INTRODUCTION TO STATISTICS

David M. Lane. *et al.* Introduction to Statistics : pp. 6–56

Next section

Prerequisites

Statistics
Statistic

(i) Statistics is the course you are studying right now, also known as statistical analysis, or statistical inference. It is a field of study concerned with summarizing data, interpreting data, and making decisions based on data.

(ii) A quantity calculated in a sample to estimate a value in a population is called a "statistic".

- *The related term data science or data analysis stands for a study of processes and systems that extract knowledge or insights from data in various forms, either structured or unstructured.*
- *Data science is a continuation of some of the fields such as statistics, data mining, and predictive analytics.*

Prerequisites

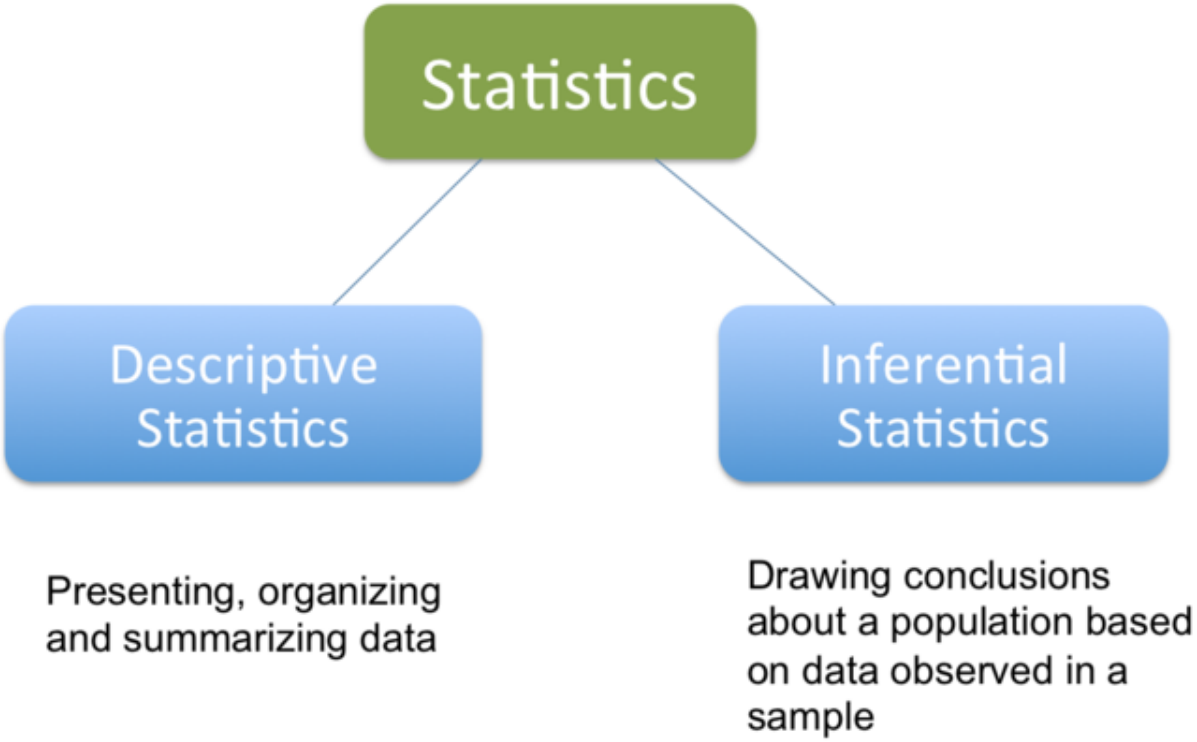
Statistics
Statistic

(i) Statistics is the course you are studying right now, also known as statistical analysis, or statistical inference. It is a field of study concerned with summarizing data, interpreting data, and making decisions based on data.

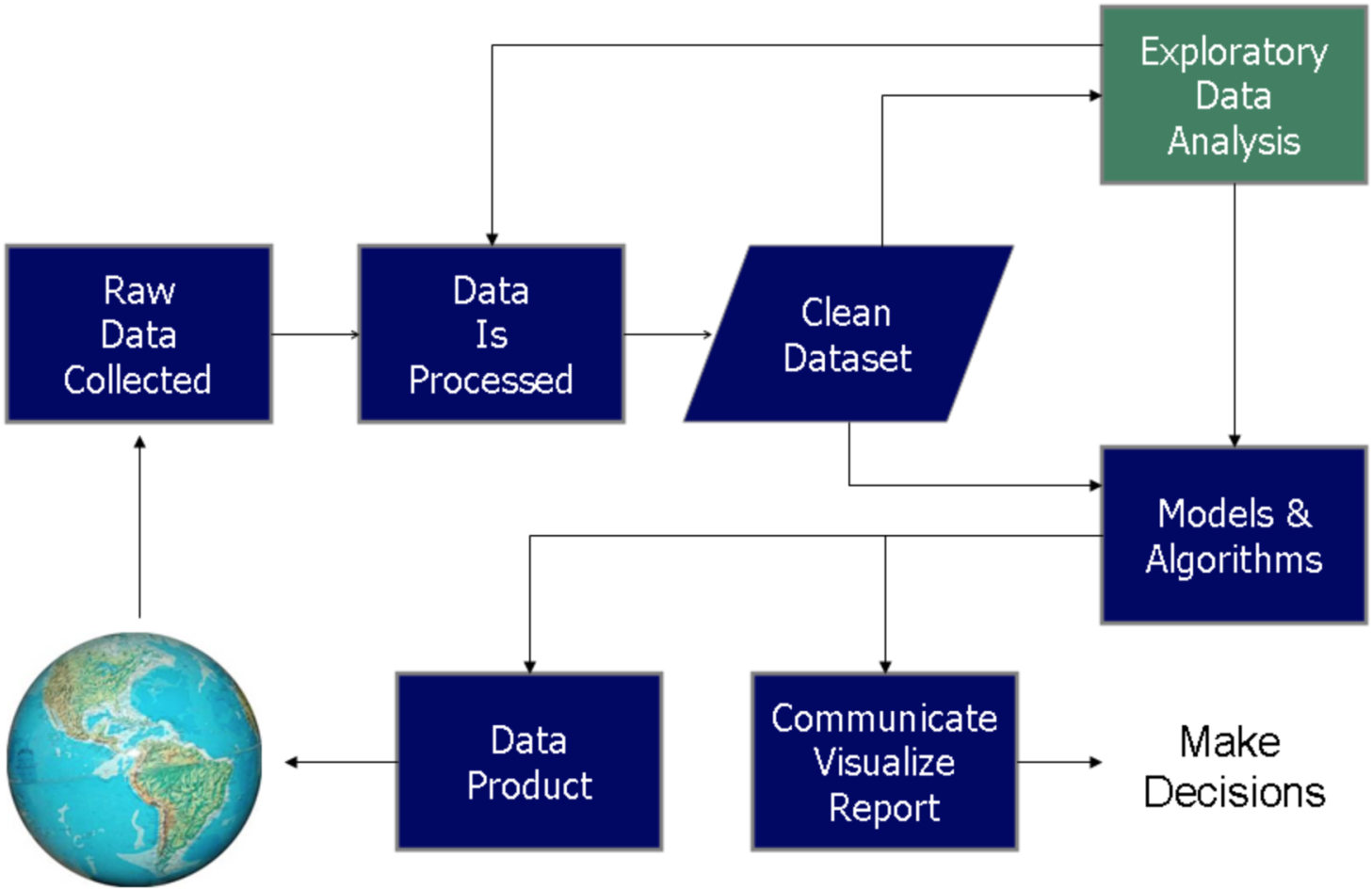
(ii) A quantity calculated in a sample to estimate a value in a population is called a "statistic".

- *The related term data science or data analysis stands for a study of processes and systems that extract knowledge or insights from data in various forms, either structured or unstructured.*
- *Data science is a continuation of some of the fields such as statistics, data mining, and predictive analytics.*

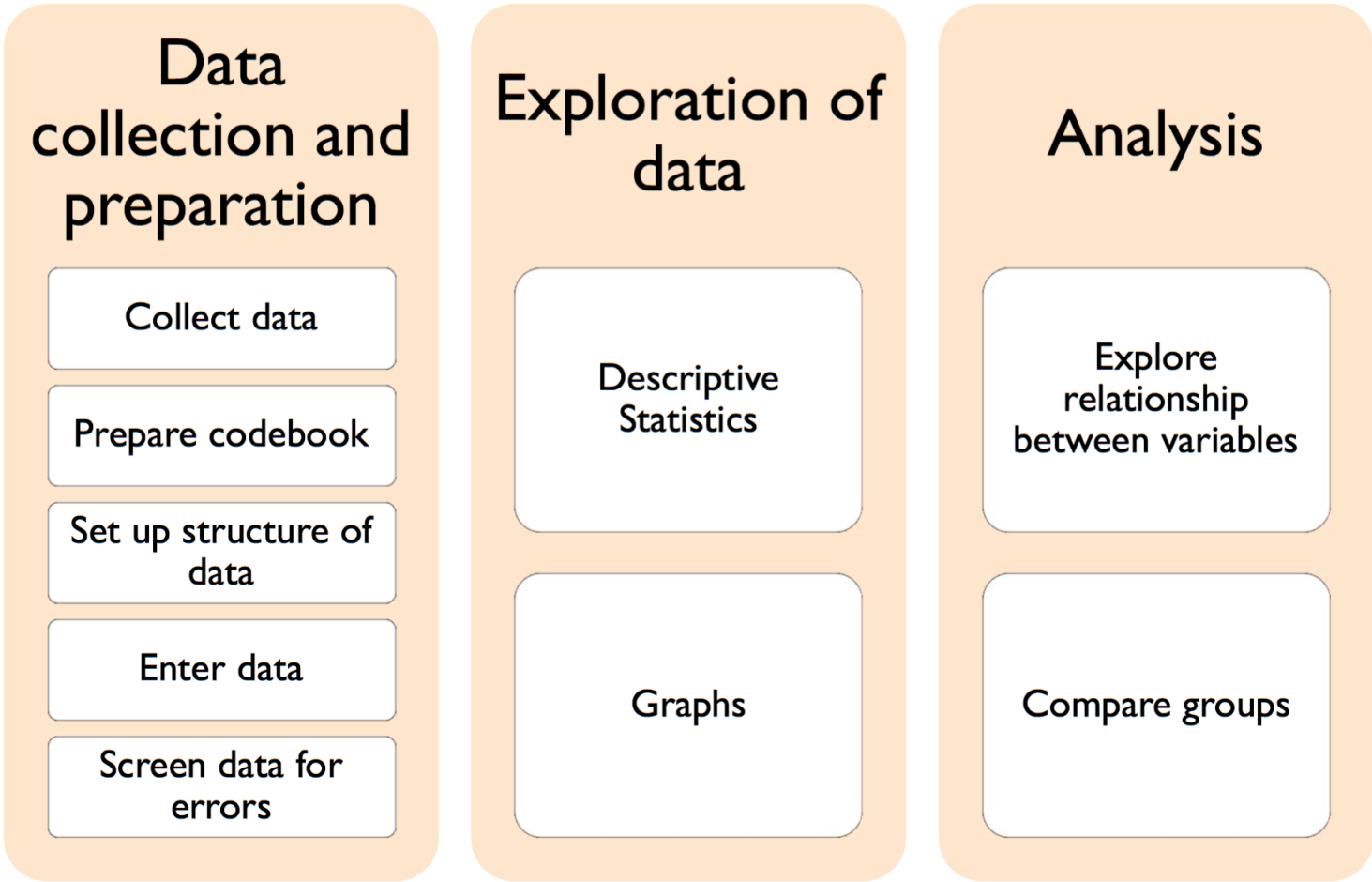
Structure of Statistics



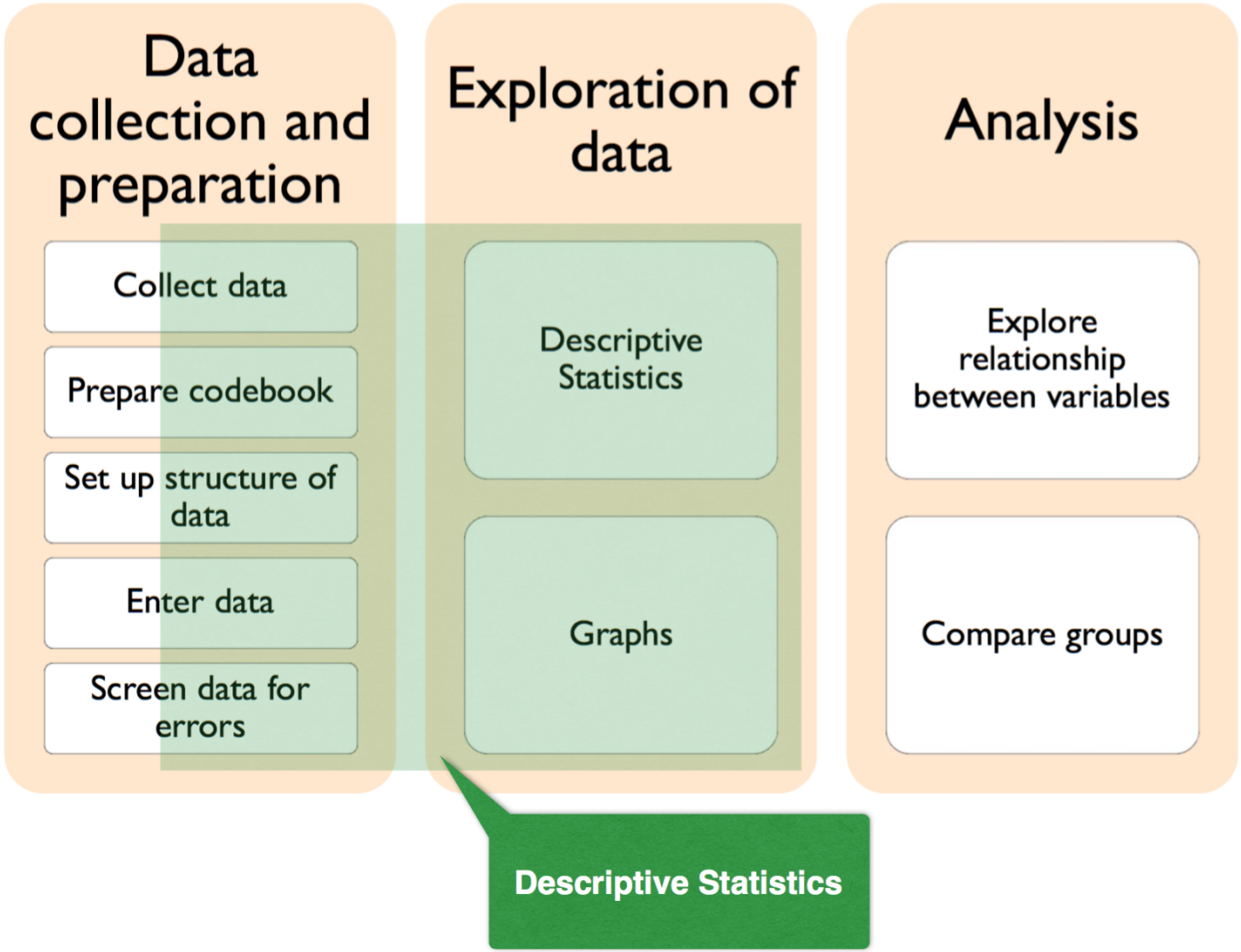
Data science process



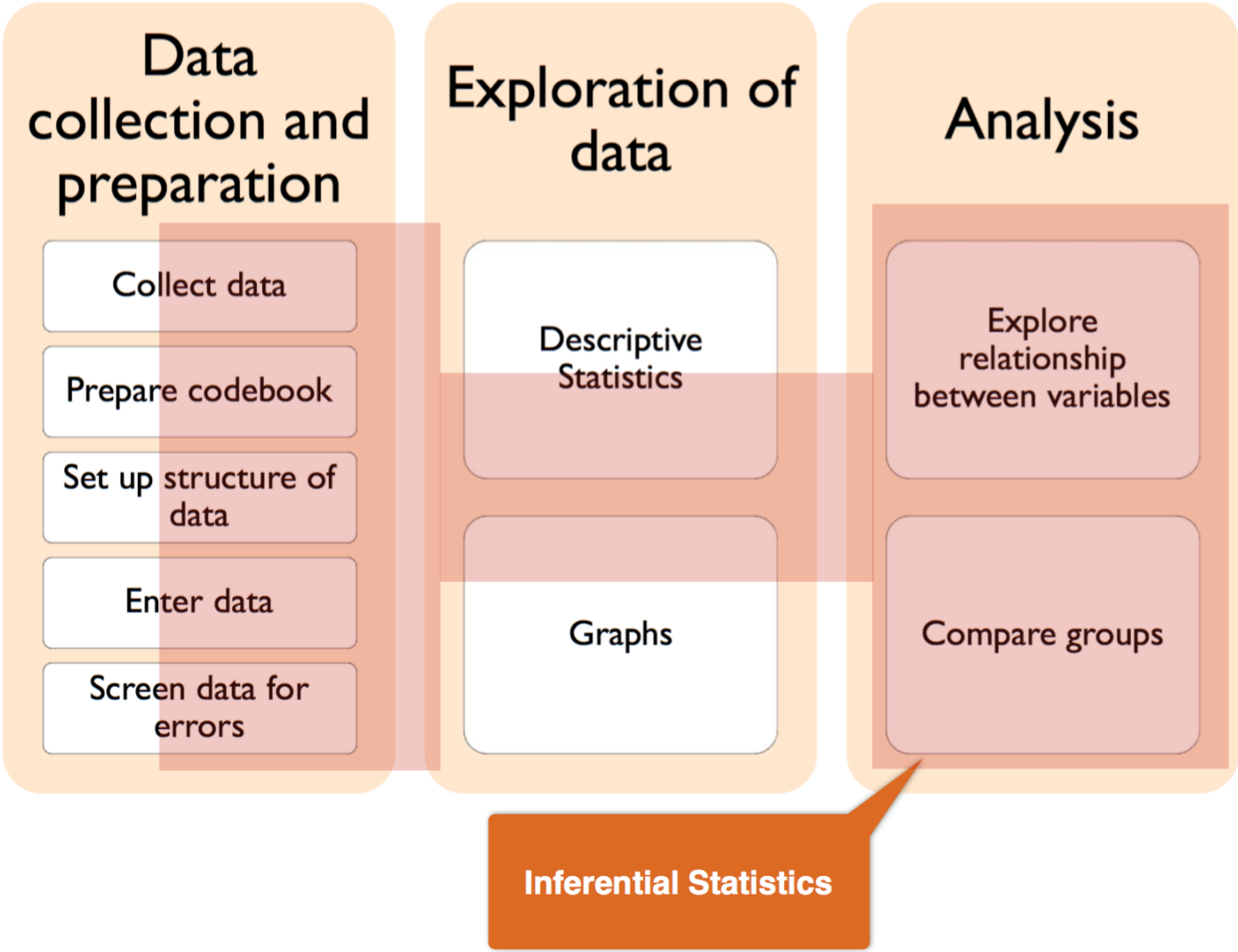
Data analysis processes



Data analysis and Descriptive statistics



Data analysis and Descriptive statistics



Descriptive Statistics (DS)

- Descriptive statistics are numbers that are used to summarize and describe data.
- Descriptive statistics are just descriptive. They do not involve generalizing beyond the data at hand.

Descriptive statistics is a collection of methods for summarizing data (e.g., mean, median, mode, range, variance, graphs).

An examle of DS: Salaries in Estonia

WS720: AVERAGE GROSS HOURLY EARNINGS OF FULL-TIME AND PART-TIME EMPLOYEES, OCTOBER by Year, Major group of occupation and Age group

	Less than 30	30–39	40–49	50–59	60 and over
2008					
Total	5.01	6.04	5.30	4.76	4.09
Legislators, senior officials and managers	7.11	8.59	7.38	7.05	6.21
Professionals	7.10	7.96	7.45	6.67	6.06
Technicians and associate professionals	5.46	6.19	5.52	4.82	4.27
Clerks	4.49	5.09	4.30	4.11	3.54
Service workers and shop and market sales workers	3.70	3.93	3.69	3.23	2.63
Skilled agricultural and fishery workers	3.20	3.46	3.52	3.56	2.80
Craft and related trades workers	4.64	5.01	4.71	4.40	3.79
Plant and machine operators and assemblers	4.38	4.68	4.28	4.02	3.74
Elementary occupations	3.70	3.49	3.07	2.77	2.21
Armed forces	..	..	..	..	..

Footnote:
Unit: euros
The data have been converted into euros on the basis of aggregated data (1 euro = 15.6466 Estonian kroons).
The data are in compliance with ISCO 88. Data are continued to be published according to the new classification in table WS640: Average gross hourly earnings of full-time and part-time employees by major group of occupation, sex and age group, October.

Example 2 of DS: Winners of Olympic marathon

	WOMEN		
Year	**Winner**	**Country**	**Time**
1984	Joan Benoit	USA	2:24:52
1988	Rosa Mota	POR	2:25:40
1992	Valentina Yegorova	UT	2:32:41
1996	Fatuma Roba	ETH	2:26:05"
2000	Naoko Takahashi	JPN	2:23:14
2004	Mizuki Noguchi	JPN	2:26:20
2008	Constantina Tomescu	ROU	2:26:44
2012	Tiki Gelana	ETH	2:23:07

	MEN						
Year	Winner	Country	Time	**Year**	Winner	Country	Time
1896	Spiridon Louis	GRE	2:58:50	**1960**	Abebe Bikila	ETH	2:15:16
1900	Michel Theato	FRA	2:59:45	**1964**	Abebe Bikila	ETH	2:12:11
1904	Thomas Hicks	USA	3:28:53	**1968**	Mamo Wolde	ETH	2:20:26
1906	Billy Sherring	CAN	2:51:23	**1972**	Frank Shorter	USA	2:12:19
1908	Johnny Hayes	USA	2:55:18	**1976**	Waldemar Cierpinski	E.Ger	2:09:55
1912	Kenneth McArthur	S. Afr.	2:36:54	**1980**	Waldemar Cierpinski	E.Ger	2:11:03
1920	Hannes Kolehmainen	FIN	2:32:35	**1984**	Carlos Lopes	POR	2:09:21
1924	Albin Stenroos	FIN	2:41:22	**1988**	Gelindo Bordin	ITA	2:10:32
1928	Boughra El Ouafi	FRA	2:32:57	**1992**	Hwang Young-Cho	S. Kor	2:13:23
1932	Juan Carlos Zabala	ARG	2:31:36	**1996**	Josia Thugwane	S. Afr.	2:12:36
1936	Sohn Kee-Chung	JPN	2:29:19	**2000**	Gezahenge Abera	ETH	2:10:10
1948	Delfo Cabrera	ARG	2:34:51	**2004**	Stefano Baldini	ITA	2:10:55
1952	Emil Ztopek	CZE	2:23:03	**2008**	Samuel Wanjiru	KEN	2:06:32
1956	Alain Mimoun	FRA	2:25:00	**2012**	Stephen Kiprotich	UGA	2:08:01

Descriptive statistics computed from the data

- The average time

  Women: 2:26:05
  Men: 2:27:18 Women have been faster!

- The average time of men

  Years 1896–1956 : 2:42:59
  Years 1960–2012 : 2:11:37 By 31 min and 22 sec. faster

- What we can conclude from these observations?
  - Does this prove that the fastest men are running faster now?
  - Or is the difference just due to chance, no more than what often emerges from chance differences in performance from year to year?
  - We can't answer this question with descriptive statistics alone.
- In 2012, Tiki Gelana was faster than any man before 1952 and in 1956. Will the gender gap close or remain constant?
- How far the times will decrease (if at all) in the next century of the Olympics?
- Might we one day witness a sub-2 hour marathon?

Descriptive statistics computed from the data

- The average time

  Women: 2:26:05
  Men: 2:27:18 Women have been faster!

- The average time of men

  Years 1896–1956 : 2:42:59
  Years 1960–2012 : 2:11:37 By 31 min and 22 sec. faster

- What we can conclude from these observations?
  - Does this prove that the fastest men are running faster now?
  - Or is the difference just due to chance, no more than what often emerges from chance differences in performance from year to year?
  - We can't answer this question with descriptive statistics alone.
- In 2012, Tiki Gelana was faster than any man before 1952 and in 1956. Will the gender gap close or remain constant?
- How far the times will decrease (if at all) in the next century of the Olympics?
- Might we one day witness a sub-2 hour marathon?

Descriptive statistics computed from the data

- The average time

  Women: 2:26:05
  Men: 2:27:18 Women have been faster!

- The average time of men

  Years 1896–1956 : 2:42:59
  Years 1960–2012 : 2:11:37 By 31 min and 22 sec. faster

- What we can conclude from these observations?
  - Does this prove that the fastest men are running faster now?
  - Or is the difference just due to chance, no more than what often emerges from chance differences in performance from year to year?
  - We can't answer this question with descriptive statistics alone.
- In 2012, Tiki Gelana was faster than any man before 1952 and in 1956. Will the gender gap close or remain constant?
- How far the times will decrease (if at all) in the next century of the Olympics?
- Might we one day witness a sub-2 hour marathon?

Descriptive statistics computed from the data

- The average time

  Women: 2:26:05
  Men: 2:27:18 Women have been faster!

- The average time of men

  Years 1896–1956 : 2:42:59
  Years 1960–2012 : 2:11:37 By 31 min and 22 sec. faster

- What we can conclude from these observations?
  - Does this prove that the fastest men are running faster now?
  - Or is the difference just due to chance, no more than what often emerges from chance differences in performance from year to year?
  - We can't answer this question with descriptive statistics alone.
- In 2012, Tiki Gelana was faster than any man before 1952 and in 1956. Will the gender gap close or remain constant?
- How far the times will decrease (if at all) in the next century of the Olympics?
- Might we one day witness a sub-2 hour marathon?

Descriptive statistics computed from the data

- The average time

  Women: 2:26:05
  Men: 2:27:18 Women have been faster!

- The average time of men

  Years 1896–1956 : 2:42:59
  Years 1960–2012 : 2:11:37 By 31 min and 22 sec. faster

- What we can conclude from these observations?
  - Does this prove that the fastest men are running faster now?
  - Or is the difference just due to chance, no more than what often emerges from chance differences in performance from year to year?
  - We can't answer this question with descriptive statistics alone.
- In 2012, Tiki Gelana was faster than any man before 1952 and in 1956. Will the gender gap close or remain constant?
- How far the times will decrease (if at all) in the next century of the Olympics?
- Might we one day witness a sub-2 hour marathon?

Descriptive statistics computed from the data

- The average time

  Women: 2:26:05
  Men: 2:27:18 Women have been faster!

- The average time of men

  Years 1896–1956 : 2:42:59
  Years 1960–2012 : 2:11:37 By 31 min and 22 sec. faster

- What we can conclude from these observations?
  - Does this prove that the fastest men are running faster now?
  - Or is the difference just due to chance, no more than what often emerges from chance differences in performance from year to year?
  - We can't answer this question with descriptive statistics alone.
- In 2012, Tiki Gelana was faster than any man before 1952 and in 1956. Will the gender gap close or remain constant?
- How far the times will decrease (if at all) in the next century of the Olympics?
- Might we one day witness a sub-2 hour marathon?

Descriptive statistics computed from the data

- The average time

  Women: 2:26:05
  Men: 2:27:18 Women have been faster!

- The average time of men

  Years 1896–1956 : 2:42:59
  Years 1960–2012 : 2:11:37 By 31 min and 22 sec. faster

- What we can conclude from these observations?
  - Does this prove that the fastest men are running faster now?
  - Or is the difference just due to chance, no more than what often emerges from chance differences in performance from year to year?
  - We can't answer this question with descriptive statistics alone.
- In 2012, Tiki Gelana was faster than any man before 1952 and in 1956. Will the gender gap close or remain constant?
- How far the times will decrease (if at all) in the next century of the Olympics?
- Might we one day witness a sub-2 hour marathon?

Let's go on with inferential statistics (IS).

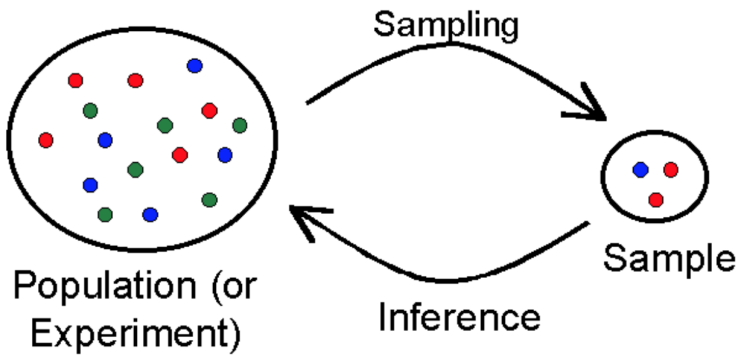
IS: Populations and samples

Gathering all data is not always possible due to barriers such as time, accessibility, or cost. Instead of that, we often gather information from a smaller subset of the population, known as a sample.

Population: The entire set of possible observations in which we are interested

Sample: A subset of the population from which information is actually collected

Inferential statistics is a collection of methods for using sample data to make conclusions about a population.



IS: Populations and samples 2

An example where sampling is required.

TV channels watched in Estonia (daily share %), by *TNS Emor*

Kanal	Detsember	August	September	Oktoober	November	Detsember
	2014	2015	2015	2015	2015	2015
ETV	16.1	12.6	16.5	15.0	15.5	14.7
Kanal 2	15.9	14.8	16.2	16.2	15.8	15.5
TV3	11.9	9.1	9.9	12.3	12.3	11.6
Kanal 11	1.6	2.3	2.3	2.3	1.9	2.3
PBK	6.3	6.1	5.8	5.3	5.5	5.6
RTR Planeta	4.5	4.5	4.1	4.0	4.1	4.2
NTV Mir	4.4	5.9	5.5	5.4	5.8	5.5
TV6	2.1	2.2	2.3	2.2	2.0	2.0
ETV2	3.2	2.8	2.7	2.5	2.5	2.9
3+	1.9	1.8	1.8	1.5	1.6	1.7
Ren TV Estonia	2.1	1.9	1.5	1.8	2.0	1.9
FOX	1.5	1.3	1.0	0.9	0.9	0.8
Fox Life	0.7	0.6	0.6	0.5	0.6	0.7
National Geographics	0.4	0.6	0.5	0.5	0.5	0.5
Sony Entertainment TV	0.5	0.5	0.4	0.5	0.3	0.3
Kanal 12	1.1	1.3	1.2	1.2	1.0	1.2
Tallinna TV	0.9	1.0	1.3	1.2	1.5	1.3
CTC	0.5	0.5	0.4	0.4	0.4	0.5
Sony Turbo	0.2	0.1	0.2	0.2	0.2	0.2
Kidzone	1.4	1.5	1.3	1.2	1.4	1.2
TLC	0.4	0.6	0.5	0.5	0.4	0.5
Discovery	0.4	0.7	0.6	0.6	0.6	0.6
ETV+				0.5	0.5	0.5
Kanal 1+					0.2	0.2
Muu vaatamine	22.3	27.3	23.5	23.5	22.6	23.5

IS: Populations and samples 3

Kinds of samples:

- A complete sample is a set of objects from a parent population that includes ALL such objects that satisfy a set of well-defined selection criteria.

  For example, a complete sample of Australian men taller than 2m would consist of a list of every Australian male taller than 2m. But it wouldn't include German males, or tall Australian females, or people shorter than 2m.

- An unbiased (representative) sample is a set of objects chosen from a complete sample using a selection process that does not depend on the properties of the objects.

  For example, an unbiased sample of Australian men taller than 2m might consist of a randomly sampled subset of 1% of Australian males taller than 2m. But one chosen from the electoral register might not be unbiased since, for example, males aged under 18 will not be on the electoral register.

Random sample

- The best way to avoid a biased or unrepresentative sample is to select a random sample.
- A **random sample** is a sample where each individual member of the population has a known, non-zero chance of being selected as part of the sample.

The types of random samples are simple random samples, systematic samples, stratified random samples, and cluster random samples.

IS: Populations and samples 3

Kinds of samples:

- A complete sample is a set of objects from a parent population that includes ALL such objects that satisfy a set of well-defined selection criteria.

  For example, a complete sample of Australian men taller than 2m would consist of a list of every Australian male taller than 2m. But it wouldn't include German males, or tall Australian females, or people shorter than 2m.

- An unbiased (representative) sample is a set of objects chosen from a complete sample using a selection process that does not depend on the properties of the objects.

  For example, an unbiased sample of Australian men taller than 2m might consist of a randomly sampled subset of 1% of Australian males taller than 2m. But one chosen from the electoral register might not be unbiased since, for example, males aged under 18 will not be on the electoral register.

Random sample

- The best way to avoid a biased or unrepresentative sample is to select a random sample.
- A **random sample** is a sample where each individual member of the population has a known, non-zero chance of being selected as part of the sample.

The types of random samples are simple random samples, systematic samples, stratified random samples, and cluster random samples.

IS: Simple Random Sampling (SRS)

SRS is a method of obtaining a sample from a population in which every member of the population has an equal chance of being selected.

Example 1: World Campus

An institutional researcher is conducting a study of World Campus students' attitudes toward community service. He takes a list of all 12,242 World Campus students and uses a random number generator to select 30 students whom he contacts to complete the survey. This researcher used **simple random sampling** because participants were selected from the overall population in a way that each individual had an equal chance of being selected.

Example 2: Language Study

A student wants to learn more about the languages spoken in her town. She has access to the census forms submitted by all 3,500 households in her town. It would take too long for her to go through all 3,500 forms, so she uses a random number generator to select 100 households. She finds those 100 census forms and records data concerning the languages spoken in those households. This is a **simple random sample** because the sample of 100 households was selected in a way that each of the 3,500 households had an equal chance of being selected.

Only a large sample size makes it likely that our sample is close to representative of the population.

IS: Stratified Sampling

Stratified Random Sampling is a method of obtaining a sample from a population in which the population is divided into important subgroups and then separate simple random samples are drawn from each subgroup which are known as strata.

Example 1: Attitudes Toward Community Service

An institutional researcher is conducting a study of World Campus students' attitudes toward community service. He thinks that there may be a difference between students who are associate's, bachelor's, and graduate students. There are 917 World Campus associate's degree students; the researcher takes a simple random sample from that population. There are 4,819 World Campus bachelor's degree students; the researcher takes a simple random sample from that population. There are 4,560 World Campus graduate students; the researcher takes a simple random sample from that population. This is a **stratified random sample** because the population of all World Campus students was divided into four strata that the researcher believes could impact the responses to the survey.

Example 2: Acceptance of Diversity

A researcher is interested in studying World Campus students' perceptions of acceptance of diversity. She knows that of the 12,242 World Campus students, 27 identify as American Indian / Alaska Native and 27 identify as Native Hawaiian / Pacific Islander. If she were to take a simple random sample from the population of all World Campus students, these groups would not likely be sufficiently represented. She chooses to use a **stratified random sample** with the strata being the different race/ethnicity groups. This ensures that there will students from each group represented in her sample.

IS: Cluster Sampling

Cluster Sampling is a method of selecting a sample from a population in which the population is divided into subgroups (i.e., clusters) and a simple random sample of those subgroups is taken; all individuals within these clusters may be sampled, or a simple random sample may be taken from the selected clusters

Example : Forestation and Population Density of Forestation and Flying Squirrel

A researcher is studying the relationship between forestation density and flying squirrel population density in Virumaa and Jõgevamaa (the counties of Estonia). His research requires him to travel to the locations that he is studying. If a simple random sample were used he may have to travel to many different locations which is not practical for him. Instead, he uses **cluster sampling**. There are 38 parishes in these counties; the researcher treats each county as a **cluster**. He randomly selects 8 parishes. Within each county, he takes a simple random sample of the townships that he will travel to for data collection.

Next section

Variables

- Variables are properties or characteristics of some event, object, or person that can take on different values or amounts;
- Constants do not vary.

Variables

- Variables are properties or characteristics of some event, object, or person that can take on different values or amounts;
- Constants do not vary.

Variables may be ...

- independent or dependent;
- discrete or continuous;
- qualitative or quantitative.

Variables

- Variables are properties or characteristics of some event, object, or person that can take on different values or amounts;
- Constants do not vary.

Variables may be ...

- independent or dependent;
- discrete or continuous;
- qualitative or quantitative.

Variables

- Variables are properties or characteristics of some event, object, or person that can take on different values or amounts;
- Constants do not vary.

Variables may be ...

- independent or dependent;
- discrete or continuous;
- qualitative or quantitative.

Variables

- Variables are properties or characteristics of some event, object, or person that can take on different values or amounts;
- Constants do not vary.

Variables may be ...

- independent or dependent;
- discrete or continuous;
- qualitative or quantitative.

Independent and dependent variables

- When conducting research, experimenters often manipulate variables.
- For example, an experimenter might compare the effectiveness of four types of antidepressants. In this case, the variable is "type of antidepressant."
- When a variable is *manipulated by an experimenter*, it is called an independent variable.
- The experiment seeks to determine the effect of the independent variable on relief from depression.
- In this example, relief from depression is called a dependent variable.

In general, the independent variable is manipulated by the experimenter and its effects on the dependent variable are measured.

Independent and dependent variables

- When conducting research, experimenters often manipulate variables.
- For example, an experimenter might compare the effectiveness of four types of antidepressants. In this case, the variable is "type of antidepressant."
- When a variable is *manipulated by an experimenter*, it is called an independent variable.
- The experiment seeks to determine the effect of the independent variable on relief from depression.
- In this example, relief from depression is called a dependent variable.

In general, the independent variable is manipulated by the experimenter and its effects on the dependent variable are measured.

Qualitative and quantitative variables

- Qualitative variables are those that express a qualitative attribute such as hair colour, eye colour, religion, favourite movie, gender, and so on.
- The values of a qualitative variable do not imply a numerical ordering.
- For example: values of the variable "religion" differ qualitatively; no ordering of religions is implied.
- Qualitative variables are also referred to as categorical variables.
- Quantitative variables are those variables that are measured in terms of numbers.
- Some examples of quantitative variables are height, weight, and shoe size.

Discrete and continuous variables

- Discrete variables can take only certain values.
- For example, a household could have three children or six children, but not 4.53 children.
- Continuous variables can take any value within the range of the scale.
- For example, "time to respond to a question" are continuous variables since the scale is continuous and not made up of discrete steps, say, the response time could be 1.64 seconds.

Questions to clarify understanding (QCU): variables

Which of the following are qualitative variables?

1. height measured in number of feet
2. weight measured in number of kilograms
3. number of days it snowed
4. hair colour
5. gender
6. average daily temperature

Questions to clarify understanding (QCU): variables

Which of the following are qualitative variables?

1. height measured in number of feet
2. weight measured in number of kilograms
3. number of days it snowed
4. hair colour
5. gender
6. average daily temperature

Answer: 4, 5.

QCU: variables

In an experiment on the effect of sleep on memory, the independent variable is

1. number of hours of sleep
2. recall score on a memory test
3. gender of the subjects
4. gender of the experimenter

QCU: variables

In an experiment on the effect of sleep on memory, the independent variable is

1. number of hours of sleep
2. recall score on a memory test
3. gender of the subjects
4. gender of the experimenter

Answer: 1.

QCU: variables

In an experiment on the effect of sleep on memory, the dependent variable is

1. number of hours of sleep
2. recall score on a memory test
3. gender of the subjects
4. gender of the experimenter

QCU: variables

In an experiment on the effect of sleep on memory, the dependent variable is

1. number of hours of sleep
2. recall score on a memory test
3. gender of the subjects
4. gender of the experimenter

Answer: 2.

Next section

Why percentiles?

- Raw data is sometimes difficult to interpret by itself. For example, you pass a shyness test:
  - and your test score is 35/50.
  - Is this good or bad result? Does this mean that you are shyer than most people?

Percentile ...

- ... shows what proportion of scores is higher that your score
- If your shyness score is higher that 65% of the population, then your score is the 65th percentile

- There are different definitions and techniques to compute percentiles,
- they may give different results when there is little data,
- neither of these definitions is explicit about how to handle rounding.

Why percentiles?

- Raw data is sometimes difficult to interpret by itself. For example, you pass a shyness test:
  - and your test score is 35/50.
  - Is this good or bad result? Does this mean that you are shyer than most people?

Percentile ...

- ... shows what proportion of scores is higher that your score
- If your shyness score is higher that 65% of the population, then your score is the 65th percentile

- There are different definitions and techniques to compute percentiles,
- they may give different results when there is little data,
- neither of these definitions is explicit about how to handle rounding.

Why percentiles?

- Raw data is sometimes difficult to interpret by itself. For example, you pass a shyness test:
  - and your test score is 35/50.
  - Is this good or bad result? Does this mean that you are shyer than most people?

Percentile ...

- ... shows what proportion of scores is higher that your score
- If your shyness score is higher that 65% of the population, then your score is the 65th percentile

- There are different definitions and techniques to compute percentiles,
- they may give different results when there is little data,
- neither of these definitions is explicit about how to handle rounding.

Definitions

Definition (1)

The i- th percentile is the lowest score that is greater than i% of all scores.

Definition (2)

The i- th percentile is the lowest score that is greater than or equal to i% of all scores.

Definition (3)

The i- th percentile of a list of N ordered values (sorted from least to greatest) is the score computed according the following method:

1. Compute the ordinal rank by the formula

$$R = \frac{i}{100} \times (N+1).$$

   If R is an integer then i is the score with the index R, otherwise take I_R and F_R as an integer part and fractional part of R respectively;
2. Let s_1 is the score with index I_R and s_2 the score with index I_R+1;
3. Compute $i = s_1 + F_R(s_2 - s_1)$.

Definitions

Definition (1)

The i- th percentile is the lowest score that is greater than i% of all scores.

Definition (2)

The i- th percentile is the lowest score that is greater than or equal to i% of all scores.

Definition (3)

The i- th percentile of a list of N ordered values (sorted from least to greatest) is the score computed according the following method:

1. Compute the ordinal rank by the formula

$$R = \frac{i}{100} \times (N+1).$$

   If R is an integer then i is the score with the index R, otherwise take I_R and F_R as an integer part and fractional part of R respectively;

2. Let s_1 is the score with index I_R and s_2 the score with index I_R+1;
3. Compute $i = s_1 + F_R(s_2 - s_1)$.

Definitions

Definition (1)

The i- th percentile is the lowest score that is greater than $i\%$ of all scores.

Definition (2)

The i- th percentile is the lowest score that is greater than or equal to $i\%$ of all scores.

Definition (3)

The i- th percentile of a list of N ordered values (sorted from least to greatest) is the score computed according the following method:

1. Compute the ordinal rank by the formula

$$R = \frac{i}{100} \times (N+1).$$

   If R is an integer then i is the score with the index R, otherwise take I_R and F_R as an integer part and fractional part of R respectively;
2. Let s_1 is the score with index I_R and s_2 the score with index I_R+1;
3. Compute $i = s_1 + F_R(s_2 - s_1)$.

Example: 25th percentile of 20 Quiz scores

Score	Rank
4	1
4	2
5	3
5	4
5	5
5	6
6	7
6	8
6	9
7	10
7	11
7	12
8	13
8	14
9	15
9	16
9	17
10	18
10	19
10	20

Def.(1) The rank $0.25 \times 20 = 5$ corresponds to the score 5 and the 25th percentile is 6.

Def.(2) The 25th percentile is 5.

Def.(3) The rank

$$R = \frac{25}{100} \times (20+1) = \frac{21}{4} = 5.25$$

and therefore $I_R = 5$, $F_R = 0.25$, $s_1 = 5$ and $s_2 = 5$.
The 25th percentile is $5 + 0.25(5-5) = 5$.

Example: 85th percentile of 20 Quiz scores

Score	Rank
4	1
4	2
5	3
5	4
5	5
5	6
6	7
6	8
6	9
7	10
7	11
7	12
8	13
8	14
9	15
9	16
9	17
10	18
10	19
10	20

- The rank

$$R = \frac{85}{100} \times (20+1) = 0.85 \times 21 = 17.85$$

and therefore $I_R = 17$, $F_R = 0.85$, $s_1 = 9$ and $s_2 = 10$.

- The 85th percentile is $9 + 0.85(10-9) = 9.85$.

Quartiles, median

Score	Rank
4	1
4	2
5	3
5	4
5	5
5	6
6	7
6	8
6	9
7	10
7	11
7	12
8	13
8	14
9	15
9	16
9	17
10	18
10	19
10	20

- **first quartile** (designated Q_1) also called the **lower quartile** $\equiv$ the 25th percentile;
- **second quartile** (designated Q2) also called the **median** $\equiv$ the 50th percentile
- **third quartile** (designated Q3) also called the **upper quartile** $\equiv$ the 75th percentile

- $Q_1 = 5$
- $Q_2 = 7 + 0.5(7-7) = 7 \qquad [R = 0.5 \times 21 = 10.5]$
- $Q_3 = 9 + 0.75(9-9) = 9 \qquad [R = 0.75 \times 21 = 15.75]$

QCU1: percentiles

For the scores 3, 5, 7, 9, 12, 21, 25, 30, calculate the 25th percentile based on "Definition 1."

QCU1: percentiles

For the scores 3, 5, 7, 9, 12, 21, 25, 30, calculate the 25th percentile based on "Definition 1."

Answer: 7.

According to Definition 1, the 25th percentile is the lowest score higher than 25% of the scores. Since there are 8 scores, this would be the lowest score higher than $(0.25)x8 = 2$ scores. The score 7 is higher than the scores 3 and 5.

QCU2: percentiles

For the scores 3, 5, 7, 9, 12, 21, 25, 30, calculate the 25th percentile based on "Definition 2."

QCU2: percentiles

For the scores 3, 5, 7, 9, 12, 21, 25, 30, calculate the 25th percentile based on "Definition 2."

Answer: 5.

According to Definition 2, the 25th percentile is the lowest number greater than or equal to 25% of the scores. Since there are 8 scores, this would be the lowest number greater than or equal to (0.25) x 8 = 2 scores. The number 5 is greater than or equal to the scores 3 and 5.

QCU3: percentiles

For the scores 3, 5, 7, 9, 12, 21, 25, 30, calculate the 25th percentile based on "Definition 3."

QCU3: percentiles

For the scores 3, 5, 7, 9, 12, 21, 25, 30, calculate the 25th percentile based on "Definition 3."

Answer: 5.5

$R = 25/100x(8+1) = 2.25$; $I_R = 2$; $F_R = 0.25$; The 25th percentile $= 0.25 \times (7-5)+5 = 5.5$.

QCU4: percentiles

For the scores 3, 5, 7, 9, 12, 21, 25, 30, calculate the 80th percentile based on "Definition 3."

QCU4: percentiles

For the scores 3, 5, 7, 9, 12, 21, 25, 30, calculate the 80th percentile based on "Definition 3."

Answer: 26

$R = 80/100x(8+1) = 7.2$; $I_R = 7$; $F_R = 0.2$; The 80th percentile $= 0.2 \times (30-25)+25 = 26$.

Next section

Scales of measurement

- We measure our dependent variables
- Different types are measured differently

Types of scales

- Four fundamental scales:
  - Nominal
  - Ordinal
  - Interval
  - Ratio

Nominal Scales

- Nominal:
  - names or categories
  - examples include:
    - ★ gender
    - ★ handedness
    - ★ favourite colour
    - ★ religion
  - lowest level of measurement

Ordinal Scales

- Ordinal:
  - names or categories and order is meaningful
  - examples include:
    - ★ consumer satisfaction ratings
    - ★ military rank
    - ★ class ranking

Limitations of ordinal scales

- We can't assume the differences between adjacent scale values are equal;
- We can't make this assumption even if the labels are numbers, not words.

Ordinal Scales

- Ordinal:
  - names or categories and order is meaningful
  - examples include:
    - ★ consumer satisfaction ratings
    - ★ military rank
    - ★ class ranking

Limitations of ordinal scales

- We can't assume the differences between adjacent scale values are equal;
- We can't make this assumption even if the labels are numbers, not words.

Interval Scales

- Interval:
  - names or categories, the order is meaningful, and intervals have the same interpretation
  - example:
    - ★ Celsius temperature scale
  - problem; No true zero point

Ratio Scales

- Ratio:
  - highest and most informative scale
  - contains the qualities of the nominal, ordinal, and interval scales with the addition of an absolute zero point
  - example: amount of money – zero money indicates the absence of money

Psychological Research

- Psychological variables:
  - Frequently use rating scales
  - Rating scales are ordinal

Example: favourable colours

Codes of colours:

Color	Code
Blue	1
Red	2
Yellow	3
Green	4
Purple	5

Answers:

Subject	Color	Code
1	Blue	1
2	Blue	1
3	Green	4
4	Green	4
5	Purple	5

- Does it make sense to compute the mean of these numbers measured on an ordinal scale?
- The prevailing opinion of statisticians is: YES
- However, sometimes the mean of an ordinally-measured variable can be very misleading.

QCU1: measurement scales

Identify the scale of measurement for the following: military title – Lieutenant, Captain, Major.

1. nominal
2. ordinal
3. interval
4. ratio

QCU1: measurement scales

Identify the scale of measurement for the following: military title – Lieutenant, Captain, Major.

1. nominal
2. ordinal
3. interval
4. ratio

Answer: 2.

The scale is ordinal. There is an inherent ordering in that a Major is higher than a Captain, which is higher than a Lieutenant.

QCU2: measurement scales

Identify the scale of measurement for the following categorization of clothing: hat, shirt, shoes, pants .

1. nominal
2. ordinal
3. interval
4. ratio

QCU2: measurement scales

Identify the scale of measurement for the following categorization of clothing: hat, shirt, shoes, pants .

1. nominal
2. ordinal
3. interval
4. ratio

Answer: 1.

Since clothes are categorized and have no inherent order, the scale is nominal.

QCU3: measurement scales

Identify the scale of measurement for the following: heat measured in degrees centigrade.

1. nominal
2. ordinal
3. interval
4. ratio

QCU3: measurement scales

Identify the scale of measurement for the following: heat measured in degrees centigrade.

1. nominal
2. ordinal
3. interval
4. ratio

Answer: 3.

The scale is interval because there are equal intervals between temperatures but no true zero point.

QCU4: measurement scales

A score on a 5-point quiz measuring knowledge of algebra is an example of a(n)

1. nominal scale
2. ordinal scale
3. interval scale
4. ratio scale

QCU4: measurement scales

A score on a 5-point quiz measuring knowledge of algebra is an example of a(n)

1. nominal scale
2. ordinal scale
3. interval scale
4. ratio scale

Answer: 2.

It is ordinal because higher scores are better than lower scores. However, there is no guarantee that the difference between, say, a 2 and a 3 represents the same difference in knowledge as the difference between a 4 and a 5.

QCU5: measurement scales

City of birth is an example of a(n)

1. nominal scale
2. ordinal scale
3. interval scale
4. ratio scale

QCU5: measurement scales

City of birth is an example of a(n)

1. nominal scale
2. ordinal scale
3. interval scale
4. ratio scale

Answer: 1.

The city that someone was born in has no inherent order, thus can only be a nominal scale.

QCU6: measurement scales

There is debate about the value of computing means for

1. nominal data
2. ordinal data
3. interval data
4. ratio data

QCU6: measurement scales

There is debate about the value of computing means for

1. nominal data
2. ordinal data
3. interval data
4. ratio data

Answer: 1.

Most statisticians agree that it is valid to compute means of ordinal data, although some vehemently disagree.

Next section

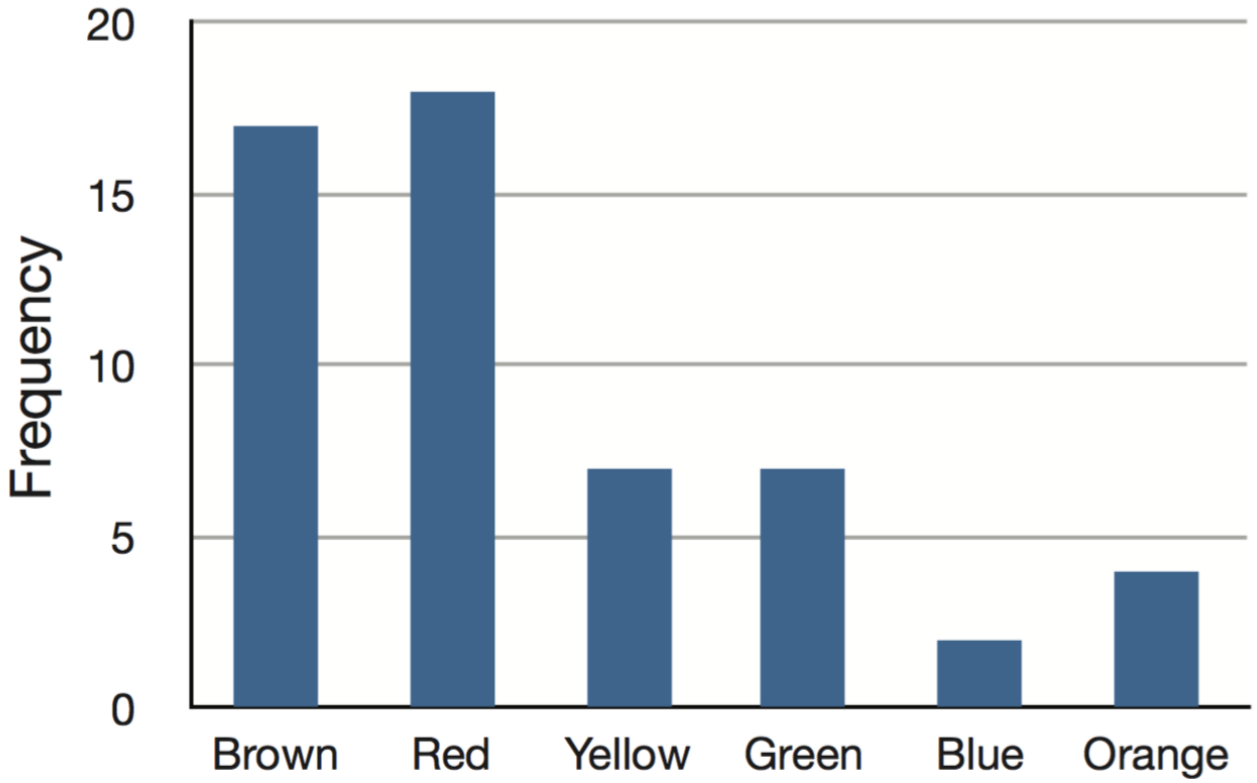
Distributions of Discrete Variables

- Frequency table
- Frequency distribution

Color	Frequency
Brown	17
Red	18
Yellow	7
Green	7
Blue	2
Orange	4

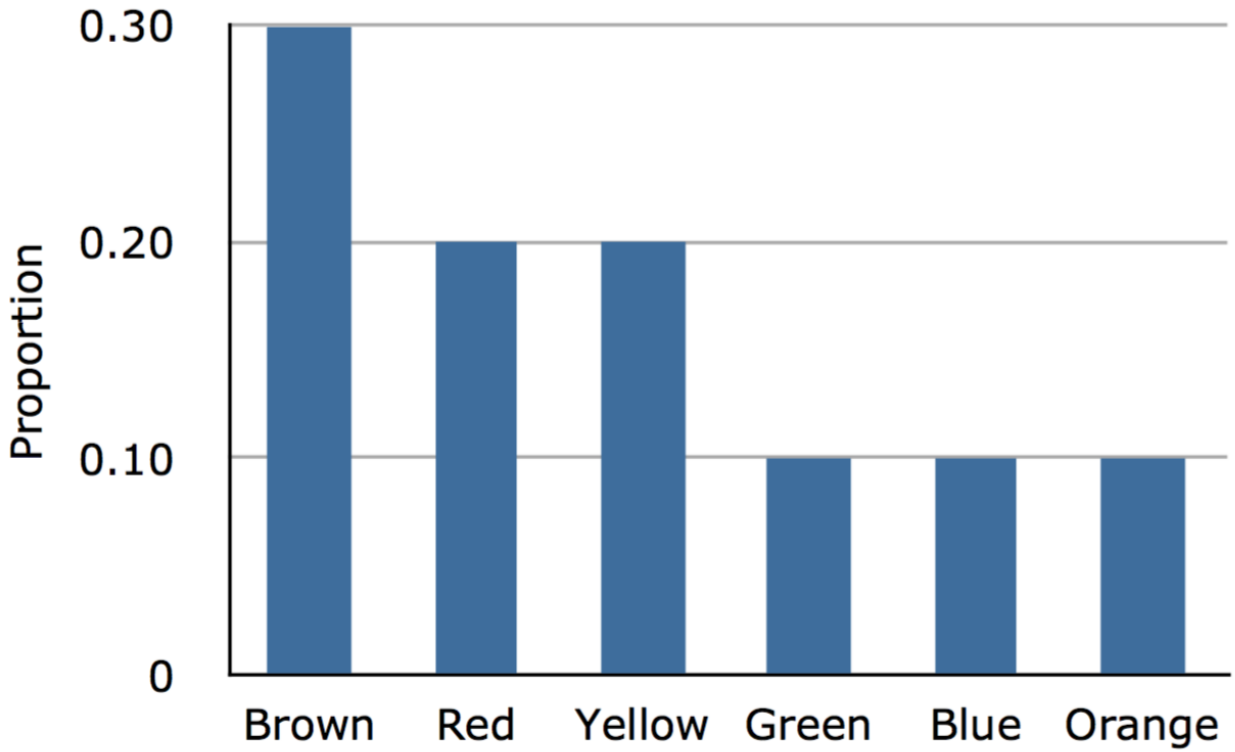
Distributions of Discrete Variables (2)

Frequency distribution is shown graphically



Distributions of Discrete Variables (3)

Probability distribution



Distributions of Continuous Variables

Example: "Response times"

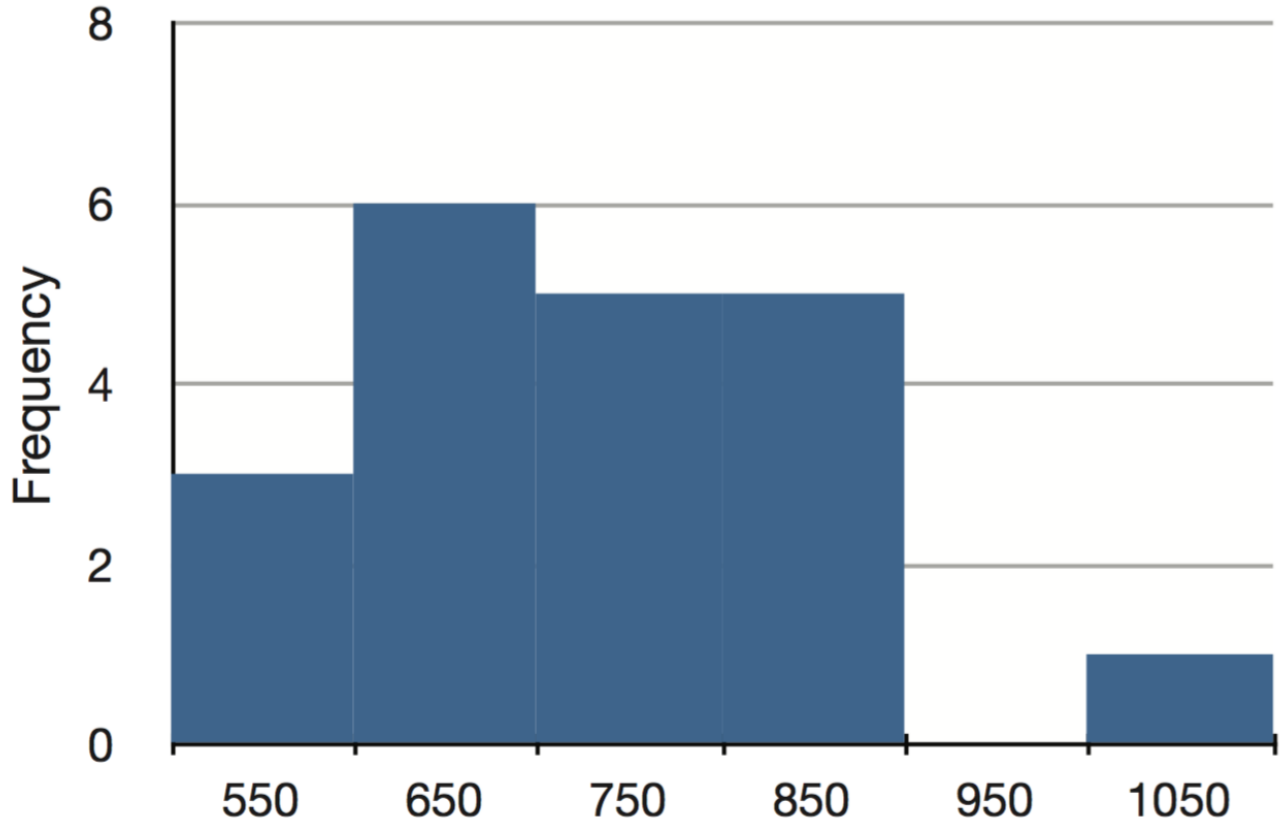
568	720
577	728
581	729
640	777
641	808
645	824
657	825
673	865
696	875
703	1007

Grouped frequency distribution;

Range	Frequency
500-600	3
600-700	6
700-800	5
800-900	5
900-1000	0
1000-1100	1

Distributions of Continuous Variables (2)

A histogram of the grouped frequency distribution:



- The labels on the X-axis are the middle values of the range they represent.

Probability Densities

Probability Density Function:

- For a discrete random variable, a probability distribution contains the probability of each possible outcome.
- For a continuous random variable, the probability of any **one** outcome is zero (if you specify it to enough decimal places).
- A probability density function is a formula that can be used to compute probabilities of a range of outcomes for a continuous random variable.
- The sum of all densities is always 1.0 and the value of the function is always greater or equal to zero.

Probability Densities

Probability Density Function:

- For a discrete random variable, a probability distribution contains the probability of each possible outcome.
- For a continuous random variable, the probability of any **one** outcome is zero (if you specify it to enough decimal places).
- A probability density function is a formula that can be used to compute probabilities of a range of outcomes for a continuous random variable.
- The sum of all densities is always 1.0 and the value of the function is always greater or equal to zero.

Probability Densities

Probability Density Function:

- For a discrete random variable, a probability distribution contains the probability of each possible outcome.
- For a continuous random variable, the probability of any **one** outcome is zero (if you specify it to enough decimal places).
- A probability density function is a formula that can be used to compute probabilities of a range of outcomes for a continuous random variable.
- The sum of all densities is always 1.0 and the value of the function is always greater or equal to zero.

Probability Densities

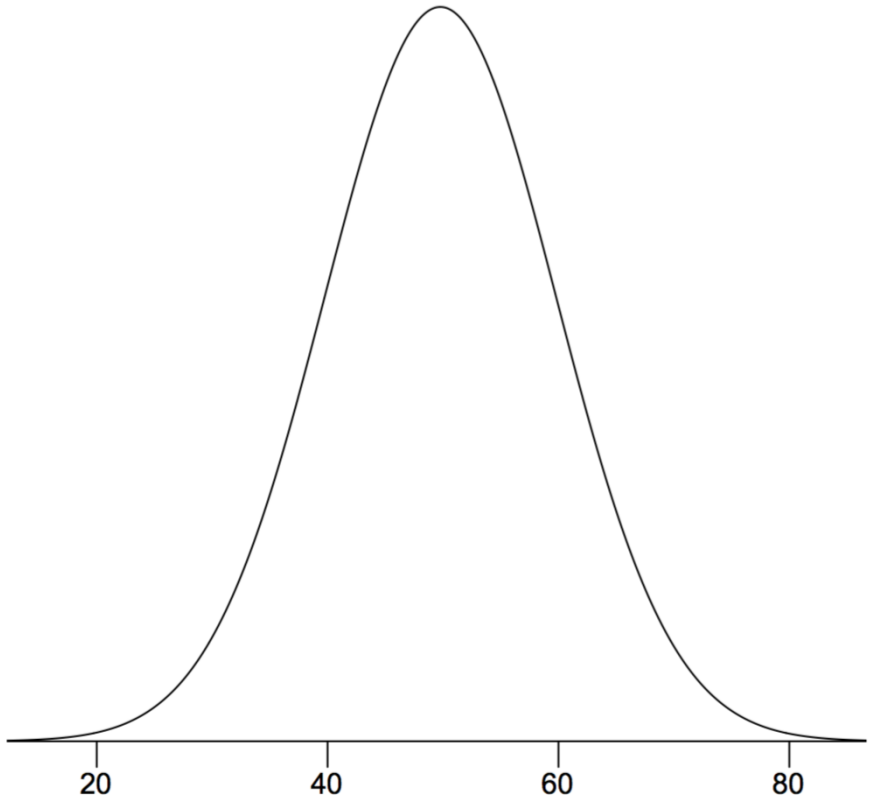
Probability Density Function:

- For a discrete random variable, a probability distribution contains the probability of each possible outcome.
- For a continuous random variable, the probability of any **one** outcome is zero (if you specify it to enough decimal places).
- A probability density function is a formula that can be used to compute probabilities of a range of outcomes for a continuous random variable.
- The sum of all densities is always 1.0 and the value of the function is always greater or equal to zero.

Example: Normal Distribution

- Also called "bell-shaped distribution"
- Density Function of normal distribution: $N(\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x+\mu)^2}{2\sigma^2}}$.

A graph of a normal distribution with a mean of $\mu = 50$ and a standard deviation of $\sigma = 10$



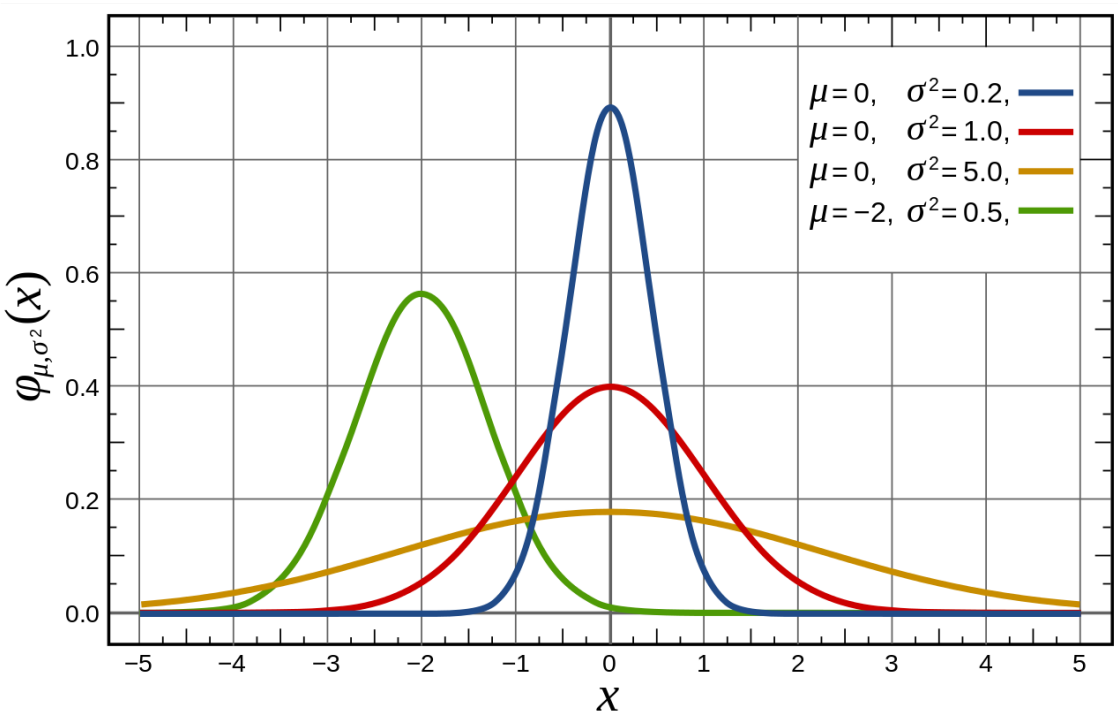
Shapes of Distributions

- Normal distributions vary
- Distributions can be asymmetric
- Distributions can have more than one peak

Shapes of Distributions

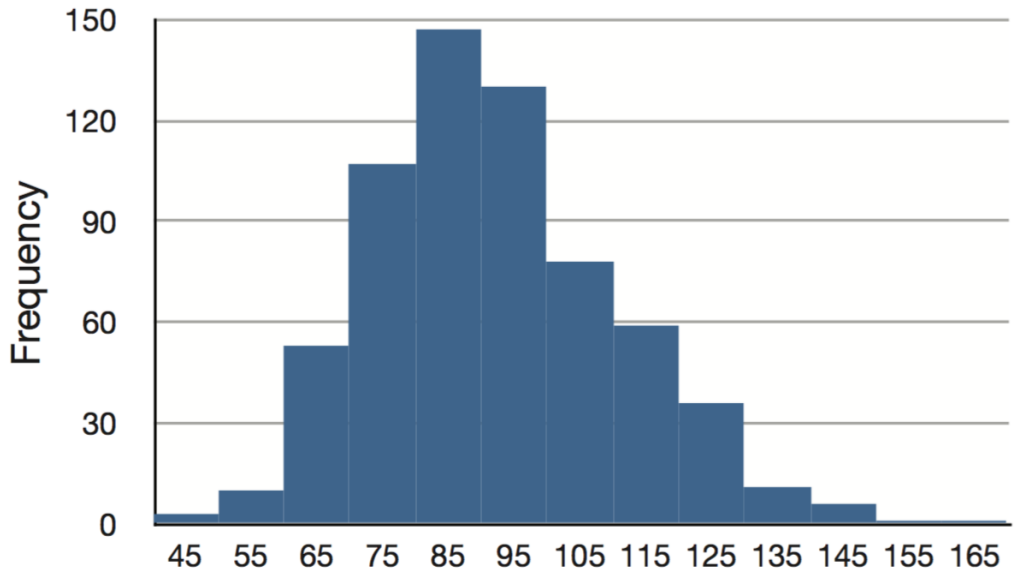
- Normal distributions vary
- Distributions can be asymmetric
- Distributions can have more than one peak

Different normal distributions:



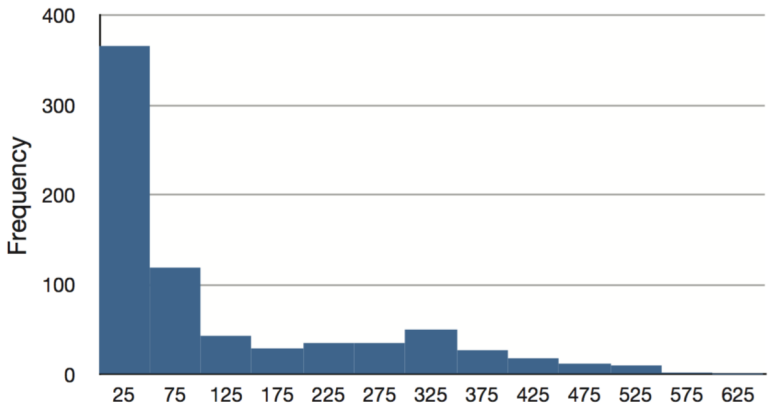
Shapes of Distributions: Skewed Distribution

Positive skew

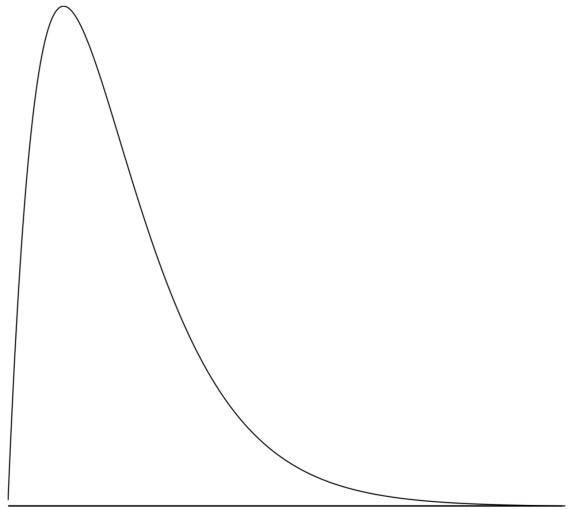


Shapes of Distributions: Skewed Distribution

A distribution with a very large positive skew



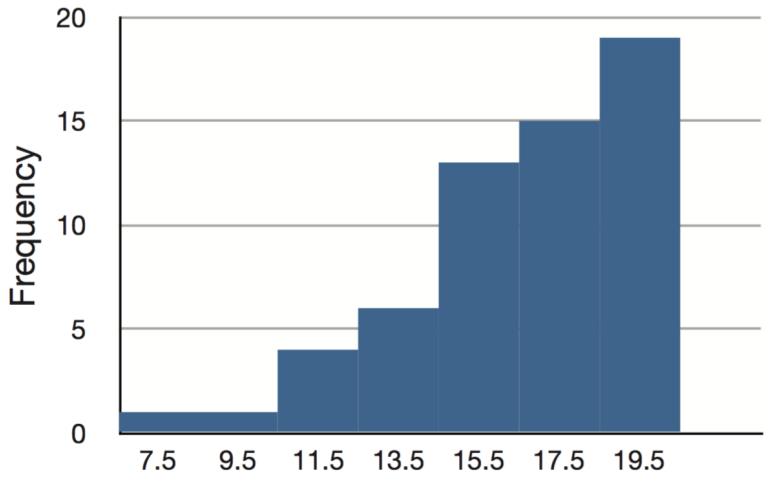
Discrete distribution



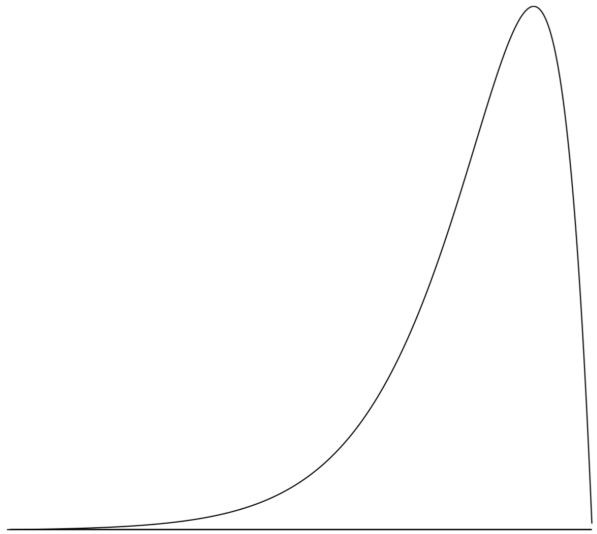
Continuous distribution

Shapes of Distributions: Skewed Distribution (2)

A distribution with a negative skew

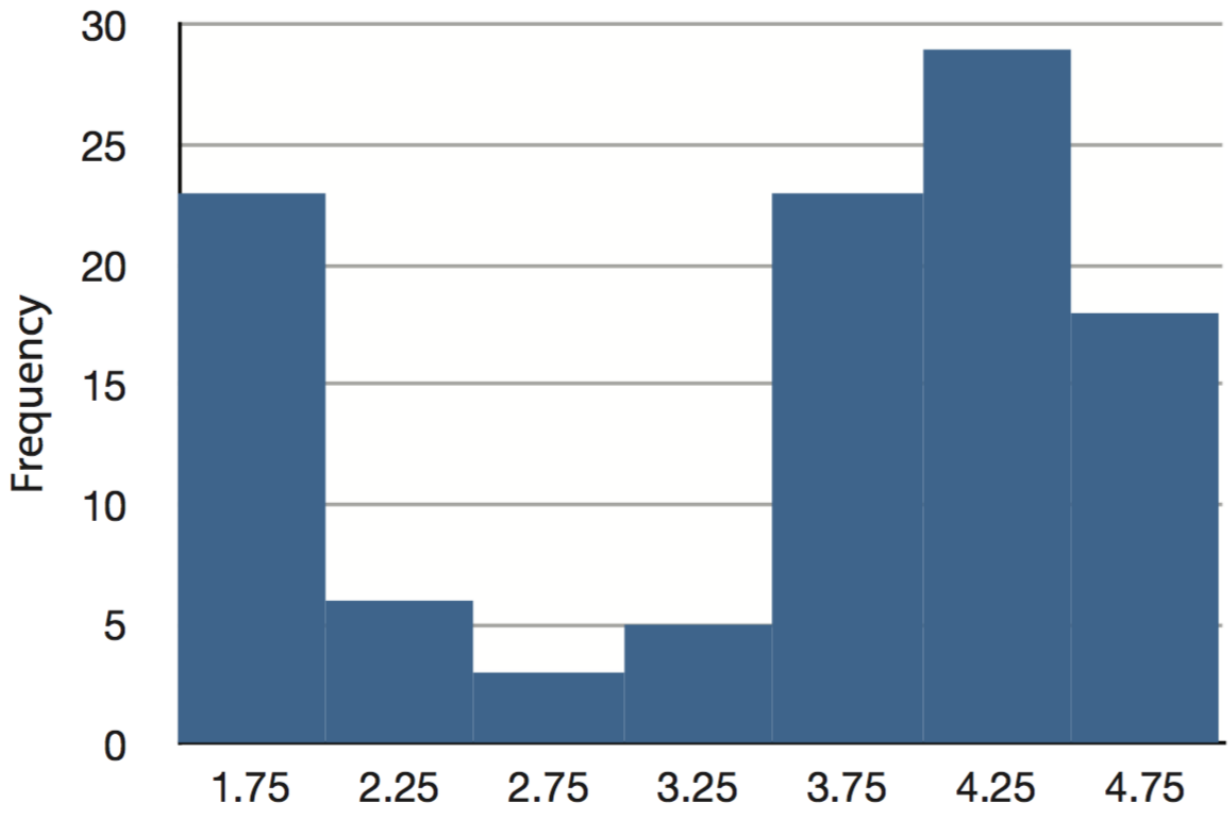


Discrete distribution



Continuous distribution

Bimodal Distribution



QCU1: distributions

A frequency distribution contains the frequency of every value in the distribution.

1. true
2. false

QCU1: distributions

A frequency distribution contains the frequency of every value in the distribution.

1. true
2. false

Answer: 1.

The distribution of empirical data is called a frequency distribution and consists of a count of the number of occurrences of each value.

QCU2: distributions

A grouped frequency distribution should be used instead of a frequency distribution when the

1. distribution is bimodal.
2. distribution is skewed.
3. variable is continuous.

QCU2: distributions

A grouped frequency distribution should be used instead of a frequency distribution when the

1. distribution is bimodal.
2. distribution is skewed.
3. variable is continuous.

Answer: 3.

When a variable is truly continuous, each value will have a frequency of 1. Therefore, grouped frequency distributions are needed with continuous variables.

QCU3: distributions

A symmetric distribution

1. has equal positive and negative skews.
2. has no skew.
3. can have either positive or negative skew, but not both.

QCU3: distributions

A symmetric distribution

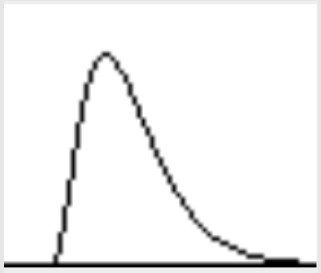
1. has equal positive and negative skews.
2. has no skew.
3. can have either positive or negative skew, but not both.

Answer: 2.

In a symmetric distribution, the tails extend equally in both directions. Therefore, there is no skew.

QCU4: distributions

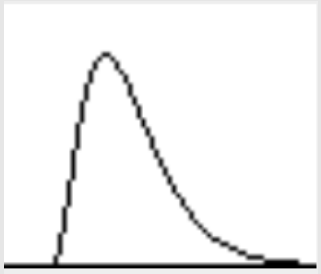
The following distribution has



1. a positive skew.
2. a negative skew.
3. no skew.

QCU4: distributions

The following distribution has



1. a positive skew.
2. a negative skew.
3. no skew.

Answer: 1.

The tail in the positive direction is longer than the tail in the negative direction, thus it has a positive skew.

QCU5: distributions

The area under the curve of a probability distribution is x.

QCU5: distributions

The area under the curve of a probability distribution is x.

Answer: 1.

The area is 1 by definition, meaning that the probability that a score chosen at random will occur under the curve is 1.

QCU6: distributions

A normal or bell-shaped distribution has its greatest probability density in its tails.

1. true
2. false

QCU6: distributions

A normal or bell-shaped distribution has its greatest probability density in its tails.

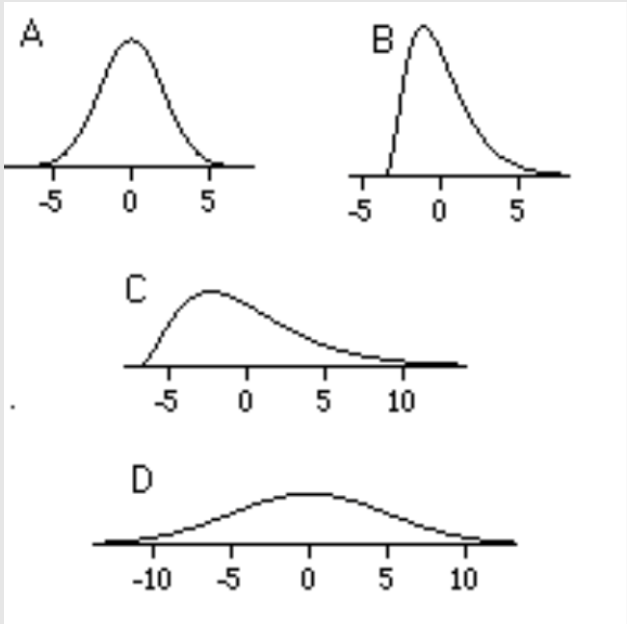
1. true
2. false

Answer: 2.

The distribution is higher and therefore denser in the middle of the distribution.

QCU7: distributions

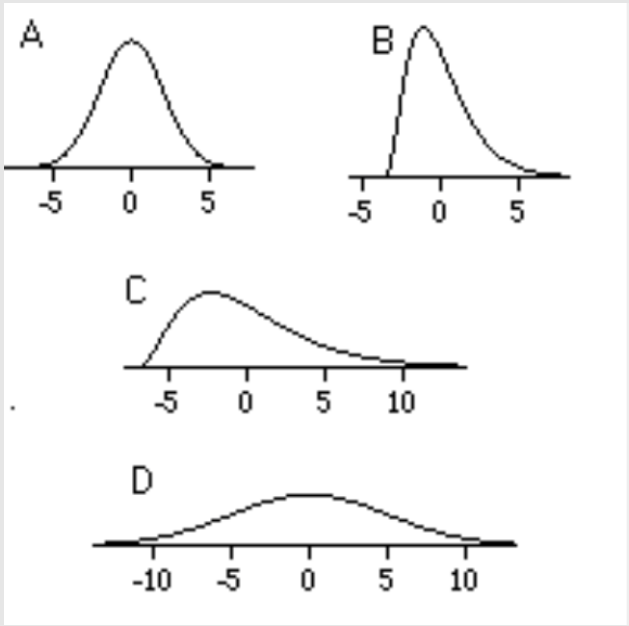
Which of the following distributions is/are symmetric?



1.
2.
3.
4.

QCU7: distributions

Which of the following distributions is/are symmetric?



1.
2.
3.
4.

Answer: 1 and 4.

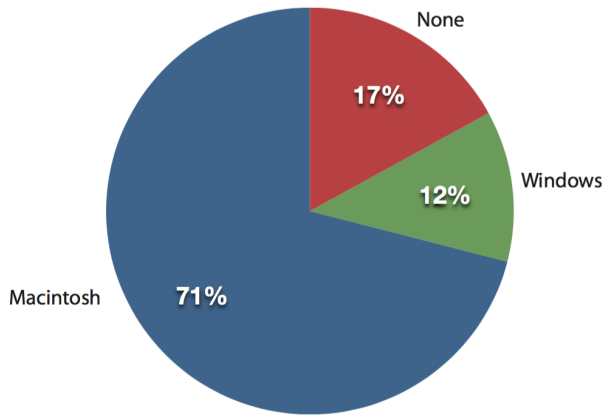
(1) and (4) are symmetric, meaning if you folded them in the middle, the two sides would match perfectly. Distributions (2) and (3) have positive skew.

Next section

Frequency Tables & Pie Charts

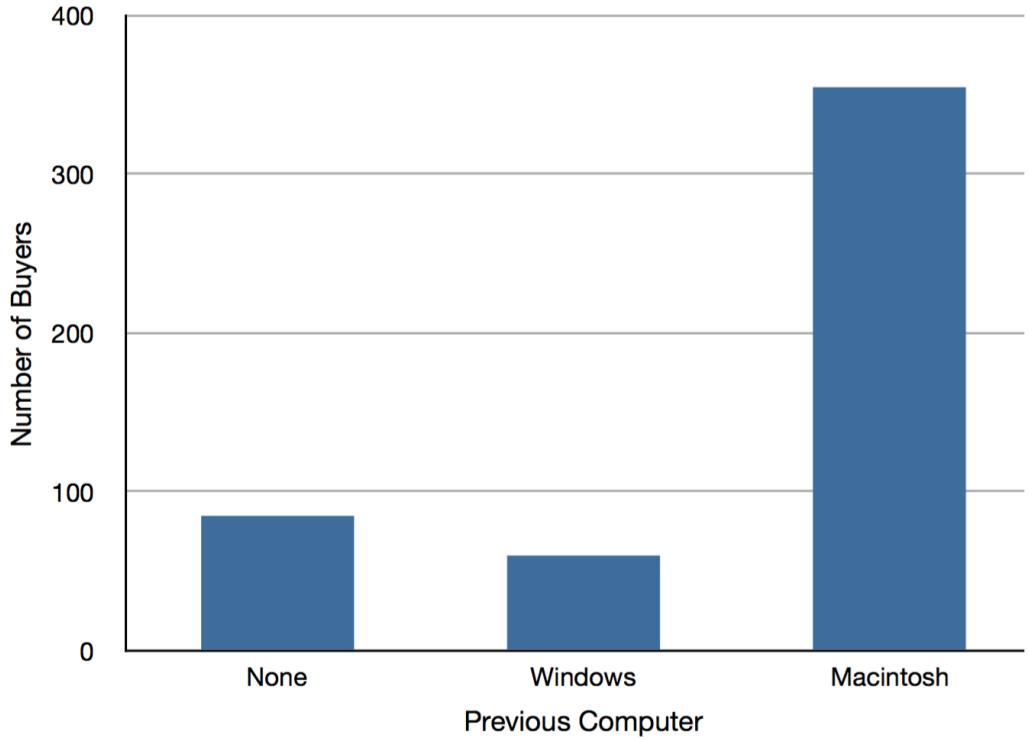
The results of Apple Computer study to learn whether the iMac was expanding Apple's market share.

Previous Ownership	Frequency	Relative Frequency
None	85	0.17
Windows	60	0.12
Macintosh	355	0.71
Total	500	1.00



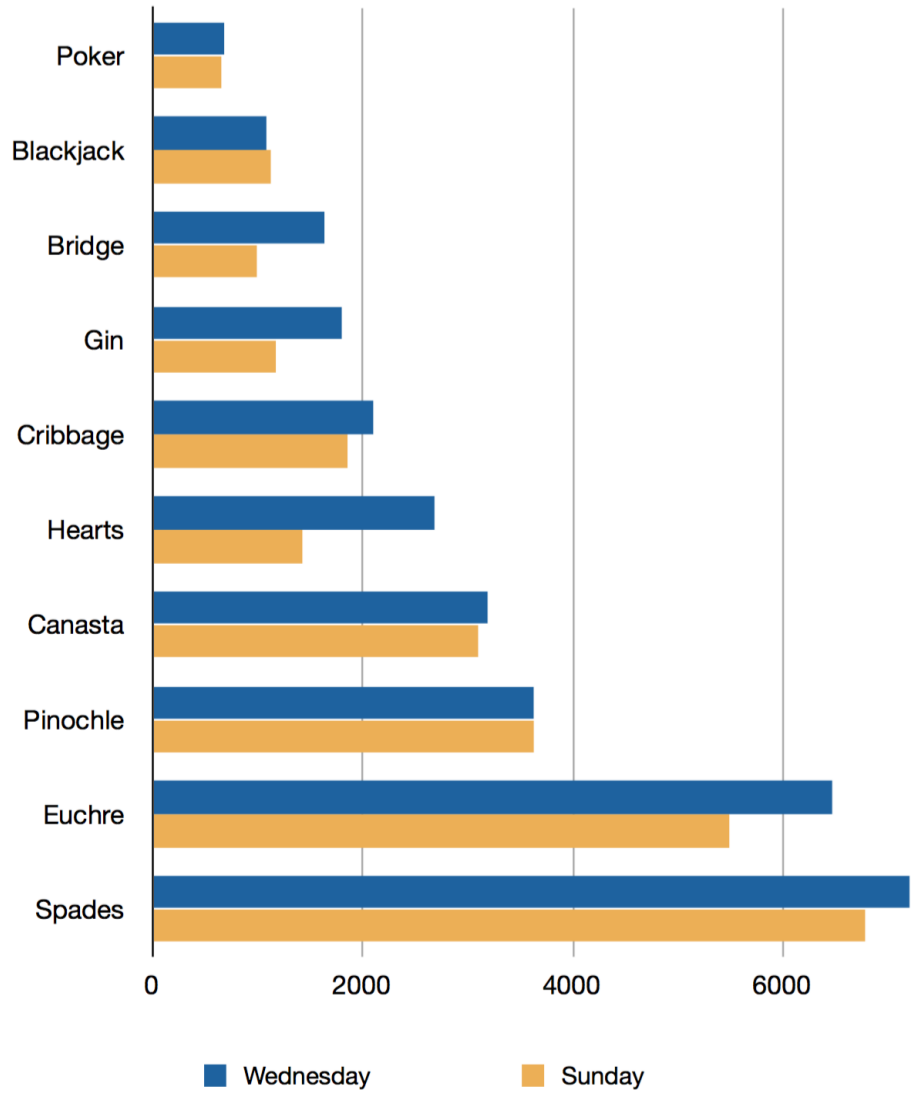
Bar charts

The results of Apple Computer study to learn whether the iMac was expanding Apple's market share.



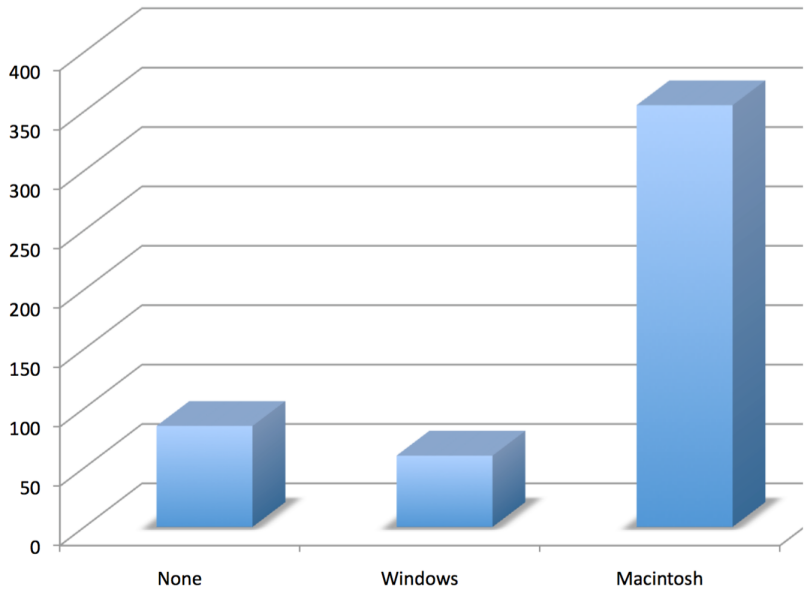
Comparing Distributions

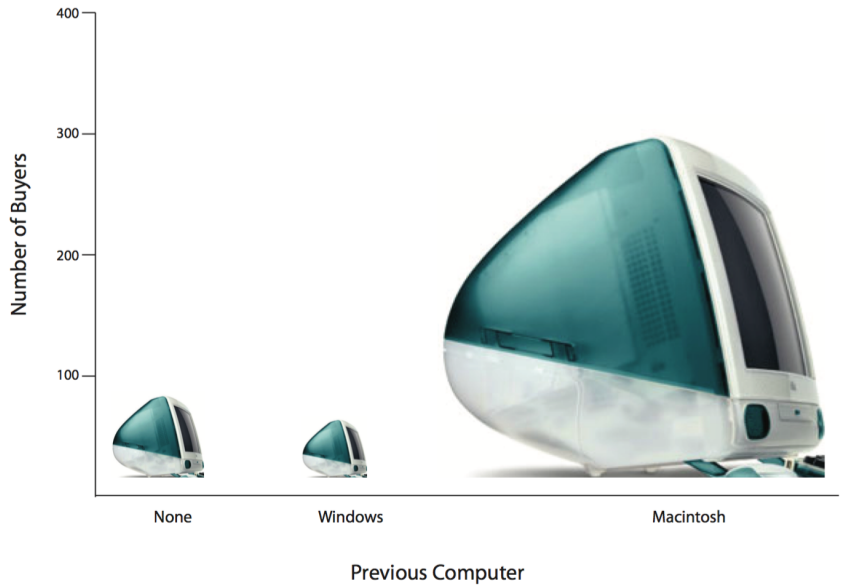
A bar chart of the number of people playing different card games on Sunday and Wednesday.



Some graphical mistakes to avoid (1)

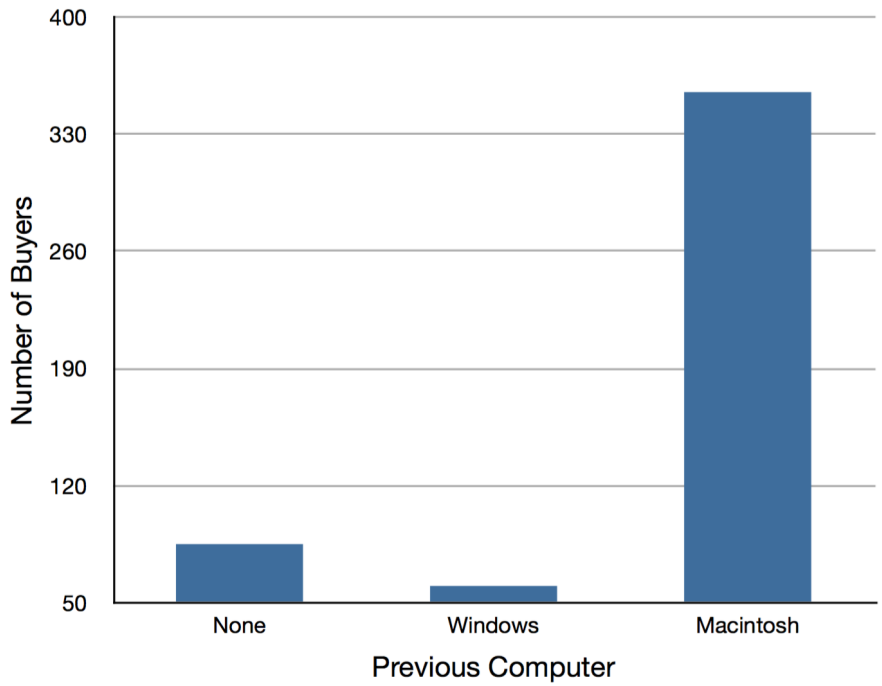
In the following examples the heights of the pictures accurately represent the number of buyers, yet they are misleading because the viewer's attention will be captured by volumes/areas.





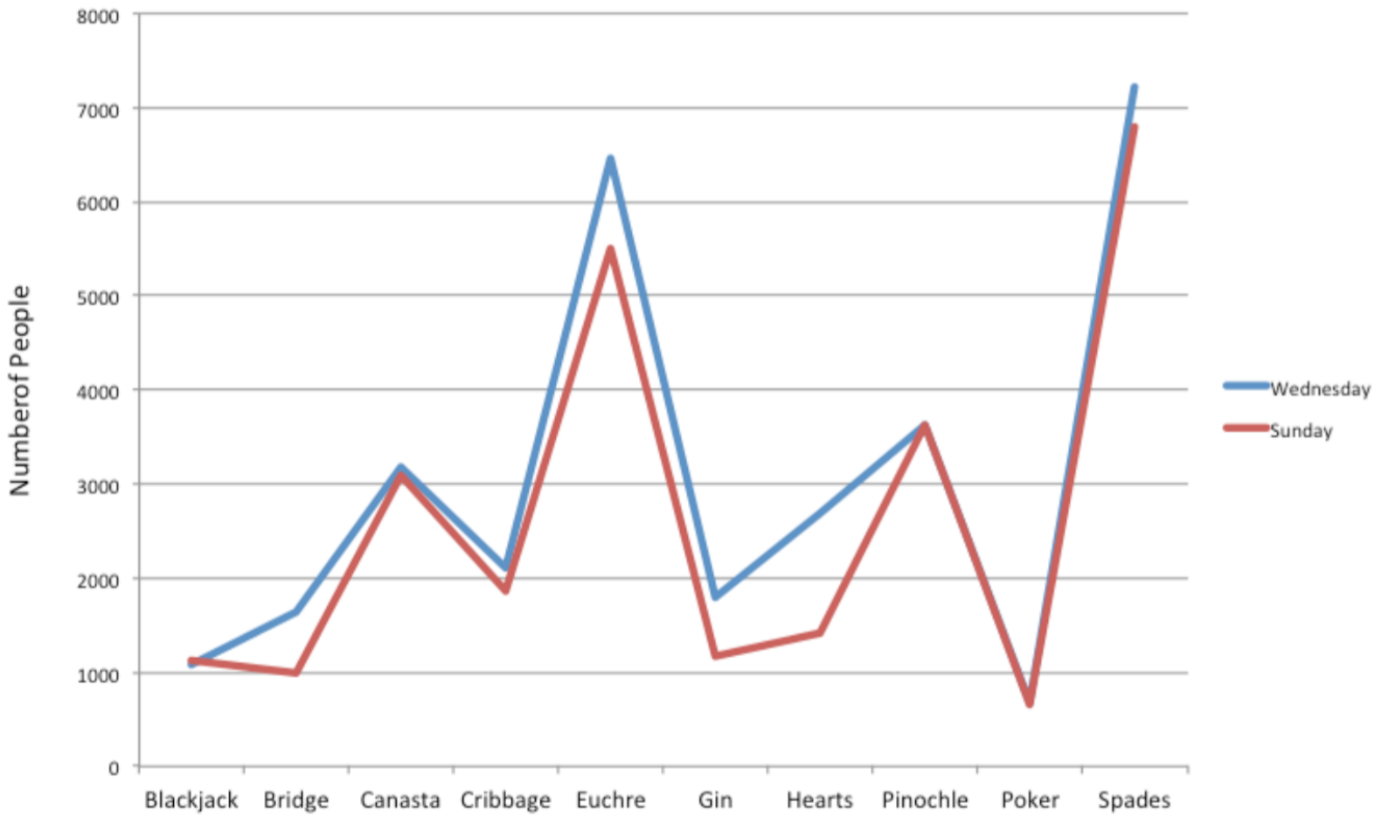
Some graphical mistakes to avoid (2)

Setting the baseline to a value other than zero may lead to misinterpretation. (Normally, but not always, this number should be zero.)



Some graphical mistakes to avoid (3)

Using a line graph when the X-axis contains merely qualitative variables that gives the false impression that the games are naturally ordered in a numerical way.



Statistical Literacy

Fox News have published the graph about the number unemployed during four quarters between 2007 and 2010.

This is misleading:

- The data show the number unemployed, Fox News' graph is titled "Job Loss by Quarter."
- The intervals on the X-axis are not equal and this gives the false impression that unemployment increased steadily (see correct graph om the next slide).

Statistical Literacy (2)