Statistics for the Biological, Environmental and Health Sciences

STAT 007

Introduction to Statistics

Chapter 1

Statistical and Critical Thinking

Section 1-1

- In this section we will:
  - Introduce some definitions, such as data, statistics, population, sample.
  - Discuss what a statistical study involves.
  - Introduce the concept of critical thinking.

Some Basic Definitions

Definition

- ***Data*** *are collections of observations, such as measurements, or survey responses.*

Data Set 9: IQ and Brain Size

Data are from 20 monozygotic (identical) twins (first five rows shown here). **PAIR** identifies the set of twins, **SEX** is the gender of the subject (1 = male, 2 = female), **ORDER** is the birth order, **IQ** is measured full IQ score, **VOL** is total brain volume (cm^3), **AREA** is total brain surface area (cm^2), **CCSA** is corpus callosum (fissure connecting left and right cerebral hemispheres) surface area (cm^2), **CIRC** is head circumference (cm), and **WT** is body weight (kg).

Data provided by M. J. Tramo, W. C. Loftus, T. A. Stukel, J. B. Weaver, M. S. Gazziniga. See "Brain Size, Head Size, and IQ in Monozygotic Twins," *Neurology*, Vol. 50.

TI-83/84 list names (IQBRAIN): PAIR, SEX, ORDER, IQ, VOL, AREA, CCSA, CIRC, BWT

PAIR	SEX (1 = M)	ORDER	IQ	VOL	AREA	CCSA	CIRC	WT
1		1	96	1005	1913.88	6.08	54.7	57.607
1	2	2	89	963	1684.89	5.73	54.2	58.968
2	2	1	87	1035	1902.36	6.22	53.0	64.184
2	2	2		1027	1860.24	5.80	52.9	58.514
3	2	1	101	1281	2264.25	7.99	57.8	63.958

(Complete data sets available at www.TriolaStats.com)

- ***Statistics*** *is the science of planning studies and experiments; obtaining data; and organizing, summarizing, presenting, analyzing, and interpreting those data and then drawing conclusions based on them.*

Some Basic Definitions

Definition

- *A **population** is the complete collection of all measurements or data that are being considered. Typically, the population is the complete collection of data that we would like to make inferences about.*
- *A **sample** is a subcollection of members selected from a population.*
- *A **census** is the collection of data from every member of the population.*

Statistical Study

- The process of conducting a statistical study involves: **prepare, analyze** and **conclude**.

Statistical Study		
Prepare	Analyze	Conclude
- what does that data describe? - what is its source? - what was the sampling method?	- type of data - graphics and descriptive summaries - statistical analysis	- statistical significance - practical significance

Statistical Study: Prepare

- Context
  - What do the data represent?
  - What is the goal of the study?
  - What hypothesis are of interest?
- Source of the data
  - Are the data from a source with a special interest so that there is pressure to obtain results that are favorable to the source?
  - What are the sources of funding of the study?
- Sampling Method
  - Were the data collected in a way that is unbiased, or were the data collected in a way that is biased.
  - Biased data can be obtained when:
    + respondents themselves decide whether to be included (voluntary response sample). Examples: internet polls, mail-in polls, telephone call-in.
    + payments may have affected participation.

Statistical Study: Analyze

- Graph the Data.
- Explore the Data:
  - Are there *outliers*?
  - What important *statistics* summarize the data?
  - How are the data *distributed*?
  - Are there *missing* data?
  - Did many selected subjects refuse to respond?
- Apply Statistical Methods: a good statistical analysis requires using common sense and paying careful attention to right statistical methods.
  - Use technology to obtain results.

Statistical Study: Conclude

- Significance
  - Do the results have statistical significance?

    Statistical significance is achieved in a study when we get a result that is very unlikely to occur by chance. A common criterion is that we have statistical significance if the *likelihood of an event occurring by chance is 5% (0.05) or less*.
  - Do the results have practical significance?

    Practical significance: It is possible that some treatment or finding is effective, but *common sense* might suggest that the treatment or finding does not make enough of a difference to justify its use or to be practical.

Statistical Study: An example

Example

Consider the data described in the following table (from Appendix B)

IQ	96	87	101	103	127	96	88	85	97	124
Brain Volume (cm^3)	1005	1035	1281	1051	1034	1079	1104	1439	1029	1160

The data in consists of measured IQ scores and measured brain volumes from 10 different randomly selected subjects. The data were provided by M. J. Tramo, W. C. Loftus, T. A. Stukel, J. B. Weaver, and M. S. Gazziniga, who discuss the data in the article "Brain Size, Head Size, and IQ in Monozygotic Twins," *Neurology*, Vol. 50.

Discuss the process of a statistical study:

a) **Prepare**: what does the data represent and what could the goal of the study be? What is the source of the data? What about the sampling method?
b) **Analyze**: explore the data by graphs and summary statistics and use statistical methods.
c) **Conclude**: do the results have statistical and/or practical significance?

Statistical Study: statistical v/s practical significance

Example

ProCare Industries once supplied a product named Gender Choice that supposedly increased the chance of a couple having a baby with the gender that they desired. In the absence of any evidence of its effectiveness, the product was banned by the Food and Drug Administration (FDA) as a "gross deception of the consumer."

Suppose that the product was tested with 10,000 couples who wanted to have baby girls, and the results consist of 5,200 baby girls born in the 10,000 births.

The likelihood of this result happening due to chance is only 0.003%.

Is this result statistically significant?

Has this result practical significance?

Analyzing Data: Potential Pitfalls

- **Misleading conclusions**: make statements that are clear and simple, and avoid making statements not justified by the statistical analysis.
  Example (Correlation does not imply causation): "in a study was concluded that there is strong *correlation* between number of people drowned in swimming pool and amount of power generated by nuclear plants. Therefore, increasing power from nuclear power plants is the cause of more deaths in swimming pools."

- **Sample Data Reported Instead of Measured**: gather data by making measurements instead of asking people to *report* the values.

- **Loaded Questions**: be careful when wording the questions in surveys.
  Example: Where do you enjoy drinking beer?

- **Nonresponse**: when someone either refuses to respond to a survey question or is unavailable. There might be information in that *nonresponse*.

- **Percentages**: some studies cite misleading or unclear percentages.
  Example: "a researcher states that a new treatment for migraine headaches reduces them in a 150%."

  6% of 1200 is 72 is obtained as follows: $\frac{6}{100} \times 1200 = 72$.

Critical Thinking

- Statistical thinking involves *critical thinking* and the ability to make sense of results.
- For the example below, discuss the sampling method, the displayed graphs, and their potential implications.
  The survey was posted on a webpage.

Critical thinking: USA Today survey: respondents were asked if they need caffeine to start up their brain for the day. Among 2,006 respondents, 74% said that they did need the caffeine.

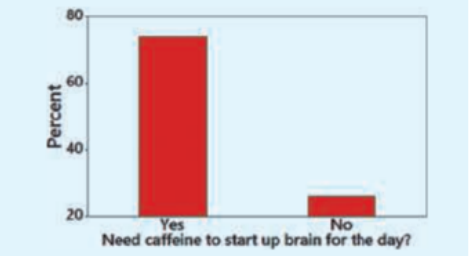
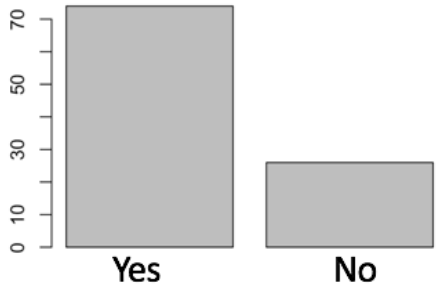


FIGURE 1-1(a) Survey Results



Practice

Look at the exercises at the end of Section 1-1 in page 9.

Specially, look at exercises:
1, 2, 3, 5, 7, 8, 10, 13, 15, 18, 19, 20, 21, 24, 25, 26, 27, and 29.

Types of Data

Section 1-2

- In this section we will:
  - define the terms *statistic* and *parameter*.
  - discuss different types of data.

- A major use of statistics is to collect and to use *sample data* to make conclusions about the *population*.
- Importance of knowing type of data: Different types of data require different statistical analysis.

Type of Data
categorical v/s quantitative
discrete v/s continuous
nominal, ordinal, interval or ratio level of measurement

- Important: Don't do computations and don't use statistical methods that are not appropriate for the data.

Parameter and Statistics

Definition

*A **parameter** is a numerical measurement describing some characteristic of the population.*

*A **statistic** is a numerical measurement describing some characteristic of a sample.*

Example

Identify a *parameter* and a *statistic* in the following statement:

"There are 17,246,372 high school students in the United States. In a study of 8505 U.S. high school students 16 years of age or older, 44.5% of them said that they texted while driving at least once during the previous 30 days."

Types of Data: Categorical v/s Quantitative

Definition

- ***Categorical*** *(or qualitative or attribute)* ***data*** *consist of names or labels or categories (not numbers that represent counts or measurements).*
- ***Quantitative*** *(or numerical)* ***data*** *consist of numbers representing counts or measurements. It should include appropriate unit of measurement.*

Types of Data: Categorical v/s Quantitative

Example

Discuss whether type of data reported below are categorical or quantitative:

a) The gender (male / female) of subjects enrolled in a clinical trial.

b) The identification numbers 1, 2, 3,..., 25 assigned randomly to 25 subjects in a clinical trial.

c) The age (in years) of the subject enrolled in a study.

d) The systolic blood pressure (in millimeters of mercury (mmHg)) of the subject enrolled in a study.

Types of Data: Discrete v/s Continuous

Definition

- ***Discrete data***: *data values are quantitative and number of values is finite or "countable" (there are infinitely many values).*
- ***Continuous data***: *infinitely many possible quantitative data values. The collection of values is not "countable" (values are on a continuous scale).*

Types of Data: Discrete v/s Continuous

Example

Discuss whether type of data reported below are discrete or continuous.

a) Number of physical examinations a subject enrolled in a study, that lasts one year, has.

b) The volume of blood drawn from a subject (between 0 mL and 50 mL).

c) Suppose that the accuracy of a blood typing test wants to be tested. To do this, a sample of the subject's blood type is submitted to the test until the test yields an error. How many times it is submitted is recorded.

Types of Data: Level of Measurement

Definition

- *The **nominal level of measurement data** type is not possible to arrange in some order.*
- *The **ordinal level of measurement data** type can be arranged in some order, but differences (obtained by subtraction) between data values either cannot be determined or are meaningless.*
- ***Interval level of measurement data** can be arranged in order, differences between data values can be found and are meaningful; but data at this level do not have a natural zero starting point at which none of the quantity is present.*
- ***Ratio level of measurement data** can be arranged in order, differences can be found and are meaningful, and there is a natural zero starting point (where zero indicates that none of the quantity is present). Also the term "twice" makes sense.*

Types of Data: Level of Measurement

Example

Discuss whether the type of data are of the nominal, ordinal, interval or ratio level of measurement.

a) Survey responses of “yes”, “no”, and “undecided”.

b) Letter grades of A, B, C, D, or F in a biostatistics class.

c) The body temperature (in degrees Farenheit) of the subject.

d) Year when the subject enrolled in the study.

e) The height (in cm) of a plant.

Practice

Look at the exercises at the end of Section 1-2 in page 22.

Specially, look at exercises:
1, 2, 3, 4, 5, 7, 9, 11, 12, 13, 15, 17, 19, 21, 23, 25, 26, 27, 28, 30, 31, 33.