Lecture Notes

For Statistics 130

Statistics and Contemporary Life

Spring 2020

by

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

Preface

A major effort is made in this course is to explain statistics in everyday language. It is hoped a student is able to get at the *concepts and ideas* of statistics without bogging down in the technicalities of the subject.

In spite of the significant attempt to make this course as user–friendly as is possible, to keep the *mathematical notation* at a minimum, the student should keep in mind the following.

- There is a substantial amount of numerical manipulation. In other words, there is a lot of adding, subtracting, multiplying, dividing and finding the square roots of lists of data.
- Although "everyday language" replaces much of the mathematical notation, the mathematics does not go away–it lurks just beneath the surface of the description and graphs.

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.

Dr. Jonathan Kuhn,
Associate Professor of Statistics,
Purdue University Northwest
November 2020.

Chapter 1

Where Do Data Come From?

An *individual* is a person, object or entity. A *variable* is a characteristic of an individual. A *data point* is a particular instance of a variable. Data may be *numerical* or not.

Data is collected in two basic ways: experiment or observational study. In an *experiment*, the *experimenter* decides who is to be given the *treatment* and who is to be the *control*, to attempt to establish a *cause* and *effect* between treatment and *response*. In an *observational study*, it is the *individual* who decides whether or not to be given the treatment, where the researcher tries to observe only, to gather data without influencing the individuals one way or the other. A *sample survey* is a special kind of observational study where a survey is used to collect information on a representative sample of individuals from a larger group. A *census* is a sample survey of *all* individuals in the population.

Both methods (experiment, observational study) use data collected, the *sample*, to infer something about the *population* from which the sample was collected.

Exercise 1.1 (Where Do Data Come From?)

1. *Milk study.* Consider table below which contains various measurements on a number of cows taken during a study on effect of a hormone, given in tablet form, on daily milk yield.

Cow ID	Test Date	Farm	Height	Health	Tablets	Before Yield	After Yield
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
17	9/11/98	M	41	poor	2	100.7	100.3
18	9/11/98	F	40	bad	1	97.8	98.1
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) Individuals in study are (choose one)
hormones / **daily milk yields** / **cow ID** / **cows**

(b) Variables in this study are (choose one)

i. 14, 15, 16, 17, 18, 19, 20

ii. Cow ID, Test Date, Farm, Height, Health, Tablets, Before Yield, After Yield

(c) Data for variable *Cow ID* is (choose one)

i. 14, 15, 16, 17, 18, 19, 20

ii. F, M, F, M, F, M, M

iii. 3, 3, 1, 2, 1, 2, 3

(d) Data for variable (name of) *Farm* is (choose one)

i. 14, 15, 16, 17, 18, 19, 20

ii. F, M, F, M, F, M, M

iii. 3, 3, 1, 2, 1, 2, 3

(e) Data for variable (number of) *Tablets* is (choose one)

i. 14, 15, 16, 17, 18, 19, 20

ii. F, M, F, M, F, M, M

iii. 3, 3, 1, 2, 1, 2, 3

(f) Data for (number of) *Tablets* (circle one) **is** / **is not** numerical.

(g) Data for (name of) *Farm* (circle one) **is** / **is not** numerical.

(h) Data for (name of) *Cow ID* (circle one) **is** / **is not** numerical.

2. *More individuals, variables and data.*

(a) **True** / **False** Data point for variable *height of a man* is 5.6 feet tall. Another data point for variable is 5.8 feet tall. Individual is a man.

(b) **True** / **False** Data point for the variable *person's country of birth* is Sweden. Another data point for this variable is U.S.A. Individual is person.

(c) **True** / **False** Data point for variable *shoulder height of a cow* is 45 inches. Another data point for variable is "Argentina". Individual is a cow.

(d) Data point for variable *person's length of time of exposure in the sun* is (circle best one) **9/24/98** / **4 hours**. Individual is a person.

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

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

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

(h) Data point “12” is a particular instance of variable (circle one or more)

i. number of eggs in a bowl
ii. number of strikes of a grandfather clock
iii. size of a dress
iv. length, in feet, of a fence

(i) Data point “silver” is a particular instance of variable (circle 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

(j) Match variable with individual.

variable	individual
(a) number of eggs in a bowl	**(A)** clock
(b) number of strikes of a grandfather clock	**(B)** fence
(c) size of a dress	**(C)** bowl
(d) length, in feet, of a fence	**(D)** dress

variable	(a)	(b)	(c)	(d)
individual				

3. *Mice ROC and temperature.* Effect of air temperature on rate of oxygen consumption (ROC) of four mice is investigated. ROC of one mouse at room temperature is 10.3 units for example.

temperature	room temperature (70^o F)	10.3	14.0
	cold temperature (-10^o F)	9.7	11.2

(a) Since experimenter (not a mouse!) decides which mouse is subjected to room temperature and which mouse is subjected to cold temperature, this is an (choose one)
observational study / **sample survey** / **experiment**.

(b) Purpose of study is to determine effect of temperature on response (choose one) **ROC** / **mice** / **temperature**

(c) Population is (choose one)
temperature / **ROC** / **all mice** / **four mice in study**

(d) Sample is (choose one)
temperature / ROC / all mice / four mice in study

4. *Traffic accidents and drinking.* 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 drinker	3	6	2
	light drinker	1	2	1

   (a) This is an observed study because (circle one)

      i. investigator decides who is going to drink
      and drive and who is not.

      ii. drivers decided who is going to drink and drive
      and who is not.

      iii. drivers are assigned to drink and drive or
      be sober drivers at random.

   (b) Purpose of study is to determine effect of drinking on response
   (choose one) **driving / driving accidents / police**

   (c) Population is (choose one)
   drinkers / drivers / Indiana drivers / six drivers in study

   (d) Sample is (choose one)
   drinkers / drivers / Indiana drivers / six drivers in study

5. *Average distance and commute distances.* At PNW, 120 students are randomly surveyed from entire 15,000 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.

   (a) This is a(n)
   (choose one) **observational study / sample survey / experiment**.

   (b) Purpose of study is to determine effect of commute distances on response
   (choose one) **students / PNW / average commute distance**

   (c) Population is (choose one)

      i. 120 students in study

      ii. all 15,000 PNW students

      iii. 9.8 miles

      iv. average commute distance

   (d) Sample is (choose one)

i. 120 students in study
ii. all 15,000 PNW students
iii. 9.8 miles
iv. average commute distance

(e) Census is survey of (choose one)

i. 120 students in study
ii. all 15,000 PNW students
iii. 9.8 miles
iv. average commute distance

(f) **True** / **False**. An appropriate analogy here would be to think of a box of 15,000 tickets where each ticket has commute distance written on it as *population.* A random sample of 120 tickets taken from this population box would be a *sample.*

"Population" may also refer to 4,500 tickets themselves, whatever is written on them.

6. *Academic achievement and instruction method.* A recent study was conducted to compare the academic achievement (measured by final examination scores) of 25 PNW online students with 28 PNW classroom students. Online students did not attend class, but received all instruction over the Internet, whereas classroom students received instruction at a fixed time every week in a specified classroom.

(a) This is an observed study because (circle one)

i. researcher decides who is a classroom student
and who is an online student.
ii. students decide to be either a classroom student or an online student.
iii. students assigned to be classroom students or online students at random.

(b) Purpose of study is to determine effect of instruction method on response (choose one) **online student** / **academic achievement** / **instructor**

(c) Population is (choose one)

i. online students
ii. PNW students
iii. students in study
iv. all students

(d) Sample is (choose one)

i. online students
ii. PNW students

iii. students in study

iv. all students

(e) **True** / **False**. In “tickets in box” analogy, 53 tickets are sampled from a box of 15,000 tickets where each ticket has “online” or “classroom” on it.

Hopefully, tickets are sampled at *random* to ensure they are representative of all tickets.

7. *Grain Yield and fertilizer.* Effect of fertilizer on grain ABX yield is investigated. Yield from one plot of grain ABX given fertilizer A is 120 kilograms for example.

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

(a) This is an experimental design because (circle one)

i. researcher decided which plot is assigned three types of fertilizer.

ii. grain plants decided which plot is assigned three types of fertilizer.

iii. plots of grain plants are assigned three types of fertilizer at random.

(b) Purpose of study is to determine effect of type of fertilizer on response (choose one) **grain plants** / **plot** / **grain ABX yield**

(c) Population is (choose one)
all plots / **all plots of grain ABX** / **plots of grain ABX in study**

(d) Sample is (choose one)
all plots / **all plots of grain ABX** / **plots of grain ABX in study**

Chapter 2

Samples, Good and Bad

On the one hand, a statistical study is *biased* if one individual in a population is systematically favored over another, often resulting in a sample which is not representative of the population. For example, bias often results from *convenience sampling* when easiest-to-reach individuals are collected. Also, bias may also occur from *voluntary response sampling* when individuals, rather than researchers, decide who is to be in the sample or not.

A *simple random sample* (SRS), on the other hand, involves selecting n units out of N population units where every distinct sample has an *equal* chance of being drawn. A SRS is collected by first assigning every individual in the population a number, then using random digits to select a sample of individuals (based on their assigned numbers) from the population. So, theoretically, a SRS produces a *representative* sample of a population, is *unbiased* (has no or zero bias); however, as will be discussed in future chapters, for "real" situations with imperfect collection procedures and "messy" data there will always be some bias even when a SRS is collected.

Exercise 2.1 (Samples, Good and Bad)

1. *Convenience or voluntary response sampling?*
   Match examples with type of (bad) sampling methods.

   (a) **convenience sampling / voluntary response sampling**
   Sample consists of radio listeners' phoned-in opinions.

   (b) **convenience sampling / voluntary response sampling**
   Sample consists of apples taken at top of truck load.

   (c) **convenience sampling / voluntary response sampling**
   Sample consists of choosing first 100 shoppers who enter a mall.

   (d) **convenience sampling / voluntary response sampling**
   Sample consists of responses to email survey.

   (e) **convenience sampling / voluntary response sampling**
   Sample consists of asking friends to fill in a survey.

2. *A First Look At Random Numbers Table.* The *first two lines* of the *random numbers table* is given below.

rows ↓								
101	19223	95034	05756	28713	96409	12531	42544	82853
102	73676	47150	99400	01927	27754	42648	82425	36290

Have you got your random numbers table out? If not, pull it out. A copy is given in blackboard. Numbers are presented in groups of five simply to help readability of table.

(a) Number (digit) in first row, first column is digit (circle one) **0** / **1** / **2** / **3**.

(b) One-digit number in first row, *sixth* column is **3** / **4** / **8** / **9**.

Numbers are presented in groups of five to help readability of table, so sixth column is first number in second group of five.

(c) One-digit number in *second* row, *fourth* column is **5** / **6** / **7** / **8**.

(d) *Two*-digit number *first* row, *first and second* column: **11** / **12** / **18** / **19**.

(e) *Three*-digit number in *first* row, and beginning in fifth column, is (choose one) **395** / **661** / **866** / **002**.

(f) Since this is a *random* numbers table and there are *ten* one digit numbers (0,1,...,9), if you close your eyes and place your finger on the random numbers table at random, the chance you place your finger on the number 2 is (circle one) **10%** / **20%** / **30%** / **40%**.

(g) If you close your eyes and place your finger on the random numbers table, the chance you place your finger on the numbers 1 *or* 2 is (circle one) **10%** / **20%** / **30%** / **40%**.

(h) Since this is a *random* numbers table and there are 100 *two* digit numbers (00,01,...,99), if you close your eyes and place your finger on the random numbers table, the chance you place your finger on the number 01 is (circle one) **1%** / **2%** / **3%** / **4%**.

(i) If you close your eyes and place your finger on the random numbers table, the chance you place your finger on the numbers 01 or 02 is (circle one) **1%** / **2%** / **3%** / **4%**.

(j) Suppose you were asked to search through a random numbers table, along one row, systematically looking at one number after another. You have a choice between finding the two numbers 1 or 2, or finding the two numbers 01 or 02. If you wanted to perform this search in the *shortest* possible time, you would search for (circle one)

i. 1 or 2

ii. 01 or 02

(k) **True / False** There are many different ways of reading a random numbers table. It is possible to read along a row, either left or right or right to left; or to read down a column, top to bottom or bottom to top; or to read along a diagonal and so on and still produce a string of random digits.

(l) **True / False** The random numbers table given in this course is the *one* (and only) possible random numbers table possible. There are no other random number tables where each one digit number appears at any location in the table at random.

There are online sites such as www.randomizer.org which generate an SRS. We will stick with the random numbers table in this course because it is possible to test this material using the table, but not the online site, to ensure everyone in class gets the same SRS on a test of random numbers.

3. *Simple Random Sample (SRS): Mice ROC and temperature.* Effect of air temperature on rate of oxygen consumption (ROC) of four mice is investigated.

temperature	room temperature (70^o F)	10.3	14.0
	cold temperature (-10^o F)	9.7	11.2

(a) A simple random sample (SRS) of 4 mice is chosen from 9 mice, numbered 1, 2, 3, 4, 5, 6, 7, 8, 9. Use random numbers table to collect SRS, starting first row, first column, reading left to right

i. 1, 9, 2, 2

ii. 7, 3, 6, 7

iii. 1, 9, 2, 3

iv. 7, 3, 6, 4

No repeated digits are allowed to occur in a SRS such as the two 2s in 1, 9, 2, 2 and so skip over, ignore, the second 2, then move to the next digit in your search.

(b) A simple random sample (SRS) of 4 mice is chosen from 9 mice, numbered 1, 2, 3, 4, 5, 6, 7, 8, 9. Use random numbers table to collect SRS, starting *second* row, first column, reading left to right.

i. 1, 9, 2, 2

ii. 7, 3, 6, 7

iii. 1, 9, 2, 3

iv. 7, 3, 6, 4

(c) A simple random sample (SRS) of 4 mice is chosen from 99 mice, numbered 01, 02, 03, . . . , 98, 99. Use random numbers table to collect SRS, starting first row, first column, reading left to right.

i. 73, 67, 64, 71

ii. 18, 22, 39, 50

iii. 19, 22, 39, 50

iv. 19, 23, 39, 50

4. *Simple Random Sample: Children's Decayed Teeth*
Number of decayed teeth for 20 children is represented by box of tickets below. Child 17, Darlene, has 3 decayed teeth for example.

$$\underbrace{\boxed{0}}_{\text{01: Sally}} \quad \underbrace{\boxed{1}}_{\text{02: Jon}} \quad \underbrace{\boxed{2}}_{\text{03: Tim}} \quad \underbrace{\boxed{9}}_{\text{04: Michelle}} \quad \underbrace{\boxed{0}}_{\text{05: Tyler}} \quad \underbrace{\boxed{4}}_{\text{06: Benjamin}} \quad \underbrace{\boxed{0}}_{\text{07: Andrew}} \quad \underbrace{\boxed{0}}_{\text{08: Violet}} \quad \underbrace{\boxed{1}}_{\text{09: Jennifer}} \quad \underbrace{\boxed{5}}_{\text{10: Betsy}}$$

$$\underbrace{\boxed{0}}_{\text{11: Patricia}} \quad \underbrace{\boxed{0}}_{\text{12: Chloe}} \quad \underbrace{\boxed{0}}_{\text{13: Paul}} \quad \underbrace{\boxed{3}}_{\text{14: Joseph}} \quad \underbrace{\boxed{1}}_{\text{15: Vijay}} \quad \underbrace{\boxed{1}}_{\text{16: Michael}} \quad \underbrace{\boxed{3}}_{\text{17: Darlene}} \quad \underbrace{\boxed{0}}_{\text{18: Samuel}} \quad \underbrace{\boxed{10}}_{\text{19: Marsha}} \quad \underbrace{\boxed{2}}_{\text{20: Thomas}}$$

(a) Use random numbers table to collect SRS of three children, starting first row, first column, reading left to right.

i. 19 (Marsha), 22 (Unknown), 39 (Unknown)

ii. 19 (Marsha), 05 (Michelle), 13 (Paul)

iii. 19 (Marsha), 05 (Tyler), 13 (Paul)

iv. 19 (Marsha), 05 (Tyler), 14 (Joseph)

If you come across a pair of two-digit numbers you cannot use (larger than 20 or repeated), skip to next two-digit number.

(b) Use random numbers table to collect SRS of three children, starting *second* row, first column, reading left to right.

i. 19 (Marsha), 17 (Darlene), 09 (Jennifer)

ii. 73 (Unknown), 67 (Unknown), 64 (Unknown)

iii. 19 (Marsha), 07 (Andrew), 13 (Paul)

iv. 19 (Marsha), 05 (Tyler), 13 (Paul)

5. *Choosing three names.*

George Frederick Marsha April May June

(a) Use random numbers table to collect SRS of three names, starting first row, first column, reading left to right.

i. 1 (George), 9 (Unknown), 2 (Frederick)

ii. 1 (George), 2 (Frederick), 2 (Frederick)

iii. 1 (George), 2 (Frederick), 3 (Marsha)

iv. 1 (George), 2 (Frederick), 5 (May)

There are different ways of numbering the six names, but one possibility is to number them 1, 2, 3, 4, 5 and 6; another possibility is 01, 02, 03, 04, 05 and 06; third possibility is 0, 1, 2, 3, 4 and 5

(b) Use random numbers table to collect SRS of three names, starting *second* row, first column, reading left to right.

i. 7 (Unknown), 3 (Marsha), 6 (June)
ii. 3 (Marsha), 3 (Marsha), 6 (June)
iii. 3 (Marsha), 6 (June), 4 (May)
iv. 3 (Marsha), 6 (June), 4 (April)