Solution *Section* 1.1 – Statistical Thinking

Exercise

Use common sense to determine whether the given event is (***a***) *impossible*; (***b***) *possible*, *but very unlikely*; (***c***) *possible and likely*.

a) Giants best the Denver Broncos in the Super Bowl by a score of 120 to 98.

b) While driving to his home in Connecticut, David was ticketed for driving 205 *mi/h* on a highway with a speed limit of 55 *mi/h*.

c) Thanksgiving Day will fall on a Monday nest year.

d) When each of 25 statistics students turns on his or her TI-84 Plus calculator, all 25 calculators operate successfully.

Solution

a) (***b***) Possible, but very unlikely.
Possible because there are no physical constraints or rules of the game that would prohibit such scores.

b) (***b***) Possible, but very unlikely.
Possible if he were driving a race car that could attain such a speed.

c) (***a***) Impossible, In the U.S., Thanksgiving Day always falls on a Thursday.

d) (***c***) Possible and likely

Exercise

Determine whether the underline value is a ***parameter*** or a ***statistic***.

a) Following the 2010 national midterm election, 18% of the governors of the 50 United States were female.

b) The average score for a class of 28 students taking a calculus midterm exam was 72%.

c) In a national survey of 1300 high school students (grades 9 to 12), 32% of respondents reported that someone has bullied them at school.

d) In a national survey on substance abuse, 10.0% of respondents aged 12 to 17 reported using illicit drugs within the past month.

e) Ty Cobb is one of major league baseball's greatest hitters of all time, with a career batting average of 0.366.

f) Only 12 men have walked on the moon. The average age of these men at the time of their moonwalks was 39 years, 11 months, 15 days.

g) A study of 6076 adults in public rest rooms (in Atlanta, Chicago, New York City, and San Francisco) found that 23% did not wash their hands before exiting.

h) Interviews of 100 adults 18 years of age or older, conducted nationwide, found that 44% could state the minimum age required for the office of U.S. president.

i) In a large sample of households, the median annual income per household for high school graduates is $19,856 (based on data from the U.S. Census Bureau).

j) Among the Senators in the current Congress, 44% are Democrats.

k) A study of all 2223 passengers aboard the Titanic found that 706 survived when it sank.
l) If the areas of the 50 states are added and the sum is divided by 50, the result is 196,533 square kilometers.
m) The average (mean) atomic weight of all elements in the periodic table is 134,355 unified mass units

Solution

a) 18% is a ***parameter*** because it describes a population (all of the governors).

b) 72% is a ***parameter*** because it describes a population (the entire class).

c) 32% is a ***statistic*** because it describes a sample (the high school students surveyed).

d) 10% is a ***statistic*** because it describes a sample (the youths surveyed).

e) 0.366 is a ***parameter*** because it describes a population (all Ty Cobb's at-bats).

f) 39 yrs, 11 mths, 15 days is a ***parameter*** because it describes a population (all Ty Cobb's at-bats).

g) 23% is a ***statistic*** because it describes a sample (the 6,076 adults studied).

h) 44% is a ***statistic*** because it describes a sample (the 100 adults interviewed).

i) ***Statistic***, since it was determined from a sample of households.

j) ***Parameter***, since it is based on the population of all Senators.

k) ***Parameter***, since it was determined from the population of all 2223 passengers.

l) ***Parameter***, since it was determined from the population of all 50 states.

m) ***Parameter***, since it was determined from the population of all elements in the periodic table.

Exercise

Classify the variable as ***qualitative*** or ***quantitative***

a) Nation of origin
b) Number of siblings
c) Grams of carbohydrates in a doughnut
d) Number on a football player's jersey
e) Number of unpopped kernels in a bag of ACT microwave popcorn
f) Assessed value of a house
g) Phone number
h) Student ID number.
i) Favorite film
j) Population of country of origin
k) Gallons of water in a swimming pool
l) Model of car driven
m) Distance in miles to nearest school
n) Time in hours that a light bulb lasts
o) Number of students at a high school

Solution

a) ***Qualitative***

b) Quantitative

c) Quantitative

d) Qualitative

e) Quantitative

f) Quantitative

g) Qualitative

h) Qualitative

i) Qualitative

j) Quantitative

k) Quantitative

l) Qualitative

m) Quantitative

n) Quantitative

o) Quantitative

Exercise

Determine whether the quantitative variable is ***discrete*** or ***continuous***

a) Goals scored in a season by a soccer player
b) Volume of water lost each day through a leaky faucet
c) Length (in minutes) os a country song
d) Number of Sequoia trees in a randomly selected acre of Yosemite National Park
e) Temperature on a randomly selected day in Memphis, Tennessee
f) Internet connection speed in Kilobytes per second
g) Points scored in an NCAA basketball game
h) Air pressure in pounds per square inch in an automobile tire
i) In the Literary Digest poll, Landon received 16,679,583 votes
j) The amount of nicotine in a Marlboro cigarette is 1.2 mg
k) The volume of cola in a can of regular coke is 12.3 oz
l) When a woman is randomly selected and measured for blood pressure, the systolic blood pressure is found to be 61 mm Hg

Solution

a) Discrete

b) Continuous

c) Continuous

d) Discrete

e) Continuous

f) Continuous

g) Discrete

h) Continuous

i) Discrete, since the number of votes received must be a whole number

j) Continuous, since the amount of nicotine could be any value on a continuum; even though it is reported to the nearest 0.1 mg.

k) Continuous, since the volume could be any value on a continuum; even though it is reported to the nearest 0.1 oz.

l) Continuous, since the systolic pressure could be any value on a continuum; even though it is reported to the nearest 61 mm Hg.

Exercise

Determine the level of measurement of each variable

a) Nation of origin
b) Movie ratings of one star through five stars
c) Volume of water used by a household in a day
d) Year of birth of college students
e) Highest degree conferred (high school, bachelor's, and so on)
f) Eye color
g) Assesses value of a house
h) Time of day measured in military time
i) Types of movies (drama, comedy, adventure, documentary, etc.)
j) Critic ratings of movies on a scale from 0 star to 4 stars
k) Ranks of cars evaluated by Consumer's Union

Solution

a) Nominal

b) Ordinal

c) Ratio

d) Interval

e) Ordinal

f) Nominal

g) Ratio

h) Interval

i) Nominal, since the data give category names only and there is no natural ordering.

j) Ordinal, since the categories have a natural ordering but the differences between the categories are not necessarily uniform.

k) Ordinal, since the ranks have a natural ordering but the differences between the ranks are not necessarily uniform.

Exercise

The Gallup Organization contacts 1026 teenagers who are 13 to 17 years of age and live in the United States and asks whether or not they had been prescribed medications for any mental disorders, such as depression or anxiety. Identify the population and sample.

Solution

The population consists of all teenagers 13 to 17 years old who live in the U.S.

The sample consists of the 1,028 teenagers 13 to 17 years old who were contacted by the Gallup Organization.

Exercise

A quality-control manager randomly selects 50 bottles of Coca-Cola that were filled on October 15 to assess the calibration of the filling machine. Identify the population and sample.

Solution

The population consists of all bottles of Coca-Cola filled by that particular machine on October 15.

The sample consists of the 50 bottles of Coca-Cola that were selected by the quality control manager.

Exercise

Each *x* value associated with the corresponding *y* value in some meaningful way? If the *x* and *y* values are not matched, does it make sense to use the difference between each *x* value and the *y* value that is the same column?

Nicotine Amounts from Menthol and King-Size Cigarettes

x	1.1	0.8	1.0	0.9	0.8
y	1.1	1.7	1.7	1.1	1.1

The *x*-values are nicotine amounts (in *mg*) in different 100 *mm* filtered, non-light menthol cigarettes; the *y*-values are nicotine amounts (in mg) in different king-size non-filtered, non-menthol, and non-light cigarettes.

Solution

There is nothing in the data to suggest that the *x* and *y* values are matched. If the *x* and *y* values are not matched as presented, it does not make sense to calculate the differences between the *x* and *y* values – moving *x* values around, for example, would yield a different set of differences.

Exercise

The Federal Trade Commission obtained the measured amounts of nicotine in the table. Is the source of the data likely to be unbiased?
Note that the table lists measured nicotine amounts from two different types of cigarette. Given these data, what issue can be addressed by conducting a statistical analysis of the values?

Nicotine Amounts from Menthol and King-Size Cigarettes

x	1.1	0.8	1.0	0.9	0.8
y	1.1	1.7	1.7	1.1	1.1

Solution

Yes, since the Federal Trade Commission receives no financial or other benefit from a particular set of results, there is no reason for the data or the data selected to be biased.

Exercise

One of Gregor Mendel's famous hybridization experiments with peas yielding 580 off spring with 152 of those peas (or 26%) having yellow pods. According to Mendel's theory, 25% of the off spring peas should have yellow pods. Do the results of the experiment differ from Mendel's claimed rate of 25% by an amount that is statistically significant?

Solution

Determining whether or not the difference between Mendel's actual results (26%) and the results predicted by his theory (25%) is statistically significant properly requires applying some techniques. Common sense suggests the 1% difference is of no practical difference. Considering the sample size, the actual difference between the observed and expected results is 152 – 145 = 7. Common sense suggests that a discrepancy of 7 (relative to an expected result of 145 plants from a total sample of 580 plants) is within the natural fluctuation inherent biological processes, and that the difference is not statistically significant.

Exercise

In a Gallup poll of 1038 randomly selected adults, 85% said that secondhand smoke is somewhat harmful or very harmful, but a representative of the tobacco industry claims that only 50% of adults believe that secondhand smoke is somewhat harmful or very harmful. Is there statistically significant evidence against the representative's claim? Why or why not?

Solution

Determining whether or not the difference between the survey results (85%) and the statement of the industry representative (50%) is statistically significant properly. But common sense suggests that this is statistically significant evidence against the representative's claim because (1) the observed result of 85% is so much greater than the stated claim of the 50% and (2) the sample of 1038 randomly selected adults appears to be properly chosen and large enough to provide reliable data.

Exercise

Determine whether the given value is parametric or a statistic

a) One of greatest baseball hitters of all time has a career batting average of 0.366
b) A sample of employees is selected and it is found that 50% own a vehicle
c) A survey of 42 out of hundreds in a dining hall showed that 17 enjoyed their meal

Solution

a) Parameter
b) Statistic because the value is a numerical measurement describing a characteristic of a sample
c) Statistic, because the data set of 42 people in a dining hall is a sample.

Exercise

Suppose a survey of 568 women in the U.S. found that more than 61% are the primary investor in their household.

a) Describe the survey represents the descriptive branch of statistic
b) Make an inference based on the results of the survey

Solution

a) 61% of women in the sample are the primary investor in their household
b) There is an association between U.S. women and being the primary investor in their household

Exercise

In the recent study, volunteers who had 8 hours of sleep were three times more likely to answer questions correctly on a math test than were sleep-deprived participaants.

a) Identify the sample used in the study
b) What is the sample's population
c) Which part of the study represents the descriptive branch of statistics
d) Make an inference based on the results of the study

Solution

a) The sample is the responses of the volunteers in the study
b) The population is the collection of the responses of all individuals who completed the math test
c) The statement "three times more likely to answer questions correctly" is an example of descriptive statistics
d) Individuals who are not sleep deprived will be more likely to answer math questions correctly than individuals who are sleep deprived.

Exercise

Determine whether the data set is a population or a sample. Explain your reasoning

The salary if each baseball player in a league

Solution

Population, because it is a collection of salaries for all baseball players in the league

Exercise

In a poll, 1,004 adults in a country were asked whether they favor or oppose the use of "federal tax dollars to find medical research using stem cells obtained from human embryos." Among the responders, 48% said that they were in favor. Describe the statistical study

a) What is the population?
b) Identify the sample

Solution

a) All adults in the country
b) The 1,004 adults selected

Exercise

A study shows that the obesity rate among boys ages 2 to 19 has increased over the past several years

a) Make an inference based on the results of this study?
b) What is wrong with this type of reasoning

Solution

a) The obesity rate among boys ages 2 to 19 is increasing
b) This inference may incorrectly imply that the trend will continue in future years.

Exercise

The newspaper USA Today published a health survey, and some readers completed the survey and returned it. Identify the (*a*) sample and (*b*) population, also determine whether the sample likely to be representative of the population.

Solution

a) The sample is the readers who completed and returned the survey.
b) The population is all people who read USA Today.
c) No, since the sample is self-selected it is not likely to be representative of the population. Only those with special interest in health matters are likely to return the survey.

Exercise

A Gallup poll of 1012 randomly surveyed adults found that 9% of them said cloning of humans should be allowed. Identify the (*a*) sample and (*b*) population, also determine whether the sample likely to be representative of the population.

Solution

a) The sample is the 1012 randomly selected adults.
b) The population is all adults. This should probably be understood to be all adults living in the U.S.
c) Yes, since the adults were selected at random by an organization with experience in polling and with no vested interest in the results, the sample is likely to be representative of the population.

Exercise

Some people responded to this request: "Dial 1-900-PRO-LIFE to participate in a telephone poll on abortion. ($1.95 per minute. Average call: 2 minutes. You must be 18 years old.)" Identify the (*a*) sample and (*b*) population, also determine whether the sample likely to be representative of the population

Solution

a) The sample is the people who responded to the request.

b) The intended population is likely all persons over 18. The actual population is actually all persons over 18 who have opportunity to receive the request. If the request went out over the internet, for example, the population would be all internet users over 18. If the request went out over several radio and/or TV stations, the population would be all persons over 18 who tune in to those stations.

c) No, since the sample is self-selected it is not likely to be representative of the population. Only those with special interest in abortion issues are likely to spend the time and money to respond to the survey.

Exercise

In the Born Loser cartoon strip by Art Sansom, Brutus expresses joy over an increase in temperature from 1° to 2°. When asked what is so good about 2°, he answers that "it's twice as warm as this morning." explain why Brutus is wrong yet again.

Solution

Temperature ratios are not meaningful because a temperature of 0° does not represent the absence of temperature in the same sense that $0 represents the absence of money. The zero temperature in the exercise (whether Fahrenheit or Centigrade) was determined by a criterion other than "the absence of temperature."

Exercise

A group of students develops a scale for rating the quality of cafeteria food, with 0 representing "neutral: not good and not bad." Bad meals are given negative numbers and good meals are given positive numbers, with the magnitude of the number corresponding to the severity of badness or goodness. The first three meals are rated as 2, 4, and –5. What is the level of measurement for such rating? Explain your choice.

Solution

This is example of ordinal data. It is not interval data because differences are not meaningful. The difference between the ratings +4 and +5 does not necessarily represent the same differential in the quality of food as the difference between 0 and +1.

Exercise

Suppose that a study based on a sample from a targeted population shows that people who own a fax machine have more money than people who do not

a) Make an inference based on the results of this study?

b) What might this inference incorrectly imply?

Solution

a) People who own a fax machine have more money than people who do not

b) This inference may incorrectly inply that if you owned a fax machine, you would have more money than if you did not.

Exercise

Determine whether the statement is true or false, rewrite it as a true statement

a) Data at the ordinal level are quantitative only

b) More types of calculations can be performed with data at the nominal level than with data at the interval level

Solution

a) False. Data at the ordinal level can be qualitative or quantitative

b) False. More types of calculations can be performed with data at the interval level than with data at the nominal level.

Exercise

The region of a country with the highest per capita income for the past six years is shown below

Northeast Southern Southwest Southeast Northern Western

a) Determine whether the data are qualitative or quantitative and identify the data set's level of measurement

b) What is the data set's level of measurement?

Solution

a) Qualitative

b) Nominal

Exercise

The region of a country with the six highest level of food production last year are shown below

1. Eastern 2. Southwest 3. Western 4. Southeast 5. Northwest 6. Southern

a) Determine whether the data are qualitative or quantitative and identify the data set's level of measurement

b) What is the data set's level of measurement?

Solution

a) Quantitative
b) Ordinal

Exercise

The region of a country with the six highest level of food production last year are shown below

22.8 26.4 24.1 22.2 21.6 21.1 25.8 21.5 24.6

a) Determine whether the data are qualitative or quantitative and identify the data set's level of measurement

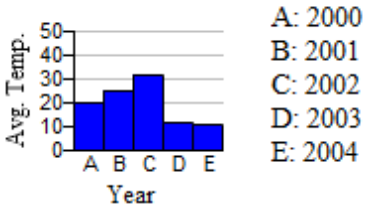
b) What is the data set's level of measurement?

Solution

a) Quantitative
b) Ratio

Exercise

The graph shows the average temperature in an artic city, in degree Fahrenheit, for certain years. Identify the level of measurement of the data listed on the horizontal axis in the graph



Solution

Interval

Exercise

Identify the level of measurement of the data:

a) Temperature
b) Age
c) Family history of illness
d) Pain level (scale of 0 to 10)

Solution

a) Interval
b) Ratio
c) Nominal
d) Ordinal

Exercise

A study was conducted in which 20,211 18-years old male military were given an exam to measure IQ. In addition, the recruits were asked to disclose their smoking status. An individual was considered a smoker if he smoked at least once cigarette per day. The goal of the study was to determine whether adolescents aged 18 to 21 who smoked have a lower IQ than nonsmokers. It was found that the average IQ of the smokers was 94, while he average IQ of the nonsmokers was 101. The researchers concluded that lower IQ individuals are more likely to choose to smoke, not that smoking makes people less intelligent.

a) What is the research objective?
b) What is the population being studied? What is the sample?
c) What are the descriptive statistics?
d) What are the conclusions of the study?

Solution

a) The research objective is to determine if adolescents who smoke have a lower IQ than nonsmokers.
b) The population is all adolescents aged 18-21. The sample consisted of 20,211 18 years old military recruits.
c) Descriptive statistics: The average IQ of the smokers was 94, and the average IQ of nonsmokers was 101.
d) The conclusion is that individuals with a lower IQ are more likely to choose to smoke.

Exercise

Determine whether the variable is qualitative, continuous, or discrete. The following represent information on smart phones.

Model	***Weight*** (oz.)	***Service Provider***	***Depth*** (in)
Motorola Droid *X*	5.47	Verizon	0.39
Motorola Droid 2	5.96	Verizon	0.53
Apple iPhone 4	4.8	ATT	0.37
Samsung Epic 4G	5.5	Sprint	0.6
Samsung Captivate	4.5	ATT	0.39

Solution

Individuals: Model (Motorola Droid X & 2, Apple, and Samsung)
Variables: Weight (oz), Service Provider, Depth (in).
Data for weight: 5.47, 5.96, 4.8, 5.5, 4.5 (oz.)
Data for service provider: Verizon, Verizon, ATT, sprint, ATT
Data for depth: 0.39, 0.53, 0.37, 0.6, 0.39 (in)

The variable weight is ***continuous***.
The variable service provider is ***qualitative***.
The variable depth is ***continuous***.

Solution *Section* 1.2 – Observational Studies vs Designed Experiments

Exercise

Researchers wanted to know if there is a link between proximity to high-tension wires and the rate of leukemia in children. To conduct the study, researchers compared the rate of leukemia for children who lived within ½ mile of high-tension wires to the rate of leukemia for children who did not live within ½ mile of high-tension wires. Determine whether the study depicts an observational study or an experiment.

Solution

This is an observational study because the researchers merely observed existing data. There was no attempt by the researchers to manipulate or influence the variable(s) of interest.

Exercise

Rats with cancer are divided into two groups. One group receives 5 milligrams (mg) of a medication that is thought to fight cancer, and the other receives 10 mg. After 2 years, the spread of the cancer is measured. Determine whether the study depicts an observational study or an experiment.

Solution

This is an experiment because the researchers intentionally changes the value of the explanatory variable (medication dose) to observe a potential effect on the response variable (cancer growth).

Exercise

Seventh-grade students are randomly divided into two groups. One group is taught math using traditional techniques; the other is taught math using a reform method. After 1 year, each group is given an achievement test to compare proficiency. Determine whether the study depicts an observational study or an experiment.

Solution

This is an experiment because the explanatory variable (teaching method) was intentionally varied to see how is affected the response variable (score on proficiency test).

Exercise

A poll is conducted in which 500 people are asked whom they plan to vote for in the upcoming election. Determine whether the study depicts an observational study or an experiment.

Solution

This is an observational study because no attempt was made to influence the variable of interest. Voting choices were merely observed.

Exercise

A survey is conducted asking 400 people. “Do you prefer Coke or Pepsi?” Determine whether the study depicts an observational study or an experiment.

Solution

This is an observational study because the survey only observed preference of Coke or Pepsi. No attempt was made to manipulate or influence the variable of interest.

Exercise

A Gallup poll surveyed 1018 adults by telephone, and 22% of them reported that they smoked cigarettes within the past year. Determine whether the description corresponds to an observation study or an experiment.

Solution

Observational study, since the poll involves collecting data from unmodified subjects.

Exercise

In a morally and criminally wrong study, 399 black men with syphilis were not given a treatment that could have cured them. The intent was to learn about the effects of syphilis on black men, The subjects were initially treated with small amounts of bismuth, neoarsphenamine, and mercury, but those treatments were replaced with aspirin. Determine whether the description corresponds to an observation study or an experiment.

Solution

Experiment, since the effect of an applied treatment (in this case a zero dose of the appropriate medicine) was measured.

Exercise

While shopping, 200 people are asked to perform a taste test in which they drink from two randomly placed, unmarked cups. They are then asked which drink they prefer. Determine whether the description corresponds to an observation study or an experiment.

Solution

This is an experiment because the researcher intentionally imposed treatments on individuals in a controlled setting.

Exercise

Conservation agents netted 250 large-mouth bass in a lake and determined how many were carrying parasites. Determine whether the description corresponds to an observation study or an experiment.

Solution

This is an observation study because the conservation agents merely observed the fish to determine which were carrying parasites. No attempt was made to manipulate or influence any variable of interest.

Exercise

Researchers wanted to determine if there was an association between the level of happiness of an individual and their risk of heart disease. The researchers studied 1739 people over the course of 10 years. During this 10-year period, they interviewed the individuals and asked questions about their daily lives and the hassles they face. In addition, hypothetical scenarios were presented to determine how each individual would handle the situation. These interviews were videotaped and studied to assess the emotions of the individuals. The researchers also determined which individuals in the study experienced any type of heart disease over the 10-year period. After their analysis, the researchers concluded that the happy individuals were less likely to experience heart disease.

a) What type of observational study is this? Explain.
b) What is the response variable?
c) What is the explanatory variable?
d) In the report, the researchers stated that "the research team also hasn't ruled out that a common factor like genetics could be causing both the emotions and the heart disease." Use the language introduced on this section to explain what this sentence means.

Solution

a) This is a cohort study because the researchers observed a group of people over a period of time.
b) The response variable is whether the individual has heart disease or not.
c) The explanatory variable is whether the individual is happy or not.
d) There may be confounding due to lurking variables. For example, happy people may be more likely to exercise, which could affect whether they will have heart disease or not.

Exercise

Researchers wanted to determine if there was an association between daily coffee consumption and the occurrence of skin cancer. The researchers looked at 93,676 women enrolled in the Women's Health Initiative Observation Study and asked them to report their coffee-drinking habits. The researchers also determined which of the women had nonmelanoma skin cancer. After their analysis, the researchers concluded that consumption of six or more cups of caffeinated coffee per day was associated with a reduction in nonmelanoma skin cancer

a) What type of observational study is this? Explain.
b) What is the response variable?

c) What is the explanatory variable?

d) In their report, the researchers stated that "After adjusting for various demographic and lifestyle variables, daily consumption of six or more cups was associated with a 30% reduced prevalence of nonmelanoma skin cancer." Why was it important to adjust for these variables?

Solution

a) This is a cross-sectional study because the researchers collected information about the individuals at a specific point in time.

b) The response variable is whether the woman has nonmelanoma skin cancer or not.

c) The explanatory variable is the daily amount of caffeinated coffee consumed.

d) It was necessary to account for these variables to avoid confounding due to lurking variables.

Exercise

Researcher Penny Gordon-Larson and her associate wanted to determine whether young couples who marry or cohabitate are more likely to gain weight than those who stay single. The researchers followed 8000 men and women for 7 years as they matured from teens to young adults. When the study began, none of the participants were married or living with a romantic partner. By the end of the study, 14% of the participants were married and 16% were living with a romantic partner. The researchers found that married or cohabiting women gained, on average, 9 pounds more than single.

a) Why is this an observation study? What type of observational study is this?

b) What is the response variable in the study?

c) What is the explanatory variable?

d) Identify some potential lurking variables in this study.

e) Can we conclude that getting married or cohabiting causes one to gain weight? Explain.

Solution

a) This is an observational study because the researchers merely observed the individuals included in the study. No attempt was made to manipulate or influence any variable of interest. This is a cohort study because the researchers identified the individuals to be included in the study, then followed them for a period of time (7 years).

b) The response variable is weight gain.

c) The explanatory variable is whether the individual is married/cohabitating or not.

d) Answers may vary. Some potential lurking variables are eating habits, exercise routine, and whether the individual has children.

e) No since this is an observational study, we can only say that being married or cohabitating is associated with weight gain.

Solution *Section* 1.3 – Sampling Methods

Exercise

A student of the author collected measurements of arm lengths from her family members. Identify what type is used: random, systematic, convenience, stratified, or cluster.

Solution

Convenience, since the sample is those who happened to be in the student's family.

Exercise

On the day of the last presidential election, ABC News organized an exit poll in which specific polling stations were randomly selected and all voters were surveyed as they left the premises. Identify what type is used: random, systematic, convenience, stratified, or cluster.

Solution

Cluster, since all the voters at randomly selected polling stations were surveyed.

Exercise

The author was an observer at a town of Poughkeepsic Police sobriety checkpoint at which every fifth driver was stopped and interviewed. (He witnessed the arrest of a former student.) Identify what type is used: random, systematic, convenience, stratified, or cluster.

Solution

Systematic, since every 5^{th} driver was stopped

Exercise

You observed professional wine taster working at the Consumer's Union testing facility in NY. Assume that a taste test involves three different wines randomly selected from each of five different wineries. Identify what type is used: random, systematic, convenience, stratified, or cluster.

Solution

Stratified, since the population of wines was subdivided into 5 different subgroups (wineries), and then samples were drawn from each subgroup.

Exercise

The U.S. Department of Corrections collects data about returning prisoners by randomly selecting five federal prisons and surveying all of the prisoners in each of the prisons. Identify what type is used: random, systematic, convenience, stratified, or cluster.

Solution

Cluster, since all the prisoners at five randomly selected prisons were surveyed.

Exercise

You instructor surveyed all of his students to obtain sample consisting of the number of credit cards students possess. Identify what type is used: random, systematic, convenience, stratified, or cluster.

Solution

Convenience, since the sample is those who happen to be students of the author.

Exercise

In a study of college programs, 820 students are randomly selected from those majoring in communications, 1463 students are randomly selected from those majoring in business, and 760 students are randomly selected from those majoring in history. Identify what type is used: random, systematic, convenience, stratified, or cluster.

Solution

Stratified, since the population was subdivided into 3 different subgroups, and then samples were drawn from each subgroup. The population of interest appears to be only communications, business and history majors – most likely because the programs being studied involved only those majors.

Exercise

Pharmacists typically fill prescriptions by scooping a sample of pills from a larger batch that is in stock. A pharmacist thoroughly mixes a large batch of Lipitor pills, then selects 30 of them. Does this sampling plan result in a random sample? Simple random sample? Explain.

Solution

Yes, it is a random sample because each pill has an equal chance of being selected.
Yes, it is a simple random sample because each sample of size 30 has an equal chance of being selected.

Exercise

A quality control engineer selects every $10{,}000^{th}$ M&M plain candy that is produced. Does this sampling plan result in a random sample? Simple random sample? Explain.

Solution

Whether the sample is a random depends on how the first selection is made. If the engineer chooses the first one at random from 1 to 10,000 and every 10,000th one thereafter, then every M&M has an equal chance of being selected (namely 1 in 10,000) and the sample is a random sample. If the engineer determines to start with #1 and choose every 10,000th one thereafter, then some M&M's have no chance of being selected (e.g., #2) and the sample is not a random sample.

No, no matter how the first selection is made the sample will not be a simple random sample of size n. All possible grouping of size n are not possible – any grouping containing #1 and #1, for example, could not occur.

Exercise

NBC News polled reactions to the last presidential election by surveying adults who were approached by a reporter at a location in N.Y. City. Does this sampling plan result in a random sample? Simple random sample? Explain.

Solution

No, this is not a random sample of all New Yorkers because persons who did not visit the location had not chance of being selected.

No it is not a simple random because it is not a random sample.

Exercise

A classroom consists of 36 students seated in six different rows, with six students in each row. The instructor rolls a die to determine a row, then rolls the die again to select a particular student in the row. This process is repeated until a sample of 6 students is obtained. Does this sampling plan result in a random sample? Simple random sample? Explain.

Solution

Yes, the results in a random sample because each student has an equal chance of being selected.

Yes, this results in a simple random sample of size 6 because each possible grouping of size 6 has an equal chance to occur.

Exercise

A computer company employs 100 software engineers and 100 hardware engineers. The personnel manager randomly selects 20 of the software engineers and 20 of the hardware engineers and questions them about career opportunities within the company. Does the sampling plan result in a random sample? Simple random sample? Explain.

Solution

Yes, the sample is random because all employees have the same chance of being selected.

No, it is not a simple random sample because some samples are possible, such as a sample consisting of 30 software engineers and 10 hardware engineers.

Exercise

A polling company obtains an alphabetical list of names of voters in a precinct. They select every 20^{th} person from the list until a sample of 100 is obtained. They then call these 100 people. Does the sampling plan result in a random sample? Simple random sample? Explain.

Solution

No. The sample is not random because not all voters have the same chance of being selected.
No, the second person on the list has no chance of being selected. It is not a simple random sample because some samples are not possible, such as a sample containing the second person on the list.

Exercise

What is an inherent zero? Describe three examples of data sets that have inherent zeroes and three that do not.

Solution

An inherent zero is a zero that implies none.

1. Average age of college students in years
2. Maximum wing speed during a hurricane
3. Average monthly precipitation in inches

Exercise

What is the different between a random sample and a simple random sample?

Solution

With a random sample, each individual has the same chance of being selected. With a simple random sample, all samples of the same size have the same chance of being selected.

Exercise

Determine whether the statement is true or false. If false, rewrite it as a true statement

a) In a randomized block design, subjects with similar characteristics are divided into blocks, and then, within each block, randomly assigned to treatment groups.
b) Using a systematic sample guarantees that members of each group within a population will be sampled.
c) The method for selected a stratified sample is to order a population in some way and then select members of the population at regular intervals.

Solution

a) True
b) False. Using a stratified sample guarantees that members of each group within a population will be sampled.
c) False. The method for selecting systematic sample is to order a population in some way and then select members of the population at regular intervals.

Exercise

Which method of data collection should be used to collect data for the following study

a) A study of the health of 148 kidney transplant patients at a hospital.
b) A study of the effect on the taste of a snack food made with a sugar substitute
c) A study of how fast a virus would spread in a herd of cattle.

Solution

a) Census
b) Experiment
c) Simulation

Exercise

A pharmaceutical company wants to test the effectiveness of a new allergy drug. The company identifies 250 females 30-35 years old who suffer from severe allergies. The subjects are randomly assigned into two groups. One group is given the new allergy drug and the other is given a placebo that looks exactly like the new allergy drug. After six months, the subjects' symptoms are studied and compared

a) Identify the experimental units and treatment used in this experiment.
b) Identify a potential problem with the experiment design being used and suggest a way to improve it.
c) How could this experiment be designed to be a double-blind?

Solution

a) The experiment units are the 30- to 35- year old females being given the treatment is the new allergy drug.
b) There may be a bias on the part of the researcher knows which patients were given the real drug.
c) The study would be a double-blind study if both the researcher and the patient did not know which patient received the real drug or the placebo

Exercise

What type of sampling is used: random, stratified, convenience, cluster, systematic, in the following?

a) To estimate the percentage of defects in a recent manufacturing batch, a quality-control manager at Intel selects every 8^{th} chip that comes off the assembly line starting with the 3^{rd} until she obtains a sample of 140 chips.
b) To determine the prevalence of human growth hormone (HGH) use among high school varsity baseball players, the State Athletic Commission randomly selects 50 high schools. All members of the selected high schools' varsity baseball teams are tested for HGH.
c) To determine customer opinion of its boarding policy. Southwest Airlines randomly selects 60 lights during a certain week and surveys all passengers on the flights.
d) A member of Congress wishes to determine her constituency's opinion regarding estate taxes. She divides her constituency into three income classes: low-income households, middle-income households, and upper-income households. She then takes a simple random sample of households from each income class.

e) In an effort to identify whether an advertising campaign has been effective, a marketing firm conducts a nationwide poll by randomly selecting individuals from a list of known users of the product.

f) A radio station asks it listeners to call in their opinion regarding the use of U.S forces in peacekeeping missions.

g) A farmer divides his orchard into 50 subsections, randomly selects 4, and samples all the trees within the 4 subsections to approximate the yield of this orchard.

h) A college official divides the student population into five classes: freshman, sophomore, junior, and graduate student. The official takes a simple random sample from each class and asks the members' opinions regarding student services.

i) Toyota wants to administer a satisfaction survey to its current customers. Using their customer database, the company randomly selects 80 customers and asks them about their level of satisfaction with the company

j) To determine her power usage, Dan divides up his day into three parts: morning, afternoon, and evening. He then measures his power usage at 3 randomly selected times during each part of the day.

k) A newspaper asks its readers to call in their opinion regarding the number of books they have read this month.

l) Toshiba wants to administer a satisfaction survey to its current customers. Using their customer database, the company randomly selects 80 customers and asks them about their level of satisfaction with the company.

m) An education researcher randomly selects 48 middle schools and interviews all the teachers at each school.

n) A market researcher selects 500 drivers under 30 years of age and 500 drivers over 30 years of age.

o) To avoid working late, a quality control analyst simply inspects the first 100 items produced in a day.

Solution

a) ***Systematic sampling***. The quality control manager is sampling every 8^{th} chip.

b) ***Cluster sampling***. The commission tests all members of the selected teams (clusters).

c) ***Cluster sampling***. The airline surveys all passengers on selected flights (clusters).

d) ***Stratified sampling***. The congresswoman samples some individuals from each of three different income brackets (strata).

e) ***Simple random sampling***. Each known user of the product has the same chance of being included in the sample.

f) ***Convenience sampling***. The radio station is relying on voluntary response to obtain the sample data.

g) ***Cluster sampling***. The farmer samples all trees within the selected subsection (clusters).

h) ***Stratified sampling***. The school official takes a sample of students from each of the four grades (strata).

i) ***Simple random***

j) ***Stratified sampling***

k) ***Convenience sampling***

l) ***Simple random***

m) ***Cluster sampling***

n) Stratified sampling

o) Convenience sampling

Exercise

Determine whether you would take a census or use a sampling to collect data for the study described:

a) The average credit card debt of the 65 employees of a company

b) The most popular grocery store among the 40,000 employees of a company

Solution

a) Census

b) Sampling

Exercise

Determine if the survey question is biased. If the question is biased, suggest a better wording

a) Why drinking fruit juice good for you?

b) Why is eating ice cream bad for you?

Solution

a) Yes. How do you think drinking fruit juice affects your health?

b) Yes. How do you think eating ice cream affects your health?

Exercise

A company has been rating television programs for more than 60 years. It uses several sampling procedures, but its main one is to track the viewing patterns of 20,000 households. These contain more than 45,000 people and are chosen to form a cross section of the overall population. The households represent various locations, ethnic groups, and income brackets. The data gathered from the sample of 20,000 households are used to draw inferences about the population of all households in the U.S.

a) What strata are used in the sample?

b) Why is it important to have a stratified sample for these ratings?

c) Observation studies are sometimes referred to as natural experiments. Explain what this means

Solution

a) The various locations, ethnic groups, and income brackets that are represented.

b) Stratified sampling ensures that each segment of the population is represented.

c) In an observation study, a researcher measures characteristics of interest of a part of a population but does not change existing conditions.

Exercise

Some polling agencies ask people to call a telephone number and give their response to a question

a) What is an advantage of this type of survey?
b) What is disadvantage of this type of survey?
c) Identify the sampling technique used.

Solution

a) This usually results in a savings in the survey cost.
b) There tends to be a lower response rate and this may introduce a bias into the sample. Only a certain segment of the population might respond.
c) Convenience sampling.

Exercise

A computer company employs 100 software engineers and hardware engineers. The personnel manager randomly selects 20 of the software engineers and 20 of the hardware and questions them about career opportunities within the company. Does this sampling plan result in a random sample? Simple random sample? Explain.

Solution

Yes it is random sample and it is not a simple random sample.
The sample is random because all employees have the same chance of being selected. It is not a simple random sample because some examples are not possible, such as a sample consisting of 30 software engineers and 10 hardware engineers

Exercise

Suppose you are the president of the student government. You wish to conduct a survey to determine that student body's opinion regarding student services. The administration provides you with a list of the names and phone numbers of the 19,935 registered students.

a) Discuss the procedure you would follow to obtain a simple random sample of 25 students.
b) Obtain this sample

Solution

a) The list provided by the administration serves as the frame. Number each student in the list of registered students, from 1 to 19,935. Generate 25 random numbers, without repetition, between 1 and 19,935.
b) Answers will vary

Exercise

True or False

a) When taking a systematic random sample of size n, every group of size n from the population has the same chance of being selected

b) A simple random sample is always preferred because it obtains the same information as other sampling plans but requires a smaller sample size.

c) When conducting a cluster sample, it is better to have fewer cluster with more individuals when the clusters are heterogeneous.

d) Inferences based on voluntary response samples are generally not reliable.

e) When obtaining a stratified sample, the number of individuals included within each stratum must be equal.

Solution

a) ***False***. In a systematic random sample, every kth individual is selected from the population.

b) ***False***. In many cases, other sampling techniques may provide equivalent or more information about the population with less "cost" than simple random sampling.

c) ***True***. When clusters are heterogeneous, the heterogeneity of each cluster likely resembles the heterogeneity of the population. In such cases, fewer clusters with more individuals from each cluster are preferred.

d) ***True***. Because the individuals in a convenience sample are not selected is not representative of the population.

e) ***False***. With stratified samples, the number of individuals sampled from each strata should be proportional to the size of the strata in the population.

Exercise

The human resource department at a certain company wants to conduct a survey regarding worker morale. The department has an alphabetical list of all 4502 employees at the company and wants to conduct a systematic sample.

a) Determine *k* if the sample size is 50

b) Determine the individuals who will be administered the survey. More than one answer is possible.

Solution

a) $\frac{N}{n} = \frac{4502}{50} = 90.04 \to 90$; thus $\underline{k = 90}$

b) Randomly select a number between 1 and 90. Suppose that we select 15. Then the individuals to be surveyed will be the 15^{th}, 105^{th}, 19^{th}, 285^{th}, and so on up uo the 4425^{th} employee on the company list.

Exercise

To predict the outcome of a county election, a newspaper obtains a list of all 945,035 registered voters in the county and wants to conduct a systematic sample.

a) Determine k if the sample size is 130

b) Determine the individuals who will be administered the survey. More than one answer is possible.

Solution

a) $\frac{N}{n} = \frac{945,035}{130} = 7269.5 \rightarrow 7269$; thus $\boxed{k = 7269}$

b) Randomly select a number between 1 and 7269. Suppose that we select 2000. Then we will survey the individuals numbered 2000, 9269, 16,538, 23,807, and so on up to the individual numbered 939,701.

Solution *Section* 1.4 – Design of Experiments

Exercise

A school psychologist wants to test the effectiveness of a new method for teaching reading. She recruits 500 first-grade students in District 203 and randomly divides them into two groups. Group 1 is taught by means of the new method, while group 2 is taught by traditional methods. The same teacher is assigned to teach both groups. At the end of the year, an achievement test is administered and the results of the two groups are compared.

a) What is the response variable in this experiment?
b) Think of some of the factors in the study. How are they controlled?
c) What are the treatments? How many treatments are there?
d) How are the factors that are not controlled dealt with?
e) Which group serves as the control group?
f) What type of experimental design is this?
g) Identify the subjects.

Solution

a) The response variable is the achievement test scores.
b) Some factors are teaching methods, grade level, intelligence, school district, and teacher.
Fixed: grade level, school district, teacher
Set at predetermined levels: teaching method.
c) The treatments are the new teaching method and the traditional method. There are 2 levels of treatment.
d) The factors that are not controlled are dealt with by random assignment into the two treatment groups.
e) Group 2, using the traditional teaching method, serves as the control group.
f) This experiment has a completely randomized design.
g) The subjects are the 500 first-grade students from District 203 recruited for the study.

Exercise

A pharmaceutical company has developed an experimental drug meant to relieve symptoms associated with the common cold. The company identifies 300 adult males 25 to 29 years old who have a common cold and randomly divides them into 2 groups. Group 1 is given the experimental drug, while group 2 is given a placebo. After 1 week of treatment, the proportions of each group that still have cold symptoms are compared.

a) What is the response variable in this experiment?
b) Think of some of the factors in the study. How are they controlled?
c) What are the treatments? How many treatments are there?
d) How are the factors that are not controlled dealt with?
e) What type of experimental design is this?
f) Identify the subjects.

Solution

a) The response variable is the proportion of subjects with a cold.
b) Some factors are gender, age, geographic location, overall health, and drug intervention,
Fixed: gender, age, location
Set at predetermined levels: drug intervention
c) The treatments are the experimental drug and the placebo. There are 2 levels of treatment
d) The factors that are not controlled are dealt with by random assignment into the two groups.
e) This experiment has a completely randomized design.
f) The subjects are the 300 adult males aged 25 to 29 who have the common cold.

Exercise

Researchers wanted to compare the effectiveness and safety of an extract of St. John's wort with placebo in outpatients with major depression. To do this, they recruited 200 adult outpatients diagnosed as having major depression and having a baseline Hamilton Rating Scale for Depression (HAM-D) score of at least 20. Participants were randomly assigned to receive either St. John's wort extract, 900 *mg* per day (*mg/d*) for a weeks, increased to 1200 *mg/d* in the absence of an adequate response thereafter, or a placebo for 8 weeks. The response variable was the change on the HAM-D over the treatment period. After analysis of the data, it was concluded that St. John's wort was not effective for treatment of major depression.

a) What type of experimental design is this?
b) What is the population that is being studied?
c) What is the response variable in this study?
d) What are the treatments?
e) Identify the experimental units.
f) What is the control group in this study?

Solution

a) This experiment has a completely randomized design
b) The population being studied is adult outpatients diagnosed as having major depression and having a baseline HAM-D score of at least 20.
c) The response variable is the change in the HAM-D over the treatment period.
d) The explanatory variable is the type of drug. The treatments are St. John's wort extract and the placebo.
e) The experimental units are the 200 adults outpatients diagnosed with depression.
f) The control group is the placebo group.

Exercise

Researchers wanted to evaluate whether ginkgo, an over-the-counter herb marketed as enhancing memory, improves memory in elderly adults as measured by objective tests. To do this, they recruited 96 men and 132 women older than 60 years and in good health. Participants were randomly assigned to receive ginkgo, 40 *mg* 3 times per day, or a matching placebo. The measure of memory improvement was determined by a standardized test of learning and memory. After 6 weeks of treatment, the data indicated that ginkgo did not increase performance on standard tests of learning, memory, attention, and concentration. These data suggest that, when taken following the manufacturer's instructions, ginkgo

provides no measurable increase in memory or related cognitive function to adults with healthy cognitive function.

a) What type of experimental design is this?
b) What is the population being studied?
c) What is the response variable in this study?
d) What is the factor that is set to predetermined levels? What are the treatments?
e) Identify the experimental units.
f) What is the control group in this study?

Solution

a) This experiment has a completely randomized design.
b) The population being studied is adults over 60 years old and in good health.
c) The response variable is the standardized test of learning and memory.
d) The factor set to predetermined levels (explanatory variable) is the drug. The treatments are 40 *mg* of ginkgo 3 times per day and the matching placebo.
e) The experimental units are the 98 men and 132 women over60 years old and in good health.
f) The control group is the placebo group.

Solution Section 1.5 – Organizing Qualitative Data

Exercise

In a study, researchers treated 570 people who smoke with either nicotine gum or a nicotine patch. Among those treated with nicotine gum, 191 continued to smoke and the other 59 stopped smoking. Among those treated with nicotine patch, 263 continued to smoke and the other 57 stopped smoking (based on data from the Center for Disease Control and Prevention). Construct the relative frequency distribution.

Solution

Obtain the relative frequencies by dividing the given frequencies by the total of 570.

Those whose smoking		***Relative Frequency***
Continued after the gum	$\frac{191}{570} = 0.335$	33.5 %
Stopped after the gum	$\frac{59}{570} \approx 0.104$	10.4 %
Continued after the patch	$\frac{263}{570} \approx 0.461$	46.1 %
Stopped after the patch	$\frac{57}{570} \approx 0.1$	10.0 %
		100.0 %

Exercise

Heights of statistics students were obtained by the author as part of a study conducted for class. The last digits of those heights are listed below. Construct a frequency distribution with 10 classes. Based on the distribution, do the heights appear to be reported or actually measured? What do you know about the accuracy of the results?

0 0 0 0 0 0 0 0 0 1 1 2 3 3 3 4 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 6 6 8 8 8 9

Solution

Digit	***Frequency***
0	9
1	2
2	1
3	3
4	1
5	15
6	2
7	0
8	3
9	1
	37

The heights are not actually measured because the data have been round off and showed unusually high numbers of 0's and 5's.

Exercise

Listed below are amounts of strontium-90 (in millibecquerels) in a simple random sample of baby teeth obtained from Pennsylvania residents born after 1979, Construct a frequency distribution with eight classes. Begin with a lower class limit of 110, and use a class width of 0. Cite a reason why such data are important.

155	142	149	130	151	163	151	142	156	133	138	161	128	144
172	137	151	166	147	163	145	116	136	158	114	165	169	145
150	150	150	158	151	145	152	140	170	129	188	156		

Solution

Lowest number: 114 Highest number: 188

level	***Frequency***
110 – 119	2
120 – 129	2
130 – 139	5
140 – 149	9
150 – 159	13
160 – 169	6
170 – 179	2
180 – 189	1
	40

Such data are important because they can be helpful in detecting potentially dangerous situations and in making recommendations for future action

Exercise

Refer to the data below and use the 40 voltage measurements from the generator. Construct a frequency distribution with seven classes. Begin with a lower class limit of 123.9 volts, and use a class width of 0.20 volt. Using a very loose interpretation of the relevant criteria, does the result appear to have a normal distribution?

124.8	124.3	125.2	124.5	125.1	124.8	125.1	125.0	124.8	124.7	124.5	125.2	124.4	124.7
124.9	124.5	124.8	124.8	124.5	124.6	125.0	124.7	124.9	124.9	124.7	124.2	124.7	124.8
124.4	124.6	124.4	124.0	124.7	124.4	124.6	124.6	124.6	124.8	124.3	124.0		

Solution

Voltage (volts)	*Frequency*
123.9 – 124.0	2
124.1 – 124.2	1
124.3 – 124.4	6
124.5 – 124.6	9
124.7 – 124.8	13
124.9 – 125.0	5
125.1 – 125.2	4

Yes;
The voltages do appear to follow a normal distribution.
There are many values near the center of the distribution, and the frequencies diminish toward either end.

Exercise

As part of the Garbage Project at the University of Arizona, the discarded garbage for 62 households was analyzed. Refers to the 62 weights from table below and construct a frequency distribution. Begin with a lower class of 1.00 lb., and use a class width of 4.00 lb. Do the weights of discarded paper appear to have a normal distribution?

2.41	11.08	9.45	5.88	7.57	12.43	12.32	8.26	9.55	6.05	20.12	12.45
8.82	13.61	7.72	10.58	8.72	6.98	6.16	5.87	6.96	14.33	7.98	8.78
6.83	13.31	9.64	11.03	11.42	3.27	8.08	12.29	16.08	6.67	10.99	20.58
6.38	17.65	13.11	12.56	13.05	12.73	3.26	9.92	11.36	9.83	1.65	3.45
15.09	16.39	10	9.09	2.8	6.33	8.96	3.69	6.44	9.19	9.46	2.61
5.86	9.41										

Solution

Weight (lbs)	*Frequency*
1.00 – 4.99	8
5.00 – 8.99	21
9.00 – 12.99	22
13.00 – 16.99	8
17.00 – 20.99	3
	62

Yes,
The weights appear to have a distribution that is approximately normal.

Exercise

a) Refer to the data below for the FICO credit rating scores. Construct a frequency distribution beginning with a lower class limit of 400, and use a class width of 50. Does the result appear to have a normal distribution? Why or why not?

708	713	781	809	797	793	711	681	768	611	698	729	829
836	768	532	657	559	741	792	701	753	745	681	594	744
598	693	743	444	502	739	755	835	714	517	787	706	752
714	497	636	637	797	568	714	618	830	579	818	722	783
751	731	850	591	802	756	689	789	654	617	849	604	630
628	692	779	756	782	760	503	784	798	611	709	661	579
591	834	694	795	660	651	696	638	697	732	796	753	782
635	795	519	682	824	603	709	777	664				

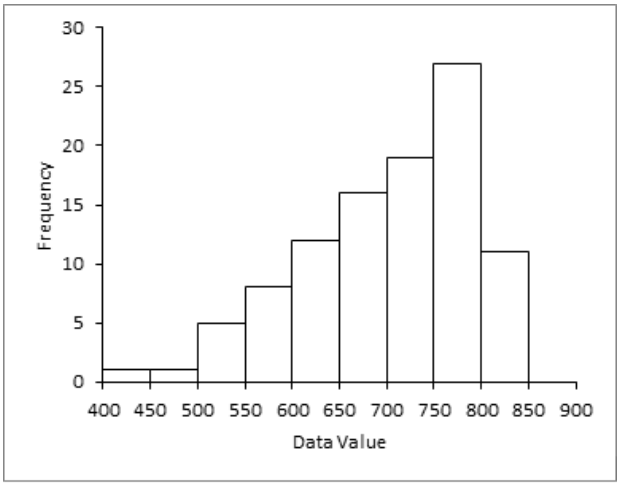
b) Use the table to construct a histogram. Does the result appear to be normal distribution? Why or why not?

Solution

a)

FICO score	***Frequency***
400 – 449	1
450 – 499	1
500 – 549	5
550 – 599	8
600 – 649	12
650 – 699	16
700 – 749	19
750 – 799	27
800 – 849	10
850 – 899	1
	100

b)



The data appear to have a distribution that is not normal. While there is a maximum value with progressively smaller frequencies on either side of the maximum, the distribution is definitely not symmetric.

Exercise

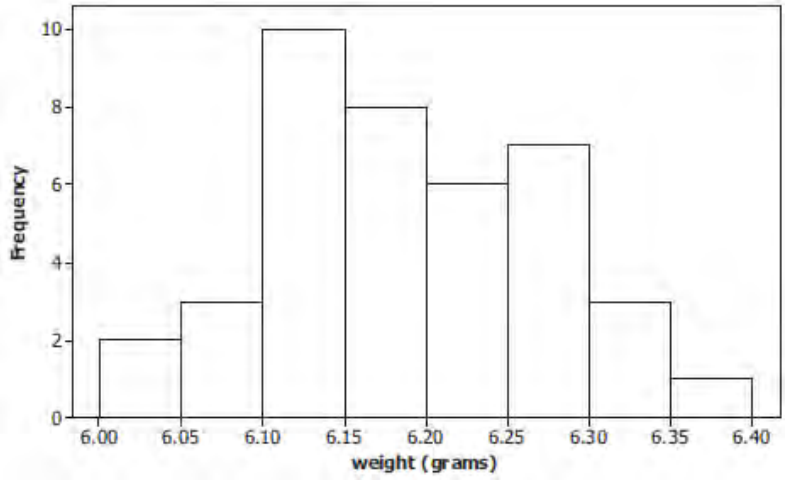
Refer to the data in the table below. Construct a frequency distribution. Begin with lower class limit of 6.0000 g, and use a class width of 0.0500 g.

6.2771	6.2371	6.1501	6.0002	6.1275	6.2151
6.2866	6.0760	6.1426	6.3415	6.1309	6.2412
6.1442	6.1073	6.1181	6.1352	6.2821	6.2647
6.2908	6.1661	6.2674	6.2718	6.1949	6.2465
6.3172	6.1487	6.0829	6.1423	6.1970	6.2441
6.3669	6.0775	6.1095	6.1787	6.2130	6.1947
6.1940	6.0257	6.1719	6.3278		

Solution

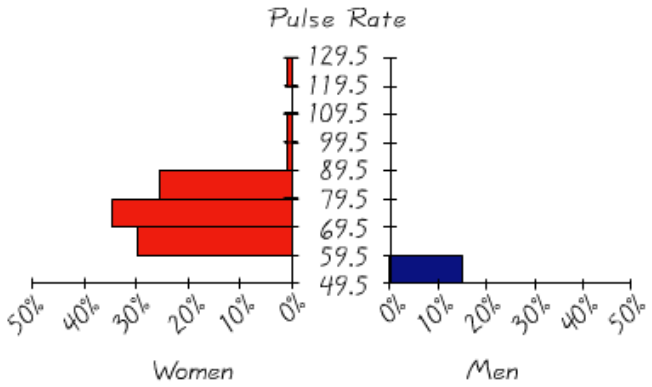
Weight (g)	*Frequency*
6.0000 – 6.04999	2
6.0500 – 6.0999	3
6.1000 – 6.1499	10
6.1500 – 6.1999	8
6.2000 – 6.2499	6
6.2500 – 6.2999	7
6.3000 – 6.3499	3
6.3500 – 6.3999	1
	40

b)



Exercise

When using histograms to compare two data sets, it is sometimes difficult to make comparisons by looking back and forth between the two histograms. A back-to-back relative frequencies histogram uses a format that makes the comparison much easier. Instead of frequencies, we should use relative frequencies (percentages or proportions) so that the comparisons are not distorted by different sample sizes. Complete the back-to-back relative frequency histograms shown below by using the data below. Then use the result to compare the two data sets.

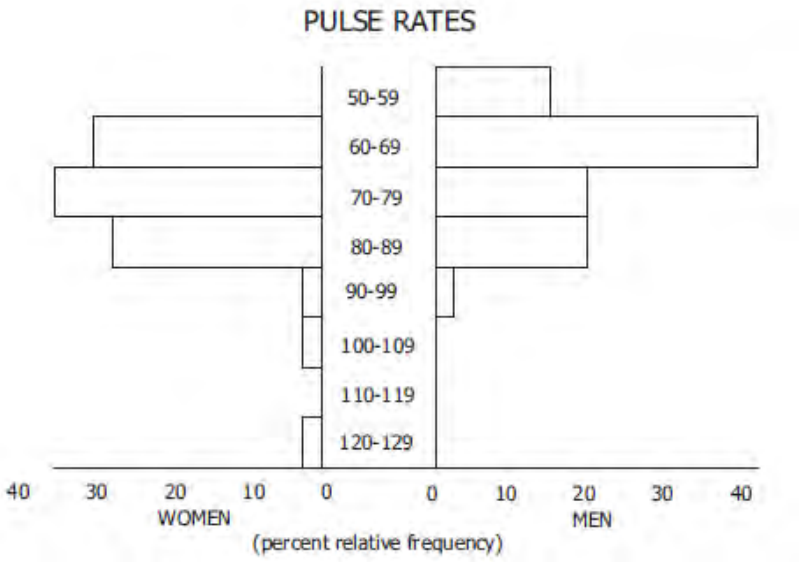


Pulse Rates (*beats per minute*) of Females and Males

Females																			
76	72	88	60	72	68	80	64	68	68	80	76	68	72	93	72	68	72	64	80
64	80	76	76	76	80	104	88	60	76	72	72	88	80	60	72	88	88	124	64
Males																			
68	64	88	72	64	72	60	88	76	60	96	72	56	64	60	64	84	76	84	88
72	56	68	64	60	68	60	60	56	84	72	84	88	56	64	56	56	60	64	72

Solution

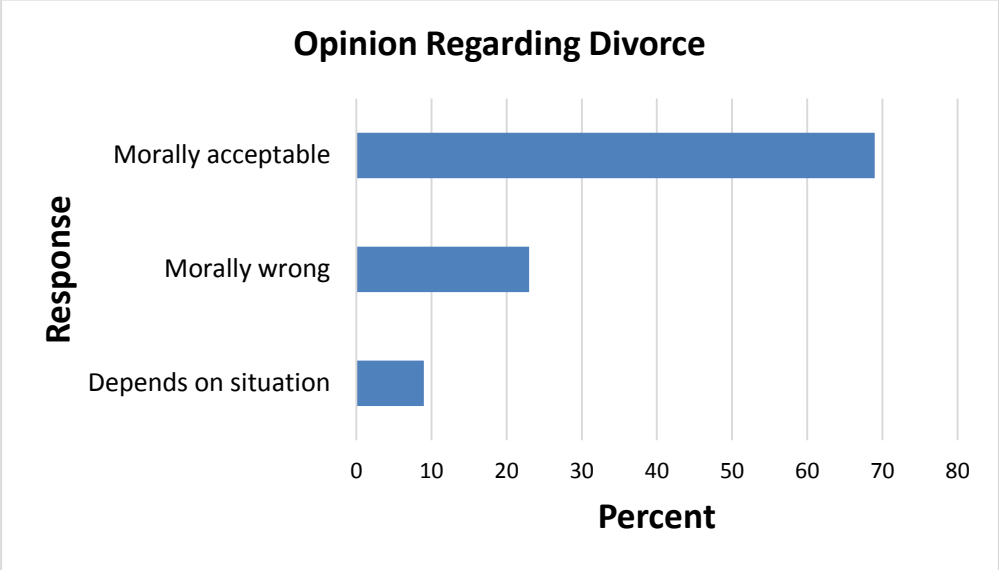
The back-to-back relative frequency histograms are given below. The pulse rates of the males tend to be lower than those of the females.



Exercise

The following graph represents the results of a survey, by Gallup in May 2010, in which a random sample of adult Americans was asked, "Please tell me whether you personally believe that in general divorce is morally accepted or morally wrong."

Response	***Percent***
Depends on situation	9
Morally wrong	23
Morally acceptable	69



a) What percent of the respondents believe divorce is morally acceptable?
b) If there were 240 million adult Americans, how many believe that divorce is morally wrong?
c) If Gallup claimed that the results of the survey indicate that 8% of adult Americans believe that divorce is acceptable in certain situations, would you say this statement is descriptive or inferential? Why?

Solution

a) 69% of the respondents believe divorce is morally acceptable
b) 23% believe divorce is morally wrong. So, 240 million * 0.23 = 55.2 million adult Americans believe divorce is morally wrong.
c) This statement is inferential, since it is a generalization based on the observed data.

Exercise

In a national survey conducted by the Centers for Disease Control to determine health-risk behaviors among college students, college students were asked, "How often do you wear a seat belt when driving a car?" The frequencies were as follow:

Response	***Frequency***
I do not drive a car	249
Never	118
Rarely	249
Sometimes	345
Most of the time	716
Always	3093

a) Construct a relative frequency distribution
b) What percentage of respondents answered "Always"?
c) What percentage of respondents answered "Never" or "Rarely"?
d) Construct a frequency bar graph.
e) Construct a relative frequency bar graph.
f) Construct a pie chart
g) Suppose that a representative from the Centers for Disease Control says, "2.5% of the college students in this survey responded that they never wear a seat belt." Is this a descriptive or inferential statement?

Solution

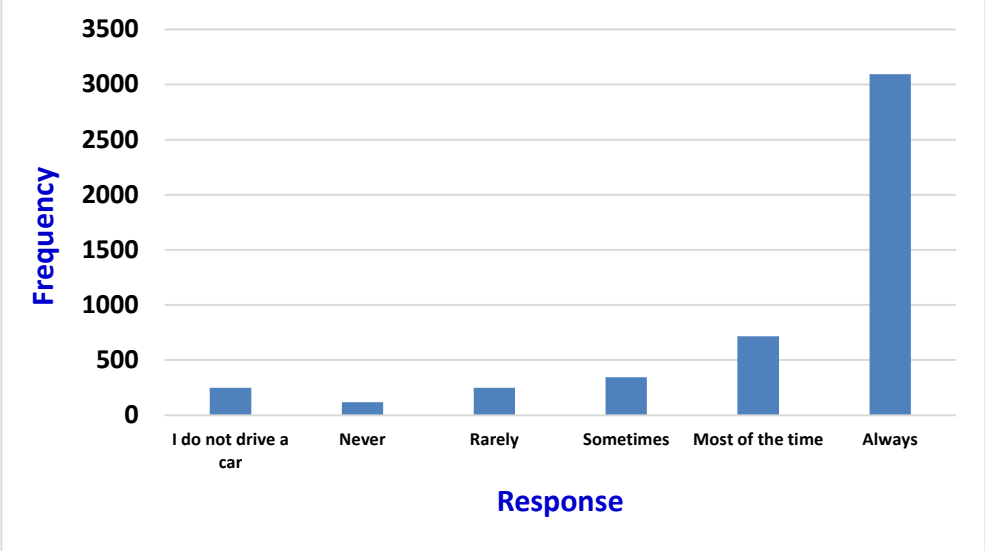
a)

Response	***Frequency***	***Relative Frequency***
I do not drive a car	249	$\frac{249}{4770} \approx 0.0522$
Never	118	$\frac{118}{4770} \approx 0.0247$
Rarely	249	$\frac{249}{4770} \approx 0.0522$
Sometimes	345	$\frac{345}{4770} \approx 0.0723$
Most of the time	716	$\frac{716}{4770} \approx 0.1501$
Always	3093	$\frac{3093}{4770} \approx 0.6484$
	4770	

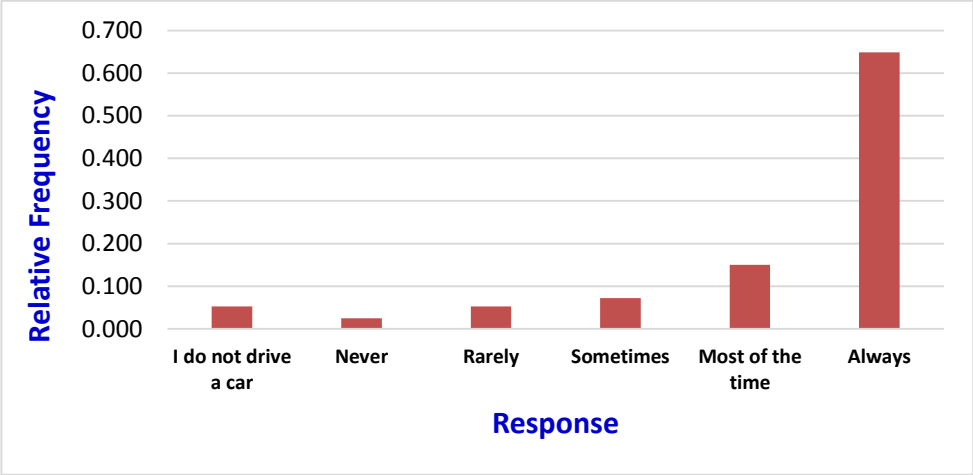
b) $0.6484 = 64.84\%$

c) $0.0247 + 0.0522 = 0.0769$ 7.69%

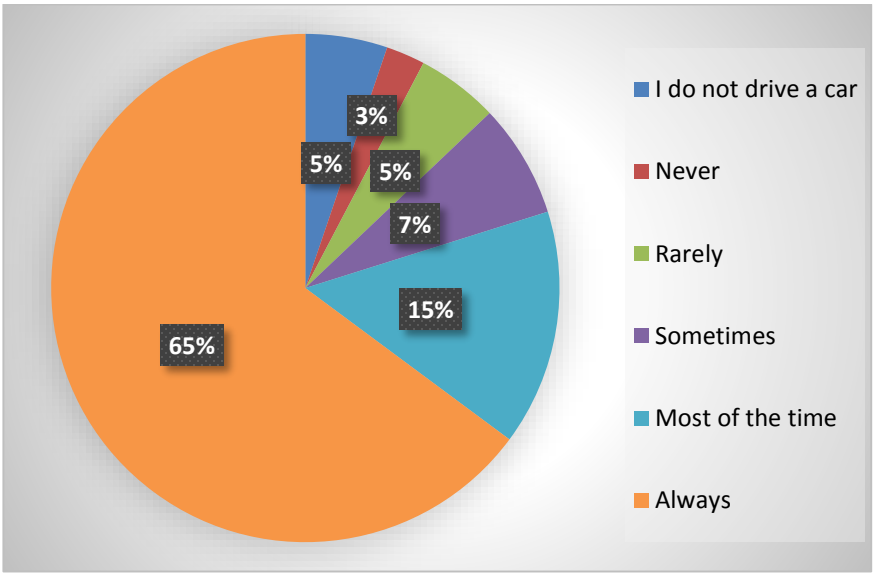
d) Frequency bar graph



e) Relative frequency bar graph



f) How often do you wear a seat belt when driving a car?



g) The statement is descriptive because it is describing the particular sample.

Exercise

A phlebotomist draws the blood of a random sample of 50 patients and determines their blood types as shown.

O	*B*	*AB*	*O*	*AB*	*O*	*O*	*O*	*O*	*O*
O	*O*	*B*	*O*	*O*	*A*	*A*	*B*	*O*	*A*
A	*B*	*A*	*A*	*A*	*A*	*O*	*A*	*O*	*O*
A	*A*	*B*	*A*	*B*	*O*	*AB*	*A*	*A*	*A*
O	*O*	*AB*	*O*	*A*	*A*	*A*	*O*	*O*	*O*

a) Construct a frequency distribution
b) Construct a relative frequency distribution
c) According to the data, which blood type is most common?
d) According to the data, which blood type is least common?
e) Use the results of the sample to conjecture the percentage of the population that has type *O* blood. Is this an example of descriptive or inferential statistics?
f) Contact a local hospital and ask them the percentage of the population that us blood type *O*. Why might the results differ?
g) Draw a frequency bar graph
h) Draw a relative frequency bar graph
i) Draw a pie chart

Solution

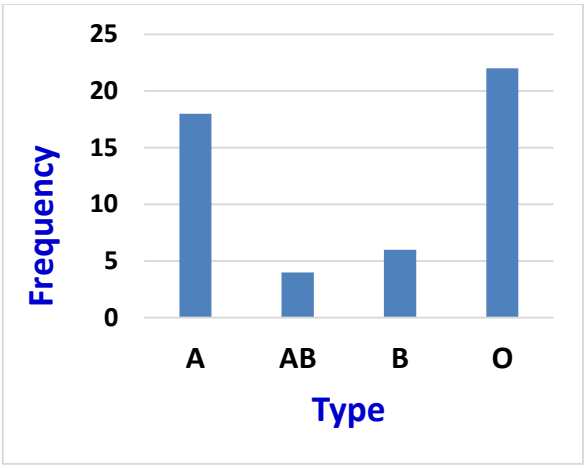
a)

b)

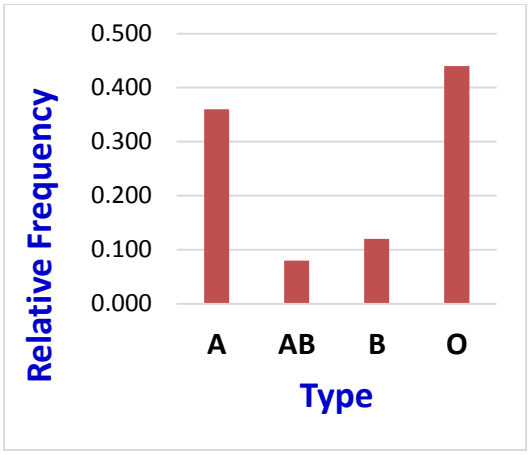
Blood Type	***Frequency***	***Relative Frequency***	***Percentage***
A	18	0.360	36%
AB	4	0.080	8%
B	6	0.120	12%
O	22	0.440	44%

c) Type *O* is the most common
d) Type *AB* is the least common
e) We estimate that 44% of the population has type *O* blood. This is considered inferential statistics because a conclusion about the population is being drawn based on sample data.
f) Answers will vary; in 2008 the Red Cross reported that 45% of the population had type *O* blood (either + or –). Results will differ because of sampling variability.

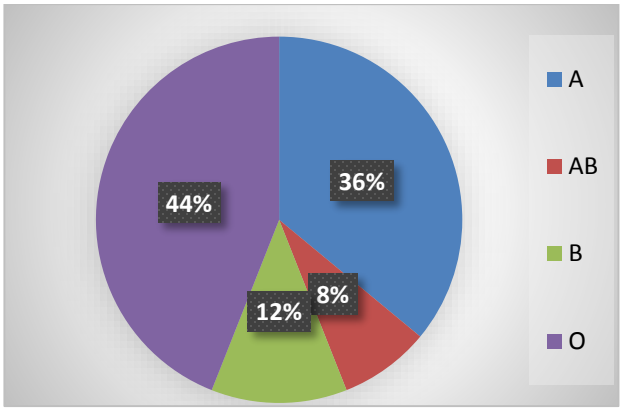
g)



h)



i) Blood Types



Solution Section 1.6 – Additional Displays

Exercise

Identify the class width, class midpoints, and class boundaries for the given frequency distribution. Then construct the cumulative frequency distribution that corresponds to the frequency distribution.

a)

Tar (mg) in Nonfiltered Cigarettes	Frequency
10 – 13	1
14 – 17	0
18 – 21	15
22 – 25	7
26 – 29	2

b)

Tar (mg) in Filtered Cigarettes	Frequency
2 – 5	2
6 – 9	2
10 – 13	6
14 – 17	15

c)

Weights (lb) of Discarded Metal	Frequency
0.00 – 0.99	5
1.00 – 1.99	26
2.00 – 2.99	15
3.00 – 3.99	12
4.00 – 4.99	4

d)

Weights (lb) of Discarded Plastic	Frequency
0.00 – 0.99	14
1.00 – 1.99	20
2.00 – 2.99	1
3.00 – 3.99	4
4.00 – 4.99	2
5.00 – 5.99	1

Solution

a) *Class width*: $14-10 = 4$ (***Subtracting the first two lower class limits***)

Class midpoints: $\frac{10+13}{2} = 11.5$ adding **4** to obtain the rest: 11.5, 15.5, 19.5, 23.5, 27.5.

Class boundaries: $\frac{13+14}{2} = 13.5$ (boundary between the first and second class). The other can be obtained by adding 4: 9.5, 13.5, 17.5, 21.5, 25.5, 29.5.

b) *Class width*: $6-2 = 4$ (***Subtracting the first two lower class limits***)

Class midpoints: $\frac{2+5}{2} = 3.5$ adding **4** to obtain the rest: 3.5, 7.5, 11.5, 15.5.

Class boundaries: $\frac{5+6}{2} = 5.5$ (boundary between the first and second class). The other can be obtained by adding 4: 1.5, 5.5, 9.5, 13.5, 17.5.

c) *Class width*: $1.00-0.00 = 1.00$ (***Subtracting the first two lower class limits***)

Class midpoints: $\frac{0.00+0.99}{2} = 0.495$ adding **1.00** to obtain the rest: 0.495, 1.495, 2.495, 3.495, 4.495.

Class boundaries: $\frac{0.99+1.00}{2} = 0.995$ (boundary between the first and second class). The other can be obtained by adding or subtracting **1.00**: -0.005, 0.995, 1.995, 2.995, 3.995, 4.995.

d) *Class width*: $1.00 - 0.00 = \underline{1.00}$ (***Subtracting the first two lower class limits***)

Class midpoints: $\frac{0.00 + 0.99}{2} = \underline{0.495}$ adding **1.00** to obtain the rest: 0.495, 1.495, 2.495, 3.495, 4.495, 5.495.

Class boundaries: $\frac{0.99 + 1.00}{2} = \underline{0.995}$ (boundary between the first and second class). The other can be obtained by adding or subtracting **1.00**: -0.005, 0.995, 1.995, 2.995, 3.995, 4.995, 5.995.

Exercise

Given listed amounts of Strontium-90 (in millibecquerels) in a simple random sample of baby teeth.

155	142	149	130	151	163	151	142	156	133	138	161	128	144	172
137	151	166	147	163	145	116	136	158	114	165	169	145	150	150
150	158	151	145	152	140	170	129	188	156					

a) Construct a dot plot of the amounts of Strontium-90. What does the dot plot suggest about the distribution of those amounts?

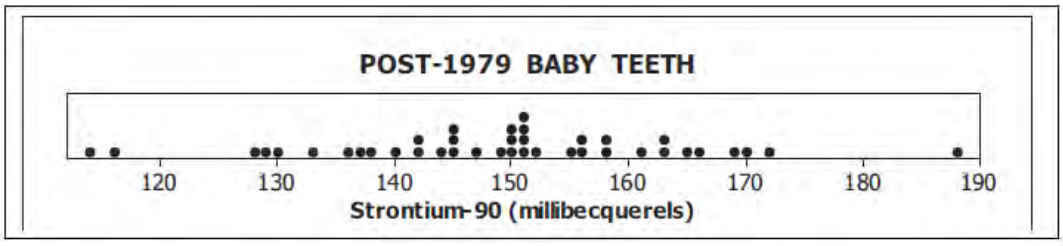
b) Construct a stemplot of the amounts of Strontium-90. What does the stemplot suggest about the distribution of those amounts?

c) Construct a frequency polygon of the amounts of Strontium-90. For the horizontal axis, use the midpoints of the class intervals in the frequency distribution: 110-119, 120-129, 130-139, …, 180-189.

d) Construct an ogive of the amounts of Strontium-90. For the horizontal axis, use the class boundaries corresponding to the class limits. How many of the amounts are below 150 millibecquerels?

Solution

a) The dotplot is given below.



The strontium-90 levels appear to have a normal distribution clusters around 150.

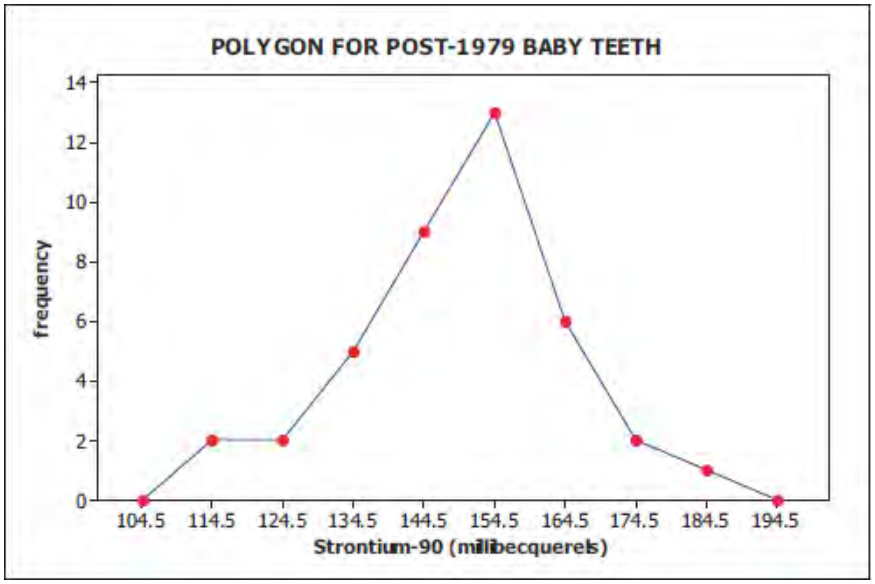
b) The stemplot is given:

Strontium – 90	
11	46
12	89
13	03678
14	022455579
15	0001111256688
16	133569
17	02

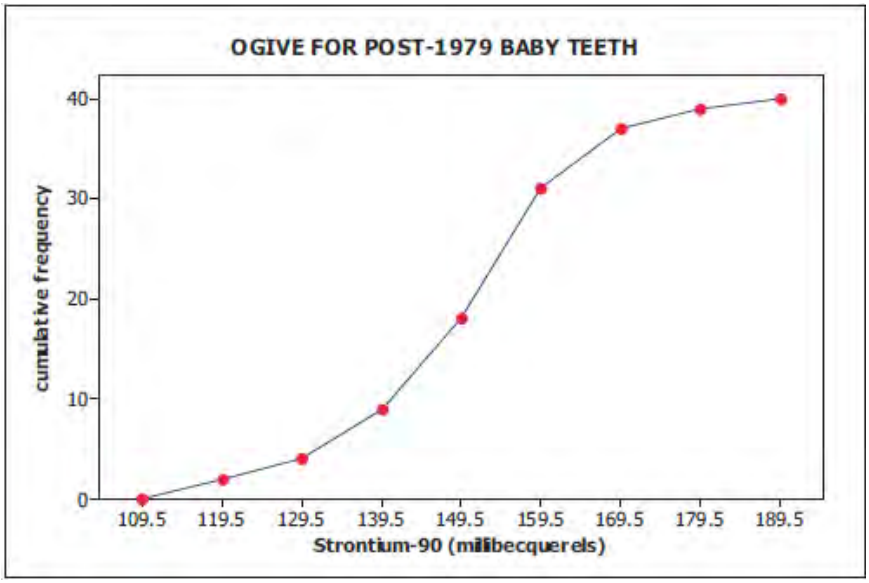
18	8

The strontium-90 levels appear to have a normal distribution clusters around 150.

c) The frequency polygon is given:



d) The ogive is given:



Using the figure: move up from 150 on the horizontal scale to intersect the graph, then move left to intersect the vertical scale at 18. This indicates there were approximately 18 data values which would have been recorded as being below 150, which agrees with the actual data values.

Exercise

Use the 62 weights if discarded plastic listed in Data set below

0.27 1.41 2.19 2.83 2.19 1.81 0.85 3.05 3.42 2.10 2.93 2.44 2.17 1.41 2.00

0.93	2.97	2.04	0.65	2.13	0.63	1.53	4.69	0.15	1.45	2.68	3.53	1.49	2.31	0.92
0.89	0.80	0.72	2.66	4.37	0.92	1.40	1.45	1.68	1.53	1.44	1.44	1.36	0.38	1.74
2.35	2.30	1.14	2.88	2.13	5.28	1.48	3.36	2.83	2.87	2.96	1.61	1.58	1.15	1.28
0.58	0.74													

a) Construct a dot plot of the weights of discarded plastic. What does the dot plot suggest about the distribution of the weights?

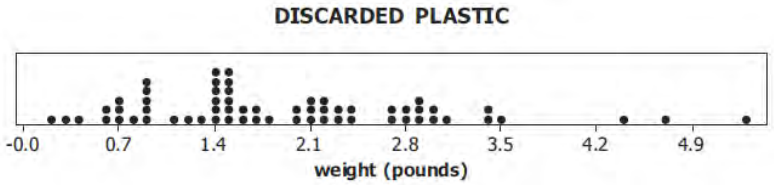
b) Construct a stemplot of the weights of discarded plastic. What does the stemplot suggest about the distribution of the weights?

c) Construct a frequency polygon of the weights of discarded plastic. For the horizontal axis, use the midpoints of the class intervals: 0.00-0.99, 1.00-1.99, 2.00-2.99, 3.00-3.99, 4.00-4.99, 5.00-5.99.

d) Construct an ogive of the weights of discarded plastic. For the horizontal axis, use these classes boundaries: –0.005, 0.995, 1.995, 2.995, 3.995, 4.995, 5.995. How many of the weights are below 4 lb.?

Solution

a) The dotplot is given below.



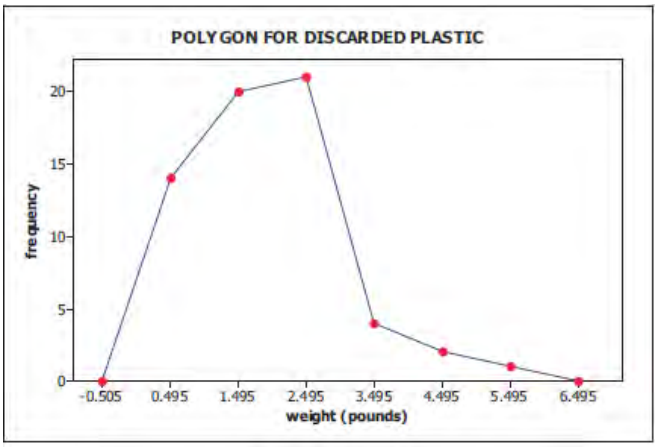
The weights appear to be approximately normally distributed, except for the presence of a few high values.

b) The stemplot is given:

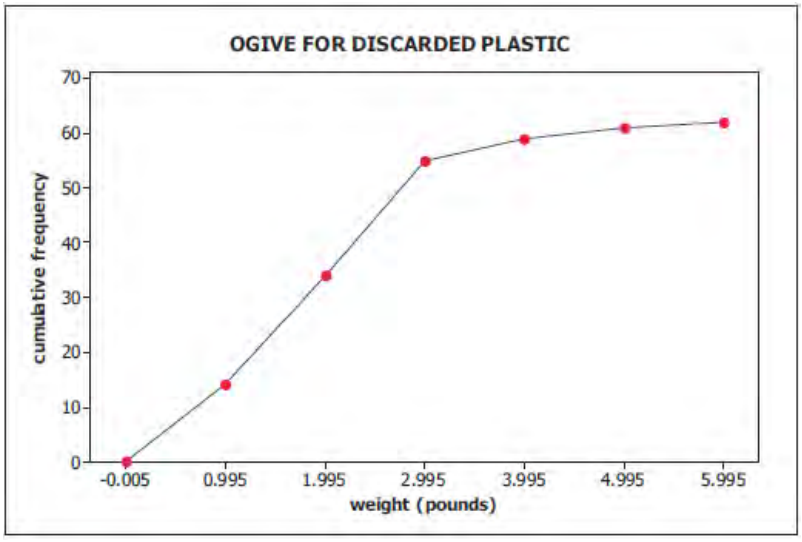
	***Weight* (*lb.*)**
0.	1256677888999
1.	11234444444445556678
2.	001111113334668888999
3.	9345
4.	36
5.	2

The weights appear to be approximately normally distributed, except for the necessary lower truncation at zero.

c) The frequency polygon is given:



d) The ogive is given



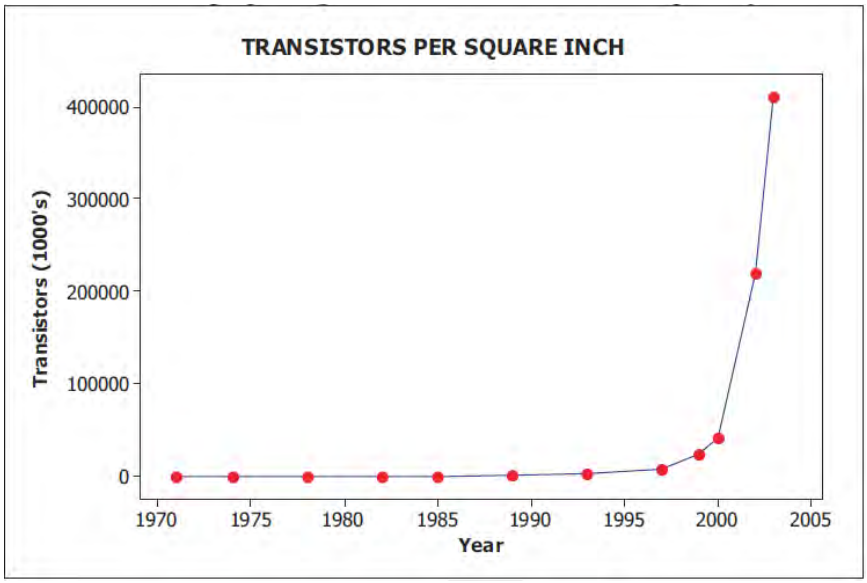
Using the figure: move up 3 on the horizontal scale to intersect the graph, then move left to intersect the vertical scale at 59. This indicates there were approximately 59 data values which would have been recorded as being below 4, which agrees with the actual data.

Exercise

In 1965, Intel cofounder Gordon Moore proposed what has since become known as Moore's law: the number of transistors per square inch on integrated circuits with double approximately every 18 months. The table below lists the number of transistors per square inch (in thousands) for several different years. Construct a time-series graph of the data.

Year	1971	1974	1978	1982	1985	1989	1993	1997	1999	2000	2002	2003
Transistors	2.3	5	29	120	275	1180	3100	7500	24,000	42,000	220,000	410,000

Solution



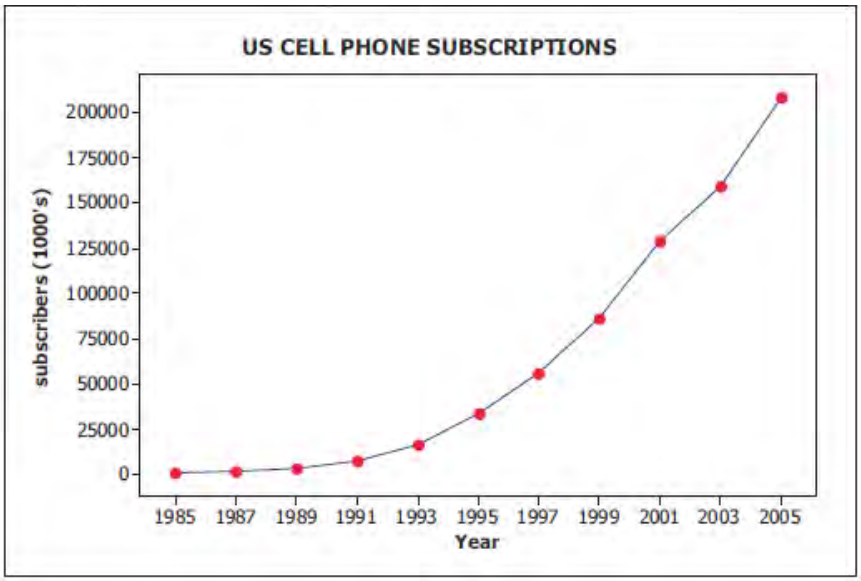
Exercise

The following table shows the numbers of cell phone subscriptions (in thousands) in the U.S. for various years. Construct a time-series graph of the data. "Linear" growth would result in a graph that is approximately a straight line. Does the time-series graph appear to show linear growth?

Year	1985	1987	1989	1991	1993	1995	1997	1999	2001	2003	2005
Number	340	1231	3509	7557	16,009	33,786	55,312	86,047	128,375	158,722	207,900

Solution

The time series graph is given:



The graph does not appear to show linear growth (constant slope) over the entire time period, but does appear that there was linear growth during certain periods (since 1999)

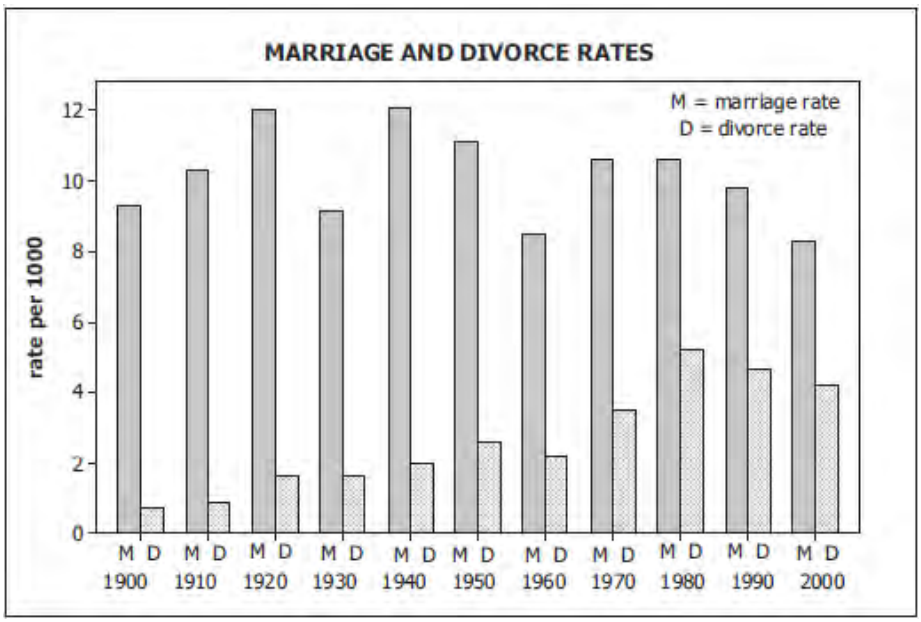
Exercise

The following table lists the marriage and divorce rates per 1000 people in the U.S. for selected years since 1900 (based on data from the Department of Health and Human Services). Construct a multiple bar graph of the data. Why do these data consist of marriage and divorce rates rather than total numbers of marriages and divorces? Comment on any trends that you observe in these rates, and give explanations for these trends.

Year	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
Marriage	9.3	10.3	12.0	9.2	12.1	11.1	8.5	10.6	10.6	9.8	8.3
Divorce	0.7	0.9	1.6	1.6	2.0	2.6	2.2	3.5	5.2	4.7	4.2

Solution

As the population increases, the *numbers* of marriages and divorces will automatically increase. To identify any change in marriage and divorce patterns, one needs to examine the *rates*. This is analogous to using percents (or relative frequencies) instead of frequencies to compare categories for two samples of different sizes. The marriage rate appears to have remained fairly constant, with possible slight decrease in recent years. The divorce rate appears to have steadily grown, with a possible slight decrease in recent years.



Exercise

A car salesman records the number of cars he sold each week for the past year. The following frequency histogram shows the results

a) What are the most frequent number of cars sold in a week?
b) For how many weeks two cars sold?
c) Determine the percentage of time two cars were sold.
d) Describe the shape of the distribution

Solution

a) 4 cars
b) There were 9 weeks in which 2 cars sold
c) Total frequency = 4+2+9+8+12+8+5+2+1+1 = 52

Percentage of time two cars were sold $= \frac{9}{52} \times 100 = 17.3\%$

d) Slightly skewed to the right

Exercise

Use the data to create a stemplot

The midterm test scores for the seventh-period typing class are listed below

85 77 93 91 74 65 68 97 88 59 74 83 85 72 63 79

Solution

5	9
6	3 5 8
7	2 4 4 7 9
8	3 5 5 8
9	1 3 7

Exercise

Use the data to create a stemplot. Twenty-four workers were surveyed about how long it takes them to travel to work each day. The data below are given in minutes

20 35 42 52 65 20 60 49 24 37 23 24
22 20 41 25 28 27 50 47 58 30 32 48

Solution

2	0 0 0 2 3 4 4 5 7 8
3	0 2 5 7
4	1 2 7 8 9
5	0 2 8
6	0 5

Exercise

Find the original data from the stemplot

a)

Stem	*Leaves*
76	2 6 7
77	2 4 9
78	1 7

b)

1	0 1 4
2	1 4 4 7 9
3	3 5 5 5 7 7 8
4	0 0 1 2 6 6 8 9 9
5	3 3 5 8
6	2

c)

24	0 4 7
25	0 2 3 9 9
26	3 4 5 8 8 9
27	0 1 1 3 6 6
28	2 3 8

Solution

a) 762, 766, 767, 772, 774, 779, 781, 787

b) 10, 11, 14, 21, 24, 24, 27, 29, 33, 35, 35, 35, 37, 37, 38, 40, 40, 41, 42, 46, 46, 48, 49, 49,53, 53, 55, 58, 62

c) 240, 244, 247, 250, 252, 253, 259, 259, 263, 264, 265, 268, 268, 269, 270, 271, 271, 273, 276, 276, 270, 271, 271, 273, 276, 276, 282, 283, 288

Solution Section 1.7 – Misrepresentations of Data

Exercise

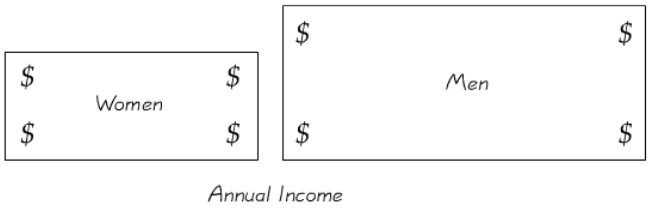
Assume that, as a newspaper reporter, you must graph data showing that increased smoking causes an increased risk of lung cancer. Given that people might be helped and lives might be saved by creating a graph that exaggerates the risk of lung cancer, is it ethical to construct such a graph?

Solution

No. Results should be presented in a way that is fair and objective so that the reader has the reliable information necessary to reach his own conclusion.

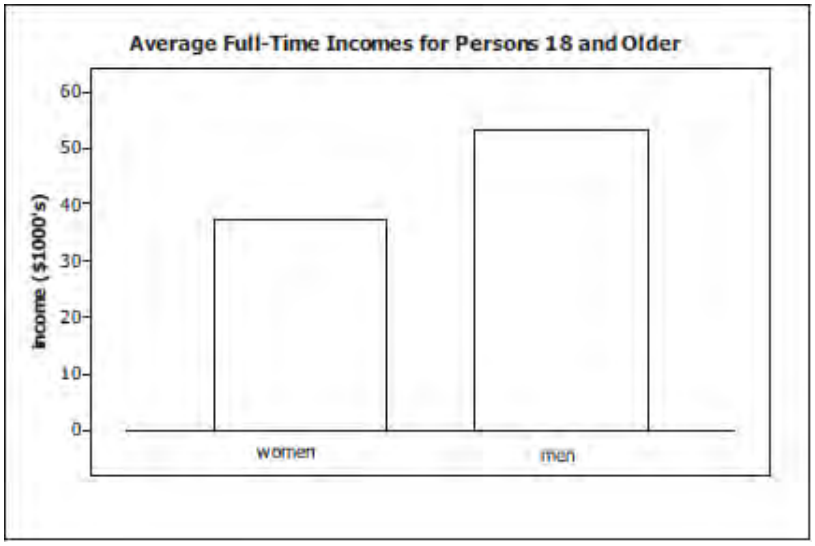
Exercise

The accompanying graph depicts average full-time incomes of women and men aged 18 and over. For a recent year, those incomes were \$37,197 for women and \$53,059 for men (based on data from the U.S. Census Bureau). Does the graph make a fair comparison of the data? Why or why not? If the graph distorts the data, construct a fair graph.



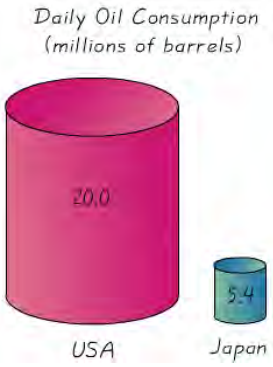
Solution

The average income for men is about 1.4 times the average income for women. Making the men's pictograph 1.4 times as wide and 1.4 times as high as the women's produces a men's image with $1.4^2 = 1.96$ *times* the areas of the women's image. Since it is the area that gives the visual impression in a two-dimensional figure, the men's average income appears to be almost twice that of the women's average income.



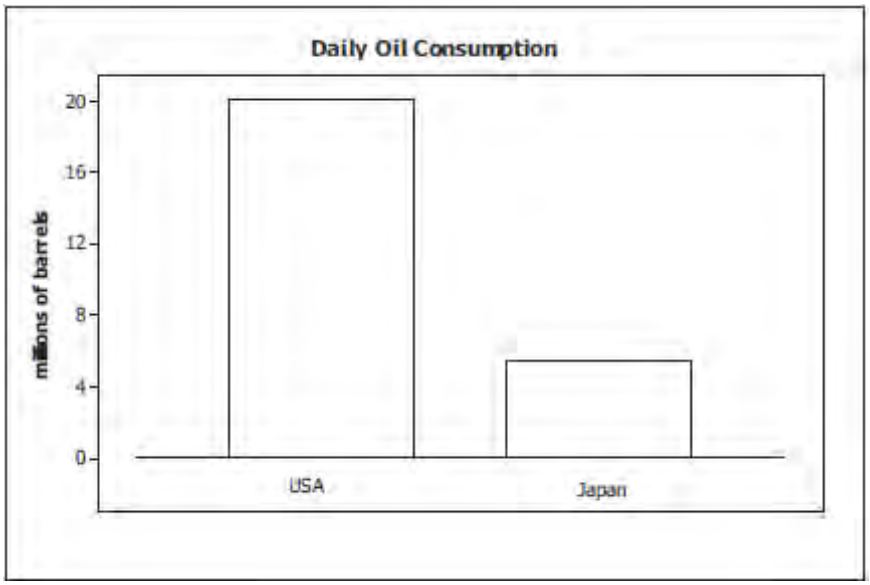
Exercise

The accompanying graph uses cylinders to represent barrels of oil consumed by the U.S. and Japan. Does the graph distort the data or doe it depict the data fairly? Why or why not? If the graph distorts the data, construct a graph that depicts the data fairly.



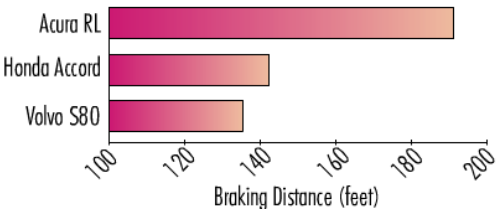
Solution

The oil consumption for the USA is about 3.7 times the oil consumption for Japan. Making the USA's pictograph 3.7 times larger than Japan's in three dimensional produces an image for the US with $3.7^3 = 50$ times the volume of the image for Japan. Since it is perceived volume that gives the visual impression in the figure, the consumption for US appears 50 times that for Japan.



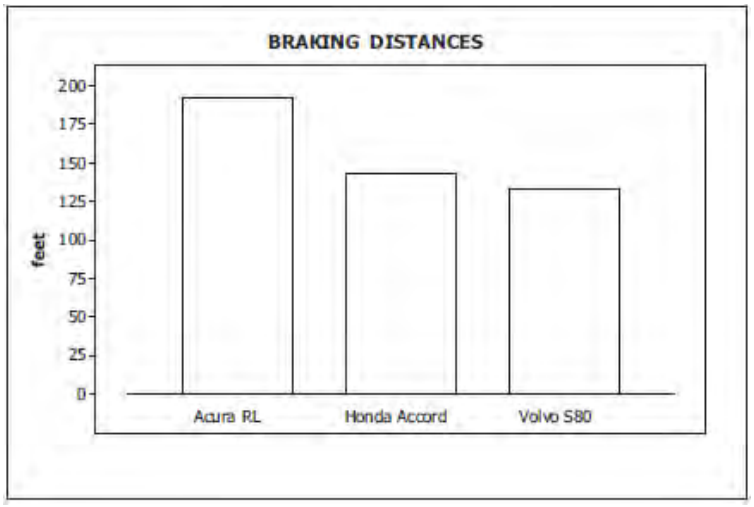
Exercise

The accompanying graph shows the braking distances for different cars measured under the same conditions. Describe the ways in which this graph might be deceptive. How much greater is the braking distance of the Acura RL than the braking distance of the Volvo S80? Draw the graph in a way that depicts the data more fairly.



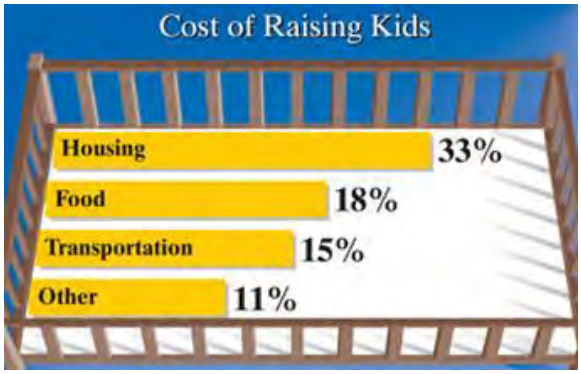
Solution

It appears that the braking distance for the Acura RL is more than twice that of the Volvo S80. The actual difference is about 60 ft., and the Acura RL distance is about $\frac{192}{133} = 1.44$ times that of the Volvo. The exaggeration of differences is caused by the fact that the distance scale dies not start at zero.



Exercise

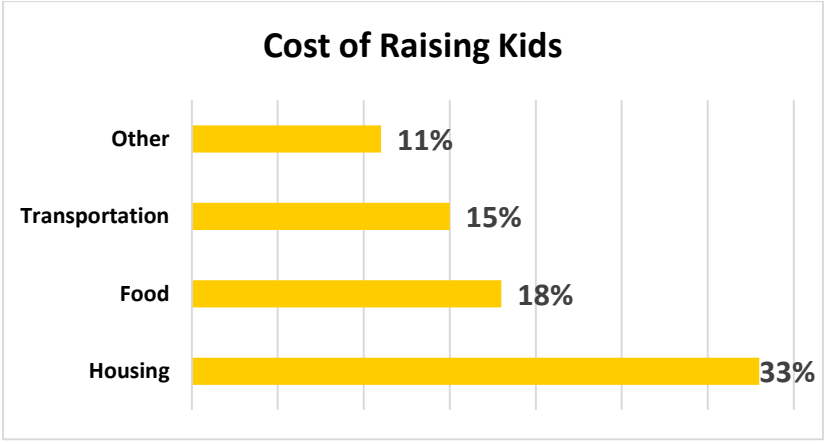
The graph represents the percentage of income a middle-income family will spend on their children



a) How is the graphic misleading?
b) What could be done to improve the graphics?

Solution

a) The graphic is misleading because the bars are not proportional. The bar for housing should be a little more than twice the length of the bar for transportation, but it is not.
b) The graphic could be improved by adjusting the bars so that their lengths are proportional



Exercise

In what sense are the mean, median, mode and midrange measures of "center"?

Solution

The mean, median, mode, and midrange are measures of "center" in the sense that they each attempt to determine (by various criteria – i.e., by using different approaches) what might be designated as a typical or representative value.

Exercise

A headline in USA Today stated that "Average family income drops 2.3%." What is the role of the term average in statistics? Should another term be used in place of average?

Solution

The term *average* is not used in statistics because it is imprecisely used by the general public as a synonym for *typical* – as in, the average American had blue eyes. When referring to the result obtained by dividing a sum by the number of values contributing to that sum, the term ***mean*** should be used.

Exercise

In an editorial, the Poughkeepsie Journal printed this statement: "The median price – the price exactly in between the highest and lowest -- …" Does that statement correctly describes the median? Why or why not?

Solution

No. The price exactly in between the highest and the lowest would be the mean of the highest and lowest values – which is the midrange, and not the median.

Exercise

A simple random sample of pages from Merriam-Webster's Collegiate Dictionary, 11th edition, was obtained. Listed below are the numbers of words defined on those pages. Given that this dictionary has 1459 pages with defined words, estimate the total number of defined words in the dictionary.

51 63 36 43 34 62 73 39 53 79

Find the

a) mean *b*) median *c*) mode *d*) midrange

e) Is that estimate likely to be an accurate estimate of the number of words in the English language?

Solution

34 36 39 43 51 53 62 63 73 79

a) *Mean*: $\bar{x} = \dfrac{\sum x}{n} = \dfrac{51+63+36+43+34+62+73+39+53+79}{10} = \dfrac{533}{10}$ $= 53.3 \text{ words}$

b) *Median*: $\tilde{x} = \dfrac{51+53}{2}$ $= 52.0 \text{ words}$

c) *Mode*: None

d) *Midrange*: $\dfrac{34+79}{2}$ $= 56.5 \text{ words}$

e) Using the mean of 53.3 words per page, a reasonable estimate for the total number of words in the dictionary is $(53.3)(1459) = 77{,}765$. Since the sample is a simple random sample, it should be representative of the population and the estimate of 77,675 is a valid estimate for the number of words in the dictionary – but not for the total number of words in the English language, since the dictionary does not claim to contain every word.

Exercise

The National Highway Traffic Administration conducted crash tests of child booster seats for cars. Listed below are results from those teats, with the measurements given in *hic* (standard head injury condition units.

774 249 1210 546 431 612

Find the

a) mean *b*) median *c*) mode *d*) midrange

e) According to the safety requirement, the *hic* measurement should be less than 1000 *hic*. Do the results suggest that all of the child booster seats meet the specified requirement?

Solution

431 546 612 649 774 1210

a) *Mean*: $\bar{x} = \dfrac{\sum x}{n} = \dfrac{431+546+612+649+774+1210}{6} = \dfrac{4222}{6}$

$= 703.7 \text{ hic}$

b) *Median*: $\tilde{x} = \dfrac{612+649}{2}$ $= 630.5 \text{ hic}$

c) *Mode*: None

d) *Midrange*: $\dfrac{431+1210}{2}$ $= 820.5 \text{ hic}$

e) No. Even though all four measures of center fall within with accepted guidelines, all the individual values do not. Since one result exceeds the guidelines, it is clear that all child booster seats do not meet the requirement.

Exercise

The insurance Institution for Highway Safety conducted tests with crashes of new cars traveling at 6 mi/h. The total cost of the damages was found for a simple random sample of the tested cars and listed below

$7448 $4911 $9051 $6374 $4277

Find the

a) mean *b*) median *c*) mode *d*) midrange

e) Do the different measures of center differ very much?

Solution

4277 4911 6374 7448 9051

a) *Mean*: $\bar{x} = \frac{\sum x}{n} = \frac{4277 + 4911 + 6374 + 7448 + 9051}{5} = \$6412.2$

b) *Median*: $= \$6,374$

c) *Mode*: None

d) *Midrange*: $\frac{4277 + 9051}{2} = \$6,664.00$

e) No. Even though the sample values cover a fairly wide range, the measures of center do not differ very much.

Exercise

Listed below are the durations (in hours) of a simple random sample of all flights (as of this writing) of NASA's Space Transport System (space shuttle).

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

Find the

a) mean *b*) median *c*) mode *d*) midrange

e) How might that duration time be explained?

Solution

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

a) *Mean*: $\bar{x} = \frac{\sum x}{n}$

$= \frac{0 + 73 + 95 + 165 + 191 + 192 + 221 + 235 + 235 + 244 + 259 + 262 + 331 + 376 + 381}{15}$

$= 217.3\ hrs$

b) *Median*: $= 235\ hrs$

c) *Mode*: 235 ***hrs***.

d) *Midrange*: $\frac{0 + 381}{2} = 190.5\ \ hrs$

e) The duration of time of 0 appears to be very unusual. It likely represents Challenger disaster of Jan 1986, when the mission ended in an explosion shortly after takeoff.

Exercise

Listed below are the playing times (in seconds) of songs that were popular at the time of this writing.

448 242 231 246 246 293 280 227 213 262 239 213 258 255 257 244

Find the

a) mean *b*) median *c*) mode *d*) midrange

e) Is there on time that is very different from the others?

Solution

213 213 227 231 239 242 244 246 246 255 257 258 262 280 293 448

a) *Mean*: $\bar{x} = \dfrac{\sum x}{n}$

$= \dfrac{213 + 213 + 227 + 231 + 239 + 242 + 244 + 246 + 246 + 255 + 257 + 258 + 262 + 280 + 293 + 448}{16}$

$= 259.6 \text{ sec}$

b) *Median*: $= \dfrac{246 + 246}{2} = 246 \text{ sec}$

c) *Mode*: 213, 246 ***sec***.

d) *Midrange*: $\dfrac{213 + 448}{2} = 330.5 \text{ sec}$

e) Yes the time of 448 seconds appears to be very different from the others.

Exercise

Listed below are numbers of satellites in orbit from different countries.

158 17 15 17 7 3 5 1 8 3 4 2 4 1 2 3 1 1 1 1 1 1 1 1

Find the

a) mean *b*) median *c*) mode *d*) midrange

e) Does on country have an exceptional number of satellites?

f) Can you guess which country has the most satellites?

Solution

1 1 1 1 1 1 1 1 1 1 2 2 3 3 3 4 4 5 7 8 15 17 18 158

a) *Mean*: $\bar{x} = \dfrac{\sum x}{n}$

$= \dfrac{1+1+1+1+1+1+1+1+1+1+2+2+3+3+3+4+4+5+7+8+15+17+18+158}{24}$

$= 10.8 \text{ satellites}$

b) *Median*: $= \dfrac{2+3}{2} = 2.5 \text{ satellites}$

c) *Mode*: 1 *satellite*.

d) *Midrange*: $\dfrac{1+158}{2} = 79.5 \text{ satellites}$

e) Yes the country with 158 satellites has an exceptional number of them.

f) That country is most likely the United States.

Exercise

Listed below are costs (in dollars) of roundtrip flights from JFK airport in NY City to San Francisco. (All flights involve one stop and a two-week stay.) The airlines are US Air, Continental, Delta, United, American, Alaska, and Northwest.

30 Days in Advance	244	260	264	264	278	318	280
1 Day in Advance	456	614	567	943	628	1088	536

a) Find the mean and median for each then compare the two sets of results.

b) Does it make much difference if the tickets are purchased 30 days in advance or 1 day in advance?

Solution

30 Days in Advance	244	260	264	264	278	280	318
1 Day in Advance	456	536	567	614	628	943	1088

a) *Mean*: $\left.\bar{x}\right|_{30\ days} = \dfrac{\sum x}{n} = \dfrac{244+260+264+278+280+318}{7} = \$272$

Median: $= \$264$

Mean: $\left.\bar{x}\right|_{1\ day} = \dfrac{456+536+567+614+628+943+1088}{7} = \$690.3$

Median: $= \$614$

b) The tickets purchased 30 days in advance are considerably less expensive – they appear to cost less than half of the tickets purchased one day in advance.

Exercise

The trend of thinner Miss America winners has generated charges that the contest encourages unhealthy diet habits among young women. Listed below are body mass indexes (BMI) for Miss America winners from two different periods.

BMI (1920 – 1930)	20.4	21.9	22.1	22.3	20.3	18.8	18.9	19.4	18.4	19.1
BMI – (from recent winners)	19.5	20.3	19.6	20.2	17.8	17.9	19.1	18.8	17.6	16.8

Find the mean and median for each then compare the two sets of results.

Solution

BMI (1920 – 1930)	18.4	18.8	18.9	19.1	19.4	20.3	20.4	21.9	22.1	22.3
BMI – (*from recent winners*)	16.8	17.6	17.8	17.9	18.8	19.1	19.5	19.6	20.2	20.3

a) *Mean*: $\bar{x}\Big|_{past} = \frac{18.4+18.8+18.9+19.1+19.4+20.3+20.4+21.9+22.1+22.3}{10} = 20.16$

Median: $= \frac{19.4+20.3}{2} = 19.85$

Mean: $\bar{x}\Big|_{present} = \frac{16.8+17.6+17.8+17.9+18.8+19.1+19.5+19.6+20.2+20.3}{10} = 18.76$

Median: $= \frac{18.8+19.1}{2} = 18.95$

b) The recent winners appear to have lower measures of BMI than do the former winners.

Exercise

Find the mean of the data summarized in the given frequency distribution.

a)

Tar (mg) in Nonfiltered Cigarettes	***Frequency***
10 – 13	1
14 – 17	0
18 – 21	15
22 – 25	7
26 – 29	2

b)

Pulse Rates of Females	***Frequency***
60 – 69	12
70 – 79	14
80 – 89	11
90 – 99	1
100 – 109	1
110 – 119	0
120 – 129	1

Solution

a) $\bar{x} = \frac{\sum(f \cdot x)}{\sum f} = \frac{523.5}{25} = 20.9\ mg$

x	***f***	***f · x***
11.5	1	11.5
15.5	0	0
19.5	15	292.5
23.5	7	164.5
27.5	2	55.0
	25	523.5

b)

x	***f***	***f · x***
64.5	12	774.0
74.5	14	1043.0
84.5	11	929.5
94.5	1	94.5
104.5	1	104.5
114.5	0	0.0
124.5	1	124.5
	40	3070.0

$\bar{x} = \frac{\sum(f \cdot x)}{\sum f} = \frac{3070.0}{40} = 76.8\ \ beats\ per\ minute$

Exercise

A student of the author earned grades of B, C, B, A, and D. Those courses has these corresponding numbers credit hours: 3, 3, 4, 4, and 1. The grading system assigns quality points to letter grades as follows: A = 4; B = 3; C = 2; D = 1; F =0. Compute the grade point average (GPA) and round the result with two decimal places. If the Dean's list requires a GPA 3.00 or greater, did this student make the Dean's list?

Solution

The x values are the numerical values of the letter grades, and the corresponding weights are the numbers of credit hours.

x	w	$w \cdot x$
3	3	9
2	3	6
3	4	12
4	4	16
1	1	1
	15	44

$$\bar{x} = \frac{\sum(f \cdot x)}{\sum f} = \frac{44}{15} = 2.93$$

No; since 2.93 is below 3.00, the student did not make the Dean's list.

Exercise

A student of the author earned grades of 92, 83, 77, 84, and 82 on her five regular tests. She earned grades of 88 on the final exam and 95 on her class projects. Her combined homework grade was 77. The five regular tests count for 60% of the final grade, the final exam counts for 10%, the project counts for 15%, and homework counts for 15%. What is her weighted mean grade? What letter grade did she earn? (A, B, C, D, or F)

Solution

The x values are the individual grades, and the corresponding weights are the percent of the final grade.
Since the 5 regular tests count for 60% of the final grade, each regular test counts for:

$$\frac{60\%}{5} = 12\% = 0.12 \text{ of the final grade.}$$

x	w	$w \cdot x$
92	0.12	11.04
83	0.12	9.96
77	0.12	9.24
84	0.12	10.08
82	0.12	9.84
88	0.10	8.80
95	0.15	14.25
77	0.15	11.55
	1.0	84.76

$$\bar{x} = \frac{\sum(w \cdot x)}{\sum w} = \frac{84.76}{1.00} = 84.76$$

Her weighted mean grade is 84.8. Within 90-80-70-60 grading system for *A-B-C-D*, that corresponds to *B*.

Exercise

You are taking a class in which your grade is determined from five sources: 50% from your test mean, 15% from your midterm, 20% from your final exam, 10% from your computer lab work, and 5% from your homework. Your scores are 86 (test mean), 96 (midterm), 82 (final exam), 98 (computer lab), and 100 (homework). What is the weighted mean of your scores? If the minimum average for an A is 90, did you get an A?

Solution

Source	x	w	$w \cdot x$
Test mean	86	0.50	73.0
Midterm	96	0.15	14.4
Final Exam	82	0.20	16.4
Computer lab	98	0.10	9.8
Homework	100	0.05	5.0
		1	88.6

$$\bar{x} = \frac{\sum(w \cdot x)}{\sum w} = \frac{88.6}{1.00} = 88.6$$

Your weighted mean for the course is 88.6. So you did not get an A.

Exercise

During a quality assurance check, the actual coffee contents (in ounces) of six jars of instant coffee were recorded as 6.03, 5.59, 6.40, 6.00, 5.99, and 6.02.

a) Find the mean and the median of the coffee content.

b) The third value was incorrectly measured and is actually 6.04. Find the mean and median of the coffee content again.

c) Which measure of central tendency, the mean or the median, was affected more by the data entry error?

Solution

a) $$\bar{x} = \frac{\sum x}{n} = \frac{6.03 + 5.59 + 6.40 + 6.00 + 5.99 + 6.02}{6} = \frac{36.03}{6} = 6.005$$

5.59 5.99 $\underbrace{6.0 \quad 6.02}_{median}$ 6.03 6.4

$$median = \frac{6 + 6.02}{2} = 6.01$$

b) $$\bar{x} = \frac{\sum x}{n} = \frac{6.03 + 5.59 + 6.04 + 6.00 + 5.99 + 6.02}{6} = \frac{35.67}{6} = 5.945$$

5.59 5.99 $\underbrace{6.0 \quad 6.02}_{median}$ 6.03 6.4

$$median = \frac{6 + 6.02}{2} = 6.01$$

c) The mean was affected more

Exercise

The table below shows the U.S. exports (in billions of dollars) to 19 countries for a recent year.

U.S. Exports (in billions of dollars)		
Canada: 261.1	Mexico: 151.2	Germany: 54.5
Taiwan: 24.9	Netherlands: 39.7	China: 69.7
Australia: 22.2	Malaysia: 12.9	Switzerland: 22.0
Saudi Arabia: 12.5	United Kingdom: 53.6	Japan: 65.1
South Korea: 34.7	Singapore: 27.9	France: 28.8
Brazil: 32.3	Belgium: 28.9	Italy: 15.5
Thailand: 9.1		

a) Find the mean and the median.

b) Find the mean and median without the U.S. exports to Canada. Which measure of central tendency, the mean or the median, was affected more by the elimination of the Canadian exports?

c) The U.S. Exports to India were $17.7 billion. Find the mean and median with the Indian exports added to the original data set. Which measure of central tendency was affected more by adding the Indian exports?

Solution

a) $\bar{x} = \frac{\sum x}{n} = \frac{966.6}{19} \approx 50.87$

9.1 12.5 12.9 15.5 22.0 22.2 24.9 27.9 28.8 28.9
32.3 34.7 39.7 53.6 54.5 65.1 69.7 151.2 261.1

median $= 28.9$

b) $\bar{x} = \frac{\sum x}{n} = \frac{705.5}{18} \approx 39.19$

9.1 12.5 12.9 15.5 22.0 22.2 24.9 27.9 28.8 28.9
median
32.3 34.7 39.7 53.6 54.5 65.1 69.7 151.2

$median = \frac{28.8 + 28.9}{2} = 28.85$

The mean was affected more

c) $\bar{x} = \frac{\sum x}{n} = \frac{984.3}{20} \approx 49.22$

9.1 12.5 12.9 15.5 *17.7* 22.0 22.2 24.9 27.9 28.8 28.9
median
32.3 34.7 39.7 53.6 54.5 65.1 69.7 151.2

$median = \frac{28.8 + 28.9}{2} = 28.85$

The mean was affected more

Solution Section 1.9 – Measures of Dispersion

Exercise

In statistics, how do variation and variance differ?

Solution

Variation is a general descriptive term that refers to the fact that all the items being measured are not identical. In Statistics, variance is a specific and well-defined measure of variation the has a particular mathematical formula.

Exercise

Collegiate Dictionary has 1459 pages of defined words. Listed below are the numbers of defined words per page for a simple random sample of those pages. If we use this sample as a basis for estimating the total number of defined words in the dictionary.

51 63 36 43 34 62 73 39 53 79

a) Find the range, variance, and standard deviation.

b) How does the variation of these numbers affect our confidence on the accuracy of the estimate?

Solution

a) $\bar{x} = \frac{\sum x}{n} = \frac{533}{10} = 53.3$

Range = 79 – 34 = 45 ***words***

Variance: $= \frac{\sum (x - \bar{x})^2}{n-1} = \frac{2206.10}{9}$

$= 245.12$ *words*

Standard deviation: $\sqrt{245.12} = 15.7$ *words*

x	$x - \bar{x}$	$(x - \bar{x})^2$	x^2
34	-19.3	372.49	1156
36	-17.3	299.29	1296
39	-14.3	204.49	1521
43	-10.3	106.09	1849
51	-2.3	5.29	2601
53	-0.3	0.09	2809
62	8.7	75.69	3844
63	9.7	94.09	3969
73	19.7	388.09	5329
79	25.7	660.49	6241
533	0.0	2206.10	30615

b) There seems to be considerable variation from page to page. For small samples with n = 10, there could be considerable variation in the sample mean and, therefore, considerable variation in the projected totals for the entire dictionary. It appears that there would be a question about the accuracy of an estimate based on this sample for the total number of words.

Exercise

The National Highway Traffic Administration conducted crash tests of child booster seats for cars. Listed below are results from those teats, with the measurements given in hic (standard head injury condition units.

774 649 1210 546 431 612

a) Find the range, variance, and standard deviation

b) According to the safety requirement, the hic measurement should be less than 1000 hic. Do the results suggest that all of the child booster seats meet the specified requirement?

Solution

x	x^2
431	185761
546	298116
612	374544
649	421201
774	599076
1210	1464100
4222	3342798

a) $\bar{x} = \dfrac{\sum x}{n} = \dfrac{774+649+1210+546+431+612}{6} = 703.67$

Range = 1210 – 431 = 779 ***hic***

Variance: $\dfrac{n\left(\sum x^2\right)-\left(\sum x\right)^2}{n(n-1)} = \dfrac{6(3342798)-(4222)^2}{6(5)}$

$= 74383.5 \ \ hic^2$

Standard deviation: $\sqrt{74383.5} = 272.7 \ \ hic$

b) Yes, there seems to be much variation – mainly due to the largest value, which appears to be substantially different from the others.

Exercise

The insurance Institution for Highway Safety conducted tests with crashes of new cars traveling at 6 mi/h. The total cost of the damages was found for a simple random sample of the tested cars and listed below

$7448 $4911 $9051 $6374 $4277

a) Find the range, variance, and standard deviation

b) Do the different measures of center differ very much?

Solution

x	x^2
4277	18292729
4911	24117921
6374	40627876
7448	55472704
9051	81920601
32061	220431831

a) *Range* = 9051 – 4277 = $4744

Variance: $\dfrac{n\left(\sum x^2\right)-\left(\sum x\right)^2}{n(n-1)} = \dfrac{5(220431831)-(32061)^2}{5(4)}$

$= 3712571.7 \ \ dollars^2$

Standard deviation: $\sqrt{3712571.7} = \$1926.8$

b) A value is considered unusual if it differs from the mean by more than two standard deviations. Since \$10,000 differs from the mean by $\frac{10{,}000-6{,}412.2}{1{,}926.8}=1.86$ standard deviations, in this context it would not be considered an unusual value.

Exercise

Listed below are the durations (in hours) of a simple random sample of all flights (as of this writing) of NASA's Space Transport System (space shuttle).

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

a) Find the range, variance, and standard deviation
b) How might that duration time be explained?

x	x^2
0	0
73	5329
95	9025
165	27225
191	36481
192	36864
221	48841
235	55225
235	55225
244	59536
256	65536
262	68644
331	109561
376	141376
381	145161
3260	865574

Solution

a) *Range* = 381 – 0 = 381 ***hrs***.

Variance: $\frac{n\left(\sum x^2\right)-\left(\sum x\right)^2}{n(n-1)}=\frac{15(865574)-(3260)^2}{15(14)}$

$=11219.1\ hrs^2$

Standard deviation: $\sqrt{11219.1}=105.9\ hrs$

$\bar{x}=\frac{\sum x}{n}=\frac{3260}{15}=217.3$

b) A value is considered unusual if it differs from the mean by more than two standard deviations. Since 0 hours differs from the mean by $\frac{0-217.3}{105.9}=-2.05$ standard deviations, in this context it would not be considered an unusual value.

Exercise

Listed below are the playing times (in seconds) of songs that were popular at the time of this writing.

448 242 231 246 246 293 280 227 213
262 239 213 258 255 257 244

a) Find the range, variance, and standard deviation
b) Is there on time that is very different from the others?

x	x^2
213	45369
213	45369
227	51529
231	53361
239	57121
242	58564
244	59536
246	60516
246	60516
255	65025
257	66049
258	66564
262	68644
280	78400
293	85849
448	200704
4154	1123116

Solution

a) *Range* = 448 – 213 = 235 ***sec***

$$\textit{Variance: } s^2 = \frac{n\left(\sum x^2\right) - \left(\sum x\right)^2}{n(n-1)}$$

$$= \frac{16(1123116) - (4154)^2}{16(15)} \qquad \left(16*1123116 - 4154^2\right) / (16*15)$$

$$= 2975.6 \ \text{sec}^2$$

Standard deviation: $\sqrt{2975.6} = 54.5 \ sec$

$$\bar{x} = \frac{\sum x}{n} = \frac{4154}{16} = 259.6$$

b) If the highest playing time is omitted, then

$$\bar{x} = \frac{3706}{15} = 247.1 \qquad s = \frac{n\left(\sum x^2\right) - \left(\sum x\right)^2}{n(n-1)} = \sqrt{\frac{15(922412) - (3706)^2}{15(14)}} = 22.0 \ \text{sec}.$$

The change is substantial.

Exercise

Listed below are numbers of satellites in orbit from different countries.

158 17 15 17 7 3 5 1 8 3 4 2 4 1 2 3 1 1 1 1 1 1 1 1

a) Find the range, variance, and standard deviation

b) Does on country have an exceptional number of satellites?

x	x^2
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	1
2	4
2	4
3	9
3	9
3	9
4	16
4	16
5	25
7	49
8	64
15	225
17	289
18	328
158	24964
259	26017

Solution

a) *Range* = 158 – 1 = 157 ***satellites***

$$\text{Variance: } s^2 = \frac{n\left(\sum x^2\right) - \left(\sum x\right)^2}{n(n-1)}$$

$$= \frac{24(26017) - (259)^2}{24(23)} \qquad \left(24*26017 - 259^2\right) / (24 * 23)$$

$$= 1009.7 \ \ satellites^2$$

$$\text{Standard deviation: } s = \sqrt{1009.7} = 31.8 \ \ satellites$$

$$\bar{x} = \frac{\sum x}{n} = \frac{259}{24} = 10.8$$

b) A values is considered unusual if it differs from the mean by more than two standard deviations. Since 0 satellites differs from the mean by $\frac{0 - 10.8}{31.8} \approx -0.34$ standard deviations, in this context it would not be considered an unusual value.

There are many countries with 0 satellites, and since the original data included no 0's, it appears that the data was gathered only from countries that had at least one satellite.

Exercise

Listed below are costs (in dollars) of roundtrip flights from JFK airport in NY City to San Francisco. (All flights involve one stop and a two-week stay.) The airlines are US Air, Continental, Delta, United, American, Alaska, and Northwest.

30 Days in Advance	244	260	264	264	278	318	280
1 Day in Advance	456	614	567	943	628	1088	536

Find the coefficient of variation for each of the two sets of data, then compare the variation.

Solution

x	x^2	x	x^2
244	59536	456	207936
260	67600	536	287296
264	69696	567	321489
264	69696	614	376996
278	77284	628	394384
280	78400	943	889249
318	101124	1088	1183744
1908	523336	4832	3661094

For 30 days:

$$\bar{x} = \frac{\sum x}{n} = \frac{1908}{7} = \$272.6$$

$$s^2 = \frac{n\left(\sum x^2\right) - \left(\sum x\right)^2}{n(n-1)} = \frac{7(523336) - (1908)^2}{7(6)} = 545.0 \ \ dollars^2$$

$$s = \sqrt{545} = \$23.3$$

Coefficient of variation: $CV = \frac{s}{\bar{x}} = \frac{23.3}{272.6} \approx 0.086 = 8.6\%$

For 1 day:

$$\bar{x} = \frac{\sum x}{n} = \frac{4832}{7} = \$690.3$$

$$s^2 = \frac{n\left(\sum x^2\right) - \left(\sum x\right)^2}{n(n-1)} = \frac{7(3661094) - (4832)^2}{7(6)} = 54272.2 \ \ dollars^2$$

$$s = \sqrt{54272.2} = \$233.0$$

Coefficient of variation: $CV = \frac{s}{\bar{x}} = \frac{233.0}{690.3} \approx 0.337 = 33.7\%$

There is considerably more variation among the costs for tickets purchased one day in advance.

Exercise

The trend of thinner Miss America winners has generated charges that the contest encourages unhealthy diet habits among young women. Listed below are body mass indexes (BMI) for Miss America winners from two different periods.

BMI (1920 – 1930)	20.4	21.9	22.1	22.3	20.3	18.8	18.9	19.4	18.4	19.1
BMI – (*from recent winners*)	19.5	20.3	19.6	20.2	17.8	17.9	19.1	18.8	17.6	16.8

Find the coefficient of variation for each of the two sets of data, then compare the variation.

Solution

x	x^2		x	x^2
18.4	338.56		16.8	282.24
18.8	353.44		17.6	309.76
18.9	357.21		17.8	316.84
19.1	364.81		17.9	320.41
19.4	376.36		18.8	353.44
20.3	412.09		19.1	364.81
20.4	416.16		19.5	380.25
21.9	479.61		19.6	384.16
22.1	488.41		20.2	408.04
22.3	497.29		20.3	412.09
201.6	4083.94		187.6	3532.04

For BMI (1920 – 1930):

$$\bar{x} = \frac{\sum x}{n} = \frac{201.6}{10} = 20.16$$

$$s^2 = \frac{n\left(\sum x^2\right) - \left(\sum x\right)^2}{n(n-1)} = \frac{10(4083.94) - (201.6)^2}{10(9)} = 2.19$$

$$s = \sqrt{2.19} = 1.48$$

Coefficient of variation: $CV = \frac{s}{\bar{x}} = \frac{1.48}{20.16} \approx 0.073 = 7.3\%$

For BMI – (from recent winners):

$$\bar{x} = \frac{\sum x}{n} = \frac{187.6}{10} = 18.76$$

$$s^2 = \frac{n\left(\sum x^2\right) - \left(\sum x\right)^2}{n(n-1)} = \frac{10(3532.04) - (187.6)^2}{10(9)} = 1.41$$

$$s = \sqrt{1.41} = 1.19$$

Coefficient of variation: $CV = \frac{s}{\bar{x}} = \frac{1.19}{18.76} \approx 0.063 = 6.3\%$

The BMI values for the two time period appear to exhibit about the same amount of variation.

Exercise

Find the Standard Deviation from the frequency distribution and find the standard deviation of sample summarized in a frequency distribution table by using the formula $s = \sqrt{\dfrac{n\left[\sum\left(f \cdot x^2\right)\right] - \left[\sum(f \cdot x)\right]^2}{n(n-1)}}$, where x represents the class midpoint, f represents the class frequency, and n represents the total number of sample values.

a)

Tar (mg) in Nonfiltered Cigarettes	***Frequency***
10 – 13	1
14 – 17	0
18 – 21	15
22 – 25	7
26 – 29	2

b)

Pulse Rates of Females	***Frequency***
60 – 69	12
70 – 79	14
80 – 89	11
90 – 99	1
100 – 109	1
110 – 119	0
120 – 129	1

Solution

a) The x values are the class midpoints from the given frequency table.

x	f	$f \cdot x$	$f \cdot x^2$
11.5	1	11.5	132.25
15.5	0	0	0
19.5	15	292.5	5703.75
23.5	7	164.5	3867.75
27.5	2	55.0	1512.50
	25	523.5	11214.25

$$s^2 = \frac{n\left(\sum x^2\right) - \left(\sum x\right)^2}{n(n-1)} = \frac{25(11214.25) - (523.5)^2}{25(24)} = 10.507$$

$$s = \sqrt{10.507} = 3.2 \ mg$$

This is the same as the true value of 3.2 mg.

b) The x values are the class midpoints from the given frequency table.

x	f	$f \cdot x$	$f \cdot x^2$
64.5	12	774.0	49923.00
74.5	14	1043.0	77703.5
84.5	11	929.5	78542.75
94.5	1	94.5	8930.25
104.5	1	104.5	10920.25
114.5	0	0	0
124.5	1	124.5	15500.25
	40	3070.0	241520.00

$$s^2 = \frac{n\left(\sum x^2\right) - \left(\sum x\right)^2}{n(n-1)} = \frac{40(241520) - (3070)^2}{40(39)} = \underline{151.218|}$$

$$s = \sqrt{151.218} \underline{= 12.3 \quad beats/min|}$$

This is close to the true value of 12.5 beats/min.

Exercise

Heights of women have a bell-shaped distribution with a mean of 161 cm and a standard deviation of 7 cm. Using the empirical rule, what is the approximate percentage of women between

a) 154 cm and 168 cm?
b) 147 cm and 175 cm?

Solution

a) The range from 154 to 168 is $\bar{x} \pm 1s$ $(161 - 7 = 154 \quad 161 + 7 = 168)$. The empirical rule suggests that about 68% of the data values should fall within those limits.

b) The range from 147 to 175 is $\bar{x} \pm 2s$ $(161 - 14 = 147 \quad 161 + 14 = 175)$. The empirical rule suggests that about 95% of the data values should fall within those limits.

Exercise

The author's Generac generator produces voltage amounts with a mean of 125.0 volts and standard deviation of 0.3 volts, and the voltages have a bell-shaped distribution. Using the empirical rule, what is the approximate percentage of voltage amounts between

a) 124.4 volts and 125.6 volts?
b) 124.1 volts and 125.9 volts?

Solution

a) The range from 124.4 to 125.6 is $\bar{x} \pm 2s$ $(125.0 - 2(.3) = 124.4 \quad 125.0 + 2(.3) = 125.6)$. The empirical rule suggests that about 95% of the data values should fall within those limits.

b) The range from 124.1 to 125.9 is $\bar{x} \pm 3s$ $(125.0 - 3(.3) = 124.3 \quad 125.0 + 3(.3) = 125.9)$. The empirical rule suggests that about 99.7% of the data values should fall within those limits.

Exercise

The mean value of land and buildings per acre from a sample of farms is \$1,500, with a standard deviation of \$200. Using the empirical rule, estimate the percent of farms whose land and building values per acre are between \$1,300 and \$1,700. (Assume the data set has a bell-shaped distribution.)

Solution

$$(1300,\ 1700) \to (1500 - 1(200),\ \ 1500 + 1(200))$$

$$\Rightarrow\ 1500 \pm 200 = \bar{x} \pm \delta$$

68% of the farms have values between \$1,300 and \$1,700 per acre

Exercise

The mean value of land and buildings per acre from a sample of farms is \$2,400, with a standard deviation of \$450. Using the empirical rule, between what two values do about 95% of the data lie? (Assume the data set has a bell-shaped distribution.)

Solution

$$95\% \ \ \to \ \ x \pm 2\delta$$

$$2400 \pm 2(450) \ \ \to \ \begin{cases} 2400 - 900 = 1500 \\ 2400 + 900 = 3300 \end{cases}$$

95% of the farms have values between \$1,500 and \$3,300 per acre

Exercise

Heights of women have a bell-shaped distribution with a mean of 161 cm and a standard deviation of 7 cm. Using Chebyshev's Theorem, what do we know about the percentage of women with heights that are within 2 standard deviations of the mean? What are the minimum and maximum heights that are within 2 standard deviations of the mean?

Solution

Chebyshev's Theorem states that for any set of data there must be at least $1 - \left(\frac{1}{2}\right)^2 = 1 - \frac{1}{4} = \frac{3}{4}$ (75%) of the data values within 2 standard deviations of the mean. In this context, the limits are $161 \pm 2(7)$: there must be at least 75% of the heights between 147 cm and 175 cm.

Exercise

The author's Generac generator produces voltage amounts with a mean of 125.0 volts and standard deviation of 0.3 volts. Using Chebyshev's theorem, what do we know about the percentage of voltage amounts that are within 3 standard deviations of the mean? What are the minimum and maximum voltage amounts that are within 3 standard deviations of the mean?

Solution

Chebyshev's Theorem states that for any set of data there must be at least

$$1-\frac{1}{k^2}=1-\frac{1}{3^2}=1-\frac{1}{9}=\underline{\left|\frac{8}{9}\right.}$$

(89%) of the data values within 3 standard deviations of the mean.

In this context, the limits are $125\pm 3(0.3)$: there must be at least 75% of the heights between 124.1 volts and 125.9 volts.

Exercise

The mean time in a women's 400-meter dash is 57.07 seconds, with a standard deviation of 1.05 seconds. Apply Chebyshev's Theorem to the data using $k=2$. Interpret the results.

Solution

$$1-\frac{1}{k^2}=1-\frac{1}{2^2}=1-\frac{1}{4}=\underline{\left|0.75\right.}$$

At least 75% of the 400-meter dash within 2 standard deviations of the mean

The limits are $x\pm 2\delta \rightarrow 57.07\pm 2(1.05)\Rightarrow \begin{cases} 57.07-2.1=54.97 \\ 57.07+2.1=59.17 \end{cases}$

There must be at least 75% of the 400-meter dash between 54.97 and 59.17 seconds.

Exercise

The number of gallons of water consumed per day by a small village are listed. Make a frequency distribution (using five classes) for the data set. Then approximate the population mean and the population standard deviation of the data set.

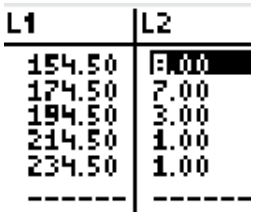
167 180 192 173 145 151 174 175 178 160
195 224 244 146 162 146 177 163 149 188

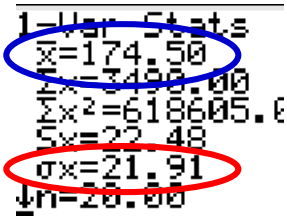
Solution

$$\text{Class width} = \frac{Max - Min}{5}$$

$$= \frac{244 - 145}{5}$$

$$= 19.8 \approx 20$$





Class	x	f	$f \cdot x$
145 – 164	154.5	8	1236.0
165 – 184	174.5	7	1221.5
185 – 204	194.5	3	583.5
205 – 224	214.5	1	214.5
225 – 244	234.5	1	234.5
		20	3490.0

$$\mu = \frac{\sum xf}{N} = \frac{3490}{20} = 174.5$$

$x - \mu$	$(x-\mu)^2$	f	$(x-\mu)^2 f$
–20	400	8	3200
0	0	7	0
20	400	3	1200
40	1600	1	1600
60	3600	1	3600
			9600

$$\delta = \sqrt{\frac{\sum (x-\mu)^2 f}{N}}$$

$$= \sqrt{\frac{9600}{20}}$$

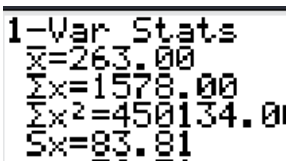
$$\approx 21.9$$

Exercise

To get the best deal on a microwave oven, Jeremy called six appliance stores and asked the cost of a specific model. The prices he was quoted are listed below:

$325 $384 $156 $210 $219 $284

Find the variance for the given data.



Solution

$\delta \approx 83.81$

$variance = 83.81^2 = \$7{,}024.00$

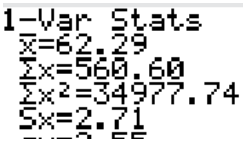
Exercise

Compare the variation in heights to the variation in weights of thirteen-year old girls. The heights (in inches) and weights (in pounds) of nine randomly selected thirteen-year old girls as listed below

Heights (inches):	59.3	61.2	62.6	64.7	60.1	58.3	64.6	63.7	66.1
Weights (pounds):	87	96	91	119	96	90	123	98	139

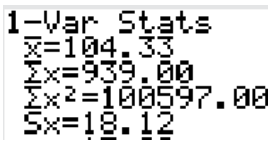
Find the coefficient of variation for each of the two sets of data, then compare the variation

Solution



Heights: $\bar{x} = 62.29, \quad s = 2.71$

Coefficient of variation: $CV = \frac{s}{\bar{x}} = \frac{2.71}{62.29} \approx 0.043 = 4.3\%$



Weights: $\bar{x} = 104.33, \quad s = 18.12$

Coefficient of variation: $CV = \frac{s}{\bar{x}} = \frac{18.12}{104.33} \approx 0.1736 = 17.4\%$

There is substantially more variation in the weights than in the heights of the girls

Exercise

The amount of Jen's monthly phone bill is normally distributed with a mean of $56 and a standard deviation of $9. What percentage of her phone bills are between $29 and $83? Use the empirical rule to solve.

Solution

$83 = 56 + 9 + 9 + 9$

$29 = 56 - 9 - 9 - 9$

29 and 83 are each exactly 3 standard deviation away from the mean 56.

The empirical rule tells us that about 99.7% of all values are within 3 standard deviation of the mean, so about 99.7% of her phone bills are between 70 and 130.

Solution *Section* 1.10 – Measures of Position, Outliers, and Boxplots

Exercise

When Reese Witherspoon won an Oscar as Best Actress for the movie *Walk the Line*, her age was converted to a z-score of -0.61 when included among the ages of all other Oscar-winning Best Actress at the time of this writing. Was her age above the mean or below the mean? How many standard deviations away from the mean is her age?

Solution

For a z score of -0.61, the negative sign indicates that her age is below the mean and the numerical portion indicates that her age is 0.61 standard deviations away from the mean.

Exercise

Hoffman was 38 years of age when he won a Best Actor Oscar for his role in Capote. The Oscar-winning Best Actors have a mean age of 43.8 years and a standard deviation of 8.9 years.

a) What is the difference between Hoffman's age and the mean age?

b) How many standard deviations is that (the difference found in part (a))?

c) Convert Hoffman's age to a z-score.

d) If we consider "usual" ages to be those that convert to z-scores between -2 and 2, is Hoffman's age usual or unusual?

Solution

a) $|x - \bar{x}| = |38 - 43.8|$

$= |-5.8|$

$= 5.8 \text{ years}$

b) $\frac{5.8}{8.9} = 0.65$

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

$= \frac{38 - 43.8}{8.9}$ $\quad (38 - 43.8) / 8.9$

$= -0.65$

d) Since $-2.00 < -0.65 < 2.00$, Hoffman's age is not considered unusual in this context.

Exercise

Eruptions of the Old Faithful geyser have duration times with a mean of 245.0 sec and a standard deviation of 36.4 sec. One eruption had a duration time of 110 sec.

a) What is the difference between a duration time of 110 sec and the mea?

b) How many standard deviations is that (the difference found in part (a))?

c) Convert duration time of 110 sec to a z-score.

d) If we consider "usual" ages to be those that convert to z-scores between –2 and 2, is a duration time of 110 sec usual or unusual?

Solution

a) $|x-\bar{x}| = |110-245.0| = |-135.0| = 135.0 \ sec$

b) $\frac{135.0}{36.4} = 3.71$

c) $z-\frac{x-\bar{x}}{s} = \frac{110-245}{36.4} = -3.71$ $(110-245)/36.4$

d) Since $-3.71 < -2.00$, a duration time of 110 seconds is considered unusual in this context.

Exercise

Human body temperatures have a mean of 98.20°F and a standard deviation of 0.62°F. Convert each given temperature to a z-score and determine whether it is usual and unusual.

a) 101.00°F *b*) 96.90°F *c*) 96.98°F

Solution

a) $z-\frac{x-\bar{x}}{s} = \frac{101.0-98.20}{0.62} = 4.52$; since $4.52 > 2.00$ it is unusual. $(101-98.2)/.62$

b) $z-\frac{x-\bar{x}}{s} = \frac{96.9-98.20}{0.62} = -2.10$; since $-2.10 < 2.00$ it is unusual.

c) $z-\frac{x-\bar{x}}{s} = \frac{96.98-98.20}{0.62} = -1.97$; since $-2.00 < -1.97 < 2.00$ it is usual.

Exercise

Scores on SAT test have a mean of 1518 and a standard deviation of 325. Scores on the ACT test have a mean of 21.1 and standard deviation of 4.8. Which is relatively better: a score of 1840 on the SAT test or a score of 26.0 on the ACT test? Why?

Solution

SAT: $z-\frac{x-\bar{x}}{s} = \frac{1840-1518}{325} = 0.99$ $(1840-1518)/325$

ACT: $z-\frac{x-\bar{x}}{s} = \frac{26.0-21.1}{4.8} = 1.02$ $(26-21.1)/4.8$

Since 1.02 > 0.99, the ACT score of 26.0 is the relatively better score.

Exercise

Scores on SAT test have a mean of 1518 and a standard deviation of 325. Scores on the ACT test have a mean of 21.1 and standard deviation of 4.8. Which is relatively better: a score of 1190 on the SAT test or a score of 16.0 on the ACT test? Why?

Solution

SAT: $z - \frac{x-\bar{x}}{s} = \frac{1190-1518}{325} = \underline{-1.01}$ $(1190-1518)/325$

ACT: $z - \frac{x-\bar{x}}{s} = \frac{16.0-21.1}{4.8} = \underline{-1.06}$ $(16-21.1)/4.8$

Since $-1.01 > -1.06$, the SAT score of 119 is the relatively better score.

Exercise

Use the given sorted values, which are the numbers of points scored in the Super Bowl for a recent period of 24 years. Find the percentile corresponding to the given number of points

36 37 37 39 39 41 43 44 44 47 50 53 54 55 56 56 57 59 61 61 65 69 69 75

a) 47 *b*) 65 *c*) 54 *d*) 41

Solution

Let b to the number of scores below x; n is the total number of scores.

The percentile score of x is $\frac{b}{n} \cdot 100$

a) The percentile score of 47 is $\frac{9}{24} \cdot 100 = \underline{38}$ 36 37 37 39 39 41 43 44 44 $\rightarrow b = 9$

b) The percentile score of 65 is $\frac{20}{24} \cdot 100 = \underline{83}$

c) The percentile score of 54 is $\frac{12}{24} \cdot 100 = \underline{50}$

d) The percentile score of 41 is $\frac{5}{24} \cdot 100 = \underline{21}$

Exercise

For the given data, find the indicated percentile or quartile

36 37 37 39 39 41 43 44 44 47 50 53 54 55 56 56 57 59 61 61 65 69 69 75

a) P_{20} *c)* P_{50} *e)* P_{25} *g)* Q_1

b) P_{80} *d)* P_{75} *f)* P_{95} *h)* Q_3

Solution

$N = 24$

a) $L=\frac{20}{100}\cdot 24=4.8\approx 5$. Since the 5th score is 39, then $P_{20}=39$

b) $L=\frac{80}{100}\cdot 24=19.2\approx 20$. Since the 20th score is 61, then $P_{80}=61$

c) $L=\frac{50}{100}\cdot 24=12$ (it is a whole number).

The mean of the 12th and 13th score, then $P_{50}=\frac{53+54}{2}=53.5$

d) $L=\frac{75}{100}\cdot 24=18$ (it is a whole number).

The mean of the 18th and 19th score, then $P_{75}=\frac{59+61}{2}=60$

e) $L=\frac{25}{100}\cdot 24=6$ (it is a whole number).

The mean of the 6th and 7th score, then $P_{25}=\frac{41+43}{2}=42$

f) $L=\frac{95}{100}\cdot 24=22.8\approx 23$. Since the 23th score is 69, then $P_{95}=69$

g) $Q_1=P_{25}$, $L=\frac{25}{100}\cdot 24=6$ (it is a whole number).

The mean of the 6th and 7th score, then $Q_1=\frac{41+43}{2}=42$

h) $Q_3=P_{75}$, $L=\frac{75}{100}\cdot 24=18$ (it is a whole number).

The mean of the 18th and 19th score, then $Q_3=\frac{59+61}{2}=60$

Exercise

The number of hours of television watched per day by a sample of 28 people

2 4 1 5 7 2 5 4 4 2 3 6 4 3 5 2 0 3 5 9 4 5 2 1 3 6 7 2

a) Find the data set's first, second, and third quartiles.
b) Draw a box-and-whisker plot that represents the data set.
c) About 75% of the people watched no more than now many hours of television per day?
d) What percent of the people watched more than 4 hours of television per day?
e) If you randomly selected one person from the sample, what is the likelihood that the person watched less than 2 hours of television per day? Write your answer as a percent.

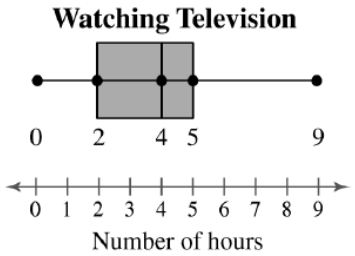
Solution

$N=24\rightarrow$ $\frac{1}{4}(28)=7$ $\frac{1}{2}(28)=14$ $\frac{3}{4}(28)=21$

0 1 1 2 2 2 **2** 2 2 3 3 3 3 **4** 4 4 4 4 5 5 5 **5** 5 6 6 7 7 9

a) $Q_1=2;\quad Q_2=4;\quad Q_3=5$

b)



c) $75\% \rightarrow Q_3$ 5

d) The percentage of the people watched more than 4 hours of television per day is 50%

e) Less than 2 hours of television per day is 25%

Exercise

The hourly earnings (in dollars) of a sample of 25 railroad equipment manufacturers

15.6 18.75 14.6 15.8 14.35 13.9 17.5 17.55 13.0 14.2 19.05 15.35 15.2
19.45 15.95 16.5 16.3 15.25 15.05 19.1 15.2 16.22 17.75 18.4 15.25

a) Find the data set's first, second, and third quartiles.

b) Draw a box-and-whisker plot that represents the data set.

c) About 75% of the manufacturers made less than \$15.80 per hour?

d) What percent of the manufacturers made more than \$15.80 per hour?

e) If you randomly selected one manufacturer from the sample, what is the likelihood that the manufacturer made less than \$15.80 per hour? Write your answer as a percent.

Solution

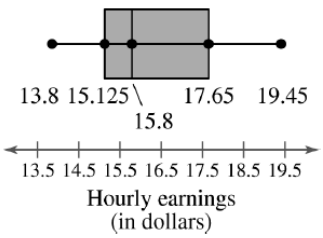
13.0 13.9 14.2 14.35 14.6 **15.05** **15.2** 15.2 15.25 15.25 15.35 15.6 **15.8**
15.95 16.22 16.3 16.5 17.5 **17.55** **17.75** 18.4 18.75 19.05 19.1 19.45

a) $Q_1 = P_{25}$, $L = \frac{25}{100} \cdot 25 = 6.25$, then $Q_1 = \frac{15.05 + 15.2}{2}$ $= 15.125$

$Q_2 = P_{50}$, $L = \frac{50}{100} \cdot 25 = 12.5 \approx 13$, then Q_2 $= 15.8$

$Q_3 = P_{75}$, $L = \frac{75}{100} \cdot 25 = 18.75$, then $Q_3 = \frac{17.55 + 17.75}{2}$ $= 17.65$

b) Railroad equipment manufacturers



c) $75\% \rightarrow Q_3$ 17.65

d) The percentage of the manufacturers made more than \$15.80 per hour is 50%

e) The manufacturer made less than \$15.80 per hour is 50%

Exercise

A certain brand of automobile tire has a mean life span of 35,000 miles, with a standard deviation of 2250 miles. (Assume the life spans of the tires have a bell-shaped distribution)

a) The life spans of three randomly selected tires are 34,000 miles, 37,000 miles, and 30,000 miles. Find the *z*-score that corresponds to each life span. According to the *z*-scores, would the life spans of any of these tires be considered unusual?

b) The life spans of three randomly selected tires are 30,500 miles, 37,250 miles, and 35,000 miles. Using the Empirical Rule, find the percentile that corresponds to each life span.

Solution

a) $x = 34{,}000 \quad z - \frac{x - \bar{x}}{s} = \frac{34{,}000 - 35{,}000}{2{,}250} \approx -0.44$

$x = 37{,}000 \quad z - \frac{x - \bar{x}}{s} = \frac{37{,}000 - 35{,}000}{2{,}250} \approx 0.89$

$x = 30{,}000 \quad z - \frac{x - \bar{x}}{s} = \frac{30{,}000 - 35{,}000}{2{,}250} \approx -2.22$

The tire with a life span of 30,000 miles has an unusual short life span.

b) $x = 30{,}500 \quad z - \frac{x - \bar{x}}{s} = \frac{30{,}500 - 35{,}000}{2{,}250} = -2 \Rightarrow 2.5^{\text{th}}$ percentile

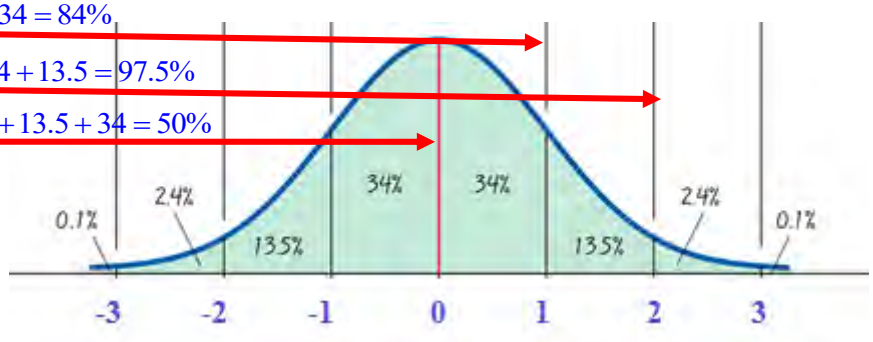
$x = 37{,}250 \quad z - \frac{x - \bar{x}}{s} = \frac{37{,}250 - 35{,}000}{2{,}250} = 1 \Rightarrow 84^{\text{th}}$ percentile

$x = 35{,}000 \quad z - \frac{x - \bar{x}}{s} = \frac{35{,}000 - 35{,}000}{2{,}250} = 0 \Rightarrow 50^{\text{th}}$ percentile

$1 \rightarrow 50 + 34 = 84\%$

$2 \rightarrow 50 + 34 + 13.5 = 97.5\%$

$0 \rightarrow 0.1 + 2.4 + 13.5 + 34 = 50\%$



Exercise

The life spans of species of fruit fly have a bell shaped distribution, with mean of 33 days and a standard deviation of 4 days.

a) The life spans of three randomly selected fruit flies are 34 days, 30 days, and 42 days. Find the *z*-score that corresponds to each life span and determine if any of these life spans are unusual.

b) The life spans of three randomly selected fruit flies are 29 days, 41 days, and 25 days. Using the Empirical Rule, find the percentile that corresponds to each life span.

Solution

a) $x = 34 \quad z - \frac{x-\mu}{\delta} = \frac{34-33}{4} \approx 0.25$

$x = 30 \quad z - \frac{x-\mu}{\delta} = \frac{30-33}{4} \approx -0.75$

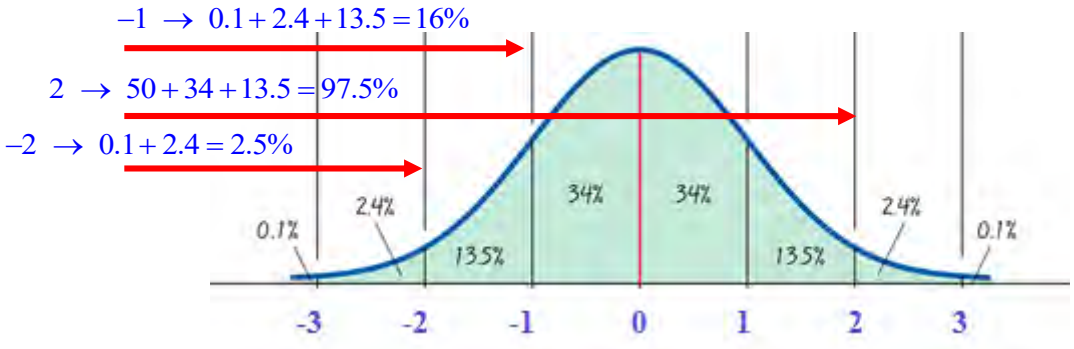
$x = 42 \quad z - \frac{x-\mu}{\delta} = \frac{42-33}{4} \approx 2.25$

The fruit fly with a life span of 42 days has an unusual short life span.

b) $x = 29 \quad z - \frac{x-\mu}{\delta} = \frac{29-33}{4} \approx -1 \quad \Rightarrow 16^{\text{th}}$ percentile

$x = 41 \quad z - \frac{x-\mu}{\delta} = \frac{41-33}{4} \approx 2 \quad \Rightarrow 97.5^{\text{th}}$ percentile

$x = 25 \quad z - \frac{x-\mu}{\delta} = \frac{25-33}{4} \approx -2 \quad \Rightarrow 2.5^{\text{th}}$ percentile



Exercise

Find the Q_1 and Q_3 for the given data: 49 52 52 52 74 67 55 55

Solution

49 52 52 52 55 55 67 74

$Q_1 = 52$

$Q_3 = \frac{55+67}{2} = 61$

Exercise

Find the Q_1 and Q_3 for the given weights (in pounds) of 30 newborn babies listed below:

5.5 5.7 5.8 6.0 6.1 6.1 6.3 6.4 6.5 6.6
6.7 6.7 6.7 6.9 7.0 7.0 7.0 7.1 7.2 7.2
7.4 7.5 7.7 7.7 7.8 8.0 8.1 8.1 8.3 8.7

Solution

$Q_1 = 6.4 \ lb$

$Q_3 = (L_{23}) = 7.7 \ lb$

Exercise

Find the percentile for the data value:

113 125 117 111 119 121 111 109 116 113 117 127 109 113 115 110

Data value: 119

Solution

Since there is 12 number less than 119

$$percentile = \frac{12 \times 100}{16} = 75$$

Exercise

The test scores of 40 students are listed below:

30 35 43 44 47 48 54 55 56 57 59 62 63 65 66 68 69 69 71 72
72 73 74 76 77 77 78 79 80 81 81 82 83 85 89 92 93 94 97 98

Find P_{56}

Solution

$$L = \frac{56 \times 40}{100} = 22.4 \approx 23$$

$P_{56} = 74$