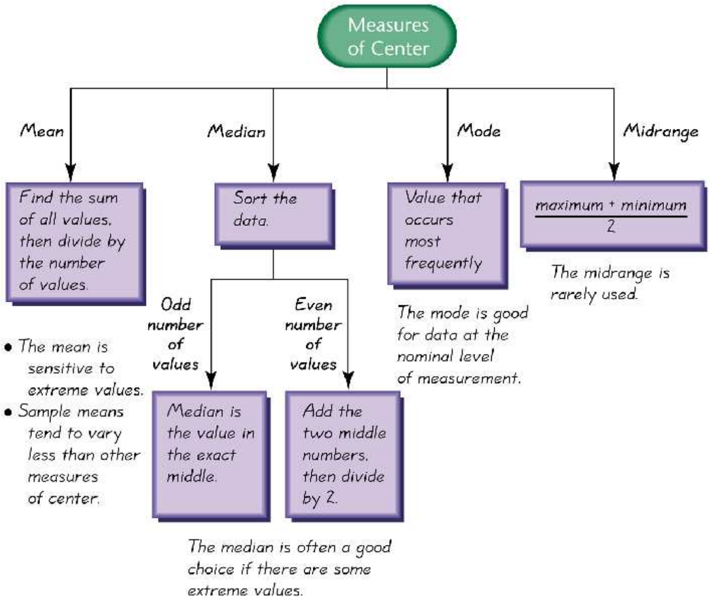
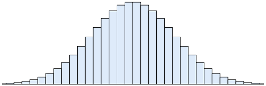


Figure : Triola [1] Flowchart.

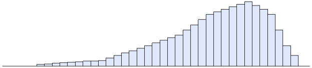
Skewness

Definition

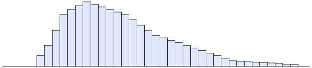
A distribution of data is **skewed** if it is not symmetric and extends more to one side than to the other. (A distribution of data is symmetric if the left half of its histogram is roughly a mirror image of its right half.) [1]



(b) Symmetric Distribution



(a) Skewed to the Left



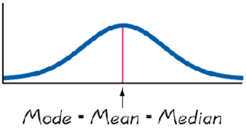
(c) Skewed to the Right

The distribution in (a) is called "skewed left" because most of the data falls to the left of the mode (the value along the x-axis associated with the largest bar in the histogram). The distribution in (c) is called "skewed right" because most of the data falls to the right of the mode.

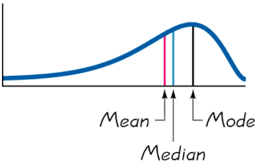
Skewness

Definition

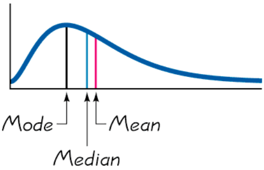
A distribution of data is **skewed** if it is not symmetric and extends more to one side than to the other. (A distribution of data is symmetric if the left half of its histogram is roughly a mirror image of its right half.) [1]



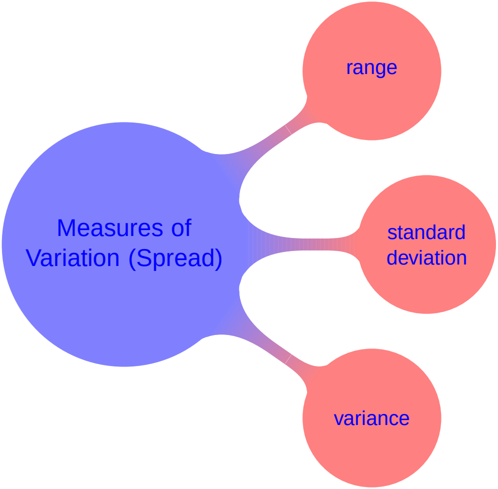
(b) Symmetric

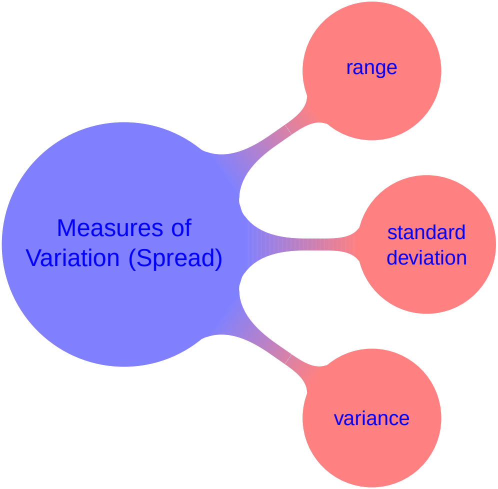


(a) Skewed to the Left (Negatively)



(c) Skewed to the Right (Positively)


Measures of Variation (Spread)
range
standard deviation
variance



These are Statistics and Parameters!

Measures of Spread

Definition

The **range** of a set of data values is the difference between the maximum data value and the minimum data value.

Range $=$ (maximum value) – (minimum value)

Definition

The **standard deviation** of a set of sample values, denoted by s, is a measure of variation of values around the mean.

Definition

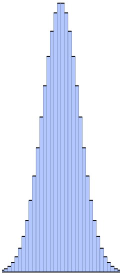
The **variance** of a set of values is a measure of variation equal to the square of the standard deviation.

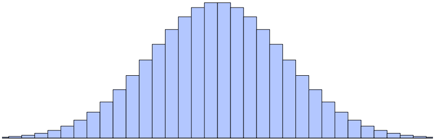
classroom worksheet

KEY

Standard Deviation

You can think of the standard deviation of a data set as being the average distance between any two consecutive data points (along the x-axis).





A low standard deviation (left figure) indicates that the data points tend to be very close to the mean; high standard deviation (right figure) indicates that the data points are spread out over a large range of measurement (x) values.

Standard Deviation

	Formula
Sample Standard Deviation	$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n-1}}$
Population Standard Deviation	$\sigma = \sqrt{\frac{\sum(x - \mu)^2}{N}}$

Standard Deviation

	Formula
Sample Standard Deviation	$s = \sqrt{\dfrac{\sum(x-\bar{x})^2}{n-1}}$
Population Standard Deviation	$\sigma = \sqrt{\dfrac{\sum(x-\mu)^2}{N}}$

Notation:

s symbol used for sample standard deviation

σ symbol used for population standard deviation

s^2 symbol used for sample variance

σ^2 symbol used for population variance

Standard Deviation

Shortcut Formula

Sample Standard Deviation $s = \sqrt{\dfrac{n\sum(x^2) - (\sum x)^2}{n(n-1)}}$

Notation:

$(\sum x)^2$ Make a list (column) of x (data entry) values. Sum these x values. Afterwards, square this sum to get the value of $(\sum x)^2$

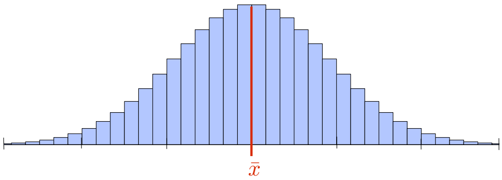
$\sum(x^2)$ Make a list (column) of x^2 values. $\sum(x^2)$ is the sum these x^2 values.

Standard Deviation Properties

- The units of the standard deviation are the same as the units of the original data values.
- The standard deviation is sensitive to outliers—meaning that extreme values (unusually low or high data entries) significantly contribute to the value of the standard deviation.
- The value of the standard deviation is usually positive.

The Empirical Rule

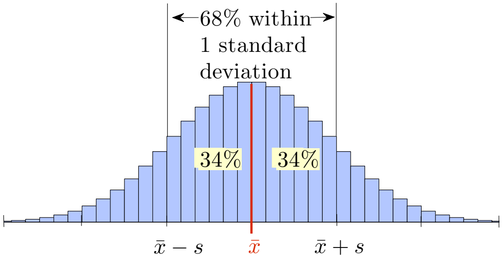
A data set is **normally distributed** when it's associated histogram has a bell shape. The Empirical Rule is a rule which holds only for data sets that follow a normal distribution.



The Empirical Rule

For normally distributed data, the following properties apply:

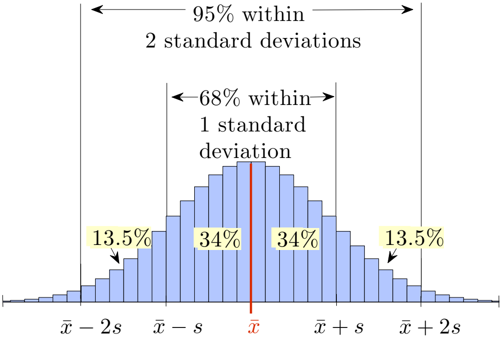
- About 68% of all values fall within 1 standard deviation of the mean.



The Empirical Rule

For normally distributed data, the following properties apply:

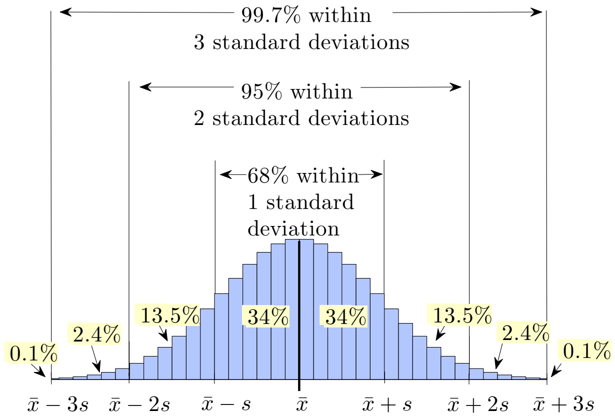
- About 95% of all values fall within 2 standard deviations of the mean.



The Empirical Rule

For normally distributed data, the following properties apply:

- About 99.7% of all values fall within 3 standard deviations of the mean.



Chebyshev's Theorem

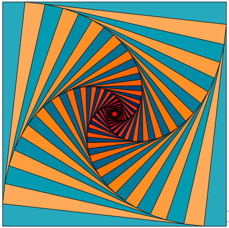
Chebyshev's Theorem is a rule similar to the empirical rule. However, it can be applied to any distribution.

Chebyshev's Theorem

Theorem (Chebyshev's Theorem)

The proportion (or fraction) of any set of data lying within K standard deviations of the mean is always at least $1 - 1/K^2$, *where K is any positive number greater than 1.*

- *For K = 2, at least 3/4 (or 75%) of all values lie within 2 standard deviations of the mean.*
- *For K = 3, at least 8/9 (or 89%) of all values lie within 3 standard deviations of the mean.*

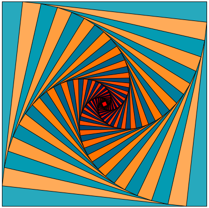


Range Rule of Thumb

To roughly estimate the standard deviation from a collection of known sample data use

$$s \approx \frac{\text{range}}{4}$$

where range = (maximum value) − (minimum value)



Range Rule of Thumb

To roughly estimate the standard deviation from a collection of known sample data use

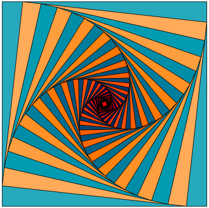
$$s \approx \frac{\text{range}}{4}$$

where range = (maximum value) − (minimum value)

Example: The heights, in feet, of people who work in an office are as follows:

6.0 5.5 5.9 5.4 5.8 5.6 5.7 6.2 5.6 5.6

Use the range rule of thumb to estimate the standard deviation.



Range Rule of Thumb

To roughly estimate the standard deviation from a collection of known sample data use

$$s \approx \frac{\text{range}}{4}$$

where range = (maximum value) – (minimum value)

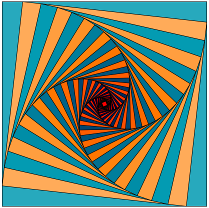
Example: The heights, in feet, of people who work in an office are as follows:

6.0 5.5 5.9 5.4 5.8 5.6 5.7 6.2 5.6 5.6

Use the range rule of thumb to estimate the standard deviation.

Answer:

$$s \approx \frac{6.2 - 5.4}{4} = 0.2 \text{ feet}$$



Range Rule of Thumb for Interpreting a Known Value of the Standard Deviation

We can find rough estimates of the minimum and maximum "usual" sample values as follows:

$$\text{Minimum "usual" value} = (\text{mean}) - 2 \times (\text{standard deviation})$$

$$\text{Maximum "usual" value} = (\text{mean}) + 2 \times (\text{standard deviation})$$

Range Rule of Thumb for Interpreting a Known Value of the Standard Deviation

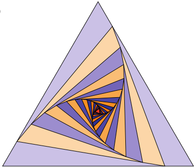
We can find rough estimates of the minimum and maximum "usual" sample values as follows:

Minimum "usual" value = (mean) − 2 × (standard deviation)

Maximum "usual" value = (mean) + 2 × (standard deviation)

Example: Environmental scientists measured the greenhouse gas emissions of a sample of cars. The amounts listed below are in tons (per year), expressed as CO_2 equivalents.

7.2 7.1 7.4 7.9 6.5 7.2 8.2 9.3



Range Rule of Thumb for Interpreting a Known Value of the Standard Deviation

We can find rough estimates of the minimum and maximum "usual" sample values as follows:

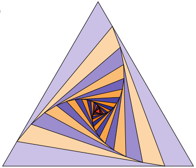
Minimum "usual" value = (mean) − 2 × (standard deviation)

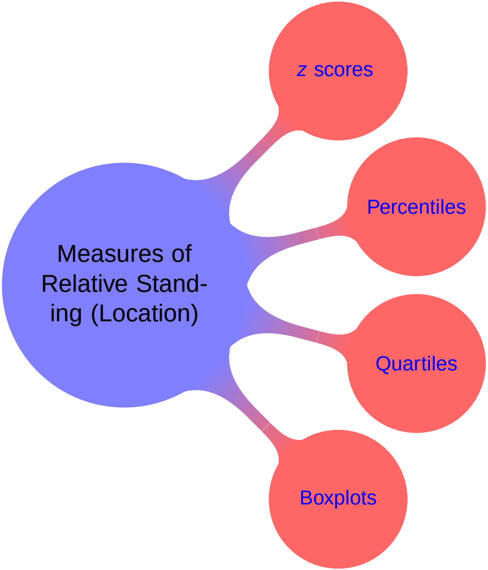
Maximum "usual" value = (mean) + 2 × (standard deviation)

Example: Environmental scientists measured the greenhouse gas emissions of a sample of cars. The amounts listed below are in tons (per year), expressed as CO_2 equivalents.

7.2 7.1 7.4 7.9 6.5 7.2 8.2 9.3

Is the value of 9.3 tons unusual?

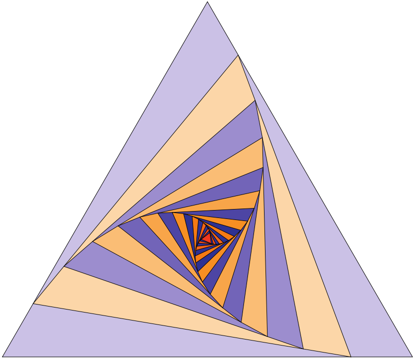



Measures of Relative Standing (Location)
z scores
Percentiles
Quartiles
Boxplots

Measures of Relative Standing

Definition (Measures of Relative Standing)

Numbers showing the location of data values relative to the other values within a data set.

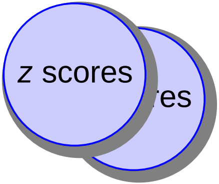


z scores

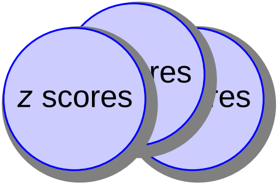
z scores

z scores

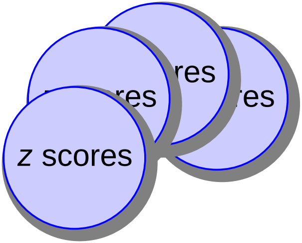
z scores



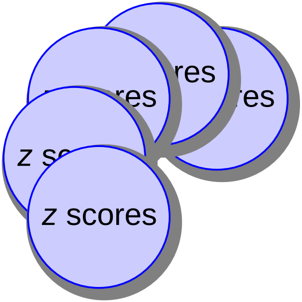
z scores



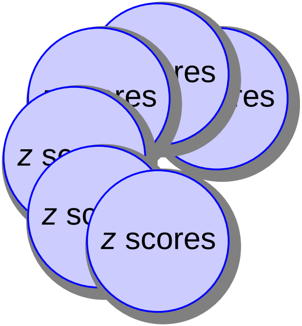
z scores



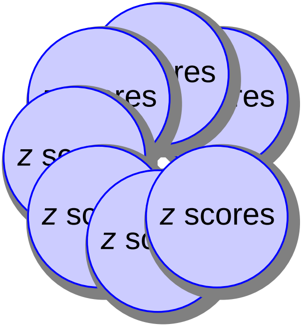
z scores



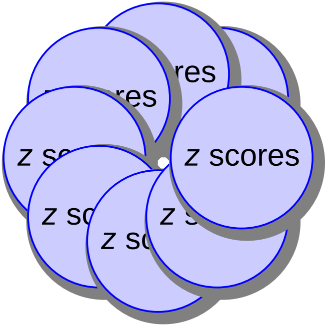
z scores



z scores



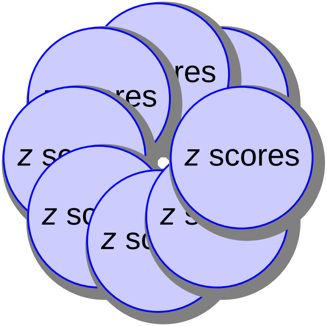
z scores



Definition (*z* score)

The number of standard deviations that a given data entry value, *x*, is above or below the mean.

z scores

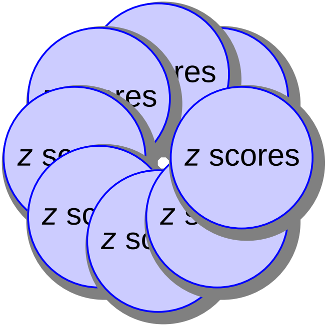


Definition (*z* score)

The number of standard deviations that a given data entry value, x, is above or below the mean.

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

z scores



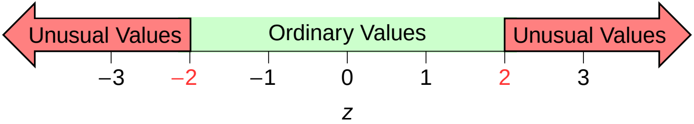
Definition (*z* score)

The number of standard deviations that a given data entry value, *x*, is above or below the mean.

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

Always round z scores to 2 decimal places!

Interpreting z Scores



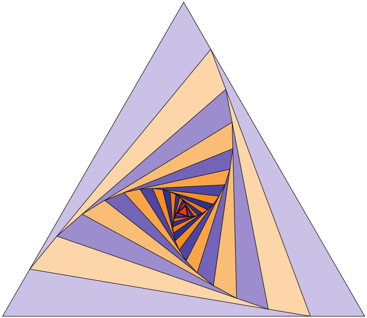
- If a data value (x) is less than the mean, then its corresponding z score is negative
- **Ordinary values**: $-2 \leq z$ score ≤ 2
- **Unusual Values**: z score < -2 or z score > 2

Example: Find the z-score corresponding to the given value and use the z-score to determine whether the value is unusual. A test score of 83.0 on a test having a mean of 66 and a standard deviation of 10.

Percentiles

Definition

Percentiles Percentiles (denoted $P_1, P_2, \ldots P_{99}$) divide a set of data into 100 groups with about 1% of the values in each group. The **percentile rank** of a data value, x, is the percentage of the data values that fall at or below a x.

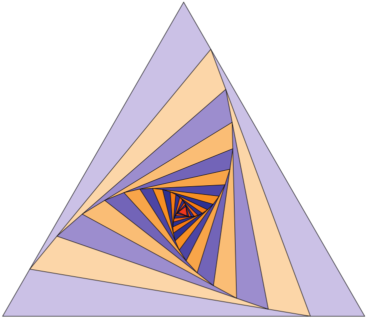


Percentiles

Definition

Percentiles Percentiles (denoted $P_1, P_2, \ldots P_{99}$) divide a set of data into 100 groups with about 1% of the values in each group. The **percentile rank** of a data value, x, is the percentage of the data values that fall at or below a x.

$$\text{Percentile Rank of value } x = \frac{(\text{number of values less than } x) + 0.5}{\text{total number of values}}$$



Percentiles

Definition

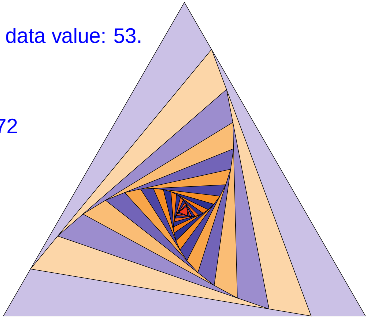
Percentiles Percentiles (denoted $P_1, P_2, \ldots P_{99}$) divide a set of data into 100 groups with about 1% of the values in each group. The **percentile rank** of a data value, x, is the percentage of the data values that fall at or below a x.

$$\text{Percentile Rank of value } x = \frac{(\text{number of values less than } x) + 0.5}{\text{total number of values}}$$

Example: Find the percentile rank for the data value: 53.

Data set:

39 44 45 53 66 69 72



Percentiles

Definition

Percentiles Percentiles (denoted $P_1, P_2, \ldots P_{99}$) divide a set of data into 100 groups with about 1% of the values in each group. The **percentile rank** of a data value, x, is the percentage of the data values that fall at or below a x.

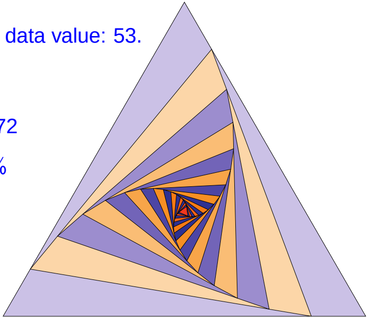
$$\text{Percentile Rank of value } x = \frac{(\text{number of values less than } x) + 0.5}{\text{total number of values}}$$

Example: Find the percentile rank for the data value: 53.

Data set:

39 44 45 53 66 69 72

Percentile of value 53 $= \frac{3.5}{7} = 0.50 = 50\%$



Percentiles

Definition

Percentiles Percentiles (denoted $P_1, P_2, \ldots P_{99}$) divide a set of data into 100 groups with about 1% of the values in each group. The **percentile rank** of a data value, x, is the percentage of the data values that fall at or below a x.

$$\text{Percentile Rank of value } x = \frac{(\text{number of values less than } x) + 0.5}{\text{total number of values}}$$

Example: Find the percentile rank for the data value: 53.

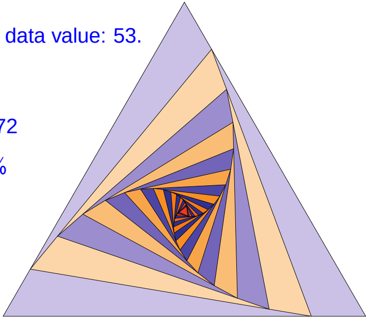
Data set:

39 44 45 53 66 69 72

$\text{Percentile of value } 53 = \frac{3.5}{7} = 0.50 = 50\%$

Interpretation: 53 is the 50th percentile.

53 separates the lowest 50% of the data from the highest 50%



Percentiles

We will also need to *locate* values associated with a certain percentile.

Example: Consider again the sample data (below) measuring space shuttle flight duration times (in hours). What flight duration time is associated with the 42nd percentile (denoted as P_{42})?

0 73 95 165 191 192 221 235 235 244 259 262 331 376 381

To answer the question we need to use the "Locator Formula,"

$$L = p \cdot n$$

Percentiles

Locator Formula

$$L = p \cdot n$$

L the location, L, that gives the position of a value in the sorted data (example: the 4th value in a sorted list, $L = 4$)

p percentile being used (as a decimal) (example: the 42nd percentile, $p = 0.42$)

n total number of values in the data set

Percentiles

Example: Consider again the sample data (below) measuring space shuttle flight duration times (in hours). What flight duration time is associated with the 42nd percentile (denoted as P_{42})?

0 73 95 165 191 192 221 235 235 244 259 262 331 376 381

Locator Formula

$$L = p \cdot n$$

L	is the location, L, that gives the position of the value in the sorted data that is associated with the p^{th} percentile.
$p = 0.42$	is the percentile as a decimal
$n = 15$	is the total number of values in the data set.

Percentiles

Example: Consider again the sample data (below) measuring space shuttle flight duration times (in hours). What flight duration time is associated with the 42nd percentile (denoted as P_{42})?

0 73 95 165 191 192 221 235 235 244 259 262 331 376 381

Locator Formula

$$L = p \cdot n = 0.42 \cdot 15$$

L is the location, L, that gives the position of the value in the sorted data that is associated with the p^{th} percentile.

$p = 0.42$ is the percentile as a decimal

$n = 15$ is the total number of values in the data set.

Percentiles

Example: Consider again the sample data (below) measuring space shuttle flight duration times (in hours). What flight duration time is associated with the 42nd percentile (denoted as P_{42})?

0 73 95 165 191 192 221 235 235 244 259 262 331 376 381

Locator Formula

$$L = p \cdot n = 0.42 \cdot 15 = 6.2$$

L is the location, L, that gives the position of the value in the sorted data that is associated with the p^{th} percentile.

$p = 0.42$ is the percentile as a decimal

$n = 15$ is the total number of values in the data set.

Since $L = 6.2$ is not a whole number, we round L up to 7—and the 7th value in the sorted data is 221. Thus, the 42nd percentile, P_{42}, is 221 hours.

How to Locate the x value corresponding to the p^{th} Percentile

Sort the data in ascending order.

How to Locate the x value corresponding to the p^{th} Percentile

Sort the data in ascending order.

↓

Compute

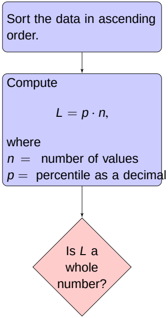
$$L = p \cdot n,$$

where

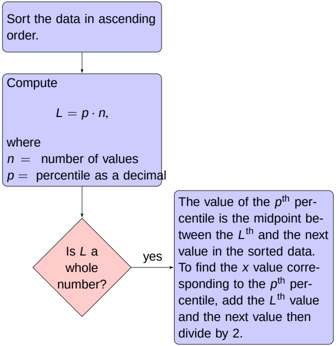
$n =$ number of values

$p =$ percentile as a decimal

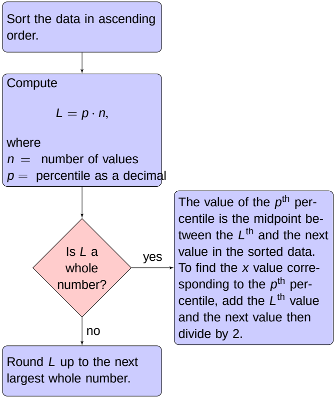
How to Locate the x value corresponding to the p^{th} Percentile



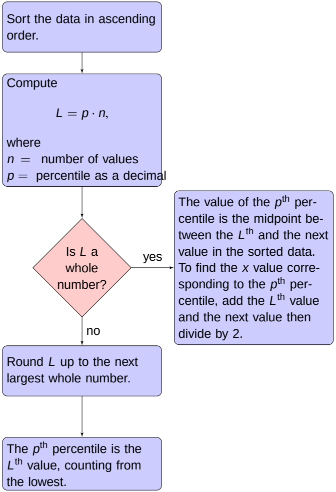
How to Locate the x value corresponding to the p^{th} Percentile



How to Locate the x value corresponding to the p^{th} Percentile



How to Locate the x value corresponding to the p^{th} Percentile



Percentiles

Example: Consider again the sample data (below) measuring space shuttle flight duration times (in hours). What flight duration time is associated with the 80th percentile (denoted as P_{80})?

0 73 95 165 191 192 221 235 235 244 259 262 331 376 381

Locator Formula

$$L = p \cdot n$$

L	the location, L, that gives the position of a value in the sorted data
$p = 0.80$	percentile as a decimal
$n = 15$	total number of values in the data set

Percentiles

Example: Consider again the sample data (below) measuring space shuttle flight duration times (in hours). What flight duration time is associated with the 80th percentile (denoted as P_{80})?

0 73 95 165 191 192 221 235 235 244 259 262 331 376 381

Locator Formula

$$L = p \cdot n = 0.80 \cdot 15$$

L	the location, L, that gives the position of a value in the sorted data
$p = 0.80$	percentile as a decimal
$n = 15$	total number of values in the data set

Percentiles

Example: Consider again the sample data (below) measuring space shuttle flight duration times (in hours). What flight duration time is associated with the 80th percentile (denoted as P_{80})?

0 73 95 165 191 192 221 235 235 244 259 262 331 376 381

Locator Formula

$$L = p \cdot n = 0.80 \cdot 15 = 12$$

L	the location, L, that gives the position of a value in the sorted data
$p = 0.80$	percentile as a decimal
$n = 15$	total number of values in the data set

$L = 12$, and the 12th value in the sorted data is 262. Since L was a whole number, we have to take the midpoint between 262 and the next value, 331, as being the flight time associated with the 80th percentile, P_{80}. So,

$$P_{80} = \frac{262 + 331}{2} = 296.5 \; hrs.$$

Quartiles

Definition

Quartiles are measures of location, denoted Q_1, Q_2, and Q_3, which divide sorted data set 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%. (*i.e.*, $Q_1 = P_{25}$)
- Q_2 **(Second Quartile)** same as the median; separates the bottom 50% of sorted values from the top 50%. (*i.e.*, $Q_2 = P_{50}$)
- Q_3 **(Third Quartile)** separates the bottom 75% of sorted values from the top 25%. (*i.e.*, $Q_3 = P_{75}$)
- **Interquartile Range (IQR)** $Q_3 - Q_1$

Box and Whisker Plot

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.

Definition

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 whose vertical edges are drawn at the first quartile, Q_1; the median; and the third quartile, Q_3.

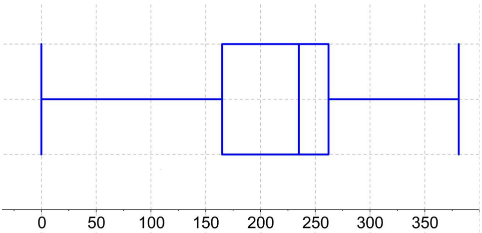


Figure : Flight Duration Times (in hours)

Modified Boxplot

Definition

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

Definition

A **modified boxplot** is a boxplot with the following modifications:

- The whiskers of the boxplot (the dotted horizontal line in the figure) extend only as far as the minimum data value that is not an outlier (defined as $Q_1 - 1.5 \cdot (\text{IQR})$), and the maximum data value that is not an outlier (defined as $Q_3 + 1.5 \cdot (\text{IQR})$).
- A special symbol (such as an cross) is used to identify outliers.
- A data value, x, is considered an outlier if

$$x > Q_3 + 1.5 \cdot (\text{IQR}), \quad \text{OR} \quad x < Q_1 - 1.5 \cdot (\text{IQR})$$

where $\text{IQR} = Q_3 - Q_1$.

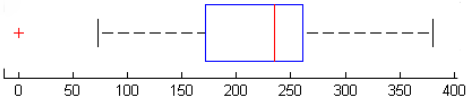


Figure : Flight Duration Times (in hours)

M. F. TRIOLA, *Essentials of Statistics*, Addison-Wesley, fourth ed., 2011.