

# Ma 322: Biostatistics

## Solutions to Homework Assignment 1

Prof. Wickerhauser

Due Friday, January 27th, 2012

Begin by obtaining access to the R software package, either by downloading a copy onto your computer or else by finding a computer with a working installation. The current version number is 2.12.

Read Chapter 6, pages 60–79, of our e-text to review basic principles of probability. Consult Chapters 1-5 as needed to find function names and syntax to solve the computation problems below.

1. Let  $\{1.3, 1.6, 1.8, 3.1, 3.2, 2.2, 3.5, 3.8\}$  be the values of a variable  $X$  over an 8-member population.
  - (a) Find the frequency distribution table for groupings into intervals  $\{[0, 1), [1, 2), [2, 3), [3, 4)\}$ . Plot the resulting frequency bar graph (histogram).
  - (b) Find the cumulative frequency distribution table and the relative cumulative frequency distribution table for this variable and population. Plot the results on a single frequency bar graph.
  - (c) Calculate the mean of the population.
  - (d) Estimate the mean of the population using the frequency distribution table from part b, with the midpoints of the intervals as the representative values.
  - (e) Calculate the median of the population using the 8 values.
  - (f) Estimate the median of the population using interpolation with the cumulative frequency distribution table from part b.
  - (g) Find the mode of the grouped values of the population using the frequency distribution table from part b. Use the midpoint of each interval as the representative value.

**Solution:** (a) Frequency distribution table:

$i$	interval $i$	$f_i$
1	[0, 1)	0
2	[1, 2)	3
3	[2, 3)	1
4	[3, 4)	4

R commands to plot the frequency histogram:

```
x<-c(1.3, 1.6, 1.8, 3.1, 3.2, 2.2, 3.5, 3.8)
hist(x,breaks=c(0,1,2,3,4))
```

See the figure below for the graphical output:

(b) Cumulative frequency distribution table:

$i$	interval $i$	Cum. $f_i$	Rel. Cum. $f_i$
1	[0, 1)	0	0.000
2	[1, 2)	3	0.375
3	[2, 3)	4	0.500
4	[3, 4)	8	1.000

R command to plot the relative cumulative frequency distribution:

```
x=c(1.3, 1.6, 1.8, 3.1, 3.2, 2.2, 3.5, 3.8)
plot(ecdf(as.integer(x)))
```

See the figure below for the graphical output:

(c) There are 8 values  $X_i$ , and  $\sum_{i=1}^8 X_i = 20.5$ , so  $\mu = \frac{1}{8} \sum_{i=1}^8 X_i = \frac{20.5}{8} = 2.5625$ . Keeping just one extra significant digit, we write  $\mu = 2.56$ .

(d) There are 4 intervals with representative values 0.5, 1.5, 2.5 and 3.5, and frequencies 0,3,1,4, respectively. By the formula, the mean is estimated as  $(\sum_{i=1}^4 X_i f_i) / (\sum_{i=1}^4 f_i) = 21/8 = 2.625$

(e) The increasing rearrangement of the 8 values is  $\{1.3, 1.6, 1.8, 2.2, 3.1, 3.2, 3.5, 3.8\}$ , so the median  $\mathcal{M}$  is the midpoint between 2.2 and 3.1, or  $\mathcal{M} = (2.2 + 3.1)/2 = 2.65$ .

(f) Since there are 8 total values, the median should be at value  $(8 + 1)/2 = 4.5$ , which falls in the fourth interval [3, 4). That interval contains 4 tied values, so use the interpolation formula to get the estimate

$$(3) + \left( \frac{0.5 * 8 - 4}{4} \right) (4 - 3) = 3.0.$$

Notice that since the cumulative frequency of the previous classes is  $4 = 0.5 * 8$ , the median falls exactly on the boundary between groups, namely the left endpoint of the median interval containing  $X_{4.5}$ .

R commands to find the mean, mean from bin midpoints, median, and median from bin midpoints:

```
x<-c(1.3, 1.6, 1.8, 3.1, 3.2, 2.2, 3.5, 3.8)
mean(x) # Arithmetic mean of original values
mean(as.integer(x)+0.5) # Arithmetic mean using bin midpoints
median(x) # Median using original 8 values
median(as.integer(x)+0.5) # Median using bin midpoint values
```

(g) Interval  $[3, 4)$  contains the largest number (4) of values, so the mode is the midpoint of that interval: 3.5.  $\square$

2. Pick a random sample of 3 values, without replacement, from the population in Problem 1, keeping the original ordering. Compute the sample mean and sample median for those 3 values. Then perform a similar sampling with replacement. Set the random number generator seed to 1029384 before each sampling to get reproducible results.

**Solution:** Keep the original ordering  $\{1.3, 1.6, 1.8, 3.1, 3.2, 2.2, 3.5, 3.8\}$  of the values. Use the following R commands:

```
x<-c(1.3, 1.6, 1.8, 3.1, 3.2, 2.2, 3.5, 3.8)
set.seed(1029384) # Set the random number seed, for reproducibility
mean(sample(x,3)) # Sample-of-3 arithmetic mean, without replacement
[2.333333]
set.seed(1029384) # Set the random number seed, for reproducibility
mean(sample(x,3,replace=TRUE)) # ... with replacement [2.533333]
set.seed(1029384) # Set the random number seed, for reproducibility
median(sample(x,3)) # Sample-of-3 median, without replacement [2.2]
set.seed(1029384) # Set the random number seed, for reproducibility
median(sample(x,3,replace=TRUE)) # ...with replacement [2.2]
```

$\square$

3. Consider the following five sample observations:  $X = \{63.2, 72.5, 65.7, 61.7, 68.3\}$ .
- (a) Compute the range, mean deviation, variance, standard deviation and coefficient of variation of  $X$ .
- (b) Suppose one additional observation, with value 70.0, is appended to  $X$ . Compute the range, mean deviation, variance, standard deviation and coefficient of variation of the six samples.

**Solution:** (a) Range: increasing rearrangement  $\{61.7, 63.2, 65.7, 68.3, 72.5\}$  gives sample range  $72.5 - 61.7 = 10.8$ .

Mean deviation: first compute the mean  $\bar{X} = 66.28$ . Then compute the deviations:

$$\{-4.58, -3.08, -.58, 2.02, 6.22\}$$

Finally, average the absolute values of these deviations to get 3.296.

Variance: use the sum of squares formula:

$$\sum_{i=1}^5 X_i = 331.4; \quad \sum_{i=1}^5 X_i^2 = 22,039; \quad s^2 = \frac{1}{5-1} \left[ \sum_{i=1}^5 X_i^2 - \frac{1}{5} \left( \sum_{i=1}^5 X_i \right)^2 \right] = 18.392$$

Standard deviation: find the square root  $s = 4.2886$  of  $s^2$ .

Coefficient of variation: using the previously calculated mean, compute  $V = s/\bar{X} = .0647 = 6.47\%$ .

(b) Range: the new value 70.0 is inside the original interval [61.7, 72.5], so the new sample has the same range  $72.5 - 61.7 = 10.8$ .

Mean deviation: first compute the new mean  $\bar{X} = 66.90$ . Then compute the deviations:

$$\{-5.20, -3.70, -1.20, 1.40, 5.60, 3.10\}$$

Finally, average the absolute values of these deviations to get 3.3667.

Variance: update the sum and the sum of squares:

$$\sum_{i=1}^6 X_i = 331.4 + 70.0 = 401.4; \quad \sum_{i=1}^6 X_i^2 = 22,039 + (70.0)^2 = 26,939;$$

$$s^2 = \frac{1}{6-1} \left[ \sum_{i=1}^6 X_i^2 - \frac{1}{6} \left( \sum_{i=1}^6 X_i \right)^2 \right] = 17.020$$

Standard deviation: find the square root  $s = 4.1255$  of  $s^2$ .

Coefficient of variation: using the previously calculated mean, compute  $V = s/\bar{X} = .0617 = 6.17\%$ .

Notice that the mean deviation of the larger sample is greater than the original sample's, while the variance, standard deviation, and coefficient of variation have all decreased. The range of the larger sample is unchanged.

```
x <- c(63.2, 72.5, 65.7, 61.7, 68.3)
length(x)
range(x) # Least and greatest values in x
x-mean(x) # Deviations from the mean
abs(x-mean(x)) # Absolute deviations from the mean
mean(abs(x-mean(x))) # Mean deviation = mean of absolute deviations
var(x) # Variance
sd(x) # Standard deviation
sd(x)/mean(x) # Coefficient of variation
x <- c(x, 70.0)
length(x) # Check that we now have 6 elements
```

```

range(x) # Least and greatest values in x
x-mean(x) # Deviations from the mean
abs(x-mean(x)) # Absolute deviations from the mean
mean(abs(x-mean(x))) # Mean deviation = mean of absolute deviations
var(x) # Variance
sd(x) # Standard deviation
sd(x)/mean(x) # Coefficient of variation

```

□

4. Consider the following table of tree species in a random sample from a forest:

Species	Frequency
White Oak	45
Red Oak	3
Shagbark hickory	27
Black walnut	12
Basswood	2
Slippery Elm	8

(a) Use the Shannon index to express the tree species diversity. Compute the maximum Shannon diversity possible for this number of species, and then calculate the Shannon evenness for this table.

(b) Compute the Brillouin diversity index for the frequency table in the previous problem. Find the maximum Brillouin diversity, then calculate the Brillouin evenness.

**Solution:** (a) Shannon index from frequency table: compute this with

$$H' = \frac{n \log(n) - \sum_{i=1}^6 f_i \log f_i}{n} \approx 1.3642,$$

where  $n = 97$  is the total number of trees in the six species. The R commands for this computation are:

```

tf<-c(45, 3, 27, 12, 2, 8); n<-sum(tf);
hp<-(n*log(n)-sum(tf*log(tf)))/n

```

Maximum Shannon diversity: for 6 species, this is  $H'_{\max} = \log 6 \approx 1.7918$ .

Evenness: this is  $J' = H'/H'_{\max} \approx 0.7613 \approx 76.1\%$ .

(b) Brillouin index from frequency table: compute this with

$$H = \frac{\log(n!) - \sum_{i=1}^6 \log f_i!}{n} \approx 1.27,$$

where  $n = 97$  is the total number of trees in the six species. The R commands for this computation are:

```
tf<-c(45, 3, 27, 12, 2, 8); n<-sum(tf);
h<-(sum(log(1:n))-sum(lfactorial(tf)))/n
```

Maximum Brillouin diversity: for 6 species, this is

$$\begin{aligned}
 H_{\max} &= \frac{\log n! - (k - d) \log c! - d \log(c + 1)!}{n} \\
 &= \frac{349.9541 - (5)(30.6718598) - (1)(33.5050755)}{97} \approx 1.681337,
 \end{aligned}$$

using  $c = 16$  and  $d = 1$  from  $n = ck + d$ , since there are  $n = 97$  trees distributed among  $k = 6$  species. The R commands for this computation are:

```
tf<-c(45, 3, 27, 12, 2, 8); n<-sum(tf); k<-length(tf); c<-16; d<-1;
hmax<-(lfactorial(n) - (k-d)*lfactorial(c) - d*lfactorial(c+1))/n
```

NOTE: The `lfactorial()` function in R computes the natural logarithm of the factorial of its argument, which is much more sensible for large values. If you use common (base 10) logarithms, then you must multiply the numbers by  $\log(10) \approx 2.3036$  to get the natural (base  $e$ ) logarithms used in Brillouin's index.

Evenness: this is  $J = H/H_{\max} \approx 0.755387 \approx 75.5\%$ . □

5. A tennis team has 6 boys and 7 girls.

(a) How many distinct mixed doubles pairs (one boy and one girl) can be formed using members of the team?

(b) How many distinct practice matchups of two mixed doubles pairs can be formed using members of the team?

**Solution:** (a) There are  $(6)(7) = 42$  distinct pairs containing one boy and one girl.

(b) There are  $(6)(6 - 1)(7)(7 - 1) = 1260$  ways to pick two mixed doubles pairs from 6 boys and 7 girls, but this counts each matchup twice, so there are  $1260/2=630$  distinct matchups. □

6. A DNA modeling kit contains 15 base units: 4 A's, 4 C's, 4 G's, and 3 T's.

(a) How many distinct sequences of length 15 can be formed from this kit?

(b) How many distinct sequences of length 3 can be formed from this kit?

(c\*) How many distinct sequences of length 6, 9, or 12 can be formed from this kit?

**Solution:** (a) There are  ${}_{15}P_{4,4,4,3}$  distinct ways to form a sequence of 15 letters from the set, where

$${}_{15}P_{4,4,4,3} = \frac{15!}{4!4!4!3!} = 5 * 7 * 13 * 11 * 5 * 9 * 2 * 7 * 5 = 15,765,750.$$

(b) Since there are at least 3 of each base unit, the number of distinct sequences of length 3 is  $4^3 = 64$ .

(c) To count sequences of length 6, 9, and 12, a computer program such as the one described in class is useful. An implementation in R may be found at the following URL:

`http://www.math.wustl.edu/~victor/classes/ma322/deduct.R`.

For comparison, an implementation in Standard C may be found at the following URL:

`http://www.math.wustl.edu/~victor/classes/ma322/deduct.c`.

Running either program with  $m = 6, 9,$  or  $12$  gives 3885, 180,600, and 4,354,350, respectively. As a check, note that  $m = 3$  gives 64 and  $m = 15$  gives 15,765,750, in agreement with the other calculations.  $\square$

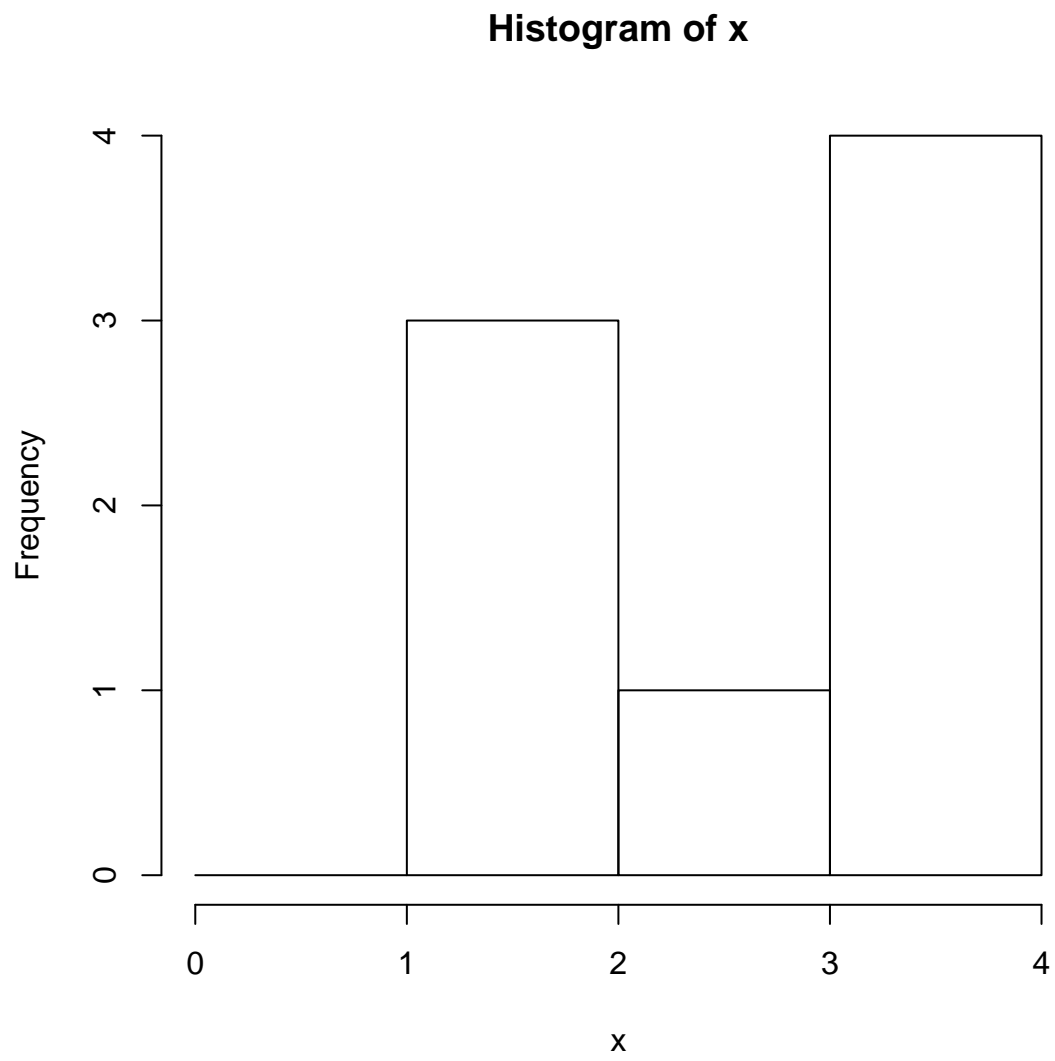


Figure 1: Histogram for Problem 1(a)

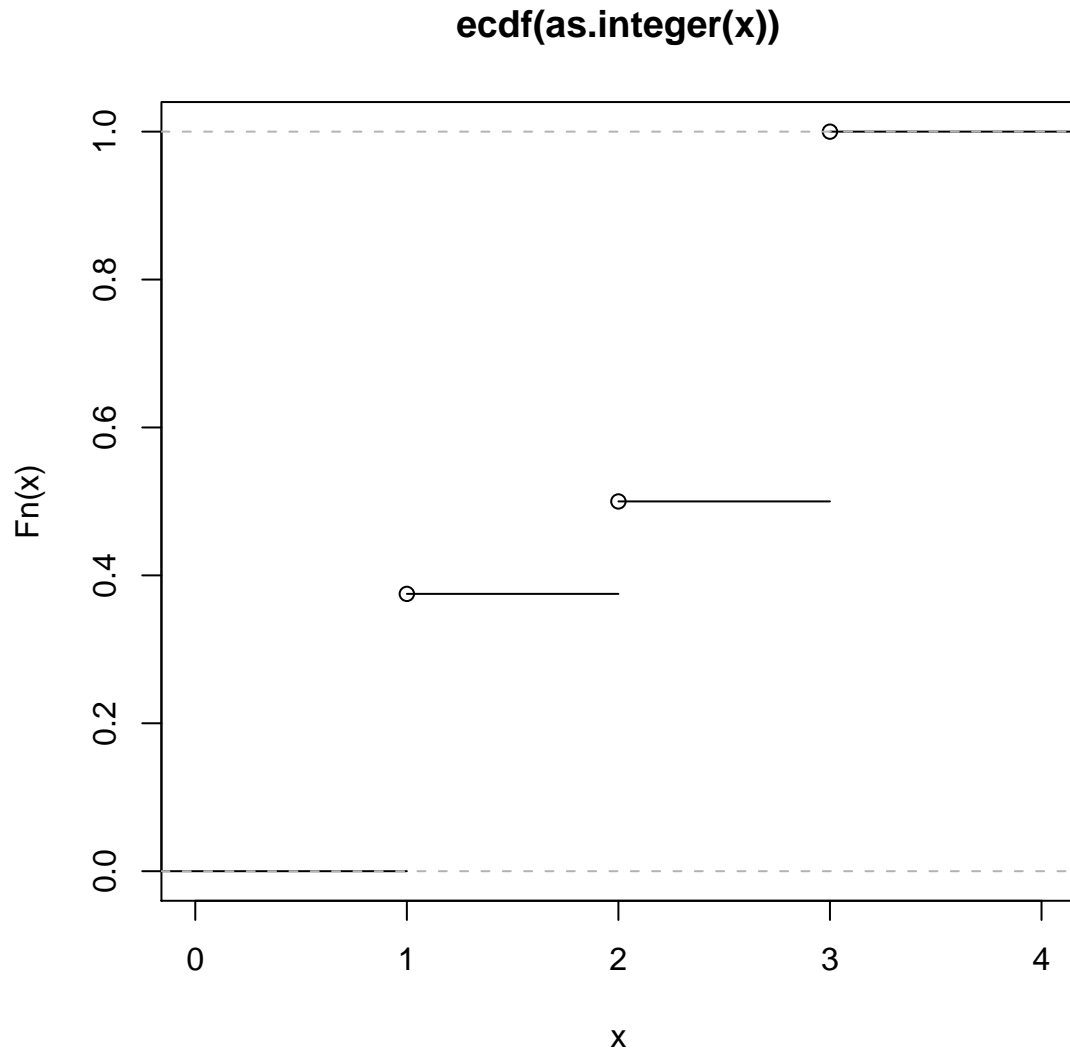


Figure 2: Histogram for Problem 1(b)