Lecture Notes
For Statistics 311

Introductory Probability

Fall 2016

by

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

Preface

The point of this course is to introduce the mathematical theory of probability. Some familiarity with both differentiation and integration is a necessary prerequisite to this course.

These lecture notes are a necessary component for a student to successfully complete this course. Without the lecture notes, a student will not be able to participate in the course. Without the text, a student will not have complete information about the course material. More than this,

- The lecture notes are *based* on the text.
- Although the material covered in each is very similar, the *presentation* of the material in the workbook 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, consists *solely* of exercises to be worked out by the student.
- The overheads presented during each lecture are based *exclusively* on the lecture notes. A student is to use these lecture notes to follow along with during a lecture.
- The lecture notes have a number of fill–in–the–blank, multiple choice, true/false and other kinds of interactive exercises which a student completes during lecture time. Rather than spend most time writing down what is given on the overhead during the lecture, a student can simply fill in the lecture notes.
- Many of the exercises given in the lecture notes are based on the text. A student should try the exercises out of both the lecture notes and text.

On the one hand, these 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. In short, these lecture notes aspire 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 North Central
July 2016.

Chapter 1

What Is Statistics?

In this chapter, we briefly introduce the field of statistics, as well as the important role that probability plays in this field.

1.1 Introduction

Statistics is to do with the idea of gathering together a sample, calculating a statistic and using this statistic to not only infer something about a parameter of the population from which this sample is taken but also to indicate how "good" this inference is[1].

Exercise 1.1 (Population, Sample, Statistic and Parameter)

1. *Proportion Of Democrats.* Since 345 of one thousand randomly chosen Americans are Democrats, we can infer that approximately $\frac{345}{1000}$ths or 34.5% of **all** Americans are Democrats. Assume the political preferences of Americans are either Democratic, Republican or Independent.

   (a) The *variable* of interest is, in this case (choose one)

      i. height of an American.
      ii. marital status of an American.
      iii. political preference of an American.
      iv. Republican, political preference of a particular American, "Susan", say.
      v. {Democrat, Democrat, Republican, Independent, ..., Republican}, the set of political preferences for the one thousand randomly selected Americans.

[1] Although the text talks about the notions of "population" and "sample" in this chapter, it leaves the discussion about "parameter" and "statistic" until a later chapter.

(b) A possible *value* of the variable of *interest* is (choose one)

i. 6 feet tall.

ii. a particular American, "Susan", say.

iii. political preference of an American.

iv. Republican.

v. {Democrat, Democrat, Republican, Independent, ..., Republican}, the set of political preferences for the one thousand randomly selected Americans.

(c) The *population* is (choose one)

i. *all* Americans.

ii. political preferences of *all* Americans.

iii. the one thousand Americans, selected at random.

iv. political preferences of the one thousand Americans, selected at random.

(d) The population of the political preferences of all Americans is an example of a (choose one)

i. *real* finite population.

ii. *conceptual* population.

(e) The *sample* is (choose one)

i. *all* Americans.

ii. political preferences of *all* Americans.

iii. the one thousand Americans, selected at random.

iv. political preferences of the one thousand Americans, selected at random.

(f) Although, loosely speaking, the population is "all Americans" and the sample is "the 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, the population is "political preferences of all Americans" and the sample is "political preferences of one thousand Americans".
(i) True (ii) False

(g) Both a *statistic* and a *parameter* are numerical values, but a *statistic* summarizes the *sample* in some way, whereas the *parameter* summarizes the *population* in some way. In this case, the statistic of *interest* is,

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

ii. proportion of Democrats, among the one thousand randomly chosen Americans.

(h) The value of the statistic of interest is (choose one)
(i) 14.5% (ii) 24.5% (iii) 34.5%

(i) The parameter of *interest* is,

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

ii. proportion of Democrats, among the one thousand randomly chosen Americans.

(j) Statistical inference involves using the known value of the statistic, 34.5%, to estimate the unknown value of the parameter: proportion of Democrats, among all Americans.
(i) True (ii) False

(k) Probability plays an important role in statistical inference. For example, we might answer how "well" 34.5% estimates the population proportion of Democrats by saying we are 95% "confident" that the population proportion is given by 34.5% ± 1.5%.
(i) True (ii) False

2. *Distance To Travel.* At Purdue North Central, 120 students from all students currently enrolled are randomly selected and asked the distance of their commute to campus. From this group, an average of 9.8 miles is computed. Match terms with travel example. (*All* of the items in the first column will be used up in the matching procedure; however, two items in the second column will be left unmatched.)

terms	travel example
(a) variable	**(a)** average commute distance for 120 students
(b) value of variable	**(b)** all students at PNC
(c) parameter	**(c)** commute distances for all students at PNC
(d) population	**(d)** commute distance
(e) sample	**(e)** average commute distance for all students
(f) statistic	**(f)** 120 students
	(g) 120 commute distances
	(h) 8 mile commute

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

(a) The population of the current commute distances for all students at PNC is an example of a (choose one)

i. real finite population.

ii. conceptual population.

(b) If we were interested not only in the current commute distances but also the commute distances for the next three years, then we have an example of a (choose one)

i. real finite population.

ii. conceptual population.

(c) Probability plays an important role in this example. We might answer how "well" 9.8 miles estimates the population average commute distance for all students at PNC by saying we are 95% "confident" that the population average is given by 9.8 ± 2.5 miles.
(i) True (ii) False

1.2 Characterizing a Set of Measurements: Graphical Methods

We continue our brief introduction to statistics by looking at *relative frequency histograms*, which are often used to describe samples, and *relative frequency distributions*, which are probability models for the population.

Exercise 1.2 (Characterizing a Set of Measurements: Graphical Methods)

1. *Freshwater bryozoans.*

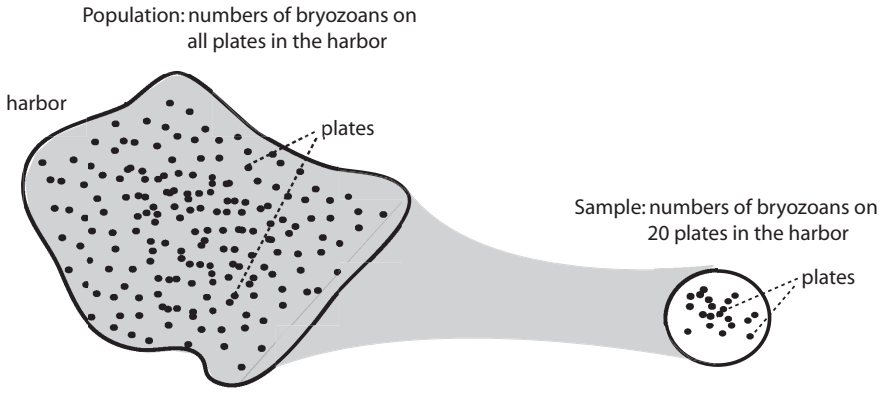

Figure 1.1: Freshwater bryozoans: sample from a population

Twenty one-meter-by-one-meter square plastic plates are attached at varying depths on randomly chosen dock posts in one harbor in Lake Ontario. The number of bryozoans on each of the twenty plates is given in Table 1.1. A relative frequency histogram is given in Figure 1.2.

(a) Table 1.1 is an example of a (choose one)
(i) sample (ii) statistic (iii) population (iv) parameter

plate	number	plate	number
1	82	11	51
2	71	12	81
3	84	13	76
4	67	14	66
5	67	15	59
6	89	16	86
7	64	17	76
8	68	18	87
9	64	19	68
10	70	20	69

Table 1.1: Numbers of bryozoans on twenty plates

(b) The observed number of bryozoans on plate 15 is y_{15} = (choose one)
(i) 58 (ii) 59 (iii) 76 (iv) 77

(c) The y_i ("little y–i") is a (choose one)
(i) value (ii) realization (iii) observation (iv) all of the above
of the number of bryozoans on plate i.

(d) The y is a (choose one)
(i) value (ii) realization (iii) observation (iv) all of the above
of the number of bryozoans on a plate, without specifying the plate.

(e) According to the relative frequency histogram in Figure 1.2[2], the propor-

[2]Use your calculator:

- Remember to turn off all the STAT PLOTs and Y = plots, and type the number of bryozoans per plate into L_1 list: STAT ENTER 82 ENTER 71 ENTER ... 68 ENTER 69 ENTER.
- To display the relative frequency histogram, press,
  - 2nd STAT PLOT ENTER
  - On ENTER
  - Type: histogram figure first row, far right ENTER
  - Xlist: L1 (for L_1 values) ENTER
  - Freq: 1

  and then hitting ZOOM ZoomStat. The TRACE key can be used to see the upper and lower bound of each class as well as the number in each class of the histogram.
- "Clean" up the histogram by pressing WINDOW, then
  - Xmin=50
  - Xmax=90
  - Xscl=5
  - Ymin=-2

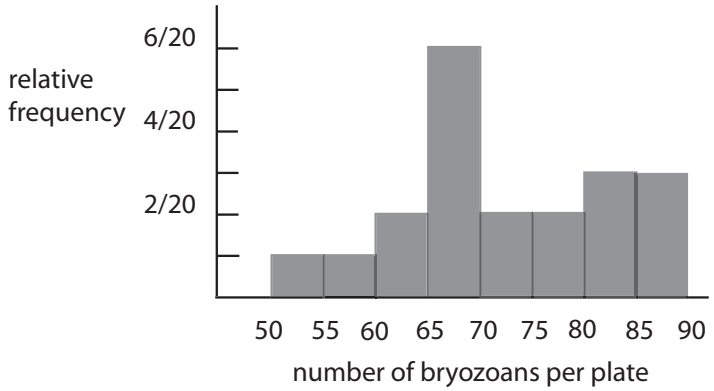

Figure 1.2: Relative frequency histogram: freshwater bryozoans

tion of plates with fewer than 65 bryozoans per plate is (choose one)
(i) $\frac{1}{20}$ (ii) $\frac{2}{20}$ (iii) $\frac{3}{20}$ (iv) $\frac{4}{20}$

(f) According to Figure 1.2, the proportion of plates with 75 or more bryozoans per plate is (choose one)
(i) $\frac{6}{20}$ (ii) $\frac{7}{20}$ (iii) $\frac{8}{20}$ (iv) $\frac{9}{20}$

(g) According to Figure 1.2, the proportion of plates with between 65 and 75 bryozoans per plate is (choose one)
(i) $\frac{6}{20}$ (ii) $\frac{7}{20}$ (iii) $\frac{8}{20}$ (iv) $\frac{9}{20}$

(h) Two *guesses* for the *relative frequency distribution* of the *population* of the number of bryozoans per plate in the harbor are given in Figure 1.3. These are the *only* possible population probability models.
(i) True (ii) False

(i) Population distribution 1 in Figure 1.3 tells us the number of bryozoans per plate ranges from (choose one)
(i) 50 to 90 (ii) 63 to 82 (iii) 67 to 75 (iv) 50 to 82
whereas the sample histogram has a range 50 to 90.

(j) Population distribution 2 in Figure 1.3 tells us the number of bryozoans per plate ranges from less than 63 to more than 82 which does correspond better than population distribution 1 to the sample histogram.
(i) True (ii) False

(k) Population distribution 2 in Figure 1.3 is bell–shaped whereas the sample

– Ymax=6
– Yscl=1
– Xres=1

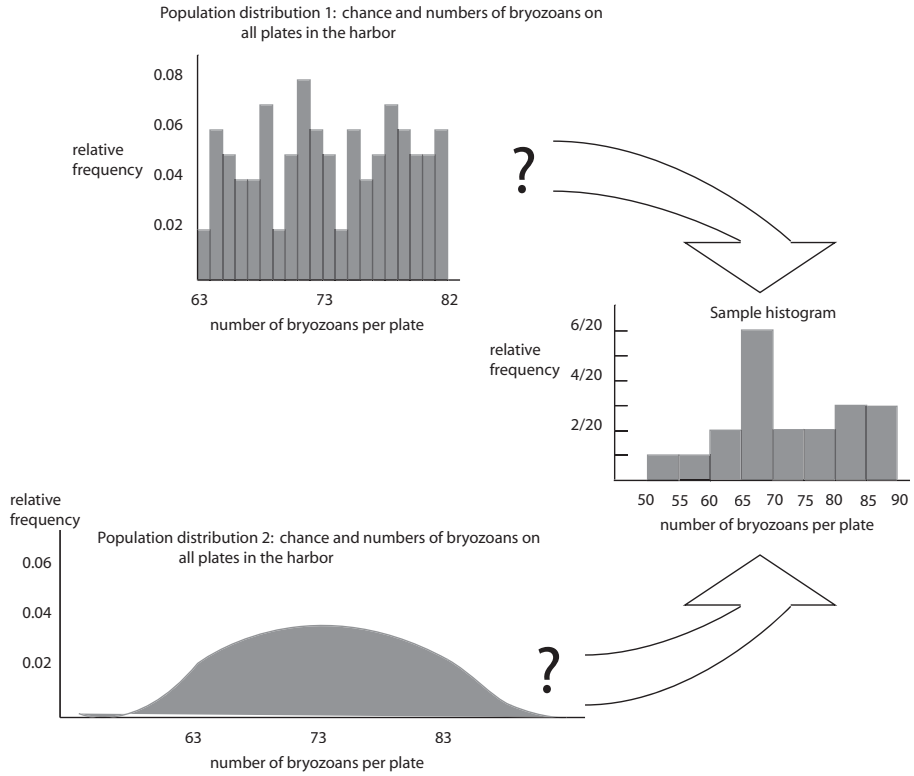

Figure 1.3: Relative frequency distribution: freshwater bryozoans

histogram is not.
(i) True (ii) False

2. *Patient ages.*
The ages of twenty patients in a blood pressure study are listed below.

47, 47, 49, 50, 51, 44, 45, 45, 45, 46,
41, 41, 42, 42, 43, 32, 37, 39, 40, 41

(a) Construct a relative frequency histogram for this data which has a range of 30 to 55 years with class intervals of width 5 years each. The proportion of patients aged 40 years or less is (choose one)
(i) $\frac{2}{20}$ (ii) $\frac{3}{20}$ (iii) $\frac{4}{20}$ (iv) $\frac{5}{20}$
Type patient ages into L_2 list. Then 2nd STAT PLOT 1 ENTER and replace L_1 with L_2 (overwrite bryozoan plot). Then ZOOM ZoomStat ENTER. Clean up histogram by WINDOW 30 55 5 -2 8 1 1 GRAPH. Trace.

(b) The proportion of patients aged 40 to 50 years is (choose one)
(i) $\frac{13}{20}$ (ii) $\frac{14}{20}$ (iii) $\frac{15}{20}$ (iv) $\frac{16}{20}$

(c) The relative frequency histogram is more or less bell–shaped.
(i) True (ii) False

1.3 Characterizing a Set of Measurements: Numerical Methods

We briefly look at three important statistics, including the sample *mean* (also called an *average*), sample *variance* and sample *standard deviation*:

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}, \qquad s^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}, \qquad s = \sqrt{s^2}.$$

We also look at parameters associated with these statistics: population mean, μ, population variance, σ^2, and population standard deviation, σ. The relationship between the various statistics (and parameters) is explored using the *Empirical rule* for the normal (bell–shaped) relative frequency distribution,

- 68% of measurements are in $\mu \pm \sigma$ (or, approximately, $\bar{y} \pm s$)
- 95% of measurements are in $\mu \pm 2\sigma$ (or, approximately, $\bar{y} \pm 2s$)
- 97.7% of measurements are in $\mu \pm 3\sigma$ (or, approximately, $\bar{y} \pm 3s$).

Exercise 1.3 (Characterizing a Set of Measurements: Numerical Methods)

1. *Freshwater bryozoans.*

plate	number	plate	number
1	82	11	51
2	71	12	81
3	84	13	76
4	67	14	66
5	67	15	59
6	89	16	86
7	64	17	76
8	68	18	87
9	64	19	68
10	70	20	69

Table 1.2: Numbers of bryozoans on twenty plates

Reconsider the bryozoan data given in Table 1.2.

(a) The sample mean number, $\bar{y}$ ("y–bar"), of bryozoans per plate in Table 1.2 is

$$\begin{aligned} \bar{y} &= \frac{\sum_{i=1}^{20} y_i}{n} \\ &= \frac{y_1 + y_2 + \cdots + y_{20}}{20} \\ &= \frac{82 + 71 + \cdots + 69}{20} \approx \end{aligned}$$

(i) 70.2 (ii) 71.7 (iii) 72.0 (iv) 72.3

Type bryozoan counts into L_1. STAT CALC 1-Var Stats ENTER 2nd L_1 ENTER. Read $\bar{x} = 72.25$.

(b) The sample standard deviation, s, in the number of bryozoans per plate is

$$\begin{aligned} s &= \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}} \\ &= \sqrt{\frac{(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_{25} - \bar{y})^2}{n-1}} \\ &= \sqrt{\frac{(82 - 72.25)^2 + (71 - 72.25)^2 + \cdots + (69 - 72.25)^2}{25-1}} \approx \end{aligned}$$

(i) 10.1 (ii) 10.4 (iii) 11.3 (iv) 11.7

After STAT CALC 1-Var Stats ENTER 2nd L_1 ENTER, read $Sx \approx 10.09$.

(c) The sample variance, s^2, in the number of bryozoans per plate is

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} \approx 10.09^2 \approx$$

(i) 77.0 (ii) 89.4 (iii) 99.3 (iv) 101.9

VARS ENTER Statistics ENTER Sx ENTER x^2 ENTER

(d) The value one (1) standard deviation below the mean is (choose one):

$$\bar{y} - s = 72.3 - 10.1 =$$

(i) 62.2 (ii) 63.2 (iii) 69.2 (iv) 72.2

(e) The value one (1) standard deviation above the mean is (choose one):

$$\bar{y} + s = 72.3 + 10.1 =$$

(i) 62.2 (ii) 73.2 (iii) 82.4 (iv) 92.2

(f) Then

$$\begin{aligned} \bar{y} \pm s &= 72.3 \pm 10.1 \\ &= (72.3 - 10.1, 72.3 + 10.1) = \end{aligned}$$

(i) (62.2, 82.4) (ii) (60.2, 82.4) (iii) (59.2, 82.4) (iv) (62.2, 80.4)

(g) According to the Empirical rule, what percentage of the twenty plates *should* contain between 62.2 and 82.4 bryozoans per plate? Choose one.
(i) 68% (ii) 95% (iii) 97.7% (iv) none of these

(h) What percentage of the twenty plates *actually* contain between 62.2 and 82.4 bryozoans per plate? Choose one.
(i) $\frac{14}{20} = 70\%$ (ii) $\frac{15}{20} = 75\%$ (iii) $\frac{16}{20} = 80\%$ (iv) none of these
STAT SortA L_1 ENTER STAT ENTER. Count number between 62.2 and 82.4 in L_1. Divide by 20.

(i) According to the Empirical rule, what percentage of the twenty plates *should* contain the following number of bryozoans per plate:

$$\begin{aligned} \bar{y} \pm 2s &= 72.3 \pm 2 \times 10.1 \\ &= (72.3 - 2 \times 10.1, 72.3 + 2 \times 10.1) \\ &= (52.1, 92.5)? \end{aligned}$$

Choose one.
(i) 68% (ii) 95% (iii) 97.7% (iv) none of these

(j) What percentage of the twenty plates *actually* contain between 52.1 and 92.5 bryozoans per plate? Choose one.
(i) $\frac{17}{20} = 85\%$ (ii) $\frac{18}{20} = 90\%$ (iii) $\frac{19}{20} = 95\%$ (iv) $\frac{20}{20} = 100\%$

(k) According to the Empirical rule, what percentage of the twenty plates *should* contain between 42 and 102.6 bryozoans per plate? Choose one.
(i) 68% (ii) 95% (iii) 97.7% (iv) none of these

(l) What percentage of the twenty plates *actually* contain between 42 and 102.6 bryozoans per plate? Choose one.
(i) $\frac{17}{20} = 85\%$ (ii) $\frac{18}{20} = 90\%$ (iii) $\frac{19}{20} = 95\%$ (iv) $\frac{20}{20} = 100\%$

(m) The Empirical rule *assumes* the relative frequency histogram for the bryozoan measurements is approximately normal (bell–shaped) with sample mean $\bar{y} = 72.3$ and sample standard deviation $s = 10.1$. This explains why there is some discrepancies between the Empirical rule and the actual number of bryozoans per plate.

interval	Empirical	actual
62.6 to 82.4	68%	70%
52.1 to 92.5	95%	95%
42 to 102.6	99.7%	100%

(i) True (ii) False

2. *Patient ages.*
The ages of twenty patients in a blood pressure study are listed below.

47, 47, 49, 50, 51, 44, 45, 45, 45, 46,
41, 41, 42, 42, 43, 32, 37, 39, 40, 41

(a) The sample mean age of patients is $\bar{y} \approx$
(i) 40.4 (ii) 42.1 (iii) 43.4 (iv) 44.6
Type patient ages into L_2. STAT CALC 1-Var Stats ENTER 2nd L_2 ENTER.

(b) The sample standard deviation in age of patients is $s \approx$
(i) 4.6 (ii) 4.8 (iii) 5.3 (iv) 6.6

(c) According to the Empirical rule, the percentage of ages in the interval 38.8 to 48 should be (choose one)
(i) 68% (ii) 95% (iii) 97.7% (iv) none of these

(d) According to the Empirical rule, the *number* of patients with ages in the interval 38.8 to 48 should be (choose one)
(i) 12.3 (ii) 13.6 (iii) 14.7% (iv) none of these
$0.68 \times 20 \approx$?

(e) The actual *number* of patients with ages in interval 38.8 to 48 (choose one)
(i) 14 (ii) 15 (iii) 16 (iv) none of these
STAT SortA L_2 ENTER STAT ENTER. Count number between 38.8 and 48 in L_2.

(f) The actual percentage of ages in the interval 38.8 to 48 is (choose one)
(i) 70% (ii) 75% (iii) 90% (iv) none of these
Divide by 20.

3. *Ph levels in soil.*
Assume the measurements for Ph levels in soil follow *exactly* a normal relative frequency distribution with population mean $\mu = 7$ and population standard deviation $\sigma = 3.4$. Use the Empirical rule.

(a) The percentage of Ph levels in the interval 3.6 to 10.4 is (choose one)
(i) 68% (ii) 95% (iii) 97.7% (iv) none of these

(b) The percentage of Ph levels in the interval 0.2 to 13.8 is (choose one)
(i) 68% (ii) 95% (iii) 97.7% (iv) none of these

(c) The percentage of Ph levels in the interval 7 to 13.8 is (choose one)
(i) 47.5% (ii) 50% (iii) 65% (iv) none of these
This half of what?

1.4 How Inferences Are Made

Covered in previous sections.

1.5 Theory and Reality

Covered in previous sections.

1.6 Summary

We briefly introduced some definitions in the field of statistics, particularly related to sample, statistic, population and parameter. Relative frequency histograms and relative frequency distributions were discussed. Finally, the mean, variance and standard deviation statistics (and parameters) were considered, as well as the Empirical rule which shows how these quantities relate to one another.