Attendance Workbook

For Statistics 503

Statistical Methods in Biology

Fall 2004

by

Jonathan Kuhn, Ph.D.
Associate Professor of Statistics,
Mathematics and Physics Section,
Purdue University North Central

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 as applied to the life sciences. Numerical computation, algebra, graphs and mathematical notation are used; calculus is *not* used.

This workbook is a necessary component for a student to successfully complete this course. Without the workbook, a student will not be able to participate in the course.

- This attendance workbook is *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; this workbook, 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 workbook. A student is to use this workbook to follow along with during a lecture.
- There are different kinds of exercises, including multiple choice, true/false, matching and fill–in–the–blank.
- Each week, I recommend you read the text, answer the questions given here in the attendance workbook and then do either the quiz or homework assignment, in that order.

On the one hand, the workbook is, 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, this workbook is 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. Jon Kuhn
Associate Professor of Statistics,
Mathematics and Physics Section,
Purdue University North Central,
April 2004

Chapter 1

Statistics: Its Objectives and Scope

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 to be used to infer something about a variable 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.

In this chapter, after describing the difference between a variable and a measurement, we will look at four important terms used in statistical inference: population, sample, statistic and parameter. We find out that clearly identifying these four components in any study helps us to clarify the statistical problem that needs to be solved. The statistician, rather than the data, decides what statistical problem needs to solved by identifying these four items.

Exercise 1.1 (Variable and Measurement) A *variable* is a characteristic of a person, object or entity and a *measurement* is a particular instance of a variable.

1. **True** / **False** A measurement for the variable *height of a man* is 5.6 feet tall. Another measurement for this variable is 5.8 feet tall. A set of measurements is, for example, {5.6, 5.6, 5.7, 5.8}.

2. **True** / **False** A measurement for the variable *shoulder height of a cow* is 45 inches. Another measurement for this variable is 45 pounds.

3. A measurement for the variable *length of time of exposure in the sun* is (circle best one) 9/24/98 / **4 hours**.

4. **True** / **False** The measurement "45" could be a particular instance of the variable *age of a elephant*. The measurement "45" could also be a particular instance of the variable *number of bees in a bag*.

5. The measurement "12" is a particular instance of the variable (circle none, one or more)

(a) number of eggs in a nest.

(b) number of calls of a crow.

(c) volume, in cubic yards, of pond.

(d) length, in cm, of a bird.

6. *Observation.* Sometimes, it is more appropriate to use the word "observation", rather than use the word "measurement". An observation (rather than measurement) for the variable *date of exposure to the sun* is (circle best one) 9/24/98 / **4 hours**.

7. *Observation, Again.* A/n (circle one) **observation** / **measurement** for the variable *time of arrival* is 3pm. A *set of observations* for this variable is {1am, 2:30am, 12noon}.

8. *Data Point.* The word "measurement" and "observation" are both special cases of the more general and neutral word "data point". A (circle one) **measurement** / **observation** / **data point** for the variable *country of origin* is Sweden. A possible data set for this variable is {Mexico, Portugal, Zimbabwe}.

9. The data point "silver" is a particular instance of the variable (circle none, one or more)

   (a) length of fox trail.

   (b) metal mined in quarry.

   (c) wing color of a bird.

   (d) name of a horse (as in "hi–ho").

10. **True** / **False** Although there is no "formula" to know precisely which of the three words, "measurement", "observation" or "data point", to use in every case, often one of the three is most appropriate. For example, county of origin *data* collected by immigration officials at national boundaries could sensibly be called *observations* (rather than data). Height *measurements*, looked at sometime after their collection, could sensibly be called height *data*.

Exercises 1.2 (Statistical Population, Sample, Statistic and Parameter[1]: Proportion Of Grey Squirrels.) Inferential statistics can be used to say that, under certain conditions, since 345 of one thousand randomly chosen squirrels are grey squirrels, we can then infer that approximately $\frac{345}{1000}$ths or 34.5% of **all** squirrels are grey squirrels. Several special statistical terms are often used to describe the different elements in this situation. Assume the color of squirrels are either grey, red or other.

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

1. The *statistical population* is (circle one)
   (a) *all* squirrels.
   (b) colors of *all* squirrels.
   (c) the one thousand squirrels, selected at random.
   (d) colors of the one thousand squirrels, selected at random.
2. The *sample* is,
   (a) *all* squirrels.
   (b) colors of *all* squirrels.
   (c) the one thousand squirrels, selected at random.
   (d) colors of the one thousand squirrels, selected at random.
3. **True** / **False** Although, loosely speaking, the (statistical) population is "all squirrels" and the sample is "the one thousand squirrels", we are actually interested in only one particular aspect of any given squirrel; namely, their fur color. In other words, more exactly, the (statistical) population is "colors of all squirrels" and the sample is "colors of one thousand squirrels". The words "population" and "sample" both are, confusingly, used to refer to both the items (squirrels) and the characteristics of the items (color of squirrels).
4. Although there are many possible variables in this case, the variable of *interest* is, in this case,
   (a) a squirrel, without specifying which squirrel.
   (b) a particular squirrel, "Susan", say.
   (c) color of a squirrel.
   (d) red, color of a particular squirrel, "Susan", say.
   (e) {grey, grey, red, other, ..., red}, the set of colors for the one thousand randomly selected squirrels.
5. An *observation* of the variable of interest is
   (a) a squirrel, without specifying which squirrel.
   (b) a particular squirrel, "Susan", say.
   (c) color of a squirrel.
   (d) red, color of a particular squirrel, "Susan", say.
   (e) {grey, grey, red, other, ..., red}, the set of colors for the one thousand randomly selected squirrels.

6. A *set of observations* is,
   (a) a squirrel, without specifying which squirrel.
   (b) a particular squirrel, "Susan", say.
   (c) color of a squirrel.
   (d) red, color of a particular squirrel, "Susan", say.
   (e) {grey, grey, red, other, ..., red}, the set of colors for the one thousand randomly selected squirrels.

7. 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. Here, the *statistic of interest* is,
   (a) proportion of grey squirrels, among *all* squirrels.
   (b) proportion of grey squirrels, among the one thousand randomly chosen squirrels.

8. The *value* of the statistic of interest is
   (circle none, one or more) **0.345** / **34.5%** / **345**.

9. The *parameter of interest* is,
   (a) proportion of grey squirrels, among *all* squirrels.
   (b) proportion of grey squirrels, among the one thousand randomly chosen squirrels.

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

11. *Random Sample.* A sample of squirrels taken at random (circle one) **is** / **is not** representative of the population of all squirrels.

12. *Sample Size.* The (circle one) **smaller** / **larger** the random sample size, the more accurate the proportion of red squirrels statistic will be of the proportion of red squirrels parameter.

13. *0–1 Measurements.* The proportion of red squirrels statistic, 0.345, is determined in the following way.
   (a) assigning all *grey* squirrels the number "1" and assigning every other color of squirrel the number "0", so that the data set of size 1000, {grey, grey, red, other, ..., red}, becomes the 1000 numbers {1, 1, 0, 0, ..., 0}, then summing these numbers and dividing by 1000.

(b) assigning all *red* squirrels the number "1" and assigning every other color of squirrel the number "0", so that the data set of size 1000, {grey, grey, red, other, ..., red}, becomes the 1000 numbers {0, 0, 1, 0, ..., 1}, then summing these numbers and dividing by 1000.

Exercises 1.3 (Review: Average Flight Distance of Migratory Whooping Cranes) One hundred and twenty whooping cranes are selected at random and the distances of their southern migratory flight is measured (observed). From this group, an average of 2,300 miles is computed. Match columns: *All* of the items in the first column will be used up in the matching procedure; however, one item in the second column will be left unmatched.

terms	whooping crane example
(a) measurement	**(a)** average flight distance for 120 whooping cranes
(b) variable	**(b)** all whooping cranes
(c) parameter	**(c)** flight distances for all whooping cranes
(d) statistical population	**(d)** flight distance for a whooping crane
(e) sample	**(e)** average flight distance for all whooping cranes
(f) statistic	**(f)** 120
(g) sample size	**(g)** 120 flight distances
	(h) 2,225 mile flight distance for a particular whooping crane

terms	(a)	(b)	(c)	(d)	(e)	(f)	(g)
whooping crane example							

Exercises 1.4 (More on Populations and Samples) Statisticians often compare two or more populations.

1. *Wheat Yields and Fertilization.* An farmer wishes to compare yields from seven different methods of fertilizing and watering plots of wheat. In this case, there are (circle one) **one** / **two** / **seven** different populations.
   The average wheat yield (statistic) from seven plots subjected to the different fertilization and watering methods (samples) would be compared in a statistical study to infer if there is a difference in the actual average yields (parameter) in wheat under the seven different methods.

2. *Cancer Rates.* The Environmental Protection Agency (EPA) wishes to compare the cancer rate of people living in four different areas near St Clair Lake. In this case, there are (circle one) **one** / **four** / **seven** different populations.
   The cancer rate (statistic) from a sample of people from each of the four areas would be compared in a statistical study to infer is there is a difference in the actual cancer rates (parameter) of all people living in the four areas.

3. *Healthy Forest.* A biologist is interested in the health of three different 100–acre forests and so measures and compares the following items (do *not* circle anything yet!)

(a) number of animals

(b) number of trees

(c) acidity of soil

(d) number of fires per year

(e) number fish in lakes

for the three forests. In this case, there *could* be (circle *none, one or more!*)

(a) three *multi*variate populations:

health, forest 1	health, forest 2	health, forest 3

where the health of each forest consists of five variables, (animals, trees, soil, fires, fish)

(b) five *multi*variate populations

animals	tree	soil	fires	fish

where, for example, the number of animals is the total count in the three forests, (forest 1, forest 2, forest 3)

(c) fifteen *uni*variate populations

animals, forest 1	animals, forest 2	animals, forest 3	$\cdots$	fish, forest 3

where, for example, the number of animals, forest 1, is the number of animals in forest 1.

Chapter 2

Describing Statistical Populations

2.1 Introduction

This chapter introduces necessary terminology and some introductory techniques for describing statistical populations.

2.2 Types Of Populations

Statistical populations can be categorized in a number of different ways.

- univariate versus multivariate
- real versus conceptual
- finite versus infinite
- qualitative versus quantitative
- continuous versus discrete

Categorizing a population is essentially a means of forcing a statistician to *clearly* define what the population is, and, in a broader sense, clearly identifying the statistical problem.

Exercise 2.1(Categorizing Statistical Populations: Hormone Effect on Milk Yield Study) Consider the table below which contains various random measurements on a number of cows taken during a study on the effect of a hormone, given in tablet form, on daily milk yield.

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

1. *Univariate versus multivariate. Cow (identification number)*, *test date*, and the other six variables could be considered, on the one hand, to be eight separate *uni*variate populations (eight populations, each with one characteristic). On the other hand, since all of these eight variables act together to determine hormone effect on milk yield, this could also be considered to be a *multi*variate (in fact, eight–variate) population (one population with eight characteristics).

   (a) There are (circle one) **5** / **6** / **7** observations for the variable *cow (identification number)*.

   (b) The seven observations for the one variable *cow (identification number)* is a (circle one) **sample** / **(statistical) population**.

   (c) There are (circle one) **5** / **6** / **7** observations for the variable *(number of) tablets*.

   (d) Each *column* of this table describes
   (circle one) **one variable, seven observations** / **eight variables, one observation**.

   (e) Each *row* of this table describes
   (circle one) **one variable, seven observations** / **eight variables, one observation**.

   (f) The entire hormone–effect–on–milk–yield study is a (circle none, one or more)

      i. univariate study.

      ii. bivariate study.

      iii. multivariate study.

      iv. k–variate study, where $k = 8$.

   (g) If this seven–observation *sample* of the entire hormone–effect–on–milk–yield study is considered to be multivariate, the corresponding (statistical) population of the hormone–effect–on–milk–yield study (circle one)
   must also be / **could not also be** multivariate.

2. *Real Versus Conceptual.* Although the seven cows in the sample are *real* in the sense we can count and measure all of these cows, the population (in the

loose all–encompassing sense of the word "population") from which these seven cows are taken may be real or conceptual. If every single cow can actually be counted, the population is real; otherwise, it is conceptual. In this case, a real population becomes a conceptual one at the point at which all statisticians are unwilling to spend more time and effort counting cows. Although not always true, conceptual populations are often associated with experiments, whereas real populations are often associated with observed data.

(a) If the population is restricted to only the cows on the two farms in the study, then, in this case, the population is probably (circle one) **real** / **conceptual** because it is possible to measure all of the cows in this restricted population.

(b) If the population is all of the cows in the state of Indiana (where the study takes place, say), then, in this case, the population is still probably (circle one) **real** / **conceptual** because, although difficult, it is probably still possible to identify and measure all of the cows in this population.

(c) If the population is all of the cows who lived in the state of Indiana for the last ten years, then, in this case, the population is (circle one) **real** / **conceptual** because, although there may be records for all of these cows, it is most likely no longer possible to make all of the measurements required in this study for the cows (since some cows in this population are dead, for example).

3. *Finite versus infinite.* On the other hand, although there is a finite (seven) number of (shoulder) heights in the sample, the population of (shoulder) heights is infinite; that is, there are an infinite of heights between 41.234 inches and 41.235 inches. The terms "finite" and "infinite" is applied to the number of possible values an *individual* member of the population *could* take on, whereas the terms "conceptual" and "real" populations have to do with the number of items in a population.

   (a) If the population is restricted to only the cows on the two farms in the study, then, in this case, the population is (circle one) **finite** / **infinite**.

   (b) If the population is all of the cows in the state of Indiana, then, in this case, the population is (circle one) **finite** / **infinite**.

   (c) If the population is all of the cows who lived in the state of Indiana for the last ten years, then, in this case, the population is still (circle one) **finite** / **infinite**. In other words, an conceptual population need not be infinite.

   (d) If the population is all of the (actual) *(shoulder) heights* of all of the (real) cows who lived in the state of Indiana for the last ten years, then, in this case, the population is (circle one) **finite** / **infinite**.

(e) If the population is all the **possible** *(shoulder) heights* of all of the (conceptual) cows who lived in the state of Indiana for the last ten years, then, in this case, the population is (circle one) **finite** / **infinite**. (Can you list a few of the ∞ of heights between 41 and 42 inches tall, say?)

(f) If the population is all of the (actual) *milk yield* of all of the (real) cows who lived in the state of Indiana for the last ten years, then, in this case, the population is (circle one) **finite** / **infinite**.

(g) If the population is all the **possible** *milk yield* of all of the (conceptual) cows who lived in the state of Indiana for the last ten years, then, in this case, the population is (circle one) **finite** / **infinite**. (Can you list a few of the ∞ of milk yields between 100.3 and 100.4 quarts?)

4. *Qualitative versus quantitative. Qualitative* data is described by a words which, although these words may be numbers, are used only to label or identify an object or entity.
Quantitative data is measured by numbers which can be ordered and manipulated using mathematical rules.
The population (variable) associated with qualitative data are said to be qualitative; similarly, the population (variable) associated with quantitative data is said to be quantitative.

(a) The variable *cow (identification number)* is (circle one) **qualitative** / **quantitative** / **neither** because, although numbers (17, 18, ...) are used to describe the data, these words are used only to label the cows, to be able to distinguish one cow from another.

(b) The variable *test date* is (circle one) **qualitative** / **quantitative** / **neither** because, although numbers are used to describe the data, and these numbers can be ordered, they cannot be manipulated using mathematical formulas.

(c) The variable *farm* is (circle one) **qualitative** / **quantitative** / **neither** because the data (F, M) are labels used simply to tell one farm from the other.

(d) The variable *(shoulder) height (of the cows)* is (circle one) **qualitative** / **quantitative** / **neither** because numbers are used to describe the data, these numbers can be ordered and manipulated using mathematical formulas (they can be added and subtracted, say).

(e) The variable *health* is (circle one) **qualitative** / **quantitative** / **neither** because, although the data (poor, good, ...) can be ordered, it cannot be manipulated using mathematical formulas.

(f) The variable *(number of) tablets* is (circle one) **qualitative** / **quantitative** / **neither** because numbers (2, 1, ... are used to describe the data,

these numbers can be ordered and manipulated using mathematical formulas (they can be added and subtracted, say).

(g) The variable *(milk yield) before (given the hormone)* is (circle one) **qualitative** / **quantitative** / **neither** because numbers (100.7, 97.8, ... are used to describe the data, these numbers can be ordered and manipulated using mathematical formulas (they can be added and subtracted, say).

(h) The variable *(milk yield) after (given the hormone)* is (circle one) **qualitative** / **quantitative** / **neither**.

5. *Discrete versus continuous. Discrete* data is countable; each measurement is distinct and different from every other measurement; there are "gaps" between discrete measurements.
Continuous data is such, for any two different measurements, there is *always* a third data point between these two measurements; in other words, there is an *infinite* number of data points between these two measurements (why?) and this infinite number of points are "so close together", they form a continuum between the two measurements.
The population (variable) associated with discrete data are said to be discrete; similarly, the population (variable) associated with continuous data is said to be continuous.

(a) The variable *cow (identification number)* is (circle one) **discrete** / **continuous** because the cows can be counted.

(b) The variable *test date* is (circle one) **discrete** / **continuous** because the dates can be counted.

(c) The variable *farm* is (circle one) **discrete** / **continuous** the farms can be counted (there are two farms)

(d) The variable *(shoulder) height (of the cows)* is (circle one) **discrete** / **continuous** because for every two data points, there is always a third (and so an infinite number) in between; between 41 and 40 there is 40.5 (and 40.435 and 40.764653580... and so on).

(e) The variable *health* is (circle one) **discrete** / **continuous** because it can be counted (there are five health descriptions).

(f) The variable *(number of) tablets* is (circle one) **discrete** / **continuous** because it can be counted (there are 1, 2, 3, or 4 tablets).

(g) The variable *(milk yield) before (given the hormone)* is (circle one) **discrete** / **continuous** because for every two data points, there is always a third (and so an infinite number) in between; between 100.7 and 97.8 there is 100.3 (and 100.435 and 100.764653580... and so on).

(h) The variable *(milk yield) after (given the hormone)* is (circle one) **discrete** / **continuous**.

(i) **True** / **False** Continuous data can*not* be counted.

(j) **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 ∞ of values in any chosen interval.

Exercise 2.2 (More on Types of Population.)

1. A farmer wishes to compare yields from seven different methods of fertilizing and watering plots of wheat. The average wheat yields (statistics) from seven plots subjected to the different fertilization and watering methods (samples) would be compared in a statistical study to infer if there is a difference in the actual average yields (parameters) in wheat under the seven different methods. Circle all appropriate descriptions for the type of population in this case.

   - **univariate** / **multivariate** (Explain how there could be either seven univariate populations or one multivariate population.)
   - **real** / **conceptual** (Why is/are the population(s) probably conceptual?)
   - **finite** / **infinite** (How many different possible yields are there between 1.23 tons and 1.24 tons?)
   - **qualitative** / **quantitative** / **neither**
   - **continuous** / **discrete** (Although yields are probably measured discretely (1, 2, 3 ... tons), why is the underlying variable continuous?)

2. The Environmental Protection Agency (EPA) wishes to compare the cancer rate of people living in four different areas near St Clair Lake. The cancer rates (statistics) from samples of people from each of the four areas would be compared in a statistical study to infer if there is a difference in the actual cancer rates (parameters) of all people living in the four areas.

   - **univariate** / **multivariate** (Explain how there could be either four univariate populations or one multivariate population.)
   - **real** / **conceptual** (Why is/are the population(s) probably conceptual?)
   - **finite** / **infinite** (How many different possible cancer rates are there between 3.12 per 1000 and 3.13 per 1000?)
   - **qualitative** / **quantitative** / **neither**

- **continuous** / **discrete** (Although cancer rates are probably measured discretely (1, 2, 3 . . . per 1000), why is the underlying variable continuous?)

3. A biologist is interested in the health of three different 100–acre forests and so measures and compares the following items (do *not* circle anything yet!),

   (a) number of animals

   (b) number of trees

   (c) acidity of soil

   (d) number of fires per year

   (e) number fish in lakes

   for the three forests. If considered as five *multi*variate populations,

animals	tree	soil	fires	fish

   where, for example, the number of animals is the total count in the three forests, (forest 1, forest 2, forest 3), and is

   - **univariate** / **multivariate** (Explain how there could be either three univariate populations or one multivariate population.)
   - **real** / **conceptual** (Why is/are the population(s) probably conceptual?)
   - **finite** / **infinite** (Why is it probably finite?)
   - **qualitative** / **quantitative** / **neither**
   - **continuous** / **discrete** (Why is it discrete?)

   and the acidity of soil is

   - **univariate** / **multivariate** (Explain how there could be either three univariate populations or one multivariate population.)
   - **real** / **conceptual** (Why is/are the population(s) probably conceptual?)
   - **finite** / **infinite** (How many different possible acidity levels are there between 0.34 and 0.35?)
   - **qualitative** / **quantitative** / **neither**
   - **continuous** / **discrete** (Why is the underlying variable continuous?)

2.3 Describing Populations Using Distributions

Exercise 2.3 (Probability Distributions, Discrete)

See TI–83 Lab 1: discrete distribution

1. *Number of Ears of Corn.* The number of ears of corn, Y, on a typical corn plant has the following probability distribution.

y	0	2	4	6	8	10
$f(y)$	0.17	0.21	0.18	0.11	0.16	0.17

   (a) The chance a corn plant has 8 ears of corn is (circle one) **0.17** / **0.21** / **0.16** / **0.11**.

   (b) The chance a corn plant has less than 6 ears of corn is (circle one) **0.17** / **0.21** / **0.56** / **0.67**.

   (c) $\Pr\{Y \leq 4\} = f(0) + f(2) + f(4) =$ (circle one) **0.17** / **0.21** / **0.56** / **0.67**.

   (d) $\Pr\{Y = 2\} = f(2) =$ (circle one) **0.17** / **0.21** / **0.56** / **0.67**.

   (e) $\Pr\{Y = 2.1\} =$ (circle one) **0** / **0.21** / **0.56** / **0.67**.

   (f) $\Pr\{Y > 2.1\} =$ (circle one) **0.21** / **0.38** / **0.56** / **0.62**.

   (g) $f(0) + f(2) + f(4) + f(6) + f(8) + f(10) =$ (circle one) **0.97** / **0.98** / **0.99** / **1**.

   (h) **True** / **False** The probability distribution is given by the pair $(y, f(y))$ and *not* just $f(y)$ alone.

   (i) *Histogram (Graph) of Distribution.* Consider the probability histograms given in the figure below.

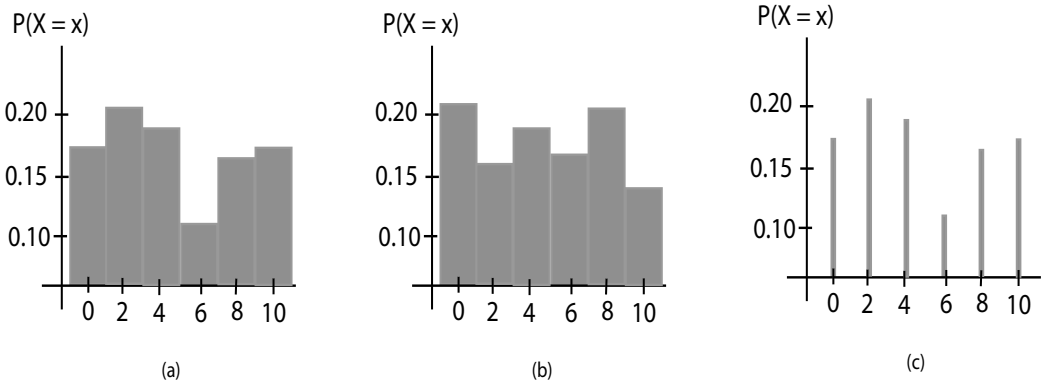


Figure 2.1 (Probability Histogram)

Which, if any, of the three probability histograms in the figure above, describe the probability distribution of the number of ears of corn? (circle none, one or more) **(a)** / **(b)** / **(c)**

2. *Number of Piglets.* The number of piglets in a litter, Y, follows the following probability distribution.

$$\Pr\{Y = y\} = \frac{1}{5}, \quad y = 5, 6, 7, 8, 9$$

   (a) The chance a litter has 8 piglets is (circle one) $\frac{1}{5}$ / $\frac{2}{5}$ / $\frac{3}{5}$ / $\frac{4}{5}$.

   (b) The chance the litter has less than 8 piglets is (circle one) $\frac{1}{5}$ / $\frac{2}{5}$ / $\frac{3}{5}$ / $\frac{4}{5}$.

   (c) $\Pr\{Y \leq 6\}$ = (circle one) $\frac{1}{5}$ / $\frac{2}{5}$ / $\frac{3}{5}$ / $\frac{4}{5}$.

   (d) $\Pr\{Y = 7\}$ = (circle one) $\frac{1}{5}$ / $\frac{2}{5}$ / $\frac{3}{5}$ / $\frac{4}{5}$.

   (e) $\Pr\{Y = 8.1\}$ = (circle one) $\frac{0}{5}$ / $\frac{1}{5}$ / $\frac{2}{5}$ / $\frac{3}{5}$.

   (f) $\mathrm{P}\{5 < Y < 8\}$ = (circle one) $\frac{0}{5}$ / $\frac{1}{5}$ / $\frac{2}{5}$ / $\frac{3}{5}$.

   (g) $\sum_{y=5}^{y=9} P\{Y = y\}$ = (circle one) $\frac{0}{5}$ / $\frac{1}{5}$ / $\frac{2}{5}$ / $\frac{5}{5}$.

   (h) **True** / **False** The Y in $\Pr\{Y = y\}$ is called a *random variable* and, in this case, represents the number of piglets in a litter.

3. *Pea Plants.* Pea plants have either yellow peas, $y = 0$ or green peas, $y = 1$, according to the following probability distribution.

$$\Pr\{Y = y\} = (0.25)^y (0.75)^{1-y}, \quad y = 0, 1$$

   (a) The chance of choosing a green pea ($y = 1$) from a pea plant is $\Pr\{Y = 1\} = (0.25)^1(0.75)^{1-1}$ = (circle one) **0** / **0.25** / **0.50** / **0.75**.

   (b) The chance of choosing a yellow pea ($y = 0$) from a pea plant is $\Pr\{Y = 0\} = (0.25)^0(0.75)^{1-0}$ = (circle one) **0** / **0.25** / **0.50** / **0.75**.

   (c) A tabular version of this probability distribution is (circle one)

      i. Distribution A.

y	0	1
$f(y)$	0.25	0.75

      ii. Distribution B.

y	0	1
$f(y)$	0.75	0.25

      iii. Distribution C.

y	0	1
$\Pr\{Y = y\}$	0.50	0.50

(d) The number of different ways of describing a probability distribution include (check none, one or more)

i. function

ii. tree diagram

iii. table

iv. graph

(e) **True** / **False** The Y in $\Pr\{Y = y\}$ is called a *random variable* and, in this case, represents whether the peas from a pea plant are either yellow or green.

(f) **True** / **False** $\Pr\{Y = y\} = f(y)$.

4. *Pea Plants, Parameter π.* Pea plants have either yellow peas, $y = 0$ or green peas, $y = 1$, according to the following probability distribution.

$$\Pr\{Y = y\} = \pi^y(1-\pi)^{1-y}, \quad y = 0, 1$$

(a) The chance of choosing a green pea ($y = 1$) from a pea plant where $\pi = 0.25$ is
$\Pr\{Y = 1\} = (0.25)^1(0.75)^{1-1} =$ (circle one) **0** / **0.25** / **0.50** / **0.75**.

(b) The chance of choosing a green pea ($y = 1$) from a pea plant where $\pi = 0.35$ is
$\Pr\{Y = 1\} = (0.35)^1(0.35)^{1-1} =$ (circle one) **0** / **0.35** / **0.50** / **0.75**.

(c) The chance of choosing a green pea ($y = 1$) from a pea plant where $\pi = 0.85$ is
$\Pr\{Y = 1\} = (0.85)^1(0.85)^{1-1} =$ (circle one) **0** / **0.35** / **0.50** / **0.85**.

(d) A tabular version of this probability distribution, for general parameter π, is (circle one)

i. Distribution A.

y	0	1
$f(y)$	π	$1-\pi$

ii. Distribution B.

y	0	1
$f(y)$	$1-\pi$	π

iii. Distribution C.

y	0	1
$\Pr\{Y = y\}$	π	π

(e) **True** / **False** The parameter π can be any value between 0 and 1.

(f) **True** / **False** The distribution $\Pr\{Y = y\} = \pi^y(1-\pi)^{1-y}$, $y = 0, 1$ is used to describe "0–1" populations.

5. *Number of Ears of Corn, Two Varieties.* The number of ears of corn on two varieties (variety A and variety B) of corn plants, Y_1 and Y_2, follows the following probability distribution.

y_1	0	0	0	1	1	1
y_2	0	1	2	0	1	2
$f(y_1, y_2)$	0.17	0.21	0.18	0.11	0.16	0.17

   (a) The chance variety A has 0 ears of corn and variety B has 1 ear of corn is (circle one) **0.17** / **0.21** / **0.16** / **0.11**.

   (b) The chance variety A has 1 ear of corn is (circle one) **0.11** / **0.16** / **0.43** / **0.67**.

   (c) The chance variety A has *fewer* ears of corn than variety B is (circle one) **0.11** / **0.16** / **0.44** / **0.56**.

   (d) $\Pr\{Y_1 < Y_2\}$ = (circle one) **0.17** / **0.21** / **0.56** / **0.67**.

   (e) $\Pr\{Y_2 = 2\} = f(2)$ = (circle one) **0.17** / **0.21** / **0.35** / **0.67**.

   (f) $\Pr\{Y_1 = 0, Y_2 = 1\}$ = (circle one) **0** / **0.21** / **0.56** / **0.67**.

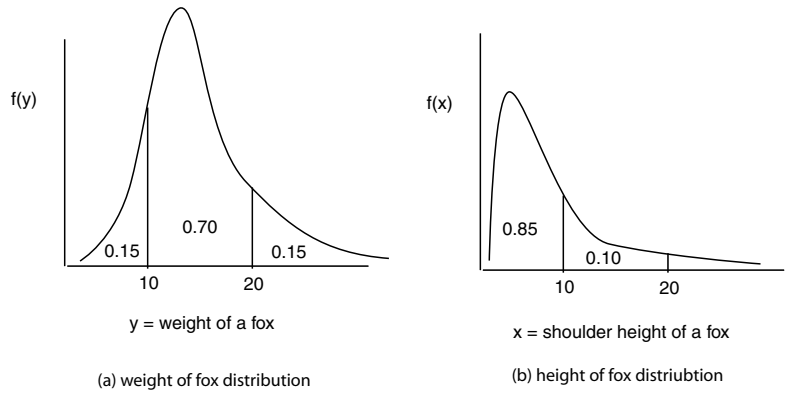
   (g) $\Pr\{Y_1 > 2.1\}$ = (circle one) **0** / **0.38** / **0.56** / **0.62**.

   (h) $f(0,0) + f(0,1) + f(0,2) + f(1,0) + f(1,1) + f(1,2) =$ (circle one) **0.97** / **0.98** / **0.99** / **1**.

   (i) **True** / **False** The (Y_1, Y_2) in $\Pr\{Y_1 = y_1, Y_2 = y_2\}$ is called a *bivariate random variable* and, in this case, represents number of ears of corn on two varieties (variety A and variety B) of corn plants.

Exercise 2.4 (Probability Distributions, Continuous)

1. *Weight and Height of Foxes.* The weight (in pounds) and (shoulder) height (in inches) of a fox, Y, taken at random from a protected wildlife preserve follows the following two continuous probability distributions.



(a) weight of fox distribution

(b) height of fox distriubtion

Figure 2.2 (Continuous Density of Weight and Height of Foxes)

(a) The chance a fox weighs less than 10 pounds is (circle one) **0.15** / **0.30** / **0.70** / **0.85**.

(b) The chance a fox weighs between 10 and 20 pounds is (circle one) **0.15** / **0.30** / **0.70** / **0.85**.

(c) $\Pr\{Y < 20\}$ = (circle one) **0.15** / **0.30** / **0.70** / **0.85**.

(d) $\Pr\{Y > 20\}$ = (circle one) **0.15** / **0.30** / **0.70** / **0.85**.

(e) $\Pr\{Y \geq 20\}$ = (circle one) **0.15** / **0.30** / **0.70** / **0.85**.

(f) $\Pr\{Y = 20\}$ = (circle one) **0** / **0.30** / **0.70** / **0.85**.

(g) **True** / **False** The Y in $\Pr\{Y = y\}$ is called a random variable and, in this case, represents the weight of a fox.

(h) The chance a fox has height between 10 and 20 inches tall is (circle one) **0.10** / **0.30** / **0.70** / **0.85**.

(i) $\Pr\{X < 10\}$ = (circle one) **0.15** / **0.30** / **0.70** / **0.85**.

(j) $\Pr\{X < 20\}$ = (circle one) **0.15** / **0.30** / **0.70** / **0.95**.

(k) $\Pr\{X \geq 20\}$ = (circle one) **0.05** / **0.30** / **0.70** / **0.85**.

(l) **True** / **False** It is *not* possible to calculate a table version of this continuous distribution (*any* continuous distribution, for that matter) such as is given, for example, by the discrete number of piglets example above.

2.4 Describing Populations Using Parameters

Exercise 2.5 (Location Parameters: Mean (Expected Value), Median, Percentile)

See TI–83 Lab 1: expected value and variance

1. *Mean (Expected Value), Discrete.* $\mu = \sum yf(y)$.

   (a) *A First Look: Number of Ears of Corn.*

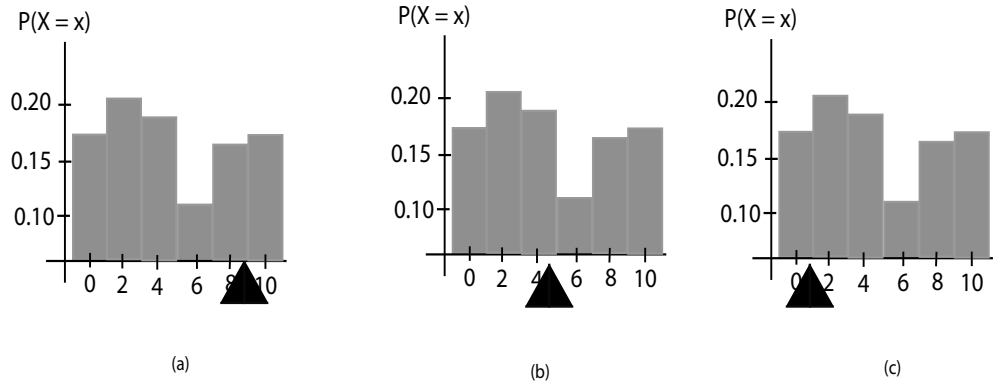


Figure 2.3 (Expected Value Is Like The Fulcrum Point of Balance)

If the expected value is like a fulcrum point which *balances* the "weight" of the probability distribution, then the expected value is most likely close to the point of the fulcrum given in which of the three graphs above?
Circle one. **(a)** / **(b)** / **(c)**
In other words, the expected value seems close to (circle one) **1** / **5** / **9**

(b) *Number of Ears of Corn.* The number of ears of corn, Y, on a typical corn plant has the following probability distribution.

y	0	2	4	6	8	10
$f(y)$	0.17	0.21	0.18	0.11	0.16	0.17

The mean (expected) number of ears of corn is given by $\mu = \sum yf(y)$,

$$\begin{aligned} \mu &= y_1f(y_1) + y_2f(y_2) + y_3f(y_3) + y_4f(y_4) + y_5f(y_5) + y_6f(y_6) \\ &= 0(0.17) + 2(0.21) + 4(0.18) + 6(0.11) + 8(0.16) + 10(0.17) \end{aligned}$$

which is equal to (circle one) **4.32** / **4.78** / **5.50** / **5.75**.
(Use your calculator: STAT ENTER; type Y, 0, 2, 4, 6 and 8, into L_1 and $P(X = x)$, 0.17, ..., 0.17, into L_2; then define $L_3 = L_1 \times L_2$; then STAT CALC ENTER 2nd L_3 ENTER; then read $\sum x = 4.78$.)

(c) *Swallows.* The number of swallows, Y, in any group of three birds is given by the following probability distribution.

Y	0	1	2	3
$P(Y = y)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

The mean (expected) number of swallows is

$$\mu = \frac{1}{8}\times 0 + \frac{3}{8}\times 1 + \frac{3}{8}\times 2 + \frac{1}{8}\times 3$$

which is equal to (circle one) **0.5** / **1.5** / **2.5** / **3.5**.

(d) *Another Distribution.* If the distribution is

x	0	1	2	3
$P(Y=y)$	$\frac{4}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	$\frac{1}{8}$

the mean is

$$\mu = \frac{4}{8}\times 0 + \frac{2}{8}\times 1 + \frac{1}{8}\times 2 + \frac{1}{8}\times 3$$

which is equal to (circle one) **1.500** / **0.875** / **1.375** / **0.625**

(e) *And Yet Another Distribution.* If the distribution is

$$P(Y=y) = \frac{3-y}{3}, \quad y = 1, 2,$$

the mean is

$$\mu = \frac{3-1}{3}\times 1 + \frac{3-2}{3}\times 2$$

which is equal to (circle one) $\mathbf{\frac{3}{3}}$ / $\mathbf{\frac{4}{3}}$ / $\mathbf{\frac{5}{3}}$ / $\mathbf{\frac{6}{3}}$

(f) *Means for Number of Ears of Corn, Two Varieties.* The number of ears of corn on two varieties (variety A and variety B) of corn plants, Y_1 and Y_2, follows the following probability distribution.

y_1	0	0	0	1	1	1
y_2	0	1	2	0	1	2
$f(y_1, y_2)$	0.17	0.21	0.18	0.11	0.16	0.17

i. $\Pr\{Y_1 = 0\}$ = (circle one) **0.17** / **0.21** / **0.38** / **0.56**.

ii. $\Pr\{Y_1 = 1\}$ = (circle one) **0.17** / **0.21** / **0.38** / **0.44**.

iii. The mean number of ears of corn for Variety A, μ_1, is

$$\mu_1 = 0 \times \Pr\{Y_1 = 0\} + 1 \times \Pr\{Y_1 = 1\} = 0 \times 0.56 + 1 \times 0.44$$

which is equal to (circle one) **0** / **0.44** / **0.56** / **1**.

iv. $\Pr\{Y_2 = 0\}$ = (circle one) **0.17** / **0.21** / **0.28** / **0.56**.

v. $\Pr\{Y_2 = 1\}$ = (circle one) **0.17** / **0.21** / **0.37** / **0.44**.

vi. $\Pr\{Y_2 = 2\}$ = (circle one) **0.17** / **0.21** / **0.35** / **0.44**.

vii. The mean number of ears of corn for Variety B, μ_2, is

$$\mu_2 = 0\times\Pr\{Y_2 = 0\}+1\times\Pr\{Y_2 = 1\}+2\times\Pr\{Y_2 = 2\} = 0\times 0.28+1\times 0.37+2\times 0.35$$

which is equal to (circle one) **0** / **0.44** / **0.56** / **1.07**.

(g) *Mean for 0–1 Population: Pea Plants.* Pea plants have either yellow peas, $y = 0$ or green peas, $y = 1$, according to the following probability distribution.

$$\Pr\{Y = y\} = \pi^y(1-\pi)^{1-y}, \quad y = 0, 1$$

and so the mean is

$$\mu = \pi^0(1-\pi)^{1-0} \times 0 + \pi^1(1-\pi)^{1-1} \times 1$$

which is equal to (circle one) $\mathbf{0}$ / $\boldsymbol{\pi}$ / $\mathbf{(1-\pi)}$ / $\boldsymbol{\pi(1-\pi)}$

2. *Median, Discrete.* $\eta = y$ where $\Pr\{Y \leq y\} = \Pr\{Y \geq y\} = 0.50$. Two examples of discrete distributions and their medians are given below.

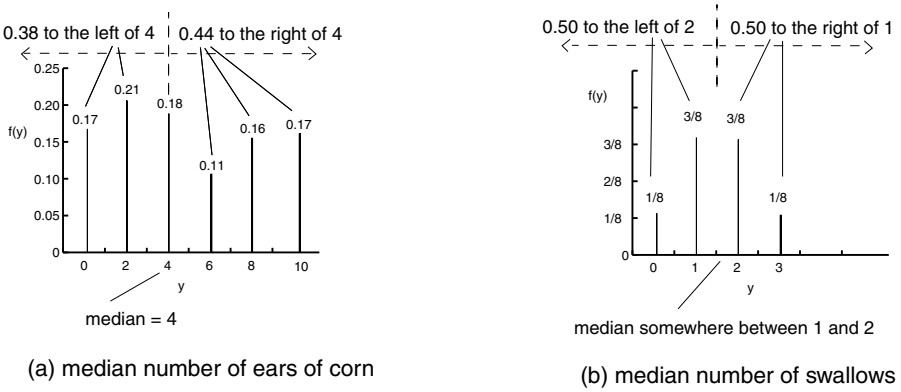


Figure 2.4 (Median for Discrete Probability Distributions)

(a) *Number of Ears of Corn.* The number of ears of corn, Y, on a typical corn plant has the following probability distribution.

y	0	2	4	6	8	10
$f(y)$	0.17	0.21	0.18	0.11	0.16	0.17

The median number of ears of corn, η, is equal to that number of ears of corn where there is a 50% chance of getting less than or equal to this number (and so a 50% of getting more than this number), in other words $\eta =$ (circle one)

i. 2 ears of corn (since there is a 17% chance of getting less than two 2 ears and a 62% chance of getting more–why?).

ii. between 2 and 4 ears of corn (since there is a 38% chance of getting less than 4 ears and a 62% chance of getting more than 2 ears of corn–why?).

iii. 4 ears of corn (since there is a 38% chance of getting less than 4 ears of corn and a 44% chance of getting more and so the 50–50 split has got be somewhere "in" 4 ears of corn)

(b) *Swallows.* The number of swallows, Y, in any group of three birds is given by the following probability distribution.

Y	0	1	2	3
$P(Y=y)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

The median number of swallows, η, is equal to that number of swallows where there is a $\frac{4}{8}$th chance of getting less than or equal to this number (and so a 50% of getting more than this number), in other words $\eta =$ (circle one)

i. 0 swallows (since there is a 0% chance of getting less than 0 swallows and a 7/8th chance of getting more than 0 swallows).

ii. 1 swallow (since there is a 1/8th chance of getting less than 1 swallow and a 4/8th chance of getting more than 1 swallow).

iii. between 1 and 2 swallows (since there is a 4/8th chance of getting less than 2 swallows and a 4/8th chance of getting more than 1 swallow).

iv. 2 swallows (since there is a 4/8th chance of getting less than 2 swallows and a 1/8th chance of getting more than 2 swallows).

(c) *Another Distribution.* If the distribution is

y	0	1	2	3
$P(Y=y)$	$\frac{4}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	$\frac{1}{8}$

The median of this distribution, η, is equal to that y where there is a $\frac{4}{8}$th chance of getting less than or equal to this number (and so a 50% of getting more than this number), in other words $\eta =$ (circle one)

i. $y = 0$

ii. between $y = 0$ and $y = 1$

iii. $y = 1$

iv. $y = 2$

3. *Percentiles, Discrete.* $\eta_p = y$ where $\Pr\{Y \leq y\} = p$.

(a) *Number of Ears of Corn.* The number of ears of corn, Y, on a typical corn plant has the following probability distribution.

y	0	2	4	6	8	10
$f(y)$	0.17	0.21	0.18	0.11	0.16	0.17

The 75th percentile (third quartile) number of ears of corn, $\eta_{0.75}$, is equal to that number of ears of corn where there is a 75% chance of getting less than or equal to this number (and so a 25% of getting more than this number), in other words $\eta_{0.75} =$ (circle one)

i. 2 ears of corn (since there is a 17% chance of getting less than 2 and 62% chance of getting more).

ii. between 2 and 4 ears of corn (since there is a 38% chance of getting less than 4 and 44% chance of getting more than 2).

iii. 6 ears of corn (since there is a 56% chance of getting less than 6 and 33% chance of getting more than 6).

iv. 8 ears of corn (since there is a 66% chance of getting less than 6 and 17% chance of getting more than 6 and so the 75–25 split must be somewhere "in" 8).

(b) *Swallows.* The number of swallows, Y, in any group of three birds is given by the following probability distribution.

Y	0	1	2	3
$P(Y = y)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

The 50th percentile (median, second quartile) number of swallows, $\eta_{0.50}$, is equal to that number of swallows where there is a $\frac{4}{8}$th chance of getting less than or equal to this number (and so a 50% of getting more than this number), in other words $\eta_{0.50} = \eta =$ (circle one)

i. 0 swallows

ii. 1 swallow

iii. between 1 and 2 swallows

iv. 2 swallows

(c) *Another Distribution.* If the distribution is

y	0	1	2	3
$P(Y = y)$	$\frac{4}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	$\frac{1}{8}$

The 25th percentile (first quartile) of this distribution, $\eta_{0.25}$, is equal to that y where there is a $\frac{2}{8}$th chance of getting less than or equal to this number (and so a 75% of getting more than this number), in other words $\eta_{0.25} =$ (circle one)

i. $y = 0$

ii. between $y = 0$ and $y = 1$

iii. $y = 1$

iv. $y = 2$

4. *Symmetry and Skewness For Mean and Median, Discrete: Number of Tablets.* Compare the mean and median for the following three discrete distribution tables of the number of tablets used in a high blood pressure experiment.

symmetric		skewed right		skewed left	
number of tablets, y	$f(y)$	number of tablets, y	$f(y)$	number of tablets, y	tablets, $f(y)$
1	2/15	1	5/15	1	1/15
2	3/15	2	4/15	2	2/15
3	4/15	3	3/15	3	2/15
4	3/15	4	2/15	4	4/15
5	2/15	5	1/15	5	6/15

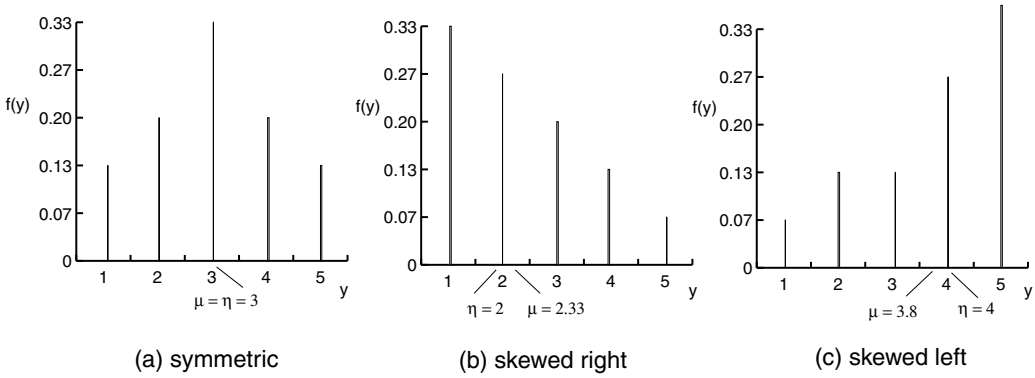


Figure 2.5 (Three Probability Distributions For Number of Tablets)

(a) **True** / **False** Data set (a) is symmetric because the histogram can be folded like a book, with its binding along a vertical centerline of the histogram, such that the left side of the histogram falls exactly on top of the right side of the histogram.

(b) Data set (b) is skewed right (or positively skewed) because there is (circle one) **less data on the right, than on the left** / **less data on the left, than on the right**.

(c) Data set (c) is skewed left (or negatively skewed) because there is (circle one) **less data on the right, than on the left** / **less data on the left, than on the right**.

(d) The mean of the symmetric data set (a) is
$\mu = 1 \times \frac{2}{15} + \cdots + 5 \times \frac{2}{15} =$ (circle one) **2.33** / **2.5** / **3.0** / **3.5**.
The median of symmetric data set (a) is equal to that y where there is a $\frac{7.5}{15}$th chance of getting less than or equal to this number (and so a 50% of getting more than this number), in other words
$\eta =$ (circle one) **2.33** / **2.5** / **3.0** / **3.5**

(e) The mean of the skewed right data set (b) is
$\mu = 1 \times \frac{5}{15} + \cdots + 5 \times \frac{1}{15} =$ (circle one) **2.33** / **2.5** / **3.0** / **3.5**.
The median of the skewed right data set (b) is equal to that y where there

is a $\frac{7.5}{15}$th chance of getting less than or equal to this number (and so a 50% of getting more than this number), in other words
$\eta =$ (circle one) **2.0** / **2.5** / **3.0** / **4.5**

(f) The mean of the skewed left data set (c) is
$\mu = 1 \times \frac{1}{15} + \cdots + 5 \times \frac{5}{15} =$ (circle one) **2.33** / **2.5** / **3.0** / **3.8**.
The median of skewed left data set (c) is equal to that y where there is a $\frac{7.5}{15}$th chance of getting less than or equal to this number (and so a 50% of getting more than this number), in other words
$\eta =$ (circle one) **2.33** / **2.5** / **3.0** / **4.0**

(g) For symmetric data, the average is (circle one) **bigger than**/ **about the same as** / **smaller than** the median.

(h) For data skewed *right*, the average is (circle one) **bigger than**/ **about the same as** / **smaller than** the median.

(i) For data skewed *left*, the average is (circle one) **bigger than**/ **about the same as** / **smaller than** the median.

5. *Percentiles, Continuous: Number of Foxes.* $\eta_p = y$ where $\Pr\{Y \leq y\} = p$.
The number of foxes, Y, taken at random from a protected wildlife preserve follows the following continuous probability distribution.

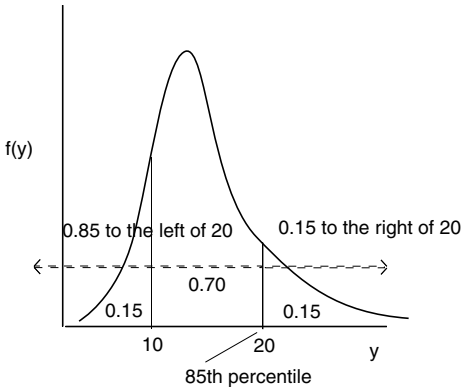


Figure 2.6 (Percentiles, Continuous Density)

(a) The 85th percentile number of foxes, $\eta_{0.85}$, is equal to that number of foxes where there is an 85% chance of getting less than or equal to this number (and so a 15% of getting more than this number), in other words $\eta_{0.85} =$ (circle one)

i. 10 foxes (since there is a 15% chance of getting less than or equal to 10 foxes and so an 85% chance of getting more–why?).

ii. between 10 and 20 foxes (since there is a 15% chance of getting less than 10 foxes and 15% of getting more than 20 foxes).

iii. 20 foxes (since there is an 85% chance of getting less than 20 foxes).

iv. more than 20 foxes (since there is less than a 15% chance of getting more than 20 foxes).

(b) The 15th percentile number of foxes, $\eta_{0.15}$, is equal to that number of foxes where there is a 15% chance of getting less than or equal to this number (and so an 85% of getting more than this number), in other words $\eta_{0.15} =$ (circle one)

i. 10 foxes

ii. between 10 and 20 foxes

iii. 20 foxes

iv. more than 20 foxes

(c) The 63rd percentile number of foxes, $\eta_{0.63}$, is equal to that number of foxes where there is a 63% chance of getting less than or equal to this number (and so a 27% of getting more than this number), in other words $\eta_{0.63} =$ (circle one)

i. 10 foxes

ii. between 10 and 20 foxes

iii. 20 foxes

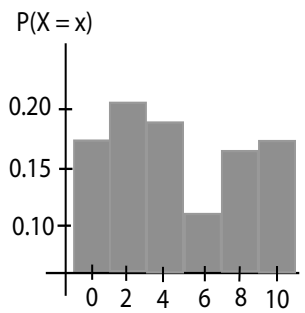
iv. more than 20 foxes

Exercise 2.6 (Dispersion Parameter: Standard Deviation, Coefficient of Variation, Empirical Rule and Covariance)

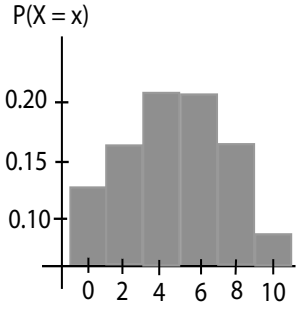
See TI–83 Lab 1: variance, covariance

1. *Variance and Standard Deviation.* $\sigma^2 = \sum(y-\mu)^2 f(y)$, $\sigma = \sqrt{\sigma^2}$

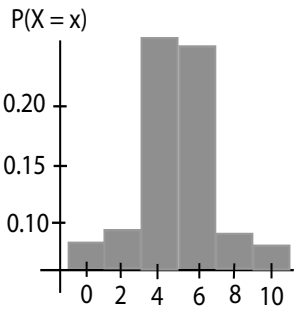
(a) *A First Look: Ears of Corn*



(a) # ears corn distribution



(b) another distribution



(c) and another distribution

Figure 2.7 (Variance Measures How "Dispersed" The Distribution Is)

If the variance (standard deviation) measures how "spread out" (or "dispersed") the distribution is, then the distribution for the number of ears of corn distribution (a) above, is (circle one) **more** / **as equally** / **less** dispersed than the other two distributions (b) and (c) above.
In other words, if ten (10) is "very" dispersed and zero (0) is not dispersed (concentrated at one point), then the variance for the ears of corn distribution seems close to (circle one) **0** / **7** / **10**

(b) *Number of Ears of Corn.* The number of ears of corn, Y, on a typical corn plant has the following probability distribution.

y	0	2	4	6	8	10
$f(y)$	0.17	0.21	0.18	0.11	0.16	0.17

and the mean (expected) number of ears of corn is given by $\mu = 4.78$, then the *variance* ($\mathrm{V}(Y)$) is given by, $\sigma^2 = \sum(y-\mu)^2 f(y)$,

$$\begin{aligned}\sigma^2 &= (y_1-\mu)^2 f(y_1) + (y_2-\mu)^2 f(y_2) + \cdots + (y_6-\mu)^2 f(y_6) \\ &= (0-4.78)^2(0.17) + (2-4.78)^2(0.21) + \cdots + (10-4.78)^2(0.17)\end{aligned}$$

which is equal to (circle one) **10.02** / **11.11** / **12.07** / **13.25**. The *standard deviation*, σ, (sometimes written as $\mathrm{SD}(Y)$ or SD) is given by

$$\sigma = \sqrt{12.07}$$

which is equal to (circle one) **3.47** / **4.11** / **5.07** / **6.25**.
(Use your calculator: as above, STAT ENTER; type Y, 0, 2, 4, 6 and 8, into L_1 and $P(X = x)$, 0.17, ..., 0.17, into L_2; then define $L_3 = (L_1 - 4.78)^2 \times L_2$; then STAT CALC ENTER 2nd L_3 ENTER; then read $\sum x = 12.07$ for the variance; $\sqrt{12.07} = 3.47$ gives the standard deviation.)

(c) *Swallows.* The number of swallows, Y, in any group of three birds is given by the following probability distribution.

Y	0	1	2	3
$P(Y=y)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

and the mean (expected) number of swallows is $\mu = 1.5$, then the *variance* is given by,

$$\sigma^2 = (0-1.5)^2\frac{1}{8} + (1-1.5)^2\frac{3}{8} + (2-1.5)^2\frac{3}{8} + (3-1.5)^2\frac{1}{8}$$

which is equal to (circle one) **0.02** / **0.41** / **0.59** / **0.75**. The *standard deviation* is given by $\sigma = \sqrt{0.75}$, which is equal to
(circle one) **0.47** / **0.87** / **1.07** / **2.25**.

(d) *Another Distribution.* Since the distribution is

Y	0	1	2	3
$P(Y=y)$	$\frac{4}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	$\frac{1}{8}$

and the mean is 0.875, then

$$\sigma^2 = (0-0.875)^2\frac{4}{8} + (1-0.875)^2\frac{2}{8} + (2-0.875)^2\frac{1}{8} + (3-0.875)^2\frac{1}{8}$$

which is equal to (circle one) **1.02** / **1.11** / **1.59** / **1.75**. The *standard deviation* (SD(Y)) is given by $\sigma = \sqrt{1.11}$, which is equal to (circle one) **1.05** / **1.77** / **2.07** / **2.25**.

(e) *And Yet Another Distribution.* Since the distribution is

$$P(Y=y) = \frac{3-y}{3}, \quad y = 1, 2,$$

and the mean is $\frac{4}{3}$, the variance is

$$\sigma^2 = \left(1-\frac{4}{3}\right)^2\frac{3-1}{3} + \left(2-\frac{4}{3}\right)^2\frac{3-2}{3},$$

which is equal to (circle one) $\mathbf{\frac{2}{9}}$ / $\mathbf{\frac{3}{9}}$ / $\mathbf{\frac{4}{9}}$ / $\mathbf{\frac{5}{9}}$. The *standard deviation* is given by $\sigma = \sqrt{\frac{2}{9}}$, which is equal to (circle one) **0.05** / **0.47** / **1.07** / **2.25**.

(f) *Standard Deviation for 0–1 Population: Pea Plants.* Since pea plants have either yellow peas, $y = 0$ or green peas, $y = 1$, according to the following probability distribution.

$$\Pr\{Y=y\} = \pi^y(1-\pi)^{1-y}, \quad y = 0, 1$$

where the mean is $\mu = \pi$, the variance is

$$\sigma^2 = (0-\pi)^2\pi^0(1-\pi)^{1-0} + (1-\pi)^2\pi^1(1-\pi)^{1-1},$$

which is equal to (circle one) $\mathbf{0}$ / $\boldsymbol{\pi}$ / $\mathbf{(1-\pi)}$ / $\boldsymbol{\pi(1-\pi)}$

(g) *Understanding the Standard Deviation.* Compare the standard deviation for the following three discrete distributions of the number of tablets used in a high blood pressure experiment.

large σ		middle σ		small σ	
number of tablets, y	$f(y)$	number of tablets, y	$f(y)$	number of tablets, y	tablets, $f(y)$
1	3/15	1	2/15	1	1/15
2	3/15	2	3/15	2	2/15
3	3/15	3	5/15	3	9/15
4	3/15	4	3/15	4	2/15
5	3/15	5	2/15	5	1/15

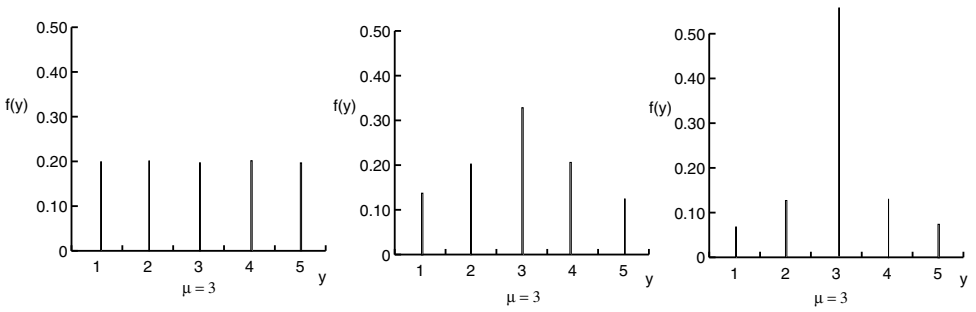


Figure 2.8 (Understanding the Standard Deviation)

As the standard deviation becomes smaller, the distribution becomes (circle one) **broader** / **narrower**. This is also true for continuous distributions.

2. *Variances and Covariance: Number of Ears of Corn, Two Varieties.* The number of ears of corn on two varieties (variety A and variety B) of corn plants, Y_1 and Y_2, follows the following probability distribution.

y_1	0	0	0	1	1	1
y_2	0	1	2	0	1	2
$f(y_1, y_2)$	0.17	0.21	0.18	0.11	0.16	0.17

(a) Since the mean number of ears of corn for Variety A is $\mu_1 = 0.44$, the variance of variety A is

$$\begin{aligned}\sigma_1^2 &= (0-0.44)^2 \times \Pr\{Y_1 = 0\} + (1-0.44)^2 \times \Pr\{Y_1 = 1\} \\ &= (0-0.44)^2 \times 0.56 + (1-0.44)^2 \times 0.44\end{aligned}$$

which is equal to (circle one) **0.25** / **0.44** / **0.56** / **1**.

(b) Since the mean number of ears of corn for Variety B is $\mu_2 = 1.07$, the variance of variety B is

$$\begin{aligned}\sigma_2^2 &= (0-1.07)^2 \times \Pr\{Y_2 = 0\} + (1-1.07)^2 \times \Pr\{Y_2 = 1\} + (2-1.07)^2 \times \Pr\{Y_2 = 2\} \\ &= (0-1.07)^2 \times 0.28 + (1-1.07)^2 \times 0.37 + (2-1.07)^2 \times 0.38\end{aligned}$$

which is equal to (circle one) **0.25** / **0.44** / **0.65** / **1**.

(c) Since $\mu_1 = 0.44$ and $\mu_2 = 1.07$, the covariance σ_{12} is

$$\begin{aligned}\sigma_{12} &= \sum (y_1 - \mu_1)(y_2 - \mu_2) f(y_1, y_2) \\ &= (0-0.44)(0-1.07) \times f(0,0) + (0-0.44)(1-1.07) \times f(0,1) + \cdots + (1-0.44)(2-1.07) \times f(1,2) \\ &= (0-0.44)(0-1.07) \times 0.17 + (0-0.44)(1-1.07) \times 0.21 + \cdots + (1-0.44)(2-1.07) \times 0.17\end{aligned}$$

which is equal to (circle one) **0.15** / **0.44** / **0.56** / **1.07**.
(To use your calculator: STAT EDIT, type y_1, y_2 and $f(y_1, y_2)$ in lists L_1, L_2 and L_3, respectively, define $L_4 = (L_1 - 0.44)(L_2 - 1.07) \times L_3$, ENTER, then STAT CALC 1:1–Var Stats L_4, ENTER, then read $\sum x$.)

(d) **True** / **False**

$$\sigma_{12} = \sigma_{21}, \quad \sigma_2^2 = \sigma_{22}$$

3. *Coefficient of Variation.* $\text{CV} = \frac{\sigma}{|\mu|}$.

(a) *Number of Ears of Corn.* Since the number of ears of corn, Y, on a typical corn plant has the following probability distribution.

y	0	2	4	6	8	10
$f(y)$	0.17	0.21	0.18	0.11	0.16	0.17

where $\mu = 4.78$ and $\sigma = 3.47$, then the coefficient of variation is given by $\text{CV} = \frac{\sigma}{|\mu|} =$ (circle one) **0.35** / **0.73** / **0.94** / **1.25**.

(b) *And Yet Another Distribution.* Since the distribution is

$$P(Y = y) = \frac{3 - y}{3}, \quad y = 1, 2,$$

where $\mu = \frac{4}{3}$ and $\sigma = 0.47$, then the coefficient of variation is given by $\text{CV} = \frac{\sigma}{|\mu|} =$ (circle one) **0.35** / **0.73** / **0.94** / **1.25**.

(c) *Understanding the coefficient of variation.* The mean, standard deviation (SD) and coefficient of variation (CV) of the price of a television and car are given below.

parameter	television	car
mean, μ	\$234	\$17,345
SD, σ	\$23	\$984
CV, $\frac{\sigma}{\lvert\mu\rvert}$	0.098	0.057

Even though the SD is *smaller* for the television than for the car, the CV is *larger* for the television than for the car. This means (circle none, one or more)

i. the television prices are more variable than the car prices.

ii. the television prices are less variable than the car prices.

iii. even though the television prices are less variable than the car prices, *relative to the amount of money involved*, the television prices are, in fact, more variable than the car prices.

4. *Empirical Rule: Salinity Values.* The rules given here specify how much (or what percentage) of the area under a *normal* distribution (discussed in greater detail a little bit latter on) is between one, two or three standard deviations away from the mean. These three rules are often, collectively, called the *Empirical Rule* (or Rule of Thumb).

   - *Roughly* 68% of the data observations *should be* within *one* SD of the mean.
   - *Roughly* 95% of the data observations should be within *two* SDs of the mean.
   - *Roughly* 99.7% of the data observations should be within *three* SDs of the mean.

   The following twenty–eight numbers are salinity values for water specimens taken from North Carolina's Pamilco Sound.

4.3	5	5.9	6.5	7.6	7.7	7.7	8.2	8.3	9.5
10.4	10.4	10.5	10.8	11.5	12	12	12.3	12.6	12.6
13	13.1	13.2	13.5	13.6	14.1	14.1	15.1		

   (a) (Review) The salinity values is probably an example of a (circle one) **population** / **sample** / **statistic** / **parameter**. If the water specimens are chosen at random, then this sample of salinity values (circle one) **is** / **is not** representative of the entire population of salinity values.

   (b) One particular salinity value from the twenty–eight above is (circle one) **10.7** / **11.8** / **12**.

   (c) **True** / **False** The average, 10.55, is determined by adding all twenty–eight salinity values together and dividing by 28, by using ave $= \frac{4.3+\cdots+15.1}{28}$. This sample average, 10.55, does *not* necessarily equal the population mean, μ, but we will approximate the mean by the average, $\mu \approx 10.55$.
   (Calculator: type STAT EDIT then type 34 sanity values in L_1, then STAT CALC 1:1–Var Stats L_1 and read off $\overline{x} = 10.55$.)

   (d) **True** / **False** The standard deviation, 3.01, is calculated using SD $= \sqrt{\frac{(4.3-10.55)^2+\cdots+(15.1-10.55)^2}{28}}$. This sample SD, 3.01, does *not* necessarily equal the population SD, σ, but we will approximate the population SD by the sample SD, $\sigma \approx 3.01$.
   (Calculator: type STAT EDIT then type 34 sanity values in L_1, then STAT CALC 1:1–Var Stats L_1 and read off $\overline{x} = 10.55$.)

   (e) The value *one SD above the average* is given by
   $\mu + \sigma = 10.55 + 3.01 =$ (circle one) **10.55** / **13.56** / **16.57**.

   (f) The value one SD above the average, 13.56, (circle one) **is** / **is not** one of the twenty–eight salinity values given above.

(g) The salinity value, "12", taken from the twenty–eight above, is (circle one) **more than** / **the same as** / **less than** the value which is one SD above the average, 13.56.

(h) The value *one SD below the average* is given by $\mu - \sigma = 10.55 - 3.01 =$ (circle one) **7.54** / **10.55** / **13.56**.

(i) The value one SD below the average, 7.54, (circle one) **is** / **is not** one of the twenty–eight salinity values given above.

(j) The salinity value, "12", taken from the twenty–eight above, is (circle one) **more than** / **the same as** / **less than** the value which is one SD below the average, 7.54.

(k) **True** / **False** Since the salinity value "12" is both more than one SD below the average, 7.54, and less than one SD above the average, 13.56, then "12" is said to be *within one SD of the average*. The salinity value "12" is between 7.54 and 13.56 or, in other words, *inside* the interval $(\mu - \sigma, \mu + \sigma) = (7.54, 13.56)$.

(l) The value "two SDs above the average" is given by $\mu + 2\sigma = 10.55 + 2(3.01) =$ (circle one) **10.55** / **13.56** / **16.57**.

(m) The value "two SDs below the average" is given by $\mu - 2\sigma = 10.55 - 2(3.01) =$ (circle one) **4.53** / **7.54** / **16.57**.

(n) The interval of values *within two SDs of the average* is given by $(\mu - 2\sigma, \mu + 2\sigma) =$
(circle one) **(4.53, 10.55)** / **(7.54, 13.56)** / **(4.53, 16.57)**.
A diagram of the values which are one, two and three SDs from the average is given below.

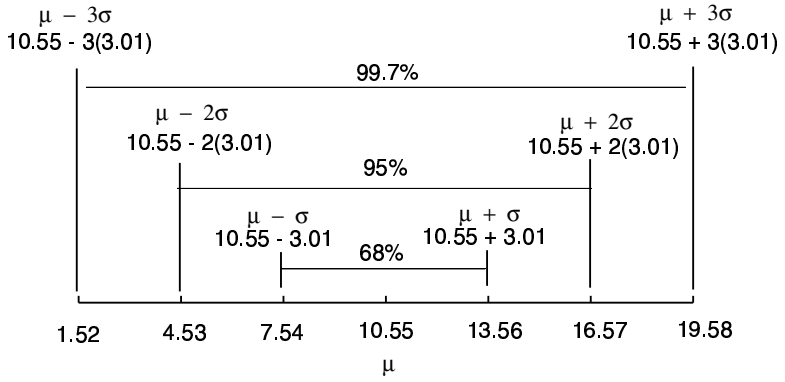


Figure 2.9 (Salinity Values 1, 2 and 3 Standard Deviations From Mean)

(o) According to the Empirical rule, about 68% or $0.68 \times 28 = 19.04 \approx 19$ values *should be* within *one* standard deviation of the average, or, in other words, in the interval (7.54,13.56). In fact, there are

(circle one) **19** / **20** / **21** salinity values in this interval, or a percentage of $\frac{20}{28} \times 100 =$ (circle one) **67%** / **71%** / **75%**.

(p) According to the Empirical rule, about 95% or $0.95 \times 28 = 26.6 \approx 27$ values *should be* within *two* SDs of the average, or, in other words, in the interval (4.53,16.57). In fact, there are
(circle one) **27** / **28** / **29** salinity values in this interval, or a percentage of $\frac{27}{28} \times 100 =$ (circle one) **95%** / **96%** / **99%**.

(q) According to the Empirical rule, about 99.7% or $0.997 \times 28 \approx 28$ values *should be* within *three* SDs of the average, or, in other words, in the interval (1.52,19.58). In fact, there are
(circle one) **27** / **28** / **29** salinity values in this interval, or a percentage of $\frac{28}{28} \times 100 =$ (circle one) **95%** / **96%** / **100%**.

(r) **True** / **False** Since 68% of data should fall within one SD of the average, a data point found *out*side this interval could be considered a "moderate" outlier; a data point found *in*side this interval would be considered "about average". A data point outside *two* SDs from the average would be considered a "strong" outlier.

5. *Population, Sample, Statistic and Parameter: Ears of Corn.* The number of ears of corn, Y, on a typical corn plant has the following probability distribution.

y	0	2	4	6	8	10
$f(y)$	0.17	0.21	0.18	0.11	0.16	0.17

where $\mu = 4.78$ and $\sigma = 3.47$.

(a) The probability distribution is what we (circle one) **expect** / **observe**.

(b) Since the probability distribution is associated with the population, the mean, $\mu = 4.78$, is a (circle one) **parameter** / **statistic**.

(c) The standard deviation, $\sigma = 3.47$, is a (circle one) **parameter** / **statistic**.

(d) If we *observed* the following data for 100 corn plants sampled at random from all corn plants,

number of ears of corn per plant	0	2	4	6	8	10
number of plants	17	21	18	11	16	17

the *average* number of ears of corn per plant would be

$$\overline{y} = \frac{17(0) + 21(2) + 18(4) + 11(6) + 16(8) + 17(10)}{100}$$

which is equal to (circle one) **4.32** / **4.78** / **5.50** / **5.75**.

(e) The average number of ears of corn per plant, $\overline{y} = 4.78$ is a (circle one) **parameter** / **statistic**.

(f) Since the observed average, $\overline{y} = 4.78$ is equal to the *expected* mean, $\mu = 4.78$, we would tend to (circle) **accept** / **reject** our guess that $\mu = 4.78$.

(g) **True** / **False**. *If* there were only 100 plants in our population, then $\overline{y} = \mu$.