Chapter 3

Statistical Inference: Basic Concepts

3.1 Introduction

We look at basic concepts necessary to understand statistical inference.

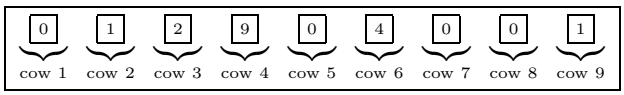
3.2 Simple Random Samples

It is important that data be sampled to be *representative* of the population, that a *random sample* be taken. We first look at a random numbers table. We use the random numbers table to perform simple random sampling.

See TI–83 Lab 3: Random Numbers Generator.

Exercises 3.1 (Simple Random Sample (SRS)) The *simple random sampling* (SRS) procedure[1] is a method of selecting n units out of N population units such that every one of the distinct samples has an *equal* chance of being drawn. It is in this way the sample is assured to be *representative* of the population and for there to be *no* bias in the sample.

1. *Milk Production, Nine Cow Population, Using the Random Numbers Table.* The box model related to the *population* of the quantity (in quarts) of daily milk produced by *nine* cows is given below.



[1]There are two types of SRS: with replacement and without replacement. The procedure described here is the more common simple random sampling with*out* replacement procedure.

(a) Cow 5 has a daily milk production of (circle one) **0** / **1** / **2** quarts.

(b) The mean quantity of daily milk production per cow for all cows is
mean $= \frac{0+1+2+9+0+4+0+0+1}{9} \approx$ (circle one) **1.7** / **1.8** / **1.9**.
The mean (1.9, in this case) is an example of a
(circle one) **statistic** / **parameter**.

(c) The mean quantity of daily milk production per cow for all cows is (circle one)

i. the sample average quantity of daily milk production per cow.

ii. the population (true or actual) average quantity of daily milk production per cow.

(d) Use the random numbers table,

rows ↓										
1	14458	66140	47281	36282	61973	36103	53684	15740	26906	77123
2	51186	01079	67704	81649	59776	70077	16643	36900	27752	16201

to draw five tickets (*out* replacement) out the box model. Start at row 1, column 1, move left to right along the first row and record your findings in the table below.

random number (cow)	1	4	_____	8	6
quantity of milk	0	9	0	_____	_____

(e) The estimated average quantity of milk production per cow for the five chosen cows, is given by,
(observed) mean $= \frac{0+9+0+0+4}{5} = \frac{13}{5} =$ (circle one) **2.3** / **2.6** / **2.9**.
This observed mean is an example of a
(circle one) **statistic** / **parameter**.

(f) Each time we draw (*out* replacement) five tickets from the box model, the expected mean (circle one) **changes** / **remains the same**, whereas the *observed* mean probably (circle one) **changes** / **remains the same**. The *observed* mean of five tickets drawn from the box model (circle one) **always equals** / **varies around** the *expected* mean of five tickets drawn from the box model.

(g) (Review.) Match the statistical terms with this milk production example.

terms	milk production example
(i) population	(i) milk production of nine cows
(ii) sample	(ii) average milk production, among few cows chosen
(iii) statistic	(iii) milk production of few cows chosen
(iv) parameter	(iv) average milk production, among nine children

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

(h) **True / False** In this case, we could have easily determined the true mean milk production of the nine cows without having to estimate this average based on a small sample. In a "real" situation, the population would be much larger, all American cows, say, and so, most likely, we would *not* be able to determine the exact average. It would be for this reason, we would be "forced" into estimating the true average based on a sample of American cows. This small population example, however, serves to not only indicate how good the SRS sampling method is, but also to demonstrate the relationships between population, sample, parameter and statistic.

2. *A SRS Using The TI–83 Calculator.* Use your calculators (with *seed* 7) to pick a simple random sample (SRS) of size 8 from a population numbered 10, 11, . . . , 550. The random numbers chosen are (circle one)

   (a) 128, 546, 322, 163, 446, 331, 313, 201

   (b) 129, 546, 322, 163, 446, 331, 313, 201

   (c) 130, 546, 322, 163, 446, 331, 313, 201

   (d) 131, 546, 322, 163, 446, 331, 313, 201

   (First input the seed 7 by typing 7 STO MATH PRB 1:rand ENTER. Generate the 8 random numbers between 10 to 550 by typing MATH PRB 5:randInt(10,550,8) ENTER.)

3. *Milk Production, Twenty Cow Population, Using The TI–83 Calculator.* A small *population* of the milk production for 20 cows is given below.

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

   (a) The mean quantity of daily milk production per cow for all cows is (expected) ave $= \frac{0+1+2+\cdots+2}{20} \approx$ (circle one) **2.1** / **2.3** / **2.5**.
   The average (2.1, in this case) is an example of a
   (circle one) **statistic** / **parameter**.

   (b) The mean quantity of daily milk production per cow for all cows is (circle one)

      i. the sample average quantity of daily milk production per cow.

      ii. the population (true or actual) average quantity of daily milk production per cow.

(c) Use your calculators (with *seed* 10!) to draw five tickets (with*out* replacement) out the box model and record your findings in the table below.

cow in SRS	10	12	_____	14	_____
milk production	5	0	_____	3	_____

(First input the seed 10 by typing 10 STO MATH PRB 1:rand ENTER. Generate the 5 random numbers between 1 to 20 by typing MATH PRB 5:randInt(1,20,5) ENTER.)

(d) The estimated average quantity of milk production per cow for the five chosen cows, is given by,
(observed) ave $= \frac{5+0+4+3+0}{5} = \frac{12}{5} =$ (circle one) **1.4** / **2.0** / **2.1** / **2.4**.
This observed average is an example of a
(circle one) **statistic** / **parameter**.

(e) Each time we draw (with*out* replacement) five tickets from the box model, the expected average (circle one) **changes** / **remains the same**, whereas the *observed* average probably (circle one) **changes** / **remains the same**. The *observed* average of five tickets drawn from the box model (circle one) **always equals** / **varies around** the *expected* average of five tickets drawn from the box model.

(f) It probably (circle one) **would** / **would not** be difficult to enumerate *all* cows in the United States and so it would be difficult to use this method to choose a simple random sample (SRS). Other methods (not covered in this course) such as multistage cluster sampling, would be required.

3.3 Describing Samples

Frequency distribution tables are tables which organize data into classes. Histograms are pictures of frequency distribution tables. Stem and leaf plots are another way of presenting data in graphical form.

See TI–83 Lab 3: Histograms and Stem and Leaf Plots.

We also look at *summary* numbers associated with frequency distributions, including the mean, median and standard deviation. When discussing the median, mention is made of percentiles and the box–plot.

Exercise 3.2 (Frequency Distribution Table and Histogram)

1. *A First Look: Age.* Twenty patients in a high blood pressure study have the following ages.

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

(a) The frequency (number) of patients between the ages of 35 and up but not including 40 (class interval 35 to 40) is (circle one) **2** / **3** / **4** / **5**.

(b) The *width* of class interval 35 to 40 is (circle one) **2** / **3** / **4** / **5** years.

(c) The *width* of class interval 40 to 45 is $45 - 40 =$
(circle one) **2** / **3** / **4** / **5** years.

(d) The *width* of class interval 41 to 44 is $44 - 41 =$
(circle one) **2** / **3** / **4** / **5** years.

(e) The *proportion of patients in the five years* in the class interval 40 to 45 is (circle one) $\mathbf{\frac{6}{20} = 0.30}$ / $\mathbf{\frac{7}{20} = 0.35}$ / $\mathbf{\frac{8}{20} = 0.40}$ / $\mathbf{\frac{9}{20} = 0.45}$.

(f) The *proportion of patients in the five years* in the class interval 50 to 55 is (circle one) $\mathbf{\frac{1}{20} = 0.05}$ / $\mathbf{\frac{2}{20} = 0.10}$ / $\mathbf{\frac{3}{20} = 0.15}$ / $\mathbf{\frac{4}{20} = 0.20}$.

(g) **True** / **False** If the proportion of patients in a class interval of width *five* years is 0.35, then the proportion of patients per *one* year in this class interval is $\frac{0.35}{5} = 0.07$.

(h) **True** / **False** If the proportion of patients in a class interval of width *ten* years is 0.35, then the proportion of patients per one year in this class interval is $\frac{0.35}{10} = 0.035$.

(i) If the proportion of patients in a class interval of width *seven* years is 0.35, then the proportion of patients per one year in this class interval is (circle one) $\mathbf{0.35 \times 7 = 2.45}$ / $\mathbf{\frac{7}{0.35} = 20}$ / $\mathbf{\frac{0.35}{7} = 0.05}$ / $\mathbf{\frac{0.35}{10} = 0.035}$.

(j) **True** / **False** If the *proportion* of patients per one year in some class interval is 0.07, then the *percentage* of patients per one year $0.07 \times 100 = 7$ percent.

2. *Frequency Distribution Table.* Consider the following incomplete distribution table for the age data,

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

class interval	frequency	relative frequency	proportion per 5 years	proportion per 1 year	% per 1 year
30 to 35	1	$\frac{1}{20} = 0.05$	$\frac{1}{20} = 0.05$	$\frac{0.05}{5} = 0.01$	1
35 to 40	2	$\frac{2}{20} = 0.10$	(c)	$\frac{0.10}{5} = 0.02$	2
40 to 45	8	(a)	$\frac{8}{20} = 0.40$	$\frac{0.40}{5} = 0.08$	8
45 to 50	7	$\frac{7}{20} = 0.35$	$\frac{7}{20} = 0.35$	$\frac{0.35}{5} = 0.07$	7
50 to 55	2	(b)	(d)	(e)	2
total	20	1.0	1.0		

(a) Complete the distribution table by filling in the following table.

(a)	(b)	(c)	(d)	(e)

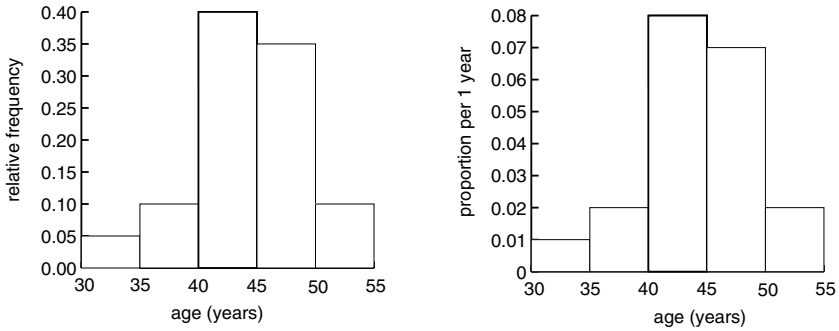
(b) The first class interval is (circle one) **30 to 35** / **30 to 40** / **40 to 45**.

(c) The number of class intervals is (circle one) **3** / **4** / **5** / **6**.

(d) The width of each class interval is (circle one) **3** / **4** / **5** / **6** years.

(e) **True** / **False** The class intervals given here are the *only* possible class intervals that could have been used for this data. For example, it would not be possible to have, instead, class intervals of *un*equal width, such as "30 to 40", "40 to 45" and "45 to 55", say.

3. *Histogram.* Use your calculators to help draw the two possible graphs given below for the age data. Notice that both the "relative frequency" graph and the "proportion per 1 year" graph have the same *shape*.



(a) not-really histogram (b) proper histogram

Figure 3.1 (Two Possible Graphs For The Age Data)

(STAT EDIT 32.5, 37.5, 42.5, 47.5, 52.5 into L_1; and 0.01, 0.02, 0.08, 0.07, 0.02 into L_2, then 2nd Y= ON, choose histogram, then WINDOW 30, 55.5, -0.05, 0.1, 0.1, 1 and then GRAPH to get the "final" histogram. Type TRACE to see the frequency in each class.)

(a) The *total* area of all the vertical bars in the "not–really" histogram (circle one) **is less than one** / **equal to one** / **greater than one**. Hint: $(35-30)\times 0.05+\cdots+(55-50)\times 0.10=5$

(b) The *total* area of all the vertical bars in the "proper" histogram is (circle one) **less than one** / **equal to one** / **greater than one**. Hint: $(35-30)\times 0.01+\cdots+(55-50)\times 0.02=1$

(c) The proportion of ages in the 30 to 40 class interval is the proportion in the 30 to 35 class interval plus the proportion in the 35 to 40 class interval,

or, in other words, equal to the total area in the two vertical bars in these two classes, or
$(5 \times 0.01) + (5 \times 0.02) =$ (circle one) **0.05** / **0.10** / **0.15**.

(d) The proportion of ages in the 35 to 50 class interval is equal to the total area in the vertical bars for the three class intervals, 35 to 40, 40 to 45 and 45 to 50, or
$(5 \times 0.02) + (5 \times 0.08) + (5 \times 0.07) =$ (circle one) **0.80** / **0.85** / **0.90**.

(e) The proportion of ages in the 35 to 37 portion of the 35 to 40 class interval is equal to the area of the portion of the vertical bar associated with 35 to 37. Since the width of the 35 to 37 portion is $37 - 35 = 2$, and the height of the 35 to 40 class interval is 0.02, the area must be
$(2 \times 0.02) =$ (circle one) **0.03** / **0.04** / **0.05**.

Exercise 3.3 (Stem and Leaf Plot)

1. *Age Data.*
The ages of the twenty patients in a high blood pressure study are given by,

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

A stem and leaf plot of this data is given as follows,

3	2	7	9*													
4	0	1	1	1**	2	2	3	4	5	5	5	6	7	7	9	stem: 10s
5	0															leaf: 1s

Table 3.1 (Stem–and–Leaf For Age Data)

(a) The starred number, **9***, represents the age (circle one) **39** / **93** / **9**. The double–starred number, 1**, represents the age (circle one) **41** / **14** / **1**.

(b) All of the numbers to the left of double line (in the first column) are called (circle one) **stems** / **leaves**; all of the numbers to the right of this double line are called (circle one) **stems** / **leaves**.

(c) The starred number 9* is a leaf of the stem (circle one) **3** / **4** / **5**.

(d) **True** / **False** The note to the right of the stem–and–leaf plot specifies the numbers used as stems are "tens" (or "10s") and the numbers used as leaves are "ones" (or "1s"). So, for instance, the stem "3" represents 3x10 = 30 and the leaf "2" represents 1x2 = 2.

(e) The stem–and–leaf plot is ordered where, in the first stem, for example, 32 is followed by (circle one) **37** / **39** / **40**.

(f) A stem–and–leaf plot is useful in identifying the center of the data set, or, where "most" data values are located. In this case, this is in the (circle one) **30**s / **40**s / **50**s stem.

2. *Split Stem and Leaf Plots: Age Data.*
The ages of the twenty patients in a high blood pressure study are given by,

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

Sometimes, to spread the data out, the stems are *split* as, for example, in the following table.

3L	2								
3H	7	9							
4L	0	1	1	1	1	2	2	3	4
4H	5	5	5	6	7	7	7	9	
5L	0								
5H									

stem: 10s
leaf: 1s

Table 3.2 (Stem and Leaf For Age Data With Split Stems)

(a) Notice how all stems have each been split into two.
Stem 3 is split into low stem 3, or (circle one) **3L** / **3H**,
and high stem 3, or (circle one) **3L** / **3H**.

(b) **True** / **False** Low stem 3L contains one half of the leaves, 0, 1, 2, 3 or 4; high stem 3H contains the other half of the leaves, 5, 6, 7, 8 and 9.

(c) **True** / **False** Stem and leaf plots can have stems split not only twice, but also three or more times. A stem–and-leaf plot with each stem split three times might, say, consist of a low stem which contains leaves 0, 1 and 2; a middle stem with leaves 3, 4, 5 and 6 and a high stem with leaves 7, 8 and 9.

(d) A stem and leaf plot with 10s as stems can be split at most
(circle one) **5** / **7** / **10** / **100** times.

(e) **True** / **False** Although there is no one "best" way of constructing a stem–and–leaf plot, most stem–and–leaf plots consist of 5 to 20 stems.

Exercise 3.4 (Sample Mean, Median, Standard Deviation, Five Number Summary and Box–Plots)

See TI–83 Lab 3: summary statistics

1. *Sample Mean and Median: Temperatures.* Consider the following ordered sample of temperatures taken from 10 different locations in Westville during the second day in January of a recent year:

$$0,\ 0,\ 0,\ 1,\ 1,\ 2,\ 2,\ 3,\ 3,\ 4.$$

   (a) The average is, since there are $n = 10$ temperatures and

$$y_1 = 0, y_2 = 0, y_3 = 0, y_4 = 1, y_5 = 1, y_6 = 2, y_7 = 2, y_8 = 3, y_9 = 3, y_{10} = 4$$

   Then:

$$\begin{aligned} \bar{y} &= \frac{y_1 + y_2 + y_3 + y_4 + y_5 + y_6 + y_7 + y_8 + y_9 + y_{10}}{n} \\ &= \frac{\sum_{i=1}^{10} y_i}{n} \end{aligned}$$

   and so $\bar{y} = \frac{0+0+0+1+1+2+2+3+3+4}{10} =$ (circle one) **1.0** / **1.6** / **1.7**.
   (Use your calculator: STAT ENTER; then type the ten temperatures, 0,0,...,4 into L_1; then calculate average ($\bar{y}$) by typing STAT CALC ENTER 2nd L_1 ENTER; read "$\bar{y} = 1.6$".)

   (b) **True** / **False** *Review.* The mean (or average) calculated for a *sample* is called a *statistic*; the mean for a *population* is called a *parameter*. The ten temperatures here could be considered a sample of all the possible locations in Westville and so, in this case, the average, 1.6 degrees, is the value of a statistic and not a parameter.

   (c) The *median* is the center value of the 10 ordered set of temperatures. Since there are an *even* number of observations, $n = 10$, in this case, there is *no* center value, or, in other words, no observation where there is the same number of observations both above and below this value. Consequently, the sample median is set equal to the average of the center *two* observations (circle one) **1.5** / **1.6** / **1.7**.
   (Use your calculator: as above, but, now, arrow down to "Med = 1.5".)

   (d) If there had been an *odd* number of observations, then the center value would have been used as the sample median. For example, if the data set had consisted of the *nine* values,

$$0, 0, 0, 1, 1, 2, 2, 3, 3,$$

   then the sample median would have been
   (circle one) **1.0** / **1.6** / **1.7**.

(e) **True** / **False** The *location* of the median can be calculated using the $\frac{n+1}{2}$ formula. For example, if there are 9 observations, then the $\frac{9+1}{2} =$ 5th ordered observation is the median. If there are 10 observations, then the $\frac{10+1}{2} = 5.5$th ordered observation (average of 5th and 6th ordered observations) is used as the median value.

2. *Average Sensitive to Outliers; Median Robust to Outliers: Temperatures.* Consider the following ordered sample of temperatures taken from 10 different locations in Minneapolis during a cold day in January where, because of a typing mistake, the last temperature, 4 degrees, is mistakenly recorded as 40 degrees:

0, 0, 0, 1, 1, 2, 2, 3, 3, 40.

(a) The average is (circle one) **1.5** / **1.6** / **5.2**.

(b) The median is (circle one) **1.5** / **1.6** / **5.2**.

(c) The present average is
(circle one) **much bigger** / **about the same** / **much smaller** than the average calculated for the ten temperatures, 0, 0, 0, 1, 1, 2, 2, 3, 3, 4.

(d) The present *median* is
(circle one) **much bigger** / **about the same** / **much smaller** than the median calculated for the ten temperatures, 0, 0, 0, 1, 1, 2, 2, 3, 3, 4.

(e) The average is said to be (circle one) **sensitive** / **robust** to outliers, whereas the median is said to be (circle one) **sensitive** / **robust** to outliers.

3. *Standard Deviation.* Consider the algae weights (in ounces) given in the following three shipments of six containers each:

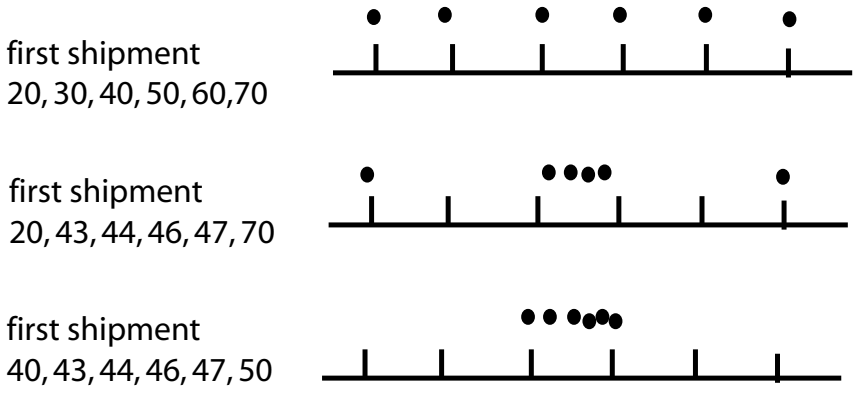


Figure 3.2 (Three Shipments of Six Containers of Algae Data)

It is often useful to have a *single* number which summarizes the fact that, for example, the weights for the first shipment are more variable (more "spread out") than the weights for the third shipment.

(a) It is fairly clear from the three diagrams above that the algae for the first shipment are (circle one) **more** / **less** spread out (or variable) than the algae for the third shipment.

(b) If a single number, called the *standard deviation*, is large when a data set was spread out and small when a data set is jammed together, then the standard deviation is largest for shipment (circle one) **one** / **two** / **three**.

(c) *Standard Deviation Using TI–83.* The *standard deviation*, denoted s, of the first shipment is given by:

$$\begin{aligned} s &= \sqrt{\frac{\sum_{i=1}^{6}(y_i - \bar{y})^2}{n-1}} \\ &= \sqrt{\frac{(20-45)^2 + (30-45)^2 + \cdots + (70-45)^2}{6-1}} \\ &= \end{aligned}$$

(circle one) **3.46** / **15.87** / **18.71**.
(Type STAT ENTER; then type the six algae weights into L_1; then STAT CALC ENTER 2nd L_1 ENTER; then read $s_x = 18.71$)
The sample standard deviation for the second shipment of weights is
(circle one) **3.46** / **15.87** / **18.71**.
and for the third shipment is
(circle one) **3.46** / **15.87** / **18.71**.

(d) Since s for the first shipment is
(circle one) **larger than** / **about the same as** / **smaller than** the s for the third shipment, this means the variability in algae weights for the first shipment is
(circle one) **larger than** / **about the same as** / **smaller than** the variability in tire weights for the third shipment.

(e) The formula for the *variance* is given by

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

and so the standard deviation is the (circle one) **square** / **square root** of the variance.

(f) *Review.* The three SDs in the algae weights for the three shipments are probably all examples of (circle one) **populations** / **samples** / **statistics** / **parameters**. The three true (or actual) SDs in weights for all of the algae in the three shipments are all examples of (circle one) **populations** / **samples** / **statistics** / **parameters**.

4. *Five Number Summary: Temperatures.* Reconsider the following sample of 10 temperatures,

$$0, 0, 0, 1, 1, 2, 2, 3, 3, 4.$$

See TI–83 Lab 2: summary statistics

(a) The 25th percentile has a special name called the lower (or first) quartile and denoted $y^{(1q)}$. Similarly, the 50th percentile is often called
(circle one or more) **middle** / **second** / **upper** / **third** quartile.
Finally, the 75th percentile is often called the
(circle one or more) **middle** / **second** / **upper** / **third** quartile and denoted $y^{(3q)}$.

(b) The *five–number summary* is given by,

minimum, lower quartile, median, upper quartile, maximum,

and so, for the 10 temperatures, the five number summary
(circle one) **is** / **is not** {0, 0, 1.5, 3, 4}.
(Use your calculator: STAT ENTER; type 0, 0, ..., 4, into L_1; then STAT CALC ENTER 2nd L_1 ENTER; then arrow down to read minX = 0, $Q_1 = 0$, Med = 1.5, $Q_3 = 3$, maxX = 4.)

(c) The *interquartile range* is equal to the upper quartile minus the lower quartile. Consequently, the interquartile range for the set of 10 temperatures is given by, $3 - 0 =$ (circle one) **1** / **2** / **3**.

(d) The interquartile range is robust to outliers whereas the variance and standard deviation are both sensitive to outliers. This means, for example, if one of the 10 temperatures above was mistyped as 40, instead of 4, the interquartile range would
(circle one) **not change much** / **would change a lot**,
but both the variance and standard deviation would
(circle one) **not change much** / **change a lot**.

5. *Box–Plot: Ph Levels Of Soil.* Consider the ordered set of the Ph levels of soil data given below.

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		

See TI–83 Lab 3: box–plot plots

(a) Six pieces of information are required to draw a box and whisker plot: median, upper and lower quartiles, the interquartile range, upper and lower fences. The lower quartile, median and upper quartile are 7.95, 11.15 and 13.05, respectively and, so, the interquartile range is
$13.05 - 7.95 =$(circle one) **4.95** / **5.10** / **6.15**.

(b) The "upper fence" is determined by adding 1.5×(interquartile range) to the upper quartile,
$13.05 + 1.5(5.1) =$(circle one) **20.95** / **21.15** / **20.7**.

(c) The "lower fence" is determined by subtracting 1.5×(interquartile range) from the lower quartile,
$7.95 - 1.5(5.1) =$(circle one) **0.1** / **0.2** / **0.3**.

(d) Use your calculator to show the box and whiskers plot is given below.

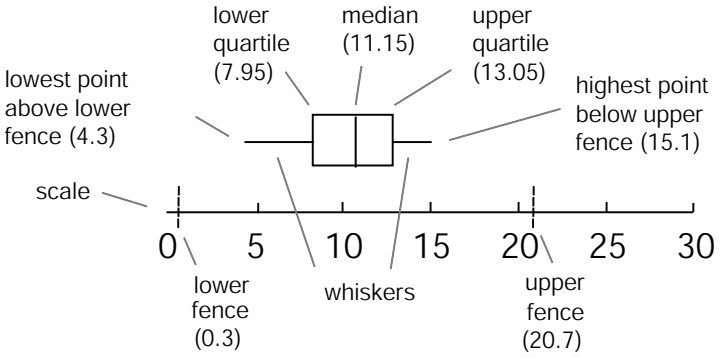


Figure 3.3 (Box and Whisker Plot For Ph Levels Of Soil Data)

The purpose of the "fences" is to determine if any data points are outliers or not. Data points outside the fences, are considered outliers. Any outliers are indicated as "dots" that are in line with, but out and away from, the whiskers.
(After STAT ENTER, type Ph levels into L_1; then 2nd STAT PLOT, choose second box plot; then ZOOM 9:ZoomStat ENTER; TRACE to see five number summary[2].)

3.4 Sampling Distributions

The sampling distribution of $\overline{Y}$ is calculated approximately, using the *Central Limit Theorem* (CLT). The CLT says that as *random* sample size n increases, the sampling distribution of $\overline{Y}$ tends to a normal distribution.

[2]The calculator *can* give a slightly different box and whisker plot which uses hinges, rather than quartiles; in this case, the calculator gives the same box and whiskers as given above. Either box and whiskers plot is acceptable.

Exercise 3.5 (Central Limit Theorem, Fishing in Montana)

1. The distributions of the average number of fish caught at a lake, $\overline{Y}$, where $n = 1, 2, 3$ are given by

$y, n = 1$	1	2	3
$P(Y = y)$	0.4	0.4	0.2

where $\mu_Y = 1.8$ and $\sigma_Y = 0.75$,

$\bar{y}, n = 2$	1	$\frac{3}{2}$	2	$\frac{5}{2}$	3
$P(\overline{Y} = \bar{y})$	0.16	0.32	0.32	0.16	0.04

where $\mu_{\overline{Y}} = 1.8$ and $\sigma_{\overline{Y}} = \frac{0.75}{\sqrt{2}} = 0.53$,

$\bar{y}, n = 3$	1	$\frac{4}{3}$	$\frac{5}{3}$	2	$\frac{7}{3}$	$\frac{8}{3}$	3
$P(\overline{Y} = \bar{y})$	0.064	0.192	0.288	0.256	0.144	0.048	0.008

where $\mu_{\overline{Y}} = 1.8$ and $\sigma_{\overline{Y}} = \frac{0.75}{\sqrt{3}} = 0.43$. The probability histograms of these three sampling distributions are given below.

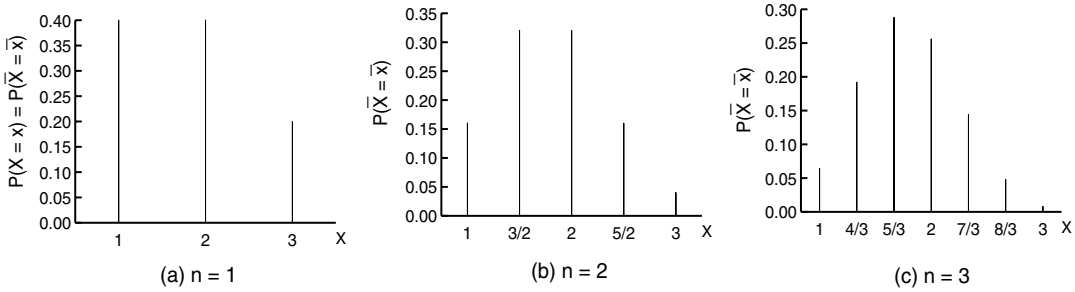


Figure 3.4 (Comparing Sampling Distributions Of Sample Mean)

As the random sample size, n, increases, the sampling distribution of the average, $\overline{Y}$, changes shape and becomes more (circle one)

(a) rectangular–shaped.

(b) bell–shaped.

(c) triangular–shaped.

In fact, the *central limit theorem* (CLT) says *no matter what the original distribution*, the sampling distribution of the average is typically normal when $n > 30$.

2. Even though the sampling distribution becomes more normal–shaped as the random sample size increases, the mean of the average, $\mu_{\overline{Y}} = 1.8$ (circle one)

   (a) decreases and is equal to $\frac{\sigma_Y^2}{n}$,

   (b) remains the same and is equal to $\mu_Y = 1.8$,

   (c) increases and is equal to $n\mu_Y$,

   and the standard deviation of the average, $\sigma_{\overline{Y}}$ (circle one)

   (a) decreases and is equal to $\frac{\sigma_Y}{\sqrt{n}}$.

   (b) remains the same and is equal to σ_Y.

   (c) increases and is equal to $n\sigma_Y$.

3. After $n = 30$ trips to the lake, the distribution in the average number of fish caught is essentially normal (why?), where (circle one)

   (a) $\mu_{\overline{Y}} = 1.8$ and $\sigma_{\overline{Y}} = \frac{0.75}{\sqrt{3}} = 0.43$.

   (b) $\mu_{\overline{Y}} = 1.8$ and $\sigma_{\overline{Y}} = \frac{0.75}{\sqrt{10}} = 0.24$.

   (c) $\mu_{\overline{Y}} = 1.8$ and $\sigma_{\overline{Y}} = \frac{0.75}{\sqrt{30}} = 0.14$

4. **True** / **False**. After $n = 30$ trips to the lake, the (approximate) chance the average number of fish caught is greater than 2.1 fish is given by (using your calculators)
   $P(\overline{Y} > 2.1) \approx 0.015$, where $\mu_{\overline{Y}} = 1.8$ and $\sigma_{\overline{Y}} = \frac{0.75}{\sqrt{30}} = 0.14$.
   (Type 2nd DISTR 2:normalcdf(2.1, 2nd EE99, 1.8, $\frac{0.75}{\sqrt{30}}$) ENTER.)

5. After 30 trips to the lake, the chance the average number of fish is *less than* 1.95 is $P(\overline{Y} < 1.95) \approx$ (circle one) **0.73** / **0.86** / **0.94**.

6. After $n = 35$ trips to the lake, the distribution in the average number of fish caught is essentially normal (why?), where (circle one)

   (a) $\mu_{\overline{Y}} = 1.8$ and $\sigma_{\overline{Y}} = \frac{0.75}{\sqrt{30}} = 0.14$.

   (b) $\mu_{\overline{Y}} = 1.8$ and $\sigma_{\overline{Y}} = \frac{0.75}{\sqrt{35}} = 0.13$.

   (c) $\mu_{\overline{Y}} = 1.8$ and $\sigma_{\overline{Y}} = \frac{0.75}{\sqrt{40}} = 0.12$

7. After 35 trips to the lake, the chance the average number of fish is *less than* 1.95 is $P(\overline{Y} < 1.95) \approx$ (circle one) **0.73** / **0.88** / **0.94**.

8. After $n = 15$ trips to the lake, the distribution in the average number of fish caught (circle one) **is** / **is not** normal.

9. The CLT is useful because (circle none, one or more):

   (a) No matter what the original distribution is, as long as a large enough random sample is taken, the average of this sample follows a normal (not a binomial or any other distribution) distribution.

   (b) In practical situations where it is not known what probability distribution to use, as long as a large enough random sample is taken, the average of this sample follows a normal distribution.

   (c) Rather than having to deal with many different probability distributions, as long as a large enough random sample is taken, the average of this sample follows *one* distribution, the normal distribution.

   (d) Many of the distributions in statistics rely in one way or another on the normal distribution because of the CLT.

10. **True** / **False** The central limit theorem requires not only that $n \geq 30$, but also that a *random sample* of size $n \geq 30$ is used.

3.5 Sampling Distribution of Sample Mean

We continue to look at the Central Limit Theorem (CLT) and compare it to Tchebycheff's Rule.

Exercise 3.6 (Central Limit Theorem)

1. *More CLT Questions.*

   (a) Suppose Y has a distribution where $\mu_Y = 2.7$ and $\sigma_Y = 0.64$. If $n = 35$, then $P(\overline{Y} > 2.75) \approx$ (circle one) **0.32** / **0.58** / **0.64**.
   (Remember that $\sigma_{\overline{Y}} = \frac{0.64}{\sqrt{35}}$.)

   (b) Suppose Y has a distribution where $\mu_Y = -1.7$ and $\sigma_Y = 1.5$. If $n = 49$, then $P(-2 < \overline{Y} < 2.75) \approx$ (circle one) **0.58** / **0.60** / **0.92**.

   (c) Suppose Y has a distribution where $\mu_Y = -1.7$ and $\sigma_Y = 1.64$. If $n = 15$, the distribution of $\overline{Y}$ (circle one) **is** / **is not** (approximately) normal.

   (d) **True** / **False** If the random sample size is large ($n \geq 30$), the distribution of $\overline{Y}$ is $N(\mu_{\overline{Y}}, \sigma^2_{\overline{Y}})$; or, equivalently,

   $$\frac{\overline{Y} - \mu_{\overline{Y}}}{\sigma_{\overline{Y}}} = \frac{\overline{Y} - \mu_Y}{\sigma_Y/\sqrt{n}} = \frac{\sqrt{n}(\overline{Y} - \mu)}{\sigma}$$

   is a standard normal, $N(0, 1)$.

2. *CLT versus Tchebycheff's Rule.* The corn yield in $n = 36$ plots has a (population) standard deviation in yield of $\sigma = 6.2$ tons. What is the probability the (observed) average corn yield, $\overline{Y}$, is within 2.5 tons of the expected average corn yield, μ?

   (a) Since $\sigma = 6.2$, $\sigma_{\overline{Y}} = \frac{\sigma}{\sqrt{n}} = \frac{6.2}{\sqrt{36}} =$ (circle one) **0.58** / **1.04** / **1.92**.

   (b) To say "$\overline{Y}$ is within 2.5 tons of μ" means $\overline{Y} - \mu \leq 2.5$; in other words, $k\sigma_{\overline{Y}} = 2.5$, or the number of standard deviations from the mean, k, is given by,
   $k = \frac{2.5}{\sigma_{\overline{Y}}} = \frac{2.5}{1.04} =$ (circle one) **1.58** / **2.40** / **3.92**.

   (c) *Tchebycheff.* Tchebycheff's rule tells us that there is a
   $1 - \frac{1}{k^2} = 1 - \frac{1}{2.40^2} =$ (circle one) **0.83** / **0.92** / **0.99**.
   chance that the average corn yield of 36 plots, $\overline{Y}$, is within 2.5 tons of the expected average corn yield, μ.

   (d) *CLT.* Since the number of standard deviations from the mean is $k = 2.40$, the CLT tells us that there is a
   $P(-2.40 < Z < 2.40) \approx$ (circle one) **0.58** / **0.58** / **0.98**
   chance that the average corn yield of 36 plots, $\overline{Y}$, is within 2.5 tons of the expected average corn yield, μ.
   (Type 2nd DISTR 2:normalcdf(-2.40, 2.40).)