CHAPTER 3
A REVIEW OF SOME BASIC CONCEPTS

3.1 Statistics has many definitions, but almost all involve the process of drawing conclusions from data. Data is subject to variability, so some would say that statistics is the study of variability with the objective of understanding its sources, measuring it, controlling whatever is controllable, and drawing conclusions in the face of it. For sample survey purposes, statistics involves a well-defined population, a sample selected according to an appropriate probabilistic design, and a methodology for making inferences from the sample to the population, usually in terms of estimation of population parameters.

3.2 A statistic is a function of (is calculated from) sample data whereas a parameter is a numerical characteristic of a population. In a common opinion poll, a sample of 500 residents may be asked whether or no they favor a certain candidate for office. The sample percentage is a statistic, but it is used to estimate the population percentage favoring that candidate, an unknown parameter.

3.3 An estimator is a statistic used to estimate a population parameter, like the sample proportion in Exercise 3.2.

3.4 A sampling distribution is a distribution of all possible values of a statistic.

3.5 The goodness of an estimator is usually measured by the standard deviation of its sampling distribution. The margin of error refers to two standard deviations of the sampling distribution of an estimator. Roughly speaking, the difference between an estimator and the true value of the parameter being estimated will be less than the margin of error with probability about .95.

3.6 An estimator should be unbiased (or nearly so) and have a small standard deviation of its sampling distribution. In other words, in repeated usage, an estimator’s values should pile up close to the value of the parameter being estimated.

3.7 An unbiased estimator is one for which the sampling distribution centers at the true value of the parameter being estimated.

3.8 The error of estimation refers to the difference between an estimator and the true value of a parameter being estimated. It is measured by the standard deviation of the sampling distribution of the estimator in question.

3.10

Summary Statistics

	Calories		Cost in Dollars	
	w/Hydra	w/o Hydra	w/ Hydra	w/o Hydra
Mean	64.78	64.62	.294	.27
Median	66.0	63.5	.260	.25
Stdev	8.51	9.09	.097	.05
Q_1	60	60	.230	.225
Q_3	70	70	.345	.320
Min	50	50	.220	.220
Max	80	80	.520	.350
Range	30	30	.300	.130

Scatterplot of cost vs. calories

```
        -                               * Hydra
    0.50+
        -
cost    -
        -
        -
    0.40+
        -
        -                                                      *
        -                                      *
        -
    0.30+
        -
        -    *                *
        -                                 *    *
        -                     2
    0.20+
          +---------+---------+---------+---------+---------+------
                                                              calories
        48.0      54.0      60.0      66.0     72.0     78.0
```

Scatterplot of cost vs. calories (without Hydra)

```
        -
   0.350+                                                      *
        -                                      *
cost      -
        -
        -
   0.300+
        -
        -
        -
        -    *                *
   0.250+
        -                                 *    *
        -
        -                     2
        -
   0.200+
          +---------+---------+---------+---------+---------+------
                                                              calories
       48.0      54.0      60.0      66.0      72.0      78.0
```

(a) The mean is a good summary number for typical calories per serving. The standard deviation is a good summary number for the variation in the calories.

Box plot of calories

```
                       ------------------
      -----------------I         +      I----------------
                       ------------------
   +---------+---------+---------+---------+---------+------
48.0      54.0      60.0      66.0      72.0      78.0
```

(b) Since there is an extreme value, the median is a good summary number for typical cost per serving, and IQR (Q_3 - Q_1) is a good summary for the variation in costs.

Box plot of cost

```
      ------------------                            Hydra
   ---I  +             I-                            *
      ------------------
------+---------+---------+---------+---------+---------+
  0.240     0.300     0.360     0.420     0.480     0.540
```

(c) Because these drinks would not generally be combined by users, the totals have little practical value here.

(d) On the average calories per serving; not much impact
On the standard deviation: slight increase
On the average cost per serving: decrease

On the standard deviation of the cost per serving: decreased

Box plot of calories (without Hydra)

```
                              ------------------
             -----------------I     +          I----------------
                              ------------------
          +---------+---------+---------+---------+---------+------
       48.0      54.0      60.0      66.0      72.0      78.0
```

Box plot of cost (without Hydra)

```
                ----------------------------
            ----I       +                  I-------------------
                ----------------------------
          ----+---------+---------+---------+---------+---------+--
          0.225     0.250     0.275     0.300     0.325     0.350
```

(e) There is no particularly influential drink on the average calories per saving, but Snappple (80 calories) has the most influence as it is furthest from the mean.

3.11 **(a)** Including the powdered drinks on the same list with the liquid drinks does not have much effect on the average calories per serving, as their calorie figures are within the range of the first data set. Including the powdered drinks lowers the average cost per serving and increases the standard deviation of cost because the new cost values are much lower than in the original set.

	calories		cost	
	mean	stdev	mean	stdev
w/o powder	64.78	8.51	.294	.097
w powder	64.82	7.93	.278	.102

Parallel box plots of calories

```
                              ------------------
w/o powder   -----------------I         +      I----------------
                              ------------------

                              ------------------
w/ powder    -----------------I         +      I----------------
                              ------------------
          +---------+---------+---------+---------+---------+------
       48.0      54.0      60.0      66.0      72.0      78.0
```

Parallel box plots of cost

```
                            -------------
w/o powder               ---I +         I--                    *
                            -------------

                           -----------
w/ powder    --------------I  +      I-----                    *
                           -----------
          --------+---------+---------+---------+---------+--------
               0.160     0.240     0.320     0.400     0.480
```

(b) Adding the light varieties to the list will not have much of an effect on the average cost and standard deviation of cost.

Mean and standard deviation for two groups (cost)

	mean	stdev
w/o lites	.294	.097
w/ lites	.286	.088

Parallel box plots of cost

```
                ------------------
w/o lites    ---I  +             I-                            *
                ------------------

                -----------
w/ lites     ---I  +      I--------                            O
                -----------
          ------+---------+---------+---------+---------+---------+
            0.240     0.300     0.360     0.420     0.480     0.540
```

(c) Adding the light varieties to the list will decrease the average calories per serving and increase the standard deviation of calories because the new cost figures are way below those of the original data set.

Mean and standard deviation for two groups (calories)

	mean	stdev
w/o lites	64.78	8.51
w/ lites	55.45	22.69

Parallel box plots of calories

```
                                                  --------
w/o lites                                  -------I   +  I------
                                                  --------

                                               ----------
w/ lites   O               *               ----I  +     I-------
                                               ----------
          +---------+---------+---------+---------+---------+------
          0        15        30        45        60        75
```

(d) Use the median because the median is not sensitive to extreme values.

3.12 Summary Statistics

```
Area            N   mean  median   stdev   Q1     Q3    Q3 - Q1
U.S            10  25.10  12.50   22.02  7.50  51.25   43.75
U.S.& Foreign  10   4.80   1.00    7.18  0.00  10.00   10.00
```

(a) As shown on the stem plot, these data are split into two groups and neither the mean nor the median are good measures of center. A more meaningful summary statistic is the total number of endangered species, 251 for those unique to the U.S. and 299 in the U.S. and foreign countries.

Box plot of U.S.

```
           -------------------------------------------
      -----I    +                                    I-------
           -------------------------------------------
     --------+---------+---------+---------+---------+--------
            10       20        30         40       50
```

Stem plot of U.S.

```
Stem-and-leaf of US
Leaf Unit = 1.0

    3    0 368
   (3)   1 023
    4    2
    4    3 7
    3    4
    3    5 057
```

(b) Again, the total number of endangered species is a more meaningful statistic than either the mean or median. For the world, this total is 791 species.

Stem plot of U.S & Foreign

```
Stem-and-leaf of US & Foreign
Leaf Unit = 1.0

    5    0 00000
    5    0 23
    3    0
    3    0
    3    0 8
    2    1
    2    1
    2    1
    2    1 6
    1    1 9
```

Box plot of U.S. & Foreign

```
     -----------------------
     I  +                  I------------------------------
     -----------------------
     +---------+---------+---------+---------+---------+------
   0.0       3.5       7.0      10.5      14.0      17.5
```

(c) No. See part (a).

3.13 **(a)** $\frac{1}{25}(3\times16+2\times4+1\times2)=\frac{58}{25}=2.32$

(b) $\mu=\sum xp(x)=3\times.64+2\times.16+1\times.08+0\times.12=2.32$

(c) $V(x)=\sum_x(x-\mu)^2p(x)=\sum_x x^2p(x)-\mu^2$

$$=3^2(.64)+2^2(.16)+1^2(.08)+0^2(.12)-2.32^2\ =6.48-5.3824=1.0976$$

$\sigma=\sqrt{V(x)}\ =1.05$

3.14 **(a)** $\mu=E(x)=\sum xp(x)$

$= 2(.443) + 3(.229) + 4(.200) + 5(.086) + 6(.028) + 7(.014)\ = 3.069$

(b) $\sigma^2\ =V(x)=\sum_x(x-\mu)^2p(x)=1.458$

$\sigma=\sqrt{V(x)}\ =1.207$

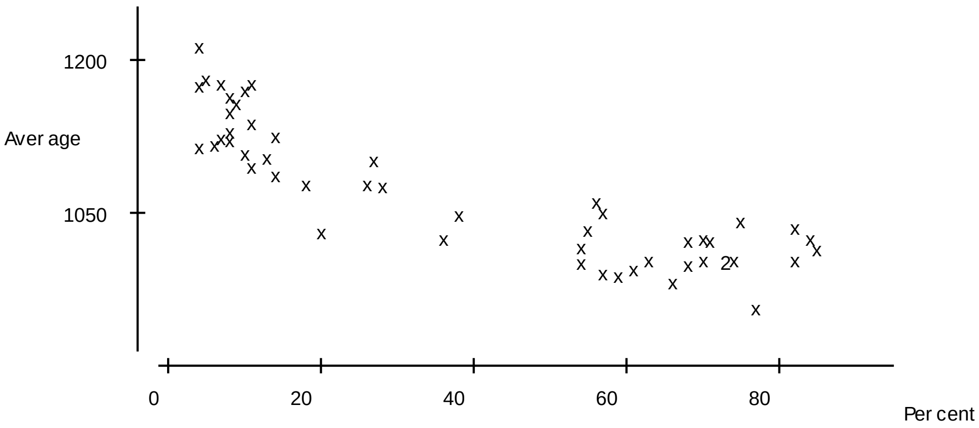
(c) The distribution of the sample data would reflect that of the population. Most of the data values would pile up around 2 and 3 , with a few larger values. The distribution of the sample would be skewed toward the larger values, with a center at approximately 3.07 and a standard deviation of approximately 1.21.

(d) The sample mean $\bar{x}$ has approximately a normal distribution with mean $\mu_{\bar{x}} = \mu_x = 3.07$ and standard deviation $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{1.21}{20} = 0.0605$

3.15 **(a)** The scatter plot shows that SAT and Percent are negatively correlated, with a curved pattern suggesting that the average score drops quickly as the percentages begin to increase and them levels off for higher percentages. The decreasing scores with increasing percentage taking the exam makes practical sense; in states with small percentages only the very best students are taking the exam.

(b) The correlation coefficient is -0.877, but this is not a good measure to use here because of the curvature in the patter. Correlation measures the strength of a linear relationship between two variables.

Scatter plot between Average Score and Percent



3.16 Tabled below are the new probabilities for samples of size 2 and estimates of the population total for the unequal probabilities of selection that favor the smaller population values.

Sample	Probability	$\hat{\tau}_{pps}$
{1.2}	0.32	3.75
{1.3}	0.08	16.25
{1.4}	0.08	21.25
{2,3}	0.08	17.50
{2,4}	0.08	22.50
{3,4}	0.02	35.00
{1,1}	0.16	2.50
{2.2}	0.16	5.00
{3,3}	0.01	30.00
{4,4}	0.01	40.00

Calculation of expectations yields:

$$E\left(\hat{\tau}_{pps}\right) = 10$$

$$V\left(\hat{\tau}_{pps}\right) = 81.25$$

3.17 The weights given in Section 3.3 for the four population values are w_1 = 4.0916, w_2 = 4.0916, w_3 = 1.3236 and w_4=1.3236. The sum of the weights for each of the six possible samples, along with the probabilities of selecting each of these samples, are shown in the accompanying table. The expected value of the sum of the weights turns out to be 4.00, the number of values in the population.

Sample	Sum of weights	Probability of sample, unequal weights
{1,2}	8.1832	.0222
{1,3}	5.4152	.1111
{1,4}	5.4152	.1111
{2,3}	5.4152	.1111
{2,4}	5.4152	.1111
{3,4}	2.6472	.5333

3.18 For samples of size n=2 taken with probabilities proportional to the populations of the states, the pertinent data and the probabilities of selection with probabilities proportional to the population, both with and without replacement, are given in the first table that follows. With replacement probabilities of selection (δ) are directly proportional to the population sizes. Without replacement probabilities of selection (π) are found by first finding the probability for each possible sample, given on the second table. Note that there are 21 with replacement samples of size 2, but only 15 without replacement samples. (Also, note that the π's sum to 2, the sample size.)

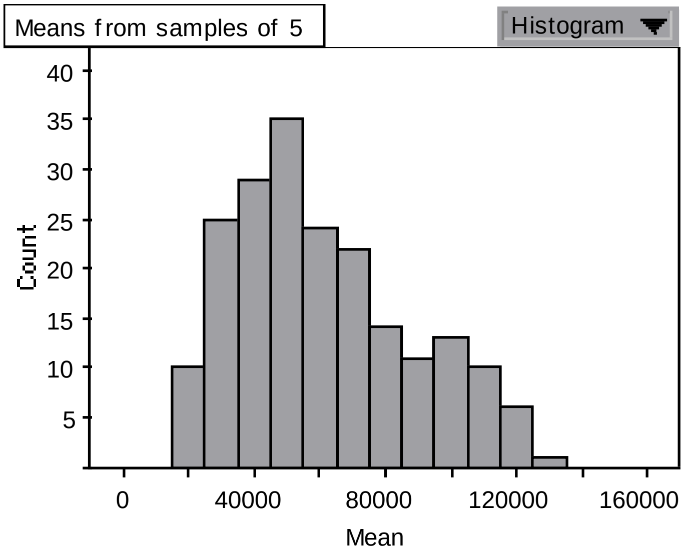
Students	Teachers	Population	With Replacement Probability of selection, δ	Without Replacement Probability of selection,π
570	42	35	0.25	0.536150
206	17	13	0.09	0.214110
973	69	64	0.45	0.746890
207	15	13	0.09	0.214110
158	11	11	0.08	0.191280
101	8	6	0.04	0.097451

The estimates of the total number of teachers are found by using the formulas of Section 3.3. The expected value of each set of estimates, with their appropriate probability distributions, is 162, the total number of teachers for New England.

Sample	Probability (with replacement)	Estimate (with replacement)	Probability (without replacement)	Estimate (without replacement)
{1.2}	0.0450	178.444	0.054725	157.735
{1,3}	0.2250	160.667	0.354545	170.719
{1,4}	0.0450	167.333	0.054725	148.394
{1,5}	0.0400	152.750	0.048406	135.844
{1,6}	0.0200	184.000	0.023750	160.429
{2,3}	0.0810	171.111	0.118142	171.781
{2,4}	0.0162	177.778	0.017802	149.456
{2,5}	0.0144	163.195	0.015738	136.906
{2,6}	0.0072	194.445	0.007706	161.491
{3.4}	0.0810	160.000	0.118142	162.441
{3,5}	0.0720	145.417	0.104585	149.890
{3,6}	0.0360	176.667	0.051477	174.476
(4,5}	0.0144	152.083	0.015738	127.565
{4,6}	0.0072	183.333	0.007706	152.150
{5,6}	0.0064	168.750	0.006812	139.600
{1,1}	0.0625	168.000		
{2,2}	0.0081	188.889		
{3,3}	0.2025	153.333		
{4,4}	0.0081	166.667		
{5,5}	0.0064	137.500		
{6,6}	0.0016	200.000		

3.19 No; the proportions of students in the various states are about the same as the proportions of the total population.

3.20 A histogram of 200 sample means from samples of size 5 each are shown in the histogram. This distribution is somewhat skewed because the population distribution of teachers per state is highly skewed. Even so, the mean of the sampling distribution is 61, 147, quite close to the population mean of 59,856. The standard deviation of the sampling distribution is 26,772, quite close to the theoretical value of 28,645.



3.21 $p(u_1) = p(u_2) = \cdot\ \cdot\ \cdot = p(u_N) = 1/N$

$$\sigma^2 = V(y) = E(y-\mu)^2 = \sum_y (y-\mu)^2 p(y) = \frac{1}{N}\sum_{i=1}^{N}(u_i-\mu)^2$$

3.22 Let a_i denote the number of times a particular y_i value from the population appears in the sample. This number could be greater than 1 because the sampling is with replacement. Then,

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{\delta_i} = \frac{1}{n}\sum_{i=1}^{N}a_i\frac{y_i}{\delta_i}$$

where n is the sample size and N the population size. Since $E(a_i) = n\delta_i$, it follows that $E(\hat{\tau}) = \tau$.

Now, $\hat{\tau}$ is the mean of n independent variables, and so its variance is given by

$$V(\hat{\tau}) = \frac{1}{n}\sum_{i=1}^{N}\delta_i\left(\frac{y_i}{\delta_i} - \tau\right)^2$$

The estimate of this variance can be rewritten as follows:

$$\begin{aligned}
\hat{\mathrm{V}}(\hat{\tau}) &= \frac{1}{n}\cdot\frac{1}{n-1}\sum_{i=1}^{n}(\frac{y_i}{\delta_i}-\hat{\tau})^2 \\
&= \frac{1}{n}\cdot\frac{1}{n-1}\sum_{i=1}^{n}[(\frac{y_i}{\delta_i}-\tau)-(\hat{\tau}-\tau)]^2 \\
&= \frac{1}{n}\cdot\frac{1}{n-1}[\sum_{i=1}^{n}(\frac{y_i}{\delta_i}-\tau)^2-\sum_{i=1}^{n}(\hat{\tau}-\tau)^2] \\
&= \frac{1}{n}\cdot\frac{1}{n-1}[\sum_{i=1}^{N}a_i(\frac{y_i}{\delta_i}-\tau)^2-\sum_{i=1}^{N}a_i(\hat{\tau}-\tau)^2] \\
&= \frac{1}{n}\cdot\frac{1}{n-1}[\sum_{i=1}^{N}a_i(\frac{y_i}{\delta_i}-\tau)^2-n(\hat{\tau}-\tau)^2]
\end{aligned}$$

Taking expected values shows that:

$$\begin{aligned}
\mathrm{E}[\hat{\mathrm{V}}(\hat{\tau})] &= \frac{1}{n}\cdot\frac{1}{n-1}[\sum_{i=1}^{N}n\delta_i(\frac{y_i}{\delta_i}-\tau)^2-nE(\hat{\tau}-\tau)^2] \\
&= \frac{1}{n}\cdot\frac{1}{n-1}[n^2V(\hat{\tau})-nV(\hat{\tau})] = V(\hat{\tau})
\end{aligned}$$

3.23

	Aspirin	Placebo
Exercise Vigorously		
Yes	7 910	7861
No	2997	3060
Total	10907	10921
Cigarette smoking		
Never	5431	5488
Past	4373	4301
Current	1213	1225
Total	11017	11014

(a) Compare the two columns we see that the counts are nearly the same across all categories. The randomization scheme did a good job in balancing these variables between the two groups.

(b) No. $\frac{5431}{11017}=.49$, $\frac{5488}{11014}=.50$ are nearly the same.

(c) No. $\frac{2997}{10907}=.27$, $\frac{3060}{10921}=.28$ are nearly the same.

3.24

Heart Attack	Aspirin	Placebo
Yes	139	239
No	10861	10761
Total	11000	11000

Yes, $\frac{139}{11000} = .012636$ $\frac{239}{10682} = .021727$ are not close.

3.25

Stroke	Aspirin	Placebo
Yes	119	98
No	10881	10902
Total	11000	11000

No, $\frac{119}{11000} = .0108$ $\frac{98}{11000} = .0089$ are close.

Comparing the two ratios $\frac{.021727}{.012636} = 1.72$, $\frac{.0089}{.0108} = .82$, we can find Aspirin is more effective as a possible prevention for heart attacks than for strokes.

3.26 The rate of heart attacks for the smokers (21/1213 = 0.0173) is greater than the rate for those who never smoked (55/5431 = 0.0101), but the effectiveness of the aspirin is about the same for both groups. One way to demonstrate the latter is to look at the ratio of the heart attack rates for aspirin and placebo treatments, as shown here.

$$\frac{21/1213}{37/1225} = .573, \quad \frac{55/5431}{96/5488} = .579$$