Lecture Notes

For Statistics 301

Elementary Statistical Methods

Spring 2020

by

Jonathan Kuhn, Ph.D.
Associate Professor of Statistics,
Mathematics, Statistics and Computer Science Department,
Purdue University Northwest

Preface

This is an introductory course in statistics. The aim of this course is acquaint a student with some of the ideas, definitions and concepts of statistics. Numerical computation, algebra and graphs are used; calculus is *not* used.

These lecture notes are a necessary component for a student to successfully complete this course. Without them, a student will not be able to participate in the course.

- These lecture notes are *based* on the text.
- Although the material covered in lecture notes and text is very similar, the *presentation* of the material in the lecture notes is quite different from the presentation given in the text. The text consists essentially of definitions, formulas, worked out examples and exercises; these lecture notes, on the other hand, consist *mostly* of exercises to be worked out by the student with some definitions and formulas.
- The overhead presentation during each lecture is based *exclusively* on these lecture notes. A student fills in these lecture notes during the lecture.
- These lecture notes essentially mimic what goes on during the lectures.
- There are different kinds of exercises in the lecture notes, including multiple choice, true/false, matching and fill–in–the–blank.
- Each week, a student reads the text, answers the questions given here in the lecture notes, looks over the StatCrunch instructions, does the online MyStatLab homework assignment and then either the MyStatLab online test or quiz, in that order.

On the one hand, the lecture notes are, as you will see, quite a bit more elaborate than typical lecture notes, which are usually a summary of what the instructor finds important in a recommended course text. On the other hand, these lecture notes are not quite a text, because although it has many exercises, it does not have quite enough exercises to qualify it as a complete text. I should also point out that this workbook, unfortunately, possesses a number of typographical errors. In short, this workbook aspires to be text and, in the next few years, when enough exercises have been collected, and when most of the typographical errors have been weeded out, it will become a text.

Dr. Jonathan Kuhn,
Associate Professor of Statistics,
Purdue University Northwest
November 2019

Chapter 1

Data Collection

Statistics is about "educated guessing". It is about drawing conclusions from incomplete information. It is about collecting, organizing and analyzing data.

One important aspect of statistics is to do with the idea of gathering together a sample, calculating a statistic and using this statistic to infer something about a parameter of the population from which this sample is taken. This is generally called an *inferential statistical analysis* and this is what we will be concentrating on in this course.

1.1 Introduction to the Practice of Statistics

After describing difference between a variable and a data point, four important terms used in statistical inference are described: population, sample, statistic and parameter. Also, a number of different categorizations of variables (data) are given:

- nominal, ordinal, interval or ratio;
- qualitative (categorical) versus quantitative;
- quantitative: discrete versus continuous.

Exercise 1.1 (Introduction to the Practice of Statistics)

1. *Variable and Data.* A *variable* is a characteristic of a person, object or entity and a *data point* is a particular (observed, measured) instance of a variable.
   (a) **True** / **False** A data point for variable *height of a man* is 5.6 feet tall. Another data point for this variable is 5.8 feet tall.
   (b) **True** / **False** A data point for variable *person's country of birth* is Sweden. Another data point for this variable is U.S.A.
   (c) **True** / **False** A data point for variable *shoulder height of a cow* is 45 inches. Another data point for this variable is "Argentina".

(d) A data point for variable *woman's length of time of exposure in the sun* is (circle best one) **9/24/98 / 4 hours**.

(e) A data point for variable *person's date of exposure to the sun* is (circle best one) **9/24/98 / 4 hours**.

(f) **True / False** A data point for variable *woman's time of arrival* is 3pm. Another data point for this variable is 4:50pm. A *data set* for this variable is {1am, 2:30am, 12noon}. Another data set for this variable is {1:30am, 2:30am, 11:30am, 2pm}.

(g) **True / False** Data point "45" could be a particular instance of the variable *age of a elephant*. Data point "45" could also be a particular instance of the variable *number of marbles in a bag*.

(h) Data point "silver" is a particular instance of variable (circle none, one or more)

i. length of football field
ii. medal achieved at a track meet
iii. color choice of a car
iv. name of a horse

2. *Population, Sample, Statistic and Parameter.*
A *population* is a set of measurements or observations of a collection of objects A *sample* is a selected subset of a population. A *parameter* is a numerical quantity calculated from a population, whereas a *statistic* is a numerical quantity calculated from a sample.
Often, "population" refers to objects themselves, rather than measurements of objects. If given both answers and only one choice on a test, measurements of objects is best answer. Often, a parameter is calculated from an approximate mathematical *model* of population, rather than population itself, when little is known about the population.

(a) *Proportion Of Democrats*
Since 345 of 1000 Americans, randomly chosen from all 100 million Americans who can vote, are Democrats, we can infer approximately $\frac{345}{1000}$ths or 34.5% of **all** Americans are Democrats. Assume political preferences are either Democratic, Republican or Independent.

i. The *population* is (choose *two*!)
   A. *all* 100 million Americans who can vote.
   B. political preferences of *all* Americans who can vote.
   C. one thousand Americans, selected at random.
   D. political preferences of 1000 Americans, selected at random.

ii. The *sample* is (choose *two*!)
   A. *all* 100 million Americans who can vote.

B. political preferences of *all* Americans who can vote.
C. one thousand Americans, selected at random.
D. political preferences of 1000 Americans, selected at random.

iii. **True** / **False**. Although, loosely speaking, population is "all Americans" and sample is "one thousand Americans", we are actually interested in only one particular aspect of any given American; namely, their political preference. In other words, more exactly, population is "political preferences of all Americans" and sample is "political preferences of one thousand Americans".

iv. Variable of *interest* is (choose one)

A. an American, without specifying which American.
B. a particular American, "Susan", say.
C. political preference of an American.
D. Republican, political preference of a particular American, "Susan".
E. {Democrat, Democrat, Republican, Independent, . . . , Republican}, the set of political preferences for the one thousand randomly selected Americans.

v. A *data point* of variable of interest is,

A. an American, without specifying which American.
B. a particular American, "Susan".
C. political preference of an American.
D. Republican, political preference of a particular American, "Susan".
E. {Democrat, Democrat, Republican, Independent, . . . , Republican}, the set of political preferences for the one thousand randomly selected Americans.

vi. The *data* (or *data set*) is,

A. an American, without specifying which American.
B. a particular American, "Susan".
C. political preference of an American.
D. Republican, political preference of a particular American, "Susan".
E. {Democrat, Democrat, Republican, Independent, . . . , Republican}, the set of political preferences for the one thousand randomly selected Americans.

vii. Both *statistic* and *parameter* are numerical values, but *statistic* summarizes *sample*, whereas *parameter* summarizes *population*. In political preference situation, *statistic of interest* is,

A. proportion of Democrats, among *all* Americans.
B. proportion of Democrats, among 1000 randomly chosen Americans.

viii. *Value* of statistic of interest is (choose one)
14.5% / 24.5% / 34.5%

ix. The *parameter of interest* is,

A. proportion of Democrats, among *all* Americans.

B. proportion of Democrats, among 1000 randomly chosen Americans.

x. **True / False**. *Value* of a statistic is *known*; in this case, value of statistic is 34.5%. On the other hand, *value* of parameter is (often, typically) *un*known. One type of inferential statistics involves using known value of statistic to *estimate* unknown value of parameter.

xi. **True / False**. An appropriate analogy here would be to think of a box of 100 million tickets where each ticket has voting preference written on it ("Republican", "Democrat" or "Independent") as *population*; proportion of all of these tickets that are Democrat would be value of *parameter*. Random sample of 1000 tickets from population box would be a *sample*; proportion of sampled tickets that are Democrat would be value of *statistic*.

"Population" may just refer to 100 million tickets themselves, whatever is written on them.

(b) *Distance To Travel.*
At PNW, 120 students are randomly selected from entire 11,500 and asked their commute distance to campus. Average of 9.8 miles is computed from 120 selected. We infer from data *all* students have 9.8 average commute.

True / False. An appropriate analogy here would be to think of a box of 11,500 tickets where each ticket has commute distance written on it as *population*; average of all of these tickets would be value of *parameter*. A random sample of 120 tickets taken from this population box would be a *sample*; average of sampled tickets would be value of *statistic*.

"Sample" may also refer to 11,500 tickets themselves, whatever is written on them.

Match columns.

terms	travel example
(a) data point	**(A)** average commute distance for 120 students
(b) variable	**(B)** all students at PNW
(c) parameter	**(C)** commute distances for all students at PNW
(d) population	**(D)** commute distance for any PNW student
(e) sample	**(E)** average commute distance for all students
(f) statistic	**(F)** 120 students
	(G) 120 commute distances
	(H) 8 mile commute distance for a particular student

terms	(a)	(b)	(c)	(d)	(e)	(f)
travel example						

Some items in first column have more than one match; for example, (d) population matches with both (b) all students at PNW and (c) commute distances for all students at PNW.

3. *Nominal, Ordinal (Ranked), Interval and Ratio: Milk Yield*

Nominal Variable (Data): Variable where data can*not* be ordered; data consists of names or labels.

Ordinal (Ranked) Variable (Data): Variable where data can be ordered but can*not* be added or subtracted.

Interval Variable (Data): Variable where data can be both ordered and added or subtracted, but can*not* be divided or multiplied.

When two data points are subtracted from one another, difference between two is an *interval*, which explains why this is called interval variable (data).

Ratio Variable (Data): Variable where data can be ordered, added or subtracted, and also divided or multiplied.

When two data points are divided, a *ratio* is formed, which explains why this is called ratio variable (data).

Measurements are given below on a number of cows taken during a study on effect of a hormone, given in tablet form, on daily milk yield. Eight variables, including "Cow", "Test Date", ..., "After Yield", are listed at top of columns in table. Seven observations (data points) are listed in seven rows below variables. Specify milk yield variables (data) as nominal, ordinal, interval or ratio.

Cow	Test Date	Farm	Height	Health	Tablets	Before Yield	After Yield
17	9/11/98	M	41	poor	2	100.7	100.3
18	9/11/98	F	40	bad	1	97.8	98.1
14	9/03/98	F	49	fair	3	98.8	99.6
15	9/01/98	M	45	good	3	100.9	100.0
16	9/10/98	F	42	poor	1	101.1	100.1
19	9/25/98	M	45	good	2	100.0	100.4
20	9/25/98	M	37	good	3	101.5	100.8

(a) **True** / **False**. Variable "Cow" is *nominal* because

- cows cannot be ordered in any meaningful way by ID (name) alone
- interval between cow names is meaningless, data cannot be subtracted: cow 14 − cow 15 = huh?
- ratio between cow names is meaningless, data cannot be divided: cow 14 ÷ cow 15 = huh?

"Cow" is actually "cow identification (ID) or name of cow".

(b) **True** / **False**. Variable "Test Date" is *interval* because

- test dates can be ordered: "9/03/98" is before "9/11/98"

- interval between test dates is meaningful, data can be subtracted: "9/11/98" − "9/03/98" = 8 days
- ratio between test dates is meaningless, data can*not* be divided: "9/11/98" ÷ "9/03/98" = huh?

"Test Date" is date when cow is subjected to hormone. Examples of other interval variables (which are not in this example) include latitudes, longitudes, compass directions, times of day and normalized scores.

(c) Variable "Farm" is **nominal** / **ordinal** / **interval** / **ratio** because

- farms cannot be ordered in any meaningful way by name alone
- interval between farm names is meaningless, data cannot be subtracted: farm M − farm F = huh?
- ratio between farm names is meaningless, data cannot be divided: farm M ÷ farm F = huh?

"Farm" is name of farm, either farm "M" or farm "F".

(d) Variable "Height" is **nominal** / **ordinal** / **interval** / **ratio** because

- heights can be ordered: "41" inches is less than "49" inches
- interval between heights is meaningful, data can be subtracted: 49 inch tall cow − 41 inch tall cow = 8 inch difference in height
- ratio between heights is meaningful, data can be divided: 40 inch tall cow $\frac{40}{20} = 2$ times as tall as 20 inch cow

"Height" is measured to shoulder of a cow.

(e) Variable "Health" is **nominal** / **ordinal** / **interval** / **ratio** because

- health levels can be ordered: "fair" is worse than "good"
- interval between health levels meaningless, data can*not* be subtracted: "good" − "fair" = huh?
- ratio between health levels meaningless, data can*not* be divided: "good" ÷ "fair" = huh?

"Health" is "health of cow at time of test" and is either "bad", "poor", "fair" or "good".

(f) Variable "Tablets" is **nominal** / **ordinal** / **interval** / **ratio** because

- number tablets can be ordered: "2" tablets is more than "1" tablet
- interval between number tablets meaningful, data can be subtracted: $2 - 1 = 1$ tablet difference
- ratio between number tablets is meaningful, data can be divided: 2 tablets is $\frac{2}{1} = 2$ times as many tablets as 1 tablet

"Tablets" is "number of tablets of hormone given to cow" and must be either 1, 2 or 3 tablets.

(g) Variable "Before Yield" is **nominal** / **ordinal** / **interval** / **ratio** because

- before yield can be ordered: 100.7 quarts is more than 97.8 quarts
- interval between before yields meaningful, data can be subtracted: $100.7 - 97.8 = 2.9$ quarts difference
- ratio between before yields is meaningful, data can be divided: 75 quarts is $\frac{75}{25} = 3$ times as many quarts as 25 quarts

"Before Yield" is milk yield, measured in quarts, before cow given hormone.

(h) *Summary.* Match columns.

milk yield example	level of measurement
(a) cow ID number	**(A)** nominal
(b) test date of cow	**(B)** ordinal
(c) cow's farm	**(C)** interval
(d) shoulder height of cow	**(D)** ratio
(e) cow's health	
(f) number of tablets given to cow	
(g) milk yield before hormone	
(h) milk yield after hormone	

example	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)
variable								

4. *Qualitative (Categorical) Versus Quantitative: Milk Yield.*
Nominal and ordinal (ranked) variables are *qualitative (categorical)* variables. Interval and ratio variables are *quantitative* variables.

Cow	Test Date	Farm	Height	Health	Tablets	Before Yield	After Yield
17	9/11/98	M	41	poor	2	100.7	100.3
18	9/11/98	F	40	bad	1	97.8	98.1
14	9/03/98	F	49	fair	3	98.8	99.6
15	9/01/98	M	45	good	3	100.9	100.0
16	9/10/98	F	42	poor	1	101.1	100.1
19	9/25/98	M	45	good	2	100.0	100.4
20	9/25/98	M	37	good	3	101.5	100.8

Specify whether the milk yield variables are either quantitative or qualitative.

milk yield example	type of variable
(a) cow ID number	**(A)** qualitative
(b) test date of cow	**(B)** quantitative
(c) cow's farm	
(d) shoulder height of cow	
(e) cow's health	
(f) number of tablets given to cow	
(g) milk yield before study	
(h) milk yield after study	

example	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)
variable								

5. *Discrete Versus Continuous: Milk Yield.*

Discrete Variable (Data): Variable where *quantitative* data is countable

Each data point is distinct and different from every other data point; there are "gaps" between discrete data points..

Continuous Variable (Data): Variable with *un*countable *quantitative* data

Variable where, for any two different data points, there is *always* a third data point between these two data points..

Cow	Test Date	Farm	Height	Health	Tablets	Before Yield	After Yield
17	9/11/98	M	41	poor	2	100.7	100.3
18	9/11/98	F	40	bad	1	97.8	98.1
14	9/03/98	F	49	fair	3	98.8	99.6
15	9/01/98	M	45	good	3	100.9	100.0
16	9/10/98	F	42	poor	1	101.1	100.1
19	9/25/98	M	45	good	2	100.0	100.4
20	9/25/98	M	37	good	3	101.5	100.8

Specify whether mild yield variables are discrete, continuous or qualitative.

(a) Variable "Cow" is **discrete** / **continuous** / **qualitative** because

- cow ID is nominal

(b) Variable "Test Date" is **discrete** / **continuous** / **qualitative** because

- test date is neither nominal or ordinal and so must be quantitative
- test dates can be counted: there are 5 different dates in study: "9/11/98", "9/03/98", "9/01/98", "9/10/98", "9/25/98"

(c) Variable "Farm" is **discrete** / **continuous** / **qualitative** because

- farm name is nominal

(d) Variable "Height" is **discrete** / **continuous** / **qualitative** because

- height is neither nominal or ordinal and so must be quantitative
- heights cannot be counted: there is always a third (and so an infinite number) in between two heights; between 41 and 40 there is 40.5 (and 40.435 and 40.764653580. . . and so on).

(e) Variable "Health" is **discrete** / **continuous** / **qualitative** because

- health level is ordinal

(f) Variable "Tablets" is **discrete** / **continuous** / **qualitative** because

- number tablets neither nominal or ordinal and so must be quantitative

- number tablets can be counted: there are 1, 2, or 3 tablets.

(g) Variable "Before Yield" is **discrete** / **continuous** / **qualitative** because

- before yield is neither nominal or ordinal and so must be quantitative
- before milk yields cannot be counted: there is always a third (and so an infinite number) in between two yields; between 100.7 and 97.8 there is 100.3 (and 100.435 and 100.764653580... and so on).

(h) **True** / **False** Continuous data is always *measured* discretely. For example, a person's age is given as 45 and not as 45.0023454304959340.... In fact, each age in this set of data can really only be one of a finite number of possibilities, say: { 1, 2, 3, ..., 120 } . So, although the age set of data *appears* to be a discrete set of data, it is really a continuous set of data, because this set belongs to a larger set where there is an infinity of values in any chosen interval.

(i) *Summary.* Match columns.

milk yield example	type of variable
(a) cow ID number	**(A)** discrete
(b) test date of cow	**(B)** continuous
(c) cow's farm	**(C)** qualitative
(d) shoulder height of cow	
(e) cow's health	
(f) number of tablets given to cow	
(g) milk yield before hormone	
(h) milk yield after hormone	

example	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)
variable								

1.2 Observational Studies versus Designed Experiments

In this section, we will describe and compare designed experiments with observational studies, learn about problem of confounding variables and describe three types of observational studies: cross-sectional, case-control and cohort.

Exercise 1.2 (Observational Studies versus Designed Experiments)

1. *Observational Study versus Designed Experiment.*
   In observational studies, *subject* who decides whether or not to be given treatment. In experimental designs, *experimenter* decides who is to be given treatment and who is to be control.

   (a) Effect of air temperature on rate of oxygen consumption (ROC) of four mice is investigated. ROC of one mouse at 0^o F is 9.7 mL/sec for example.

temperature (F^o)	0	10	20	30
ROC (mL/sec)	9.7	10.3	11.2	14.0

Since experimenter (not a mouse!) decides which mice are subjected to which temperature, this is (choose one)
observational study / designed experiment.

(b) Indiana police records from 1999–2001 on six drivers are analyzed to determine if there is an association between drinking and traffic accidents. One heavy drinker had 6 accidents for example.

drinking →	heavy	light
	3	1
	6	2
	2	1

This is an observational study because (choose one)

i. police decided who was going to drink and drive and who was not.
ii. drivers decided who was going to drink and drive and who was not.

(c) A recent study was conducted to compare academic achievement (measured by final examination scores) of Internet students with classroom students. This is an observational study because (choose one)

i. instructor assigned students to classroom or internet.
ii. students decided to attend classroom or Internet class.

(d) *Effect of drug on patient response.* Response from one patient given drug A is 120 units for example.

drug →	A	B	C
	120	97	134
	140	112	142
	125	100	129
	133	95	137

If this is a designed experiment, then (choose one)

i. experimenter assigns drugs to patients.
ii. patients assigns drugs to themselves.

2. *Explanatory variables, responses, confounding and lurking variables.*
Point of both observational studies and designed experiments is to identify variable or set of variables, called *explanatory variables*, which are thought to predict outcome or *response variable*. *Confounding* between explanatory variables occurs when two or more explanatory variables are not separated and so it is not clear how much each explanatory variable contributes in prediction of response variable. *Lurking* variable is explanatory variable not considered in study but confounded with one or more explanatory variables in study.

(a) *Effect of temperature on mice rate of oxygen consumption.*

temperature (F^o)	0	10	20	30
ROC (mL/sec)	9.7	10.3	11.2	14.0

i. Explanatory variable considered in study is (choose one)
   A. temperature
   B. rate of oxygen consumption
   C. mice
   D. mouse weight

ii. Response is (choose one)
   A. temperature
   B. rate of oxygen consumption
   C. mice
   D. room temperature

iii. Possible explanatory variable *not* considered in study (choose *two*!)
   A. temperature
   B. rate of oxygen consumption
   C. noise level
   D. mouse weight

iv. Mouse weight is lurking variable if *confounded* with temperature in, for example, following way.

temperature (F^o)	0^o	10^o	20^o	30^o
mouse weight (oz)	10	14	18	20
ROC (mL/sec)	9.7	10.3	11.2	14.0

Hotter temperatures are associated with heavier mice. Hottest temperature, 30^o F, is associated with heaviest mouse with weight (choose one) **9.7** / **14.0** / **20** ounces

(b) *Effect of drinking on traffic accidents.*
Indiana police records from 1999–2001 are analyzed to determine if there is an association between drinking and traffic accidents.

drinking	heavy drinker	3	6	2
	light drinker	1	2	1

i. Match columns.

Terminology	Example
(a) explanatory variable	**(A)** driver's age
(b) response	**(B)** amount of drinking
(c) lurking variable	**(C)** number of traffic accidents

Terminology	(a)	(b)	(c)
Example			

ii. Suppose age influences number of traffic accidents. Age is a confounding (and so lurking) variable with drinking in number of traffic accidents of Indiana drivers if (circle one)

A. young drivers had more traffic accidents than older drivers.

B. intoxicated drivers had more traffic accidents than sober drivers.

C. it was not clear at end of study whether number of traffic accidents was a consequence of being intoxicated or not, or whether it was a consequence of age.

iii. One way to eliminate confounding effect of age with drinking on number of traffic accidents (to control for age) in this observational study would be to (choose one)

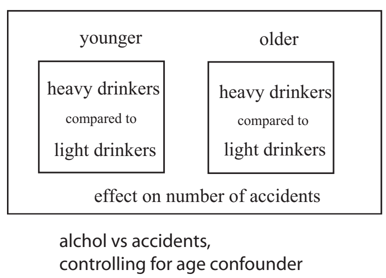

Figure 1.1 (Drinking and Driving Study, Controlling For Age)

A. assign drivers to be either drunk or sober at random (Is this possible, since the data was collected from police records?)

B. compare number of traffic accidents of drunk drivers with sober driver who both have similar ages, to *control* for age.

C. compare number of traffic accidents of drunk drivers with sober driver who both have different ages

(c) *Effect of teaching method on academic achievement.*
A recent study compares academic achievement (measured by final examination scores) of Internet students with classroom students.

i. Suppose average GPA influences academic achievement. Average student GPA is a confounding (lurking) variable with teaching method on academic achievement of students if (circle one)

A. students with high average GPAs had better final examination scores than students with low average GPAs.

B. Internet students had better final examination scores than classroom students.

C. it is not clear at end of study whether students' academic achievement is a consequence of being either an Internet students or classroom students, or is a consequence of average GPA.

ii. One way to eliminate confounding effect of average GPA with teaching method on academic achievement of students (to control for average GPA) in this observed study would be to (choose one)

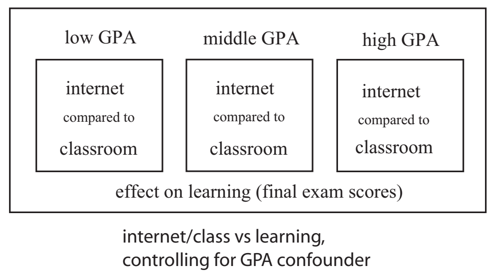

Figure 1.2 (Academic Achievement and Teaching Method, Controlling For GPA)

A. assign students to be either "classroom" or "Internet" students at random (Is this possible, since these students choose between these two options themselves?)

B. compare academic achievement of classroom students with Internet students who both have similar average GPAs, to *control* for GPA.

C. compare the academic achievement of classroom students with Internet students who both have different average GPAs

iii. **True / False** Controlling for confounder average GPA in this study does not control for any other confounder. Each confounder (lurking variable) must be controlled for separately from every other confounder in an observational study.

(This is unlike in a randomized designed experiment, where randomization takes care of all confounders all at once.)

3. *Types of observational studies: case-control, cross-sectional and cohort.*
Data collected for individuals over a short period of time is called a *cross-sectional* study. Data collected from historical records is a *retrospective* (or *case-control*) study. In this study, individuals with a characteristic are matched with individuals without this characteristic (a control). Data collected over time (into future) of a group (or cohort) of individuals is a *prospective* (or *cohort*) study.

(a) *Effect of drinking on traffic accidents.*
Indiana police records from 1999–2001 are analyzed to determine if there is an association between drinking and traffic accidents.

drinking	heavy drinker	3	6	2
	light drinker	1	2	1

If number of traffic accidents of drunk drivers is compared with sober drivers who both have similar characteristics such as age, gender, health and so on, this is a (choose one)

i. cross-sectional observational study
ii. retrospective (case-control) observational study
iii. prospective (cohort) observational study
iv. designed experiment

(b) *Explanatory variables influencing traffic accidents.*
If a large group of individuals are observed over an extended period of time to determine explanatory variables contributing to traffic accidents, this is a (choose one)

i. cross-sectional observational study
ii. retrospective (case-control) observational study
iii. prospective (cohort) observational study
iv. designed experiment

(c) *Effect of teaching method on academic achievement.*
A recent study compares academic achievement (measured by final examination scores) of Internet students with classroom students. If data is collected for one set of exams given at one time, this is a (choose one)

i. cross-sectional observational study
ii. retrospective (case-control) observational study
iii. prospective (cohort) observational study
iv. designed experiment

(d) *Effect of temperature on mice rate of oxygen consumption.*

temperature (F^o)	0	10	20	30
ROC (mL/sec)	9.7	10.3	11.2	14.0

This is a (choose one)

i. cross-sectional observational study
ii. retrospective (case-control) observational study
iii. prospective (cohort) observational study
iv. designed experiment

4. *Census.*
True / **False**. A census is a data set of specified variables for all members of a population.

1.3 Simple Random Sampling

Random sampling uses chance to choose a subset from a population. *Simple random sampling* (SRS) involves selecting n units out of N population units where every distinct sample has an *equal* chance of being drawn. SRS produce *representative* samples of population. *Frame* is list of all members in population.

Exercise 1.3 (Simple Random Sampling)

1. *Simple Random Sample: Decayed Teeth*
   Small *population* of number of decayed teeth for 20 children is represented by box of tickets below. Child 17 is ticket with 3 decayed teeth for example.

0	1	2	9	0	4	0	0	1	5
child 1	child 2	child 3	child 4	child 5	child 6	child 7	child 8	child 9	child 10
0	0	0	3	1	1	3	0	10	2
child 11	child 12	child 13	child 14	child 15	child 16	child 17	child 18	child 19	child 20

   Estimate population average number of decayed teeth per child with a sample average calculated from a simple random sample (SRS) of five children.

   (a) *Sample Average.* Use *StatCrunch*, *seed* 7, to draw SRS with*out* replacement of five tickets out box of tickets and record your findings in table below.

child in SRS	3	19	______	9	______
number of decayed teeth	2	10	______	1	______

   (StatCrunch: Relabel var1 as children, type 1, 2 ... 20 in column "children", click Data, Sample; in dialog box, click on "children", Sample size: 5, leave everything as is, until Seeding, choose "Use fixed seed" (to make sure everyone generates same sample) and Seed: replace given number with 7, then Compute!; x-out pop-up; scroll up "Sample(Child)" column to obtain top row of table (3, 19, 2, 9, 14), children chosen for sample. Look to population box of tickets for second row of table, to identify number of decayed teeth for each child chosen in sample. Click Data, Save data, "1.3 Decayed Teeth SRS".)

   Sample average of decayed teeth per child for five children is
   (observed) ave $= \frac{2+10+1+1+3}{5} = \frac{17}{5} =$ (circle one) **1.8** / **2.0** / **2.1** / **3.4**.
   Sample average (3.4, in this case) is example of **statistic** / **parameter**.

   (b) *Population Average.* Average of *all* tickets in box represents (circle one)

   i. sample average number of decayed teeth per child,
   ii. population average number of decayed teeth per child,

and is equal to: ave $= \frac{0+1+2+\cdots+2}{20} \approx$ (circle one) **2.1** / **2.3** / **2.5**.
Population average (2.1, in this case) is example of **statistic** / **parameter**.

It is possible to calculate population average in this case because population is small with only 20 individuals. It is typically much more difficult, often practically impossible, to determine average of larger populations and so necessary to rely on sample average.

(c) *Comparing Sample and Population Averages.*
True / **False**. Value of sample average, 3.4, is a "poor" estimate of population average, 2.1, because 3.4 is "far from" to 2.1 in this case.

We will discuss what it means to say 3.4 is "far from" to 2.1 later in the course.

(d) *Frame.* Frame is (circle one)

i. all tickets in box.
ii. sample of five tickets chosen from box.

2. *Simple Random Sample: Decayed Teeth Again*

0	1	2	9	0	4	0	0	1	5
child 1	child 2	child 3	child 4	child 5	child 6	child 7	child 8	child 9	child 10
0	0	0	3	1	1	3	0	10	2
child 11	child 12	child 13	child 14	child 15	child 16	child 17	child 18	child 19	child 20

Estimate population average number of decayed teeth per child with a sample average calculated from a simple random sample (SRS) of five children.

(a) *Sample Average.* Use StatCrunch, seed 11 (rather than 7, used above), to draw another SRS of five tickets out population box of tickets.

child in SRS	1	18	______	16	______
number of decayed teeth	0	0	______	1	______

(StatCrunch: click Data, Sample; in dialog box, click on "children", Sample size: 5, then Seed: replace given number with 11, then Compute!; x-out pop-up; scroll up "Sample(Child)" column to obtain top row of table (1, 18, 20, 16, 19), children chosen for sample. Look to population box of tickets for second row of table, to identify number of decayed teeth for each child chosen in sample.)

Sample average of decayed teeth per child for five children is
(observed) ave $= \frac{0+0+2+1+10}{5} =$ (circle one) **0.4** / **2.0** / **2.6** / **3.0**.
Sample average (2.6, in this case) is example of **statistic** / **parameter**.

(b) *Population Average.* Population average number of decayed teeth per child, average of *all* tickets in box, **remains same** / **changes** because box is same as before and so ave $= \frac{0+1+2+\cdots+2}{20} \approx 2.1$.

(c) *Comparing Sample and Population Averages.*
True / **False**. Value of sample average, 2.6, is a "good" estimate of population average, 2.1, because 2.6 is "close" to 2.1 in this case.

In general, each time a SRS of tickets is drawn from box of tickets, observed *sample* average probably **changes / remains the same**, whereas *population* average **changes / remains the same**. Observed sample average **always equals / varies around** population average.

(d) Match columns.

terms	children's teeth example
(a) population	(A) number of decayed teeth per child of 20 children
(b) sample	(B) average number decayed teeth, among few children chosen
(c) statistic	(C) number of decayed teeth per child of few children chosen
(d) parameter	(D) average number decayed teeth, among 20 children

terms	(a)	(b)	(c)	(d)
example				

1.4 Other Effective Sampling Methods

Other sampling techniques include stratified sampling, cluster sampling and systematic sampling. *Stratified sampling* involves dividing population into strata and choosing a simple random sample (SRS) from each strata. *Cluster sampling* involves dividing population into clusters, choosing a subset of these clusters at random, and then using all items of selected clusters. *Systematic sampling* involves selecting every kth item from population, where first item is chosen at random.

Exercise 1.4 (Other Effective Sampling Methods)

1. *Stratified Random Sample: Jerseys.* Twenty–four football jerseys are arranged according to size in three strata, large, medium and small, and identified as being made of polyester (indicated by a "1") or of a polyester–cotton (indicated as a "0") blend, as given in following (tiny) population box model.

small →	1	1	0	0	1	1
	jersey 1	jersey 2	jersey 3	jersey 4	jersey 5	jersey 6
medium →	0	1	0	0	1	1
	jersey 01	jersey 02	jersey 03	jersey 04	jersey 05	jersey 06
medium →	1	1	1	0	0	0
	jersey 07	jersey 08	jersey 09	jersey 10	jersey 11	jersey 12
large →	1	1	1	1	1	1
	jersey 1	jersey 2	jersey 3	jersey 4	jersey 5	jersey 6

Estimate population proportion of polyester-made jerseys with observed sample proportion of jerseys using stratified sampling method.

(a) *Terminology.*
If "jersey size" is a *stratum*, there are **one** / **two** / **three** / **four** strata.
Second stratum has **twelve** / **thirteen** / **sixteen** / **eighteen** jerseys.
Each stratum is **homogenous** / **heterogeneous**.

(b) *Stratified Sample.*
Possible stratified sample of *eight* jerseys is (choose one *or more*)

i. one (1) small, two (2) medium and three (3) large jerseys,
ii. two (2) small, three (3) medium and three (3) large jerseys,
iii. two (2) small, four (4) medium and two (2) large jerseys,

where jerseys are chosen using SRSs from each strata.

(c) *Population Proportion.*
Proportion of *all* tickets in box with 1s represents (circle one)

i. sample proportion of polyester-made jerseys,
ii. population proportion of polyester-made jerseys,

and equal to: (expected) proportion $= \frac{1+1+0+\cdots+1}{24} \approx$ **0.67** / **0.77** / **0.89**.
Proportion (0.67, in this case) is example of **statistic** / **parameter**.

2. *Cluster Random Sample: Jerseys Again.*

small →	1 jersey 01	1 jersey 02	0 jersey 03	0 jersey 1	1 jersey 1	1 jersey 1
medium →	0 jersey 04	1 jersey 05	0 jersey 06	0 jersey 2	1 jersey 3	1 jersey 4
medium →	1 jersey 07	1 jersey 08	1 jersey 09	0 jersey 3	0 jersey 5	0 jersey 6
large →	1 jersey 10	1 jersey 11	1 jersey 12	1 jersey 4	1 jersey 7	1 jersey 8
	cluster 1			cluster 2	cluster 3	

Once again, estimate population proportion of polyester-made jerseys with observed sample proportion of jerseys but, this time, use cluster sampling method.

(a) *Terminology.*
There are **one** / **two** / **three** / **four** clusters.
Second cluster has **four** / **nine** / **twelve** / **eighteen** jerseys.
Each cluster is **homogenous** / **heterogeneous**

(b) *Cluster Sample.*
Possible cluster sample of *two* is (choose *one or more*!)

i. clusters 1 and 2,
ii. cluster 1,
iii. clusters 1, 2, and 3,

where all jerseys in chosen clusters are used in sample.
A cluster sample does not require *all* jerseys are chosen from selected clusters. Often a SRS is chosen from each selected cluster. Only one or two clusters can be chosen from three clusters in cluster sampling; if all three clusters are chosen, sampling method becomes stratified sampling.

(c) *Population Proportion.* Population proportion of polyester-made jerseys, proportion of *all* tickets in box with 1s, **remains same** / **changes** because box is same as before and so proportion $= \frac{1+1+0+\cdots+1}{24} \approx 0.67$.

(d) *Cluster sampling versus stratified sampling.*
In cluster sampling, a *subset* of all clusters are sampled from, whereas in stratified sampling, (circle one) **some** / **all** strata are sampled from.

3. *Systematic Sampling: Jerseys Again.* Once again, estimate population proportion of polyester-made jerseys with observed sample proportion of jerseys using systematic sampling method. Jerseys are re–numbered, as given in box model.

small →	1	1	0	0	1	1
	jersey 01	jersey 02	jersey 03	jersey 04	jersey 05	jersey 06
medium →	0	1	0	0	1	1
	jersey 07	jersey 08	jersey 09	jersey 10	jersey 11	jersey 12
medium →	1	1	1	0	0	0
	jersey 13	jersey 14	jersey 15	jersey 16	jersey 17	jersey 18
large →	1	1	1	1	1	1
	jersey 19	jersey 20	jersey 21	jersey 22	jersey 23	jersey 24

(a) *Systematic Sample and Proportion.*
After choosing number between 1 and 6 at random, 5, say, this number and every 4th number after this (5, 9, 13, 17, 21) is chosen from jersey box model. Sample proportion of polyester-made for five jerseys is
(observed) proportion $= \frac{1+0+1+0+1}{5} =$ (circle one) **0.4** / **0.6** / **0.7** / **0.8**.
Sample proportion (0.6, in this case) is example of **statistic** / **parameter**.

(b) *Another Systematic Sample.*
If starting number 4 is used, systematic sample is (choose one)

i. 1, 5, 9, 13, 17, 21
ii. 2, 6, 10, 14, 18, 22
iii. 3, 7, 11, 15, 19, 23
iv. 4, 8, 12, 16, 20, 24
v. 6, 10, 14, 18, 22

(c) *Population Proportion.* Population proportion of polyester-made jerseys, proportion of *all* tickets in box with 1s, **remains same** / **changes** because box is same as before and so proportion $= \frac{1+1+0+\cdots+1}{24} \approx 0.67$.

(d) If a systematic sample is started by a number chosen at random from 1 to 6, there are **four** / **five** / **six** possible systemic samples *of size 6*.

(e) *Systematic versus SRS.* Systematic sampling is **same as** / **different from** a simple random sampling (SRS) because it is possible to choose two tickets next to one another in an SRS but not so in a systematic sample.

(f) *Systematic, Cluster and Stratified.* Systematic sampling is special case of **cluster sampling** / **stratified sampling** because both result in a choosing a subset from all possible clusters in population.

4. *Simple, Stratified, Cluster or Systematic?* Match racing horses examples with sampling technique. Recall "SRS" means "simple random sample".

(a) **Simple** / **Stratified** / **Cluster** / **Systematic**
An SRS is taken from all racing horses.

(b) **Simple** / **Stratified** / **Cluster** / **Systematic**
All racing horses are listed from lightest to heaviest. Sample consists of taking every seventh racing horse from this list.

(c) **Simple** / **Stratified** / **Cluster** / **Systematic**
All racing horses are listed alphabetically, by name. Sample consists of taking every third racing horse from this list.

(d) **Simple** / **Stratified** / **Cluster** / **Systematic**
All racing horses are classified as light, middle or heavy weight horses. Sample consists of taking SRSs from the racing horses in each weight class.

(e) **Simple** / **Stratified** / **Cluster** / **Systematic**
Horses racing occurs in many cities in the U.S.A. Sample consists of taking SRSs from the racing horses in New York, Los Angeles and Chicago.

5. *Convenience Sampling.*
This sampling procedure chooses, in a non-random way, from only easy-to-access part of population. This is a poor sampling technique. Sampling only jerseys at top of a box of jerseys **is** / **is not** a case of convenience sampling.

1.5 Bias in Sampling

Under repeated sampling, value of statistic typically *varies* around parameter value. *Bias* is difference between average value of statistic and parameter. A good sampling method has both small bias and small variability (*(random) sampling error*). Bias is reduced by using random sampling. Simple random sampling (SRS) gives an *unbiased* estimate: average value of statistic equals parameter under repeated sampling. Variability (sampling error) in SRS reduced by taking large samples.

Bias can occur as a direct consequence of a poor sampling method. In particular, *sampling bias* occurs if items in one part of population are favored over other parts of population and results from a non-SRS (such as convenience or voluntary response) sampling method.

Bias also results from errors not related to act of selecting sampling. *Nonresponse bias* occurs if individuals selected for a sample choose to not participate, to not respond to a survey in particular. *Response bias* occurs if participant answers do not reflect true feelings.

Exercise 1.5 (Bias in Sampling)

1. *Sampling Error and Bias: Darts* Describe dart hit patterns on following four dart boards.

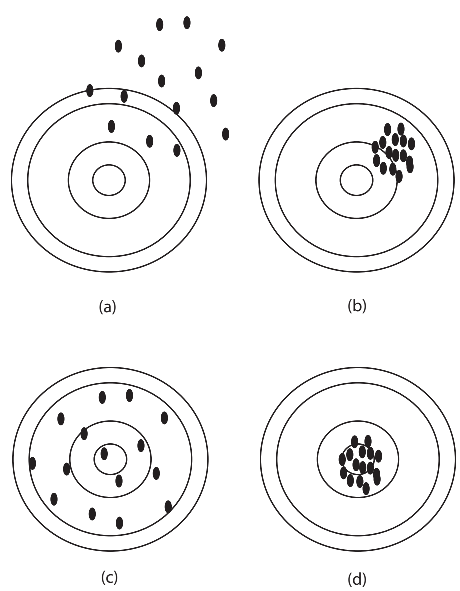

Figure 1.3 (Sampling Error and Bias)

(a) **True** / **False** Dart hits on board (a) have *large sampling error* because they are spread out and *large bias* because they are thrown in wrong place, up and away from bull's eye.

(b) Dart hits on board (b): Clustered together but thrown in wrong place, so **small** / **large** sampling error and **small** / **large** bias

(c) Dart hits on board (c): Spread out but thrown in right place, so **small** / **large** sampling error and **small** / **large** bias

(d) Dart hits on board (d): Clustered together and thrown in right place, so **small** / **large** sampling error and **small** / **large** bias

2. *More Sampling Error and Bias: Telephone Survey.* It is known 68% of registered voters in Berrien County, Michigan are registered as Democrats. To test a new telephone sampling method, we call 500 Berrien County voters and ask their party. We do this 5 times. Results are 59.2%, 58.9%, 60.5%, 57.4% and 61.3% Democratic. Sampling method *appears* to have:

   (i) large sampling error and large bias

   (ii) large sampling error and small bias

   (iii) small sampling error and large bias

   (iv) small sampling error and small bias

3. *Type of Bias: Sampling, Nonresponse or Response?*
   Match examples with type of bias.

   (a) **sampling bias** / **nonresponse bias** / **response bias**
   Only females included to determine average height of "typical" student.

   (b) **sampling bias** / **nonresponse bias** / **response bias**
   Only 32 of 130 mail surveys returned.

   (c) **sampling bias** / **nonresponse bias** / **response bias**
   The wording, "Given that you are a humane and caring individual, are you in favor of capital punishment?" might tend to give a different response than, "Given the despicable nature of today's criminal, are you in favor of capital punishment?"

   (d) **sampling bias** / **nonresponse bias** / **response bias**
   Survey conducted by email only.

   (e) **sampling bias** / **nonresponse bias** / **response bias**
   Interviewer makes interviewee feel uncomfortable.

1.6 The Design of Experiments

Unlike observational studies, designed experiments identify cause and effect between variables in a study. Three types of designed experiments are described. *Completely randomized design* investigates how one explanatory variable (with two or more levels or treatments) influences a response variable. *Randomized block design* also investigates how one explanatory variable influences a response variable, but also, to improve precision of statistical inference, groups experimental units into homogeneous blocks. An important special case of randomized block design is *matched-pair design* where there are only two experimental units per block.

Exercise 1.6 (The Design of Experiments)

1. *Terminology in Experimental Designs.*

   (a) *Effect of Temperature on Mice ROC.*

temperature →	70^o F	-10^o F
	10.3	9.7
	14.0	11.2
	15.2	10.3

i. Explanatory variable or, equivalently, factor is (choose one)
   A. temperature
   B. mice
   C. room temperature, 70^o F
   D. 70^o F and -10^o F

ii. Two treatments or, equivalently, two levels of factor are (choose one)
   A. temperature
   B. mice
   C. room temperature, 70^o F
   D. 70^o F and -10^o F

iii. Experimental units are (choose one)
   A. temperature
   B. mice
   C. room temperature, 70^o F
   D. 70^o F and -10^o F

iv. Control, or "do-nothing" treatment is (choose one)
   A. temperature
   B. mice
   C. room temperature, 70^o F
   D. 70^o F and -10^o F

   Although it makes sense to designate room temperature as control, it is possible to designate cold temperature as control. Control, then, can actually be either treatment. Also, if only two treatments and one treatment is control, the other is, confusingly, referred to as "treatment". So, if "room temperature" is control, "treatment" is "cold temperature".

v. Number of replications (mice per treatment) is **2** / **3** / **4**.

(b) *Effect of drug on patient response.*

drug →	A	B	C
	120	97	134
	140	112	142
	125	100	129
	133	95	137

i. Explanatory variable or, equivalently, factor is (choose one)
   A. drug
   B. patients
   C. drug A
   D. drug A, B and C

ii. Three treatments or, equivalently, three levels of factor are
   A. drug
   B. patients
   C. drug A
   D. drug A, B and C

iii. Experimental units or, better, subjects are (choose one)
   A. drug
   B. patients
   C. drug A
   D. drug A, B and C

iv. Number of replications (patients per treatment) is **2** / **3** / **4**.

v. If drug A is a *placebo*, a drug with no medical properties, control is
   A. drug
   B. patients
   C. drug A
   D. drug A, B and C

vi. **True** / **False**.
   This experiment is *single-blind* if patients do not know which drug is assigned to which patient and *double-blind* if both patients and experimenters do not know which drug is assigned to which patient, until after experiment is over.

2. *Experimental designs: completely randomized, randomized block, matched-pair.* Consider experiment to determine effect of temperature on mice ROC.

   (a) *Completely Randomized Design.*

temperature →	70^o F	-10^o F
	10.3	9.7
	14.0	11.2
	15.2	10.3

This is a completely randomized design with (choose *one or more*!)

i. one factor (temperature) with two treatments,
ii. three pair of mice matched by age,
iii. mice assigned to temperatures at random.

(b) *Randomized Block Design.*

age ↓ temperature →	70^o F	-10^o F
10 days	10.3	9.7
20 days	14.0	11.2
30 days	15.2	10.3

This is a randomized block design with (choose *one or more*!)

i. one factor (temperature) with two treatments,
ii. one block (age) with three levels,
iii. mice assigned to temperatures within each age block at random.

(c) *Matched-Pair Design.*

age ↓ temperature →	70^o F	-10^o F
10 days	10.3	9.7
20 days	14.0	11.2
30 days	15.2	10.3

This is a matched-pair design with (choose *one or more*!)

i. one factor (temperature) with two treatments,
ii. three pair of mice matched by age,
iii. mice assigned to temperatures within each age at random.

3. *Experimental designs: completely randomized, randomized block, matched-pair?* Consider experiment to determine effect of drug on patient response.

(a) *Study A.*

drug →	A	B	C
	120	97	134
	140	112	142
	125	100	129
	133	95	137

Since this experiment consists of one factor with three levels, no blocks and where patients are assigned to drugs at random, this is a (choose one)

i. completely randomized design
ii. randomized block design
iii. matched-pair design

(b) *Study B.*

age↓ drug →	A	B	C
20-25 years	120	97	134
25-30 years	140	112	142
30-35 years	125	100	129
35-40 years	133	95	137

Since this experiment consists of one factor (drug) with three levels, one block (age) with four levels and where patients are assigned to drugs within common ages at random, this is a (choose one)

i. completely randomized design
ii. randomized block design
iii. matched-pair design

Although a randomized block design, it is not also a matched-pair design because there are three, not two, levels of the drug factor.