STATISTICAL INFERENCES (2cr)
Chapter 8 Sampling Distributions & Data Descriptions

Zhong Guan

Math, IUSB

Outline

1. 8.1 Random Sampling
   - Basic terminology
2. 8.2 Some Important Statistics
   - Measure of Central Tendency for the Sample
   - Range and The Sample Variance
   - Quartiles & Box plot
   - Relative Frequency and Histogram

Outline

Population and Sample

- **Population:** A well-defined collection of all possible observations with which we are concerned.
- For instance, the population of U.S. registered voters as of November 1 in the most recent presidential election year.
- "Population $f(x)$" means a population whose observations are values of random variable having distribution $f(x)$.
- **Sample:** a subset of a population. It is often denoted as $X_1, X_2, \ldots, X_n$.
- For instance, a random sample of size 1000 from the above list of U.S. registered voters.

Population and Sample

- **Population:** A well-defined collection of all possible observations with which we are concerned.
- For instance, the population of U.S. registered voters as of November 1 in the most recent presidential election year.
- "Population $f(x)$" means a population whose observations are values of random variable having distribution $f(x)$.
- **Sample:** a subset of a population. It is often denoted as $X_1, X_2, \ldots, X_n$.
- For instance, a random sample of size 1000 from the above list of U.S. registered voters.

Population and Sample

- **Population:** A well-defined collection of all possible observations with which we are concerned.
- For instance, the population of U.S. registered voters as of November 1 in the most recent presidential election year.
- "Population $f(x)$" means a population whose observations are values of random variable having distribution $f(x)$.
- **Sample:** a subset of a population. It is often denoted as $X_1, X_2, \ldots, X_n$.
- For instance, a random sample of size 1000 from the above list of U.S. registered voters.

Population and Sample

- **Population:** A well-defined collection of all possible observations with which we are concerned.
- For instance, the population of U.S. registered voters as of November 1 in the most recent presidential election year.
- "Population $f(x)$" means a population whose observations are values of random variable having distribution $f(x)$.
- **Sample:** a subset of a population. It is often denoted as $X_1, X_2, \dots, X_n$.
- For instance, a random sample of size 1000 from the above list of U.S. registered voters.

Population and Sample

- **Population:** A well-defined collection of all possible observations with which we are concerned.
- For instance, the population of U.S. registered voters as of November 1 in the most recent presidential election year.
- "Population $f(x)$" means a population whose observations are values of random variable having distribution $f(x)$.
- **Sample:** a subset of a population. It is often denoted as $X_1, X_2, \ldots, X_n$.
- For instance, a random sample of size 1000 from the above list of U.S. registered voters.

Parameter and Statistic

- **Parameter:** characteristic of a population.
- For instance, p =percentage of Democratic voters in the above list of U.S. registered voters.
- **Statistic:** characteristic of a sample, an estimate of the parameter.
- For instance, $\hat{p}$ = percentage of Democratic voters in the above sample.
- **A Parameter is to a Population as a Statistic is to a Sample.**
- **In statistics, we often rely on a sample to draw inferences about the population.**

Parameter and Statistic

- **Parameter:** characteristic of a population.
- For instance, p =percentage of Democratic voters in the above list of U.S. registered voters.
- **Statistic:** characteristic of a sample, an estimate of the parameter.
- For instance, $\hat{p}$ = percentage of Democratic voters in the above sample.
- **A Parameter is to a Population as a Statistic is to a Sample.**
- **In statistics, we often rely on a sample to draw inferences about the population.**

Parameter and Statistic

- **Parameter:** characteristic of a population.
- For instance, p =percentage of Democratic voters in the above list of U.S. registered voters.
- **Statistic:** characteristic of a sample, an estimate of the parameter.
- For instance, $\hat{p}$ = percentage of Democratic voters in the above sample.
- **A Parameter is to a Population as a Statistic is to a Sample.**
- **In statistics, we often rely on a sample to draw inferences about the population.**

Parameter and Statistic

- **Parameter:** characteristic of a population.
- For instance, p =percentage of Democratic voters in the above list of U.S. registered voters.
- **Statistic:** characteristic of a sample, an estimate of the parameter.
- For instance, $\hat{p}$ = percentage of Democratic voters in the above sample.
- **A Parameter is to a Population as a Statistic is to a Sample.**
- **In statistics, we often rely on a sample to draw inferences about the population.**

Parameter and Statistic

- **Parameter:** characteristic of a population.
- For instance, p =percentage of Democratic voters in the above list of U.S. registered voters.
- **Statistic:** characteristic of a sample, an estimate of the parameter.
- For instance, $\hat{p}$ = percentage of Democratic voters in the above sample.
- **A Parameter is to a Population as a Statistic is to a Sample.**
- **In statistics, we often rely on a sample to draw inferences about the population.**

Parameter and Statistic

- **Parameter:** characteristic of a population.
- For instance, p =percentage of Democratic voters in the above list of U.S. registered voters.
- **Statistic:** characteristic of a sample, an estimate of the parameter.
- For instance, $\hat{p}$ = percentage of Democratic voters in the above sample.
- **A Parameter is to a Population as a Statistic is to a Sample.**
- **In statistics, we often rely on a sample to draw inferences about the population.**

Example 1.

Examples of population, sample, parameter, and statistic.

CONCERN	POPULATION	SAMPLE	PARAMETER/ STATISTIC
Election	All voters	Gallup poll	% of votes for candidate A
Cancer	All cancer patients	20 cancer patients	Average cancer size
Water safety	Water in the well	Water in test cube	bacterial counts in unit volume
Pseudorandom numbers generator	The complete sequence	Subsequence	Distribution of pseudorandom numbers
Unemployment	Entire labor force	One million force	rate of unemployment

Example 1.

Examples of population, sample, parameter, and statistic.

CONCERN	POPULATION	SAMPLE	PARAMETER/ STATISTIC
Election	All voters	Gallup poll	% of votes for candidate A
Cancer	All cancer patients	20 cancer patients	Average cancer size
Water safety	Water in the well	Water in test cube	bacterial counts in unit volume
Pseudorandom numbers generator	The complete sequence	Subsequence	Distribution of pseudorandom numbers
Unemployment	Entire labor force	One million force	rate of unemployment

Example 1.

Examples of population, sample, parameter, and statistic.

CONCERN	POPULATION	SAMPLE	PARAMETER/ STATISTIC
Election	All voters	Gallup poll	% of votes for candidate A
Cancer	All cancer patients	20 cancer patients	Average cancer size
Water safety	Water in the well	Water in test cube	bacterial counts in unit volume
Pseudorandom numbers generator	The complete sequence	Subsequence	Distribution of pseudorandom numbers
Unemployment	Entire labor force	One million force	rate of unemployment

Example 1.

Examples of population, sample, parameter, and statistic.

CONCERN	POPULATION	SAMPLE	PARAMETER/ STATISTIC
Election	All voters	Gallup poll	% of votes for candidate A
Cancer	All cancer patients	20 cancer patients	Average cancer size
Water safety	Water in the well	Water in test cube	bacterial counts in unit volume
Pseudorandom numbers generator	The complete sequence	Subsequence	Distribution of pseudorandom numbers
Unemployment	Entire labor force	One million force	rate of unemployment

Example 1.

Examples of population, sample, parameter, and statistic.

CONCERN	POPULATION	SAMPLE	PARAMETER/ STATISTIC
Election	All voters	Gallup poll	% of votes for candidate A
Cancer	All cancer patients	20 cancer patients	Average cancer size
Water safety	Water in the well	Water in test cube	bacterial counts in unit volume
Pseudorandom numbers generator	The complete sequence	Subsequence	Distribution of pseudorandom numbers
Unemployment	Entire labor force	One million force	rate of unemployment

Example 1.

Examples of population, sample, parameter, and statistic.

CONCERN	POPULATION	SAMPLE	PARAMETER/ STATISTIC
Election	All voters	Gallup poll	% of votes for candidate A
Cancer	All cancer patients	20 cancer patients	Average cancer size
Water safety	Water in the well	Water in test cube	bacterial counts in unit volume
Pseudorandom numbers generator	The complete sequence	Subsequence	Distribution of pseudorandom numbers
Unemployment	Entire labor force	One million force	rate of unemployment

Example 1.

Examples of population, sample, parameter, and statistic.

CONCERN	POPULATION	SAMPLE	PARAMETER/ STATISTIC
Election	All voters	Gallup poll	% of votes for candidate A
Cancer	All cancer patients	20 cancer patients	Average cancer size
Water safety	Water in the well	Water in test cube	bacterial counts in unit volume
Pseudorandom numbers generator	The complete sequence	Subsequence	Distribution of pseudorandom numbers
Unemployment	Entire labor force	One million force	rate of unemployment

Example 1.

Examples of population, sample, parameter, and statistic.

CONCERN	POPULATION	SAMPLE	PARAMETER/ STATISTIC
Election	All voters	Gallup poll	% of votes for candidate A
Cancer	All cancer patients	20 cancer patients	Average cancer size
Water safety	Water in the well	Water in test cube	bacterial counts in unit volume
Pseudorandom numbers generator	The complete sequence	Subsequence	Distribution of pseudorandom numbers
Unemployment	Entire labor force	One million force	rate of unemployment

Example 1.

Examples of population, sample, parameter, and statistic.

CONCERN	POPULATION	SAMPLE	PARAMETER/ STATISTIC
Election	All voters	Gallup poll	% of votes for candidate A
Cancer	All cancer patients	20 cancer patients	Average cancer size
Water safety	Water in the well	Water in test cube	bacterial counts in unit volume
Pseudorandom numbers generator	The complete sequence	Subsequence	Distribution of pseudorandom numbers
Unemployment	Entire labor force	One million force	rate of unemployment

Example 1.

Examples of population, sample, parameter, and statistic.

CONCERN	POPULATION	SAMPLE	PARAMETER/ STATISTIC
Election	All voters	Gallup poll	% of votes for candidate A
Cancer	All cancer patients	20 cancer patients	Average cancer size
Water safety	Water in the well	Water in test cube	bacterial counts in unit volume
Pseudorandom numbers generator	The complete sequence	Subsequence	Distribution of pseudorandom numbers
Unemployment	Entire labor force	One million force	rate of unemployment

Example 1.

Examples of population, sample, parameter, and statistic.

CONCERN	POPULATION	SAMPLE	PARAMETER/ STATISTIC
Election	All voters	Gallup poll	% of votes for candidate A
Cancer	All cancer patients	20 cancer patients	Average cancer size
Water safety	Water in the well	Water in test cube	bacterial counts in unit volume
Pseudorandom numbers generator	The complete sequence	Subsequence	Distribution of pseudorandom numbers
Unemployment	Entire labor force	One million force	rate of unemployment

Example 1.

Examples of population, sample, parameter, and statistic.

CONCERN	POPULATION	SAMPLE	PARAMETER/ STATISTIC
Election	All voters	Gallup poll	% of votes for candidate A
Cancer	All cancer patients	20 cancer patients	Average cancer size
Water safety	Water in the well	Water in test cube	bacterial counts in unit volume
Pseudorandom numbers generator	The complete sequence	Subsequence	Distribution of pseudorandom numbers
Unemployment	Entire labor force	One million force	rate of unemployment

Example 1.

Examples of population, sample, parameter, and statistic.

CONCERN	POPULATION	SAMPLE	PARAMETER/ STATISTIC
Election	All voters	Gallup poll	% of votes for candidate A
Cancer	All cancer patients	20 cancer patients	Average cancer size
Water safety	Water in the well	Water in test cube	bacterial counts in unit volume
Pseudorandom numbers generator	The complete sequence	Subsequence	Distribution of pseudorandom numbers
Unemployment	Entire labor force	One million force	rate of unemployment

Example 1.

Examples of population, sample, parameter, and statistic.

CONCERN	POPULATION	SAMPLE	PARAMETER/ STATISTIC
Election	All voters	Gallup poll	% of votes for candidate A
Cancer	All cancer patients	20 cancer patients	Average cancer size
Water safety	Water in the well	Water in test cube	bacterial counts in unit volume
Pseudorandom numbers generator	The complete sequence	Subsequence	Distribution of pseudorandom numbers
Unemployment	Entire labor force	One million force	rate of unemployment

Example 1.

Examples of population, sample, parameter, and statistic.

CONCERN	POPULATION	SAMPLE	PARAMETER/ STATISTIC
Election	All voters	Gallup poll	% of votes for candidate A
Cancer	All cancer patients	20 cancer patients	Average cancer size
Water safety	Water in the well	Water in test cube	bacterial counts in unit volume
Pseudorandom numbers generator	The complete sequence	Subsequence	Distribution of pseudorandom numbers
Unemployment	Entire labor force	One million force	rate of unemployment

Example 1.

Examples of population, sample, parameter, and statistic.

CONCERN	POPULATION	SAMPLE	PARAMETER/ STATISTIC
Election	All voters	Gallup poll	% of votes for candidate A
Cancer	All cancer patients	20 cancer patients	Average cancer size
Water safety	Water in the well	Water in test cube	bacterial counts in unit volume
Pseudorandom numbers generator	The complete sequence	Subsequence	Distribution of pseudorandom numbers
Unemployment	Entire labor force	One million force	rate of unemployment

Example 1.

Examples of population, sample, parameter, and statistic.

CONCERN	POPULATION	SAMPLE	PARAMETER/ STATISTIC
Election	All voters	Gallup poll	% of votes for candidate A
Cancer	All cancer patients	20 cancer patients	Average cancer size
Water safety	Water in the well	Water in test cube	bacterial counts in unit volume
Pseudorandom numbers generator	The complete sequence	Subsequence	Distribution of pseudorandom numbers
Unemployment	Entire labor force	One million force	rate of unemployment

Example 1.

Examples of population, sample, parameter, and statistic.

CONCERN	POPULATION	SAMPLE	PARAMETER/ STATISTIC
Election	All voters	Gallup poll	% of votes for candidate A
Cancer	All cancer patients	20 cancer patients	Average cancer size
Water safety	Water in the well	Water in test cube	bacterial counts in unit volume
Pseudorandom numbers generator	The complete sequence	Subsequence	Distribution of pseudorandom numbers
Unemployment	Entire labor force	One million force	rate of unemployment

Example 1.

Examples of population, sample, parameter, and statistic.

CONCERN	POPULATION	SAMPLE	PARAMETER/ STATISTIC
Election	All voters	Gallup poll	% of votes for candidate A
Cancer	All cancer patients	20 cancer patients	Average cancer size
Water safety	Water in the well	Water in test cube	bacterial counts in unit volume
Pseudorandom numbers generator	The complete sequence	Subsequence	Distribution of pseudorandom numbers
Unemployment	Entire labor force	One million force	rate of unemployment

Example 1.

Examples of population, sample, parameter, and statistic.

CONCERN	POPULATION	SAMPLE	PARAMETER/ STATISTIC
Election	All voters	Gallup poll	% of votes for candidate A
Cancer	All cancer patients	20 cancer patients	Average cancer size
Water safety	Water in the well	Water in test cube	bacterial counts in unit volume
Pseudorandom numbers generator	The complete sequence	Subsequence	Distribution of pseudorandom numbers
Unemployment	Entire labor force	One million force	rate of unemployment

Example 1.

Examples of population, sample, parameter, and statistic.

CONCERN	POPULATION	SAMPLE	PARAMETER/ STATISTIC
Election	All voters	Gallup poll	% of votes for candidate A
Cancer	All cancer patients	20 cancer patients	Average cancer size
Water safety	Water in the well	Water in test cube	bacterial counts in unit volume
Pseudorandom numbers generator	The complete sequence	Subsequence	Distribution of pseudorandom numbers
Unemployment	Entire labor force	One million force	rate of unemployment

Random Sample

- **Definition:** If $X_1, X_2, \ldots, X_n$ are n random observations from population $f(x)$, and are independent, that is, they have joint distribution

$$f(x_1, \ldots, x_n) = f(x_1) \cdots f(x_n)$$

then $X_1, X_2, \ldots, X_n$ is said to be a **random sample** of a **size** n from the population $f(x)$.

- When we use capital letters, we treat $X_1, X_2, \ldots, X_n$ as n independent random variables having the same distribution $f(x)$;
- When we use lower case letters, we treat $x_1, x_2, \ldots, x_n$ as the numerical values of the n independent random variables;

Random Sample

- **Definition:** If $X_1, X_2, \ldots, X_n$ are n random observations from population $f(x)$, and are independent, that is, they have joint distribution

$$f(x_1, \ldots, x_n) = f(x_1) \cdots f(x_n)$$

then $X_1, X_2, \ldots, X_n$ is said to be a **random sample** of a **size** n from the population $f(x)$.
- When we use capital letters, we treat $X_1, X_2, \ldots, X_n$ as n independent random variables having the same distribution $f(x)$;
- When we use lower case letters, we treat $x_1, x_2, \ldots, x_n$ as the numerical values of the n independent random variables;

Random Sample

- **Definition:** If $X_1, X_2, \ldots, X_n$ are n random observations from population $f(x)$, and are independent, that is, they have joint distribution

$$f(x_1, \ldots, x_n) = f(x_1) \cdots f(x_n)$$

  then $X_1, X_2, \ldots, X_n$ is said to be a **random sample** of a **size** n from the population $f(x)$.
- When we use capital letters, we treat $X_1, X_2, \ldots, X_n$ as n independent random variables having the same distribution $f(x)$;
- When we use lower case letters, we treat $x_1, x_2, \ldots, x_n$ as the numerical values of the n independent random variables;

Random Sample

- **Definition:** If $X_1, X_2, \ldots, X_n$ are n random observations from population $f(x)$, and are independent, that is, they have joint distribution

$$f(x_1, \ldots, x_n) = f(x_1) \cdots f(x_n)$$

  then $X_1, X_2, \ldots, X_n$ is said to be a **random sample** of a **size** n from the population $f(x)$.
- When we use capital letters, we treat $X_1, X_2, \ldots, X_n$ as n independent random variables having the same distribution $f(x)$;
- When we use lower case letters, we treat $x_1, x_2, \ldots, x_n$ as the numerical values of the n independent random variables;

Random Sample

Example 2. The number of traffics during weekday 11:30am to 12:30pm at an intersection of a city has Poisson distribution with mean λ.

- A random sample of size 20, $X_1, X_2, \ldots, X_{20}$ are 20 independent random variables having the same Poisson distribution $P(\lambda)$.
- The sample values $x_1, x_2, \ldots, x_{20}$ are actually observed data values in 20 weekdays.

Random Sample

Example 2. The number of traffics during weekday 11:30am to 12:30pm at an intersection of a city has Poisson distribution with mean λ.

- A random sample of size 20, $X_1, X_2, \ldots, X_{20}$ are 20 independent random variables having the same Poisson distribution $P(\lambda)$.
- The sample values $x_1, x_2, \ldots, x_{20}$ are actually observed data values in 20 weekdays.

Random Sample

Example 2. The number of traffics during weekday 11:30am to 12:30pm at an intersection of a city has Poisson distribution with mean λ.

- A random sample of size 20, $X_1, X_2, \ldots, X_{20}$ are 20 independent random variables having the same Poisson distribution $P(\lambda)$.
- The sample values $x_1, x_2, \ldots, x_{20}$ are actually observed data values in 20 weekdays.

Definition

Statistic Any function of random variables constituting a random sample is called a **statistic**, which is free of any unknown parameter.

Example 0. The number of traffics during weekday 11:30am to 12:30pm at an intersection of a city has Poisson distribution $P(\lambda)$ with mean λ.

- The mean value of a random sample of size 20, $X_1, X_2, \ldots, X_{20}$ is a statistic:

$$\bar{X} = \frac{1}{20}(X_1 + \cdots + X_{20})$$

- The mean value of the actual observed sample values $x_1, x_2, \ldots, x_{20}$ is denoted by

$$\bar{x} = \frac{1}{20}(x_1 + \cdots + x_{20})$$

Definition

Statistic Any function of random variables constituting a random sample is called a **statistic**, which is free of any unknown parameter.

Example 0. The number of traffics during weekday 11:30am to 12:30pm at an intersection of a city has Poisson distribution $P(\lambda)$ with mean λ.

- The mean value of a random sample of size 20, $X_1, X_2, \ldots, X_{20}$ is a statistic:

$$\bar{X} = \frac{1}{20}(X_1 + \cdots + X_{20})$$

- The mean value of the actual observed sample values $x_1, x_2, \ldots, x_{20}$ is denoted by

$$\bar{x} = \frac{1}{20}(x_1 + \cdots + x_{20})$$

Definition

Statistic Any function of random variables constituting a random sample is called a **statistic**, which is free of any unknown parameter.

Example 0. The number of traffics during weekday 11:30am to 12:30pm at an intersection of a city has Poisson distribution $P(\lambda)$ with mean λ.

- The mean value of a random sample of size 20, $X_1, X_2, \ldots, X_{20}$ is a statistic:

$$\bar{X} = \frac{1}{20}(X_1 + \cdots + X_{20})$$

- The mean value of the actual observed sample values $x_1, x_2, \ldots, x_{20}$ is denoted by

$$\bar{x} = \frac{1}{20}(x_1 + \cdots + x_{20})$$

Definition

Statistic Any function of random variables constituting a random sample is called a **statistic**, which is free of any unknown parameter.

Example 0. The number of traffics during weekday 11:30am to 12:30pm at an intersection of a city has Poisson distribution $P(\lambda)$ with mean λ.

- The mean value of a random sample of size 20, $X_1, X_2, \ldots, X_{20}$ is a statistic:

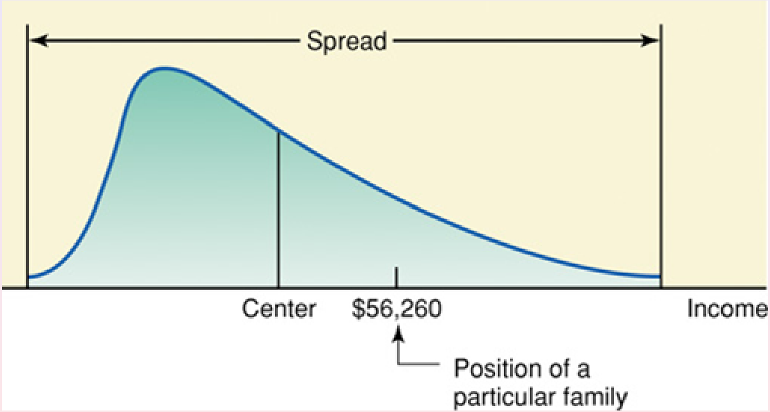
$$\bar{X} = \frac{1}{20}(X_1 + \cdots + X_{20})$$

- The mean value of the actual observed sample values $x_1, x_2, \ldots, x_{20}$ is denoted by

$$\bar{x} = \frac{1}{20}(x_1 + \cdots + x_{20})$$

Outline

Measures for a distribution



Mean is the average value

$$\text{Sample Mean} = \frac{\text{Sum of values}}{\text{Number of values}}$$

Let $X_1, X_2, \ldots, X_n$ be a sample of size n. Let $x_1, x_2, \ldots, x_n$ be the values of the sample data.

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Properties of the Mean

- Uniqueness: For any data set, there is only one arithmetic mean.
- Simplicity: Easy to calculate.
- Non-robustness: A single extreme value in a sample could cause undesirable result. Isolated extreme values are called *outliers*.

Properties of the Mean

- Uniqueness: For any data set, there is only one arithmetic mean.
- Simplicity: Easy to calculate.
- Non-robustness: A single extreme value in a sample could cause undesirable result. Isolated extreme values are called *outliers*.

Properties of the Mean

- Uniqueness: For any data set, there is only one arithmetic mean.
- Simplicity: Easy to calculate.
- Non-robustness: A single extreme value in a sample could cause undesirable result. Isolated extreme values are called *outliers*.

Properties of the Mean

- Uniqueness: For any data set, there is only one arithmetic mean.
- Simplicity: Easy to calculate.
- Non-robustness: A single extreme value in a sample could cause undesirable result. Isolated extreme values are called *outliers*.

Median is the middle value

(1) Sort the data values in increasing order.
(2) If the number of values is odd, then the middle term is the median.
(3) If the number of values is even, then the average of the two middle terms is the median.

Example 1: Find the median of each data set.

(a) Data set 1: 7, 2, -1, 5, 9, 2, 4.
(b) Data set 2: 1.2, 0.7, 3.5, 1.6, 0.3, 2.4.

Median is the middle value

(1) Sort the data values in increasing order.

(2) If the number of values is odd, then the middle term is the median.

(3) If the number of values is even, then the average of the two middle terms is the median.

Example 1: Find the median of each data set.

(a) Data set 1: 7, 2, -1, 5, 9, 2, 4.

(b) Data set 2: 1.2, 0.7, 3.5, 1.6, 0.3, 2.4.

Median is the middle value

(1) Sort the data values in increasing order.

(2) If the number of values is odd, then the middle term is the median.

(3) If the number of values is even, then the average of the two middle terms is the median.

Example 1: Find the median of each data set.

(a) Data set 1: 7, 2, -1, 5, 9, 2, 4.

(b) Data set 2: 1.2, 0.7, 3.5, 1.6, 0.3, 2.4.

Median is the middle value

(1) Sort the data values in increasing order.

(2) If the number of values is odd, then the middle term is the median.

(3) If the number of values is even, then the average of the two middle terms is the median.

Example 1: Find the median of each data set.

(a) Data set 1: 7, 2, -1, 5, 9, 2, 4.

(b) Data set 2: 1.2, 0.7, 3.5, 1.6, 0.3, 2.4.

Median is the middle value

(1) Sort the data values in increasing order.
(2) If the number of values is odd, then the middle term is the median.
(3) If the number of values is even, then the average of the two middle terms is the median.

Example 1: Find the median of each data set.

(a) Data set 1: 7, 2, -1, 5, 9, 2, 4.
(b) Data set 2: 1.2, 0.7, 3.5, 1.6, 0.3, 2.4.

Median is the middle value

(1) Sort the data values in increasing order.
(2) If the number of values is odd, then the middle term is the median.
(3) If the number of values is even, then the average of the two middle terms is the median.

Example 1: Find the median of each data set.

(a) Data set 1: 7, 2, -1, 5, 9, 2, 4.
(b) Data set 2: 1.2, 0.7, 3.5, 1.6, 0.3, 2.4.

Properties of the Median

- Uniqueness: For any data set, there is only one median.
- Simplicity: Easy to calculate.
- Robustness: It is not drastically affected by extreme values.

Properties of the Median

- Uniqueness: For any data set, there is only one median.
- Simplicity: Easy to calculate.
- Robustness: It is not drastically affected by extreme values.

Properties of the Median

- Uniqueness: For any data set, there is only one median.
- Simplicity: Easy to calculate.
- Robustness: It is not drastically affected by extreme values.

Properties of the Median

- Uniqueness: For any data set, there is only one median.
- Simplicity: Easy to calculate.
- Robustness: It is not drastically affected by extreme values.

Mode is the most frequent value

The mode of a dataset is the value that has the highest frequency.

Example 2: Find the mode of each data set.

(a) Find the mode of the 50 students status data.

Table 2.2 Status of 50 Students

J	F	SO	SE	J	J	SE	J	J	J
F	F	J	F	F	F	SE	SO	SE	J
J	F	SE	SO	SO	F	J	F	SE	SE
SO	SE	J	SO	SO	J	J	SO	F	SO
SE	SE	F	SE	J	SO	F	J	SO	SO

(b) The speeds(mph) of 8 cars stopped on I-95 for speeding: 77, 82, 74, 81, 79, 84, 74, 78.

(c) The ages of 10 randomly selected students are: 21, 19, 27, 22, 29, 19, 25, 21, 22, 30.

Mode is the most frequent value

The mode of a dataset is the value that has the highest frequency.

Example 2: Find the mode of each data set.

(a) Find the mode of the 50 students status data.

Table 2.2 Status of 50 Students

J	F	SO	SE	J	J	SE	J	J	J
F	F	J	F	F	F	SE	SO	SE	J
J	F	SE	SO	SO	F	J	F	SE	SE
SO	SE	J	SO	SO	J	J	SO	F	SO
SE	SE	F	SE	J	SO	F	J	SO	SO

(b) The speeds(mph) of 8 cars stopped on I-95 for speeding: 77, 82, 74, 81, 79, 84, 74, 78.

(c) The ages of 10 randomly selected students are: 21, 19, 27, 22, 29, 19, 25, 21, 22, 30.

Mode is the most frequent value

The mode of a dataset is the value that has the highest frequency.

Example 2: Find the mode of each data set.

(a) Find the mode of the 50 students status data.

Table 2.2 Status of 50 Students

J	F	SO	SE	J	J	SE	J	J	J
F	F	J	F	F	F	SE	SO	SE	J
J	F	SE	SO	SO	F	J	F	SE	SE
SO	SE	J	SO	SO	J	J	SO	F	SO
SE	SE	F	SE	J	SO	F	J	SO	SO

(b) The speeds(mph) of 8 cars stopped on I-95 for speeding: 77, 82, 74, 81, 79, 84, 74, 78.

(c) The ages of 10 randomly selected students are: 21, 19, 27, 22, 29, 19, 25, 21, 22, 30.

Mode is the most frequent value

The mode of a dataset is the value that has the highest frequency.

Example 2: Find the mode of each data set.

(a) Find the mode of the 50 students status data.

Table 2.2 Status of 50 Students

J	F	SO	SE	J	J	SE	J	J	J
F	F	J	F	F	F	SE	SO	SE	J
J	F	SE	SO	SO	F	J	F	SE	SE
SO	SE	J	SO	SO	J	J	SO	F	SO
SE	SE	F	SE	J	SO	F	J	SO	SO

(b) The speeds(mph) of 8 cars stopped on I-95 for speeding: 77, 82, 74, 81, 79, 84, 74, 78.

(c) The ages of 10 randomly selected students are: 21, 19, 27, 22, 29, 19, 25, 21, 22, 30.

Mode is the most frequent value

The mode of a dataset is the value that has the highest frequency.

Example 2: Find the mode of each data set.

(a) Find the mode of the 50 students status data.

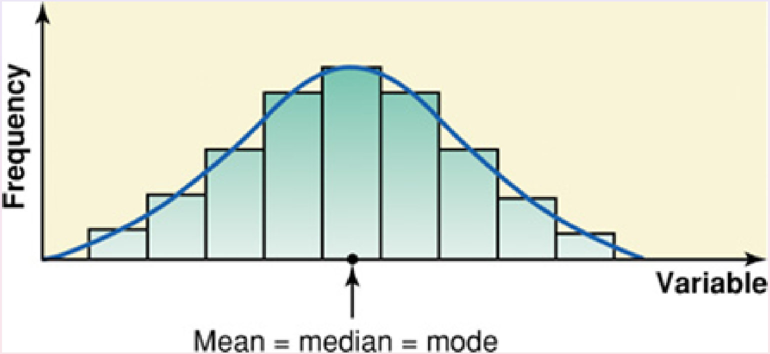
Table 2.2 Status of 50 Students

J	F	SO	SE	J	J	SE	J	J	J
F	F	J	F	F	F	SE	SO	SE	J
J	F	SE	SO	SO	F	J	F	SE	SE
SO	SE	J	SO	SO	J	J	SO	F	SO
SE	SE	F	SE	J	SO	F	J	SO	SO

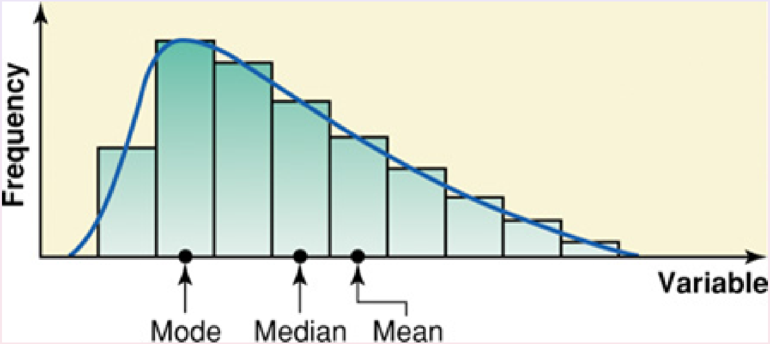
(b) The speeds(mph) of 8 cars stopped on I-95 for speeding: 77, 82, 74, 81, 79, 84, 74, 78.

(c) The ages of 10 randomly selected students are: 21, 19, 27, 22, 29, 19, 25, 21, 22, 30.

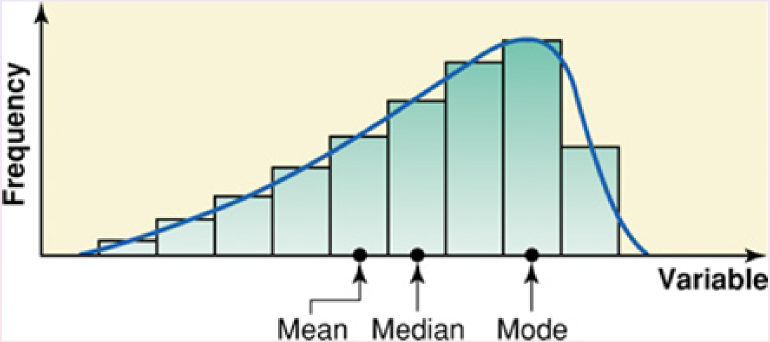
Relationships



Relationships



Relationships



Trimmed mean

Because mean value is not robust, easily impacted by outlier. If outliers present we can use either median or the following trimmed mean.

The k% trimmed mean is the mean of the data values after cutting off k% of the values from each end of the sorted data.

Example 3: The following are the money spent (in dollars) on books in 2015 by 10 randomly selected students from a small college.

890 1354 1861 1644 87 5403 1429 1993 938 2176

Find the 10% trimmed mean.

Trimmed mean

Because mean value is not robust, easily impacted by outlier. If outliers present we can use either median or the following trimmed mean.

The $k\%$ trimmed mean is the mean of the data values after cutting off $k\%$ of the values from each end of the sorted data.

Example 3: The following are the money spent (in dollars) on books in 2015 by 10 randomly selected students from a small college.

890 1354 1861 1644 87 5403 1429 1993 938 2176

Find the 10% trimmed mean.

Trimmed mean

Because mean value is not robust, easily impacted by outlier. If outliers present we can use either median or the following trimmed mean.

The $k\%$ trimmed mean is the mean of the data values after cutting off $k\%$ of the values from each end of the sorted data.

Example 3: The following are the money spent (in dollars) on books in 2015 by 10 randomly selected students from a small college.

890 1354 1861 1644 87 5403 1429 1993 938 2176

Find the 10% trimmed mean.

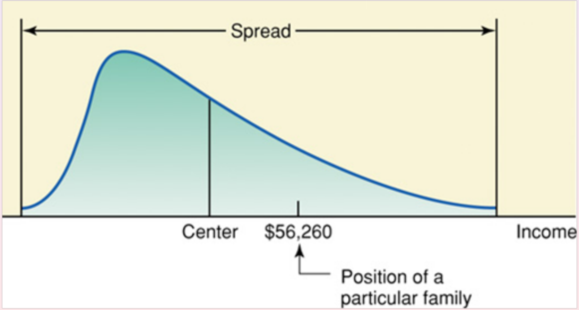
Trimmed mean

Solution The sorted data are
87 890 938 1354 1429 1644 1861 1993 2176 5403
Number of 10% value equals $n * k/100 = 10 * 10/100 = 1$. So we drop one value from each end and we have the trimmed data
890 938 1354 1429 1644 1861 1993 2176
Then the 10% trimmed mean is \$1535.625=(890 + 938+ 1354+ 1429+ 1644+ 1861 +1993+ 2176)/8.

Outline

Range

$$\text{Range} = \text{Max} - \text{Min}$$



Variance Measures Dispersion from the mean

Let $X_1, X_2, \ldots, X_n$ be a sample size n. Let $x_1, x_2, \ldots, x_n$ be the values of the sample data.

$$\text{Sample Variance } S^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$$

$$\text{The Value of Sample Variance } s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

$$\text{The Value of Sample Standard Deviation } s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

Variance Measures Dispersion from the mean

Let $X_1, X_2, \ldots, X_n$ be a sample size n. Let $x_1, x_2, \ldots, x_n$ be the values of the sample data.

$$\text{Sample Variance } S^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$$

$$\text{The Value of Sample Variance } s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

$$\text{The Value of Sample Standard Deviation } s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

Variance Measures Dispersion from the mean

Let $X_1, X_2, \ldots, X_n$ be a sample size n. Let $x_1, x_2, \ldots, x_n$ be the values of the sample data.

$$\text{Sample Variance } S^2 = \frac{\sum(X_i - \bar{X})^2}{n-1}$$

$$\text{The Value of Sample Variance } s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

$$\text{The Value of Sample Standard Deviation } s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

Variance Measures Dispersion from the mean

Let $X_1, X_2, \ldots, X_n$ be a sample size n. Let $x_1, x_2, \ldots, x_n$ be the values of the sample data.

$$\text{Sample Variance } S^2 = \frac{\sum(X_i - \bar{X})^2}{n-1}$$

$$\text{The Value of Sample Variance } s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

$$\text{The Value of Sample Standard Deviation } s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

Short-Cut Formulas

$$\text{Sample Variance } s^2 = \frac{\sum x_i^2 - n\bar{x}^2}{n-1}$$

$$\text{Sample Standard Deviation } s = \sqrt{s^2}$$

Short-Cut Formulas

Sample Variance $s^2 = \dfrac{\sum x_i^2 - n\bar{x}^2}{n-1}$

Sample Standard Deviation $s = \sqrt{s^2}$

Calculation using Short-Cut Formulas

Table 3.6

x
46.5
18.0
16.0
7.8
7.2

$n = 5$

Calculation using Short-Cut Formulas

Table 3.6

x
46.5
18.0
16.0
7.8
7.2
$\Sigma x = 95.5$

$n = 5$

Calculation using Short-Cut Formulas

Table 3.6

x
46.5
18.0
16.0
7.8
7.2
$\Sigma x = 95.5$

$n = 5$

$$\bar{x} = \frac{\sum x_i}{n} = 95.5/5 = 19.1$$

Calculation using Short-Cut Formulas

Table 3.6

x	x^2
46.5	2162.25
18.0	324.00
16.0	256.00
7.8	60.84
7.2	51.84
$\Sigma x = 95.5$	

$n = 5$

$$\bar{x} = \frac{\sum x_i}{n} = 95.5/5 = 19.1$$

Calculation using Short-Cut Formulas

Table 3.6

x	x^2
46.5	2162.25
18.0	324.00
16.0	256.00
7.8	60.84
7.2	51.84
$\Sigma x = 95.5$	$\Sigma x^2 = 2854.93$

n = 5

$$\bar{x} = \frac{\sum x_i}{n} = 95.5/5 = 19.1$$

Calculation using Short-Cut Formulas

Table 3.6

x	x^2
46.5	2162.25
18.0	324.00
16.0	256.00
7.8	60.84
7.2	51.84
$\Sigma x = 95.5$	$\Sigma x^2 = 2854.93$

n = 5

$$\bar{x} = \frac{\sum x_i}{n} = 95.5/5 = 19.1$$

$$s^2 = \frac{\sum x_i^2 - n\bar{x}^2}{n-1}$$

Calculation using Short-Cut Formulas

Table 3.6

x	x^2
46.5	2162.25
18.0	324.00
16.0	256.00
7.8	60.84
7.2	51.84
$\Sigma x = 95.5$	$\Sigma x^2 = 2854.93$

n = 5

$$\bar{x} = \frac{\sum x_i}{n} = 95.5/5 = 19.1$$

$$s^2 = \frac{\sum x_i^2 - n\bar{x}^2}{n-1}$$

= (2854.93 - 5*19.1^2)/4

Calculation using Short-Cut Formulas

Table 3.6

x	x^2
46.5	2162.25
18.0	324.00
16.0	256.00
7.8	60.84
7.2	51.84
$\Sigma x = 95.5$	$\Sigma x^2 = 2854.93$

$n = 5$

$$\bar{x} = \frac{\sum x_i}{n} = 95.5/5 = 19.1$$

$$s^2 = \frac{\sum x_i^2 - n\bar{x}^2}{n-1}$$

= (2854.93 - 5*19.1^2)/4

= 257.72

Calculation using Short-Cut Formulas

Table 3.6

x	x^2
46.5	2162.25
18.0	324.00
16.0	256.00
7.8	60.84
7.2	51.84
$\Sigma x = 95.5$	$\Sigma x^2 = 2854.93$

n = 5

$$\bar{x} = \frac{\sum x_i}{n} = 95.5/5 = 19.1$$

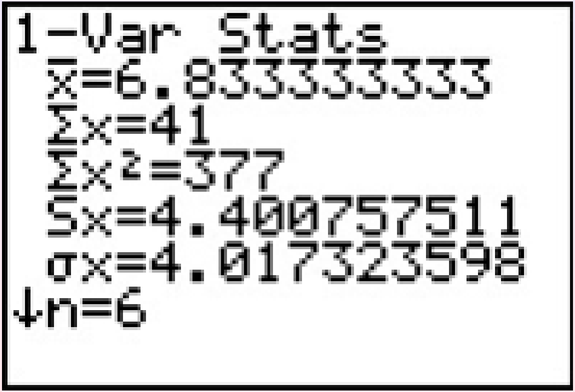
$$s^2 = \frac{\sum x_i^2 - n\bar{x}^2}{n-1}$$

= (2854.93 - 5*19.1^2)/4

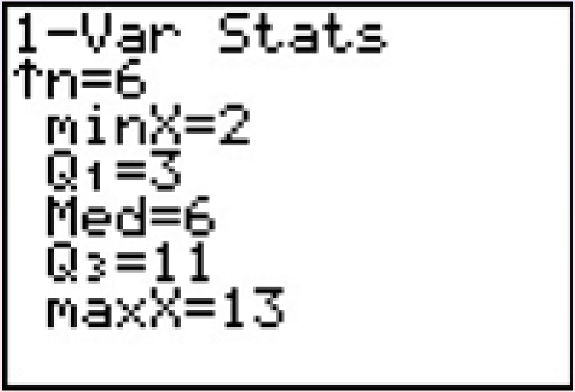
= 257.72

$$s = \sqrt{257.72} = 16.0536$$

Calculation using TI-83/84



Calculation using TI-83/84



Calculation using Excel

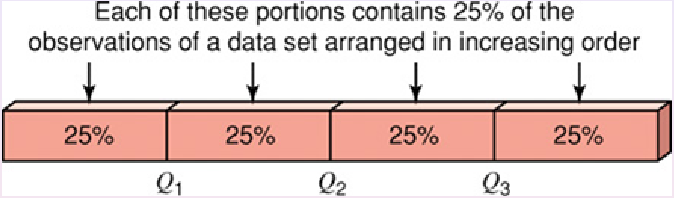
	A	B	C
1	Data	Average	
2			
3	2	=average(A3:A8)	
4	3		
5	5		
6	7		
7	11		
8	13		

Calculation using Excel

	A	B	C
1	Data	Average	
2			
3	2	6.833333	
4	3		
5	5		
6	7		
7	11		
8	13		

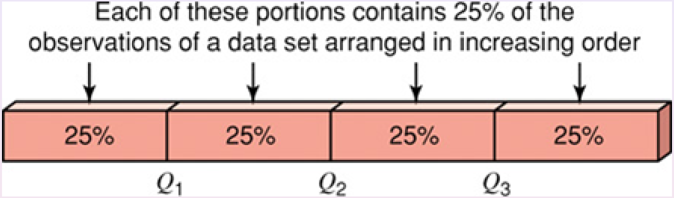
Outline

Quartiles



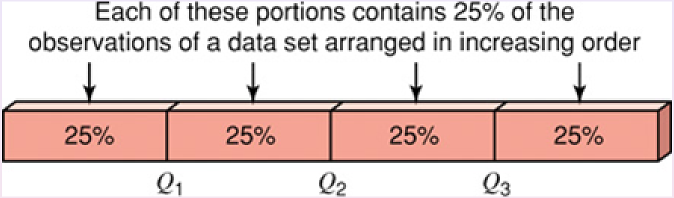
- First quartile: Q_1 the 25th percentile;
- Second quartile: Q_2 the 50th percentile, also the median;
- Third quartile: Q_3 the 75th percentile;
- 5 number summary: Min, Q_1, Q_2, Q_3, Max.

Quartiles



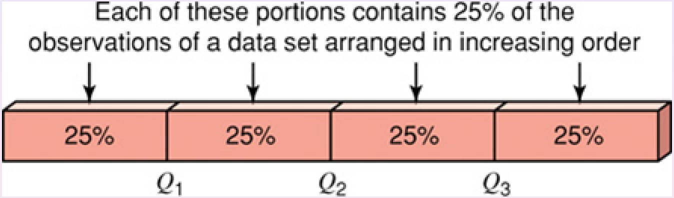
- First quartile: Q_1 the 25th percentile;
- Second quartile: Q_2 the 50th percentile, also the median;
- Third quartile: Q_3 the 75th percentile;
- 5 number summary: Min, Q_1, Q_2, Q_3, Max.

Quartiles



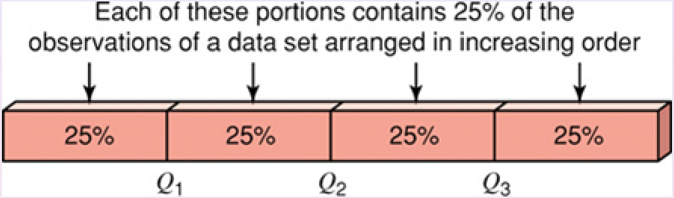
- First quartile: Q_1 the 25th percentile;
- Second quartile: Q_2 the 50th percentile, also the median;
- Third quartile: Q_3 the 75th percentile;
- 5 number summary: Min, Q_1, Q_2, Q_3, Max.

Quartiles



- First quartile: Q_1 the 25th percentile;
- Second quartile: Q_2 the 50th percentile, also the median;
- Third quartile: Q_3 the 75th percentile;
- 5 number summary: Min, Q_1, Q_2, Q_3, Max.

Quartiles



- First quartile: Q_1 the 25th percentile;
- Second quartile: Q_2 the 50th percentile, also the median;
- Third quartile: Q_3 the 75th percentile;
- 5 number summary: Min, Q_1, Q_2, Q_3, Max.

Interquartile-range:

$$IQR = Q_3 - Q_1;$$

Using TI-83

- First input the data as a list by Pressing button $\boxed{\text{STAT}}$ and select "1: Edit..." then press $\boxed{\text{ENTER}}$.
- Second, Press button $\boxed{\text{STAT}}$, then choose "CALC" and "1-Var Stats" then $\boxed{\text{ENTER}}$.
- Press $\boxed{\text{ENTER}}$ again you will get $\bar{x}$, $S_X = s$, $\sigma_X = \sigma$, and 5-number summary $minX$, Q_1, $Med = Q_2$, Q_3 and $maxX$.

Interquartile-range:

$$IQR = Q_3 - Q_1;$$

Using TI-83

- First input the data as a list by Pressing button $\boxed{\text{STAT}}$ and select “1: Edit...” then press $\boxed{\text{ENTER}}$.
- Second, Press button $\boxed{\text{STAT}}$, then choose “CALC” and “1-Var Stats” then $\boxed{\text{ENTER}}$.
- Press $\boxed{\text{ENTER}}$ again you will get $\bar{x}$, $S_X = s$, $\sigma_X = \sigma$, and 5-number summary $minX$, Q_1, $Med = Q_2$, Q_3 and $maxX$.

Interquartile-range:

$$IQR = Q_3 - Q_1;$$

Using TI-83

- First input the data as a list by Pressing button $\boxed{\text{STAT}}$ and select "1: Edit..." then press $\boxed{\text{ENTER}}$.
- Second, Press button $\boxed{\text{STAT}}$, then choose "CALC" and "1-Var Stats" then $\boxed{\text{ENTER}}$.
- Press $\boxed{\text{ENTER}}$ again you will get $\bar{x}$, $S_X = s$, $\sigma_X = \sigma$, and 5-number summary $minX$, Q_1, $Med = Q_2$, Q_3 and $maxX$.

Simplified Method for Finding Quartiles

Step 1. Sort the data in increasing order;

Step 2. Q_2 is the median of the complete data.

Step 3. Q_1 is the median of the subdataset that are smaller than or equal to Q_2, and

Step 4. Q_3 is the median of the subdataset that are greater than or equal to Q_2.

Simplified Method for Finding Quartiles

Step 1. Sort the data in increasing order;

Step 2. Q_2 is the median of the complete data.

Step 3. Q_1 is the median of the subdataset that are smaller than or equal to Q_2, and

Step 4. Q_3 is the median of the subdataset that are greater than or equal to Q_2.

Simplified Method for Finding Quartiles

Step 1. Sort the data in increasing order;

Step 2. Q_2 is the median of the complete data.

Step 3. Q_1 is the median of the subdataset that are smaller than or equal to Q_2, and

Step 4. Q_3 is the median of the subdataset that are greater than or equal to Q_2.

Simplified Method for Finding Quartiles

Step 1. Sort the data in increasing order;

Step 2. Q_2 is the median of the complete data.

Step 3. Q_1 is the median of the subdataset that are smaller than or equal to Q_2, and

Step 4. Q_3 is the median of the subdataset that are greater than or equal to Q_2.

Simplified Method for Finding Quartiles

Step 1. Sort the data in increasing order;

Step 2. Q_2 is the median of the complete data.

Step 3. Q_1 is the median of the subdataset that are smaller than or equal to Q_2, and

Step 4. Q_3 is the median of the subdataset that are greater than or equal to Q_2.

Example

City	Number of Car Thefts
Phoenix-Mesa, Arizona	40,769
Washington, D.C.	33,956
Miami, Florida	21,088
Atlanta, Georgia	29,920
Chicago, Illinois	42,082
Kansas City, Kansas	11,669
Baltimore, Maryland	13,435
Detroit, Michigan	40,197
St. Louis, Missouri	18,215
Las Vegas, Nevada	18,103
Newark, New Jersey	14,413
Dallas, Texas	26,343

Source: National Insurance Crime Bureau.

(a) Find the values of the three quartiles. Where does the number of car thefts of 40,197 fall in relation to these quartiles?

(b) Find the interquartile range.

Example

City	Number of Car Thefts
Phoenix-Mesa, Arizona	40,769
Washington, D.C.	33,956
Miami, Florida	21,088
Atlanta, Georgia	29,920
Chicago, Illinois	42,082
Kansas City, Kansas	11,669
Baltimore, Maryland	13,435
Detroit, Michigan	40,197
St. Louis, Missouri	18,215
Las Vegas, Nevada	18,103
Newark, New Jersey	14,413
Dallas, Texas	26,343

Source: National Insurance Crime Bureau.

(a) Find the values of the three quartiles. Where does the number of car thefts of 40,197 fall in relation to these quartiles?

(b) Find the interquartile range.

Example

City	Number of Car Thefts
Phoenix-Mesa, Arizona	40,769
Washington, D.C.	33,956
Miami, Florida	21,088
Atlanta, Georgia	29,920
Chicago, Illinois	42,082
Kansas City, Kansas	11,669
Baltimore, Maryland	13,435
Detroit, Michigan	40,197
St. Louis, Missouri	18,215
Las Vegas, Nevada	18,103
Newark, New Jersey	14,413
Dallas, Texas	26,343

Source: National Insurance Crime Bureau.

(a) Find the values of the three quartiles. Where does the number of car thefts of 40,197 fall in relation to these quartiles?

(b) Find the interquartile range.

Solution of Example

Values less than the median: 11,669 13,435 14,413 18,103 18,215 21,088

Values greater than the median: 26,343 29,920 33,956 40,197 40,769 42,082

$$Q_1 = \frac{14{,}413 + 18{,}103}{2} = \mathbf{16{,}258}$$

$$Q_2 = \frac{21{,}088 + 26{,}343}{2} = \mathbf{23{,}715.50}$$

Also the median

$$Q_3 = \frac{33{,}956 + 40{,}197}{2} = \mathbf{37{,}076.50}$$

(a) The three quartiles are $Q_1 = 16,258$, $Q_2 = 23,715.5$, and $Q_3 = 37,076.5$. The number of car thefts of 40,197 falls in the top 25%.

(b) The interquartile range:
$IQR = Q_3 - Q_1 = 37,076.50 - 16,258 = 20,818.50$ car thefts.

Solution of Example

Values less than the median						Values greater than the median					
11,669	13,435	14,413	18,103	18,215	21,088	26,343	29,920	33,956	40,197	40,769	42,082

$$Q_1 = \frac{14{,}413 + 18{,}103}{2} = \mathbf{16{,}258}$$

$$Q_2 = \frac{21{,}088 + 26{,}343}{2} = \mathbf{23{,}715.50}$$

Also the median

$$Q_3 = \frac{33{,}956 + 40{,}197}{2} = \mathbf{37{,}076.50}$$

(a) The three quartiles are $Q_1 = 16,258$, $Q_2 = 23,715.5$, and $Q_3 = 37,076.5$. The number of car thefts of 40,197 falls in the top 25%.

(b) The interquartile range:
$IQR = Q_3 - Q_1 = 37,076.50 - 16,258 = 20,818.50$ car thefts.

Solution of Example

Values less than the median | Values greater than the median

11,669 13,435 14,413 18,103 18,215 21,088 | 26,343 29,920 33,956 40,197 40,769 42,082

$$Q_1 = \frac{14{,}413 + 18{,}103}{2} = \mathbf{16{,}258}$$

$$Q_2 = \frac{21{,}088 + 26{,}343}{2} = \mathbf{23{,}715.50}$$

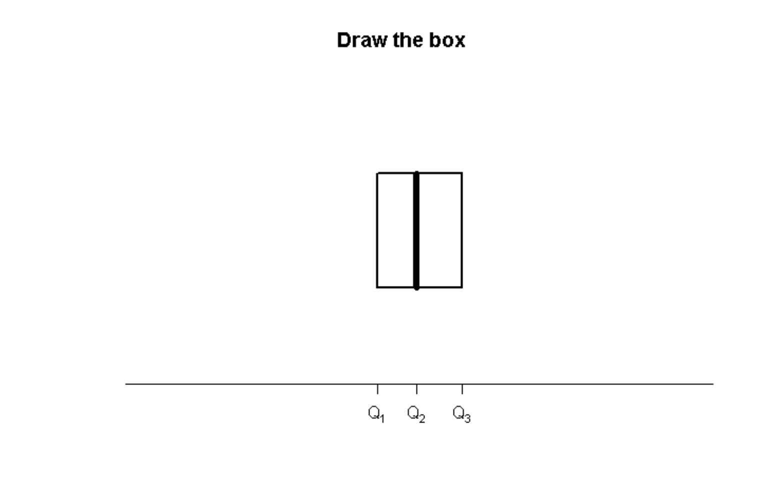
Also the median

$$Q_3 = \frac{33{,}956 + 40{,}197}{2} = \mathbf{37{,}076.50}$$

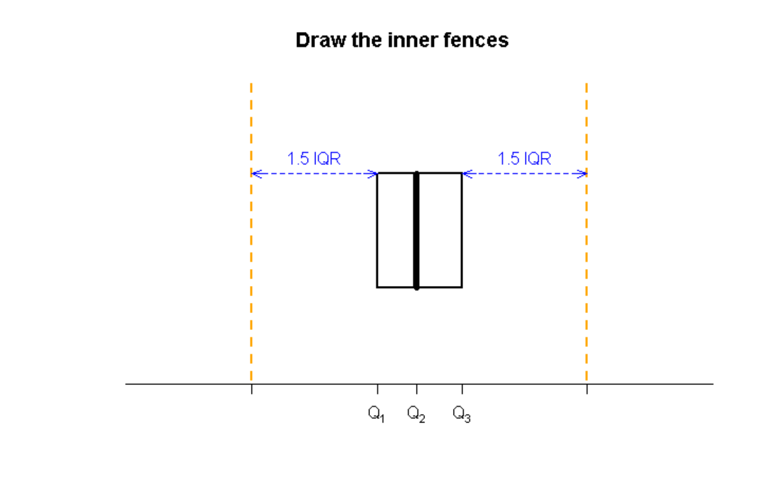
(a) The three quartiles are $Q_1 = 16,258$, $Q_2 = 23,715.5$, and $Q_3 = 37,076.5$. The number of car thefts of 40,197 falls in the top 25%.

(b) The interquartile range: $IQR = Q_3 - Q_1 = 37,076.50 - 16,258 = 20,818.50$ car thefts.

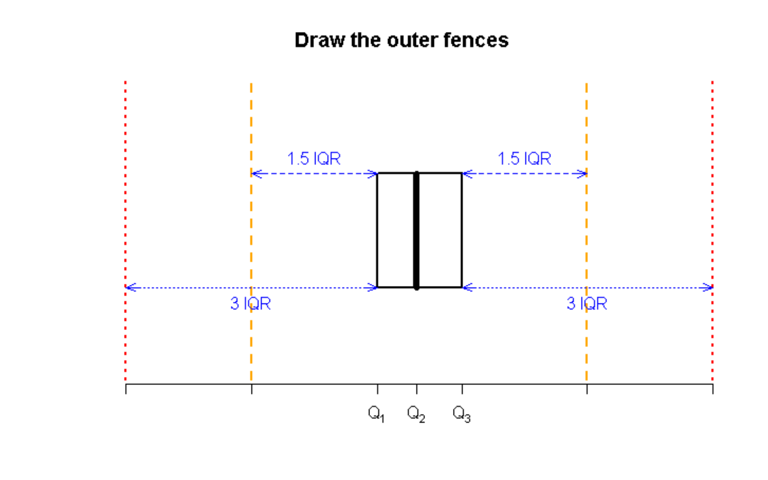
Box-and-Whisker Plot



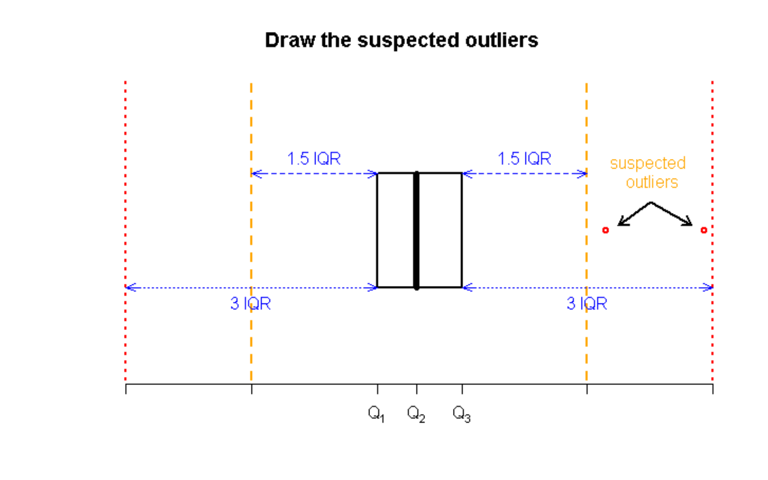
Box-and-Whisker Plot



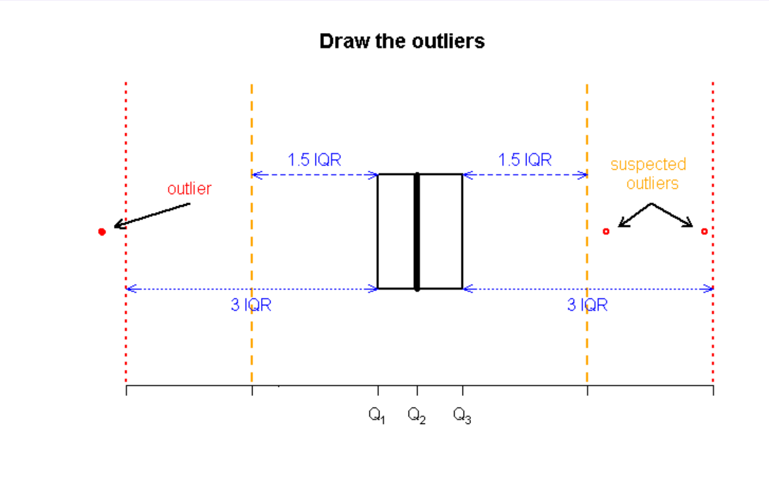
Box-and-Whisker Plot



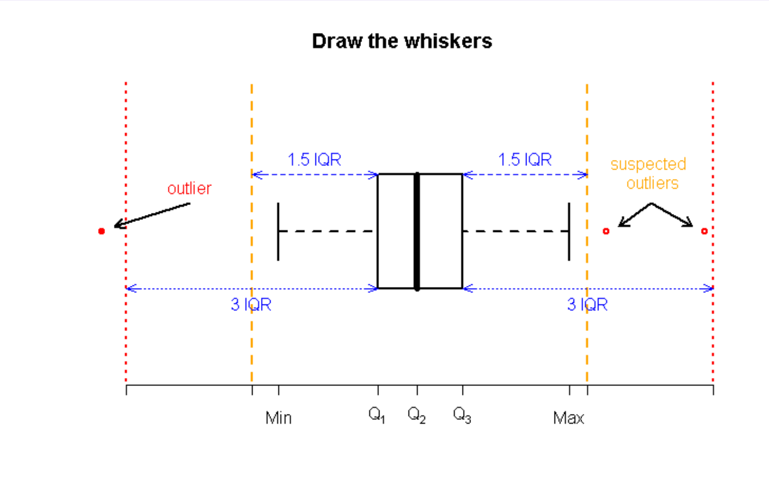
Box-and-Whisker Plot



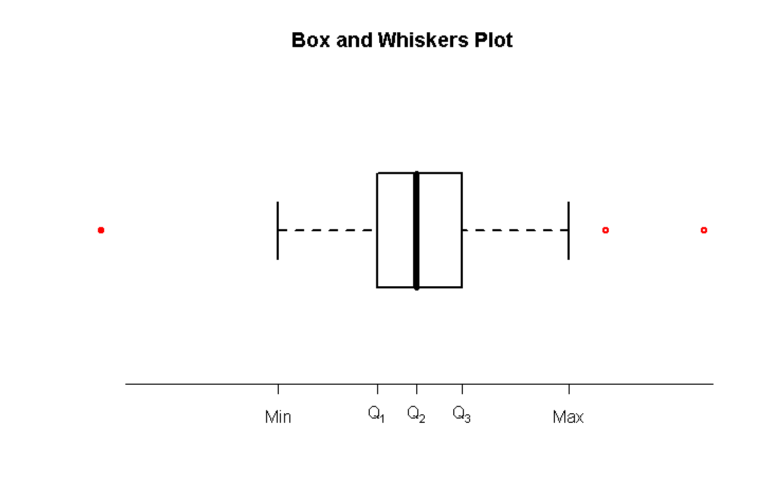
Box-and-Whisker Plot



Box-and-Whisker Plot



Box-and-Whisker Plot



Example of Box-and-Whisker Plot

Example: The following data are the incomes(in thousands of dollars) for a sample of 12 households.
35, 29, 44, 72, 34, 64, 41, 50, 54, 104, 39, 58
Construct a box-and-whisker plot.

Solution: (1) The sorted data:
29, 34, 35, 39, 41, 44, 50, 54, 58, 64, 72, 104
(2) $Q_2 = (44 + 50)/2 = 47$, $Q_1 = (35 + 39)/2 = 37$,
$Q_3 = (58 + 64)/2 = 61$
(3) $IQR = Q_3 - Q_1 = 61 - 37 = 24$.
(4) $1.5 \times IQR = 1.5 \times 24 = 36$.
(4) $3 \times IQR = 3 \times 24 = 72$.

Example of Box-and-Whisker Plot

Example: The following data are the incomes(in thousands of dollars) for a sample of 12 households.
35, 29, 44, 72, 34, 64, 41, 50, 54, 104, 39, 58
Construct a box-and-whisker plot.
Solution: (1) The sorted data:
29, 34, 35, 39, 41, 44, 50, 54, 58, 64, 72, 104
(2) $Q_2 = (44 + 50)/2 = 47$, $Q_1 = (35 + 39)/2 = 37$,
$Q_3 = (58 + 64)/2 = 61$
(3) $IQR = Q_3 - Q_1 = 61 - 37 = 24$.
(4) $1.5 \times IQR = 1.5 \times 24 = 36$.
(4) $3 \times IQR = 3 \times 24 = 72$.

Example of Box-and-Whisker Plot

Example: The following data are the incomes(in thousands of dollars) for a sample of 12 households.
35, 29, 44, 72, 34, 64, 41, 50, 54, 104, 39, 58
Construct a box-and-whisker plot.
Solution: (1) The sorted data:
29, 34, 35, 39, 41, 44, 50, 54, 58, 64, 72, 104
(2) $Q_2 = (44 + 50)/2 = 47$, $Q_1 = (35 + 39)/2 = 37$,
$Q_3 = (58 + 64)/2 = 61$
(3) $IQR = Q_3 - Q_1 = 61 - 37 = 24$.
(4) $1.5 \times IQR = 1.5 \times 24 = 36$.
(4) $3 \times IQR = 3 \times 24 = 72$.

Example of Box-and-Whisker Plot

Example: The following data are the incomes(in thousands of dollars) for a sample of 12 households.
35, 29, 44, 72, 34, 64, 41, 50, 54, 104, 39, 58
Construct a box-and-whisker plot.
Solution: (1) The sorted data:
29, 34, 35, 39, 41, 44, 50, 54, 58, 64, 72, 104
(2) $Q_2 = (44 + 50)/2 = 47$, $Q_1 = (35 + 39)/2 = 37$,
$Q_3 = (58 + 64)/2 = 61$
(3) $IQR = Q_3 - Q_1 = 61 - 37 = 24$.
(4) $1.5 \times IQR = 1.5 \times 24 = 36$.
(4) $3 \times IQR = 3 \times 24 = 72$.

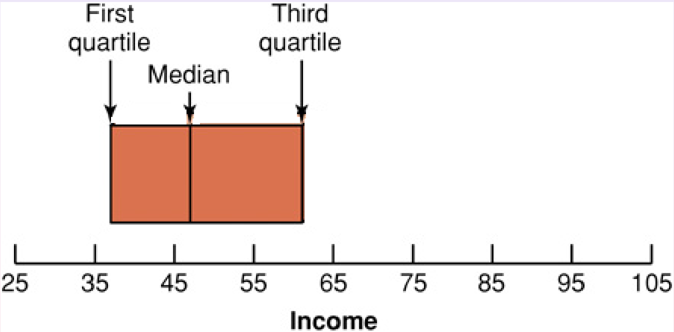
Example of Box-and-Whisker Plot

Example: The following data are the incomes(in thousands of dollars) for a sample of 12 households.
35, 29, 44, 72, 34, 64, 41, 50, 54, 104, 39, 58
Construct a box-and-whisker plot.
Solution: (1) The sorted data:
29, 34, 35, 39, 41, 44, 50, 54, 58, 64, 72, 104
(2) $Q_2 = (44 + 50)/2 = 47$, $Q_1 = (35 + 39)/2 = 37$,
$Q_3 = (58 + 64)/2 = 61$
(3) $IQR = Q_3 - Q_1 = 61 - 37 = 24$.
(4) $1.5 \times IQR = 1.5 \times 24 = 36$.
(4) $3 \times IQR = 3 \times 24 = 72$.

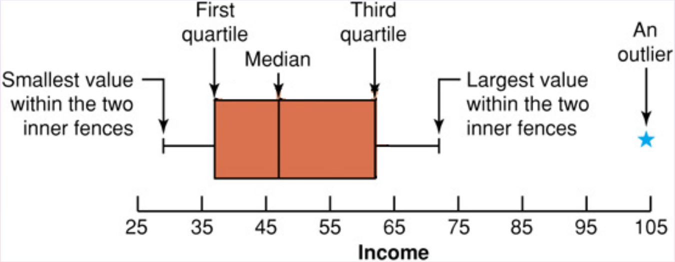
Example of Box-and-Whisker Plot

Example: The following data are the incomes(in thousands of dollars) for a sample of 12 households.
35, 29, 44, 72, 34, 64, 41, 50, 54, 104, 39, 58
Construct a box-and-whisker plot.
Solution: (1) The sorted data:
29, 34, 35, 39, 41, 44, 50, 54, 58, 64, 72, 104
(2) $Q_2 = (44 + 50)/2 = 47$, $Q_1 = (35 + 39)/2 = 37$,
$Q_3 = (58 + 64)/2 = 61$
(3) $IQR = Q_3 - Q_1 = 61 - 37 = 24$.
(4) $1.5 \times IQR = 1.5 \times 24 = 36$.
(4) $3 \times IQR = 3 \times 24 = 72$.

Example of Box-and-Whisker Plot



Example of Box-and-Whisker Plot



Outline

Relative Frequency and Histogram

To describe continuous-type data, we group the data values into classes (intervals) and count the (relative) frequency of the data values in each class.

Frequency Table and Histogram

1. Find *Min* & *Max* values and **range** $R = Max - Min$;
2. Find k non-overlapping intervals (class intervals) of equal length h by the endpoints (**class boundaries**)

$$c_0 < c_1 < c_2 < \cdots < c_{k-1} < c_k$$

   The endpoints should contain one more decimal place than the data values and $c_0 \lessapprox Min < Max \lessapprox c_k$.
3. Find the **class mark** for each class: the midpoint of the class interval: $m_i = \frac{c_{i-1}+c_i}{2}$
4. Calculate relative frequency (density) for each class

$$h(x) = \frac{f_i}{n(c_i - c_{i-1})}, \text{ for } c_{i-1} < x \leqslant c_i, \quad i = 1, 2, \ldots, k.$$

5. Suggested $k \approx R/h$, $h = 2IQR/n^{1/3}$.

Frequency Table and Histogram

1. Find *Min* & *Max* values and **range** $R = Max - Min$;
2. Find k non-overlapping intervals (class intervals) of equal length h by the endpoints (**class boundaries**)

$$c_0 < c_1 < c_2 < \cdots < c_{k-1} < c_k$$

The endpoints should contain one more decimal place than the data values and $c_0 \lessapprox Min < Max \lessapprox c_k$.

3. Find the **class mark** for each class: the midpoint of the class interval: $m_i = \frac{c_{i-1}+c_i}{2}$
4. Calculate relative frequency (density) for each class

$$h(x) = \frac{f_i}{n(c_i - c_{i-1})}, \text{ for } c_{i-1} < x \leq c_i, \quad i = 1, 2, \ldots, k.$$

5. Suggested $k \approx R/h$, $h = 2IQR/n^{1/3}$.

Frequency Table and Histogram

1. Find *Min* & *Max* values and **range** $R = Max - Min$;
2. Find k non-overlapping intervals (class intervals) of equal length h by the endpoints (**class boundaries**)

$$c_0 < c_1 < c_2 < \cdots < c_{k-1} < c_k$$

   The endpoints should contain one more decimal place than the data values and $c_0 \lessapprox Min < Max \lessapprox c_k$.
3. Find the **class mark** for each class: the midpoint of the class interval: $m_i = \frac{c_{i-1}+c_i}{2}$
4. Calculate relative frequency (density) for each class

$$h(x) = \frac{f_i}{n(c_i - c_{i-1})}, \text{ for } c_{i-1} < x \leq c_i, \quad i = 1, 2, \ldots, k.$$

5. Suggested $k \approx R/h$, $h = 2IQR/n^{1/3}$.

Frequency Table and Histogram

1. Find *Min* & *Max* values and **range** $R = Max - Min$;
2. Find k non-overlapping intervals (class intervals) of equal length h by the endpoints (**class boundaries**)
$$c_0 < c_1 < c_2 < \cdots < c_{k-1} < c_k$$
The endpoints should contain one more decimal place than the data values and $c_0 \lessapprox Min < Max \lessapprox c_k$.
3. Find the **class mark** for each class: the midpoint of the class interval: $m_i = \frac{c_{i-1}+c_i}{2}$
4. Calculate relative frequency (density) for each class
$$h(x) = \frac{f_i}{n(c_i - c_{i-1})}, \text{ for } c_{i-1} < x \leqslant c_i, \quad i = 1, 2, \ldots, k.$$
5. Suggested $k \approx R/h$, $h = 2IQR/n^{1/3}$.

Frequency Table and Histogram

1. Find Min & Max values and **range** $R = Max - Min$;
2. Find k non-overlapping intervals (class intervals) of equal length h by the endpoints (**class boundaries**)
$$c_0 < c_1 < c_2 < \cdots < c_{k-1} < c_k$$
The endpoints should contain one more decimal place than the data values and $c_0 \lessapprox Min < Max \lessapprox c_k$.
3. Find the **class mark** for each class: the midpoint of the class interval: $m_i = \frac{c_{i-1}+c_i}{2}$
4. Calculate relative frequency (density) for each class
$$h(x) = \frac{f_i}{n(c_i - c_{i-1})}, \text{ for } c_{i-1} < x \leqslant c_i, \quad i = 1, 2, \ldots, k.$$
5. Suggested $k \approx R/h$, $h = 2IQR/n^{1/3}$.

Relative Frequency

Example 6. The weights (in grams) of 50 nails:

8.05	8.31	8.51	8.56	8.66	8.76	8.85	8.90	9.20	9.34
8.24	8.36	8.51	8.57	8.69	8.79	8.85	8.93	9.21	9.40
8.27	8.38	8.51	8.58	8.69	8.79	8.85	8.98	9.21	9.41
8.27	8.41	8.55	8.58	8.71	8.82	8.88	9.08	9.25	9.42
8.29	8.43	8.56	8.59	8.73	8.82	8.88	9.15	9.26	9.63

1. $n = 50$, $Min = 8.05$, $Max = 9.63$, $R = Max - Min = 1.58$, $IQR = 0.4475$;
2. $h = \lceil 2\frac{IQR}{n^{1/3}} \rceil = 0.25$. $k = \lceil \frac{R}{h} \rceil = 7$.
3. Choose $c_0 = 7.995$. $c_i = c_0 + ih$, $i = 1, \ldots, k$. The class boundaries are 7.995 8.245 8.495 8.745 8.995 9.245 9.495 9.745
4. Calculate frequency f_i and density $h(x)$

Relative Frequency

Example 6. The weights (in grams) of 50 nails:

8.05	8.31	8.51	8.56	8.66	8.76	8.85	8.90	9.20	9.34
8.24	8.36	8.51	8.57	8.69	8.79	8.85	8.93	9.21	9.40
8.27	8.38	8.51	8.58	8.69	8.79	8.85	8.98	9.21	9.41
8.27	8.41	8.55	8.58	8.71	8.82	8.88	9.08	9.25	9.42
8.29	8.43	8.56	8.59	8.73	8.82	8.88	9.15	9.26	9.63

1. $n = 50$, $Min = 8.05$, $Max = 9.63$, $R = Max - Min = 1.58$, $IQR = 0.4475$;
2. $h = \lceil 2\frac{IQR}{n^{1/3}} \rceil = 0.25$. $k = \lceil \frac{R}{h} \rceil = 7$.
3. Choose $c_0 = 7.995$. $c_i = c_0 + ih$, $i = 1, \ldots, k$. The class boundaries are 7.995 8.245 8.495 8.745 8.995 9.245 9.495 9.745
4. Calculate frequency f_i and density $h(x)$

Relative Frequency

Example 6. The weights (in grams) of 50 nails:

8.05	8.31	8.51	8.56	8.66	8.76	8.85	8.90	9.20	9.34
8.24	8.36	8.51	8.57	8.69	8.79	8.85	8.93	9.21	9.40
8.27	8.38	8.51	8.58	8.69	8.79	8.85	8.98	9.21	9.41
8.27	8.41	8.55	8.58	8.71	8.82	8.88	9.08	9.25	9.42
8.29	8.43	8.56	8.59	8.73	8.82	8.88	9.15	9.26	9.63

1. $n = 50$, $Min = 8.05$, $Max = 9.63$, $R = Max - Min = 1.58$, $IQR = 0.4475$;
2. $h = \lceil 2\frac{IQR}{n^{1/3}} \rceil = 0.25$. $k = \lceil \frac{R}{h} \rceil = 7$.
3. Choose $c_0 = 7.995$. $c_i = c_0 + ih$, $i = 1, \ldots, k$. The class boundaries are 7.995 8.245 8.495 8.745 8.995 9.245 9.495 9.745
4. Calculate frequency f_i and density $h(x)$

Relative Frequency

Example 6. The weights (in grams) of 50 nails:

8.05	8.31	8.51	8.56	8.66	8.76	8.85	8.90	9.20	9.34
8.24	8.36	8.51	8.57	8.69	8.79	8.85	8.93	9.21	9.40
8.27	8.38	8.51	8.58	8.69	8.79	8.85	8.98	9.21	9.41
8.27	8.41	8.55	8.58	8.71	8.82	8.88	9.08	9.25	9.42
8.29	8.43	8.56	8.59	8.73	8.82	8.88	9.15	9.26	9.63

1. $n = 50$, $Min = 8.05$, $Max = 9.63$, $R = Max - Min = 1.58$, $IQR = 0.4475$;
2. $h = \lceil 2\frac{IQR}{n^{1/3}} \rceil = 0.25$. $k = \lceil \frac{R}{h} \rceil = 7$.
3. Choose $c_0 = 7.995$. $c_i = c_0 + ih$, $i = 1, \ldots, k$. The class boundaries are 7.995 8.245 8.495 8.745 8.995 9.245 9.495 9.745
4. Calculate frequency f_i and density $h(x)$

Relative Frequency

Example 6. The weights (in grams) of 50 nails:

8.05	8.31	8.51	8.56	8.66	8.76	8.85	8.90	9.20	9.34
8.24	8.36	8.51	8.57	8.69	8.79	8.85	8.93	9.21	9.40
8.27	8.38	8.51	8.58	8.69	8.79	8.85	8.98	9.21	9.41
8.27	8.41	8.55	8.58	8.71	8.82	8.88	9.08	9.25	9.42
8.29	8.43	8.56	8.59	8.73	8.82	8.88	9.15	9.26	9.63

1. $n = 50$, $Min = 8.05$, $Max = 9.63$, $R = Max - Min = 1.58$, $IQR = 0.4475$;
2. $h = \lceil 2\frac{IQR}{n^{1/3}} \rceil = 0.25$. $k = \lceil \frac{R}{h} \rceil = 7$.
3. Choose $c_0 = 7.995$. $c_i = c_0 + ih$, $i = 1, \ldots, k$. The class boundaries are 7.995 8.245 8.495 8.745 8.995 9.245 9.495 9.745
4. Calculate frequency f_i and density $h(x)$

Relative Frequency

Example 6. The weights (in grams) of 50 nails:

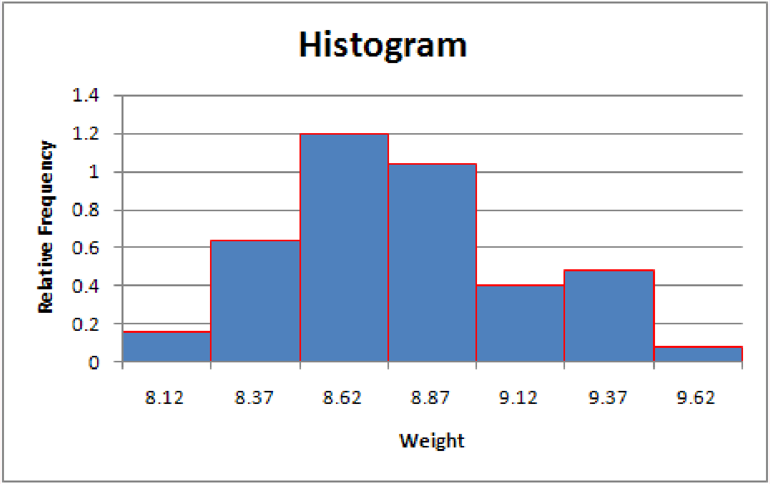
8.05	8.31	8.51	8.56	8.66	8.76	8.85	8.90	9.20	9.34
8.24	8.36	8.51	8.57	8.69	8.79	8.85	8.93	9.21	9.40
8.27	8.38	8.51	8.58	8.69	8.79	8.85	8.98	9.21	9.41
8.27	8.41	8.55	8.58	8.71	8.82	8.88	9.08	9.25	9.42
8.29	8.43	8.56	8.59	8.73	8.82	8.88	9.15	9.26	9.63

1. $n = 50$, $Min = 8.05$, $Max = 9.63$, $R = Max - Min = 1.58$, $IQR = 0.4475$;
2. $h = \lceil 2\frac{IQR}{n^{1/3}} \rceil = 0.25$. $k = \lceil \frac{R}{h} \rceil = 7$.
3. Choose $c_0 = 7.995$. $c_i = c_0 + ih$, $i = 1, \ldots, k$. The class boundaries are 7.995 8.245 8.495 8.745 8.995 9.245 9.495 9.745
4. Calculate frequency f_i and density $h(x)$

Frequency Table

Frequency Table					
	Class	Interval	Frequency	Rel. Freq.	Class
i	c(i-1)	c(i)	fi	h(x)	Mark
1	7.995	8.245	2	0.16	8.12
2	8.245	8.495	8	0.64	8.37
3	8.495	8.745	15	1.2	8.62
4	8.745	8.995	13	1.04	8.87
5	8.995	9.245	5	0.4	9.12
6	9.245	9.495	6	0.48	9.37
7	9.495	9.745	1	0.08	9.62

Histogram


Histogram
Relative Frequency
1.4
1.2
1
0.8
0.6
0.4
0.2
0
8.12
8.37
8.62
8.87
9.12
9.37
9.62
Weight

Example 7. Heights of 5000 female students

Lower Bound XL	Upper Bound XU	Frequency f	Class Mark	Rel. Freq. Density
59	60	0	59.5	0
60	61	90	60.5	0.018
61	62	170	61.5	0.034
62	63	460	62.5	0.092
63	64	750	63.5	0.15
64	65	970	64.5	0.194
65	66	760	65.5	0.152
66	67	640	66.5	0.128
67	68	440	67.5	0.088
68	69	320	68.5	0.064
69	70	220	69.5	0.044
70	71	180	70.5	0.036
71	72	0	71.5	0
Total		5000		1

Relative frequency density for class i with frequency f_i is

$$\frac{f_i}{Nw}$$

where w is the class width.

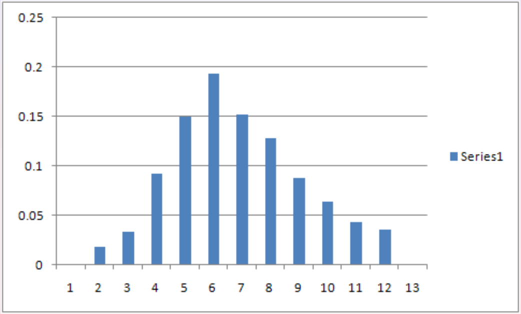
Histogram and Polygon Using Excel

Step 1: Make a frequency table in Excel:

Lower Bound XL	Upper Bound XU	Frequency f	Class Mark	Rel. Freq. Density
59	60	0	59.5	0
60	61	90	60.5	0.018
61	62	170	61.5	0.034
62	63	460	62.5	0.092
63	64	750	63.5	0.15
64	65	970	64.5	0.194
65	66	760	65.5	0.152
66	67	640	66.5	0.128
67	68	440	67.5	0.088
68	69	320	68.5	0.064
69	70	220	69.5	0.044
70	71	180	70.5	0.036
71	72	0	71.5	0
Total		5000		1

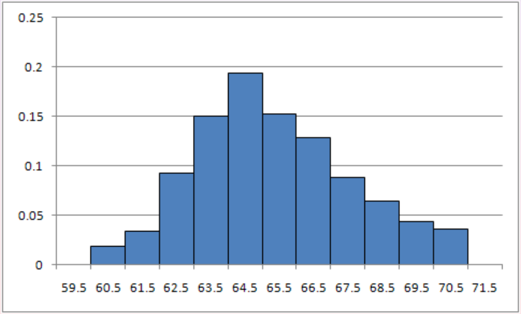
Histogram and Polygon Using Excel

Step 2: Highlight column "Relative Frequency Density", then insert a 2-d column chart:



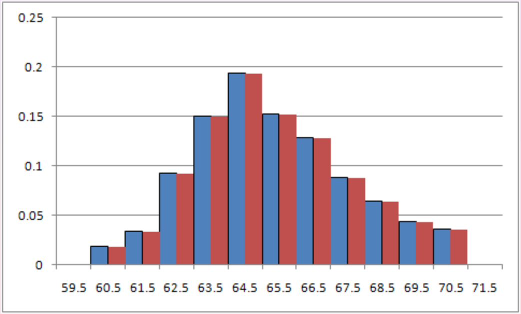
Histogram and Polygon Using Excel

Step 3: Edit the chart: reduce the gap width to 0% and replace "the Horizontal (Category) Axis Lable" with the "class mark" column.



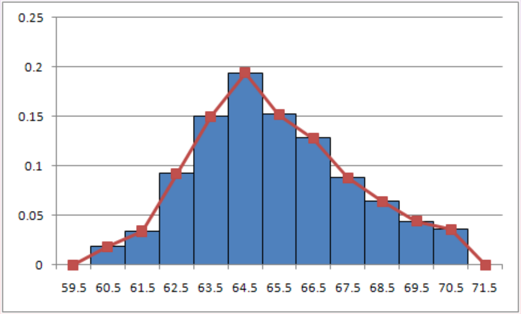
Histogram and Polygon Using Excel

Step 4: Right click the chart and choose “Select Data...”, add a new series using “Relative Frequency Density” column.

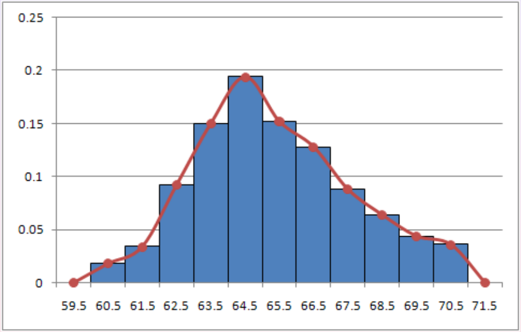


Histogram and Polygon Using Excel

Step 5: Right click the chart on the new column chart(one of the dark red bars) and choose "Change Series Chart Type...", select a line chart.

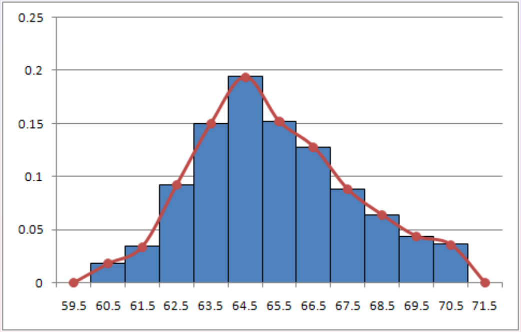


Relative Frequency Density Histogram and Polygon



Total area of all shaded rectangles equals 1.
The smooth curve is called the *probability density function* of random variable X, Height of randomly selected female student.

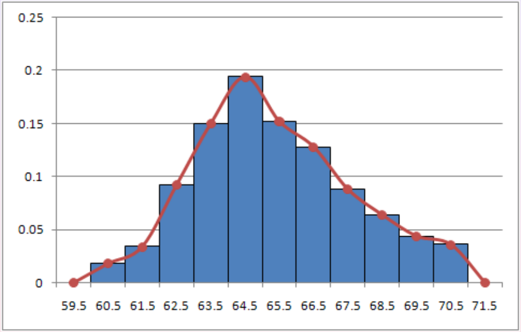
Relative Frequency Density Histogram and Polygon



Total area of all shaded rectangles equals 1.

The smooth curve is called the *probability density function* of random variable X, Height of randomly selected female student.

Relative Frequency Density Histogram and Polygon



Total area of all shaded rectangles equals 1.
The smooth curve is called the *probability density function* of random variable X, Height of randomly selected female student.