Ma 322: Biostatistics
Homework Assignment 1

Prof. Wickerhauser

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. Any version number 2.2 or greater should work. You may then also download R Studio for its convenience.

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 `SEED` be your student ID number. Generate 100 samples from a standard normal density using the following `R` commands:

   ```
   set.seed(SEED);  x <- rnorm(100)
   ```

   (a) Plot the histogram of `x` using `R` defaults.

   (b) Plot the histogram of `x` using enough integer breakpoints to include all samples, with bins that include their left endpoint but not their right endpoint. HINT: this is not the default for hist() in R. It will be necessary to check `range(x)`, and to set both the `right=` and `include.lowest=` parameters to non-default values. Read the documentation page!

   (c) Calculate the mean and median of the samples.

   (d) Calculate the standard deviation, variance, and mean absolute deviation of the samples.

   Solution: For this model solution I used `SEED<-63130`.

   (a) Plot with defaults:

   ```
   hist(x)
   ```

   Put the graphical output into a PDF file as follows:

   ```
   pdf("hw0101a.pdf"); hist(x); dev.off();
   ```

   (b) Plot with integer breakpoints, bins closed at left:

   Note: check `range(x)`, which returns `[1] -3.092222 2.109003`, so the breakpoints should be integers from -4 to 3:

```
hist(x,right=FALSE,breaks=c(-4,-3,-2,-1,0,1,2,3))
```

Put the graphical output into a PDF file as follows:

```
pdf("hw0101b.pdf"); hist(x,right=FALSE,breaks=c(-4,-3,-2,-1,0,1,2,3)); dev.off();
```

(c) `mean(x)` returns 0.03977045, `median(x)` returns 0.09007506.

(d) `sd(x)` returns 0.9937323;

`var(x)` returns 0.9875039;

`mean(abs(x-mean(x)))` returns 0.7952072. □

2. Reuse the population `x` generated in Problem 1, keeping the original ordering. Set the random number generator seed to your student ID before each sampling (that is, before each `rnorm()` and each `sample()`) to get reproducible results.

   (a) Pick a random subsample of 10 values, without replacement, and compute the sample mean and sample median for those 10 values.

   (b) Pick a random subsample of 10 values with replacement, then compute the sample mean and sample median for those 10 values.

   Solution: Use the following `R` commands:

   ```
   SEED<-63130; set.seed(SEED); x<-rnorm(100);
   ```

   (a)

   ```
   set.seed(SEED); a <- sample(x,10,replace=FALSE); mean(a); median(a);
   ```

   In `R version 3.2.3 (2015-12-10) -- "Wooden Christmas-Tree"`, these commands return 0.02611848 and −0.04974433, respectively.

   In `R version 3.6.3 (2020-02-29) -- "Holding the Windsock"`, these commands return 0.1033666 and 0.0185128, respectively.

   (b)

   ```
   set.seed(SEED); b <- sample(x,10,replace=TRUE); mean(b); median(b);
   ```

   In `R version 3.2.3 (2015-12-10) -- "Wooden Christmas-Tree"`, these commands return −0.1420805 and −0.04416747, respectively.

   In `R version 3.6.3 (2020-02-29) -- "Holding the Windsock"`, these commands return 0.1033666 and 0.0185128, respectively, the same values because the updated random number generator just happens to select the same 10 values with or without replacement. □

3. Consider the following table of tree species in a complete count from a section of forest:

Species	Frequency
White Oak	40
Red Oak	33
Shagbark hickory	17
Black walnut	12
Basswood	13
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

$$H' = \frac{n\log(n) - \sum_{i=1}^{6} f_i \log f_i}{n} \approx 1.6341,$$

where $n = 123$ is the total number of trees in the six species. The R commands for this computation are:

```
tf<-c(40, 33, 17, 12, 13, 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 $J' = H'/H'_{\max} \approx 0.912 \approx 91.2\%$.

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

$$H = \frac{\log(n!) - \sum_{i=1}^{6} \log f_i!}{n} \approx 1.546,$$

where $n = 123$ is the total number of trees in the six species. The R commands for this computation are:

```
tf<-c(40, 33, 17, 12, 13, 8); n<-sum(tf);
h<-(sum(log(1:n))-sum(lfactorial(tf)))/n
```

Maximum Brillouin diversity: for 6 species, this is

$$H_{\max} = \frac{\log n! - (k-d)\log c! - d\log(c+1)!}{n} \approx 1.6998$$

using $c = 20$ and $d = 3$ from $n = ck + d$, since there are $n = 123$ trees distributed amoong $k = 6$ species. The R commands for this computation are:

```
tf<-c(40, 33, 17, 12, 13, 8); n<-sum(tf); k<-length(tf); c<-20; d<-3;
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 $J = H/H_{\max} \approx 0.9097132 \approx 91.0\%$. □

4. A debate team has 8 Klingons and 7 Vulcans.

   (a) How many distinct mixed pairs (one Klingon and one Vulcan) can be formed using members of the team?

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

   Solution: (a) There are $(8)(7) = 56$ distinct pairs containing one Klingon and one Vulcan.

   (b) There are $\binom{8}{2}\binom{7}{2} = (28)(21) = 588$ ways to pick two (of 8) Klingons and two (of 7) Vulcans. There are two ways to make two mixed pairs from these choices, so there are $(2)(28)(21) = 1176$ distinct matchups. □

5. A DNA modeling kit contains 13 base units: 4 A's, 2 C's, 4 G's, and 3 T's.

   (a) How many distinct sequences of length 2 can be formed from this kit?

   (b) How many distinct sequences of length 13 can be formed from this kit?

   (c) How many distint sequences of length 3 can be formed from this kit?

   (d) How many distinct sequences of length 6 can be formed from this kit? Of length 9? Of length 12? (Hint: use `deduct()`.)

   Solution: (a) Since there at least 2 of each base unit, the number of distinct sequences of length is $4^2 = 16$.

   (b) There are ${}_{13}P_{4,2,4,3}$ distinct ways to form a sequence using all 13 letters from the set, where

$$ {}_{13}P_{4,2,4,3} = \frac{13!}{4!2!4!3!} = 900900. $$

   (c) This may be done by hand using a tree. There are 63 distinct sequences of length 3.

   (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:

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

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

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

Running either program with $m = 6$, 9, or 12 gives 3210, 98 700, and 900 900, respectively. As a check, note that $m = 3$ gives 63 and $m = 13$ gives 900 900, in agreement with the other calculations. □

6. Subsets A, B, C, D, E satisfy $B \subset A$, $C \subset B$, $D \subset B$, $C \cap D = \emptyset$, and $A \cap E = \emptyset$.

   (a) Depict the sets using a Venn diagram.

   (b) Is $C \cap E = \emptyset$?

   Solution: (a) See the figure below.

   (b) Yes, the conditions imply that $C \cap D = \emptyset$, since $C \subset B \subset A$ and $A \cap E = \emptyset$. □

7. A standard set of 52 playing cards is divided into 4 suits of 13 ranks each: ace, 2, 3, 4, 5, 6, 7, 8, 9, 10, jack, queen, and king. The suits are called clubs, diamonds, hearts and spades, with clubs and spades being "black" and hearts and diamonds being "red."

   (a) Taking 1 card at random, what is the probability of drawing a king of spades? a black king? a face card (jack, queen, or king)?

   (b) Taking 2 cards at random without replacement, what is the probability of drawing a pair of kings? a pair of clubs? a pair of black cards? a pair of cards of different ranks and suits?

   (c) Taking 5 cards at random without replacement, what is the probability of drawing a "full house," namely 3 cards of one rank and 2 cards of a second rank?

   Solution: (a) With 1 card:

   P(king of spades)= $1/52 \approx 0.0192$;

   P(black king)= $2/52 \approx 0.0385$;

   P(face card)= $12/52 \approx 0.231$.

   (b) With 2 cards:

   P(pair of kings)= $\binom{4}{2} / \binom{52}{2} = \frac{4}{52} \times \frac{3}{51} \approx 0.00452$, or `choose(4,2)/choose(52,2)` in R.

   P(pair of clubs)= $\binom{13}{2} / \binom{52}{2} = \frac{13}{52} \times \frac{12}{51} \approx 0.0588$, or `choose(13,2)/choose(52,2)` in R.

   P(pair of black cards)= $\binom{26}{2} / \binom{52}{2} = \frac{26}{52} \times \frac{25}{51} \approx 0.245$, or `choose(26,2)/choose(52,2)` in R.

   P(pair of cards of different ranks and suits)= $\frac{52}{52} \times \frac{36}{51} \approx 0.706$.

(c) With 5 cards:

$$\text{P(full house)} = \frac{\binom{13}{1}\binom{4}{3}\binom{12}{1}\binom{4}{2}}{\binom{52}{5}} = \frac{(13 \times 12)(4)(6)(5!)}{52 \times 51 \times 50 \times 49 \times 48} \approx 0.00144$$

The choice sequence is first rank (of 13), three suits at that rank (of 4), second rank (of the remaining 12), two suits at that second rank (of 4). Divide by the number of 5-card hands (of 52 cards). Compute the value with the R commands:

```
choose(13,1)*choose(4,3)*choose(12,1)*choose(4,2)/choose(52,5).
```

Alternatively, choose two ranks (of 13), choose the three-card rank (of 2), choose three suits at that rank (of 4), choose two suits at the second rank (of 4). Divide by the number of 5-card hands (of 52 cards). Compute the value with the R commands:

```
choose(13,2)*choose(2,1)*choose(4,3)*choose(4,2)/choose(52,5).
```

Both formulas yield the same result. □

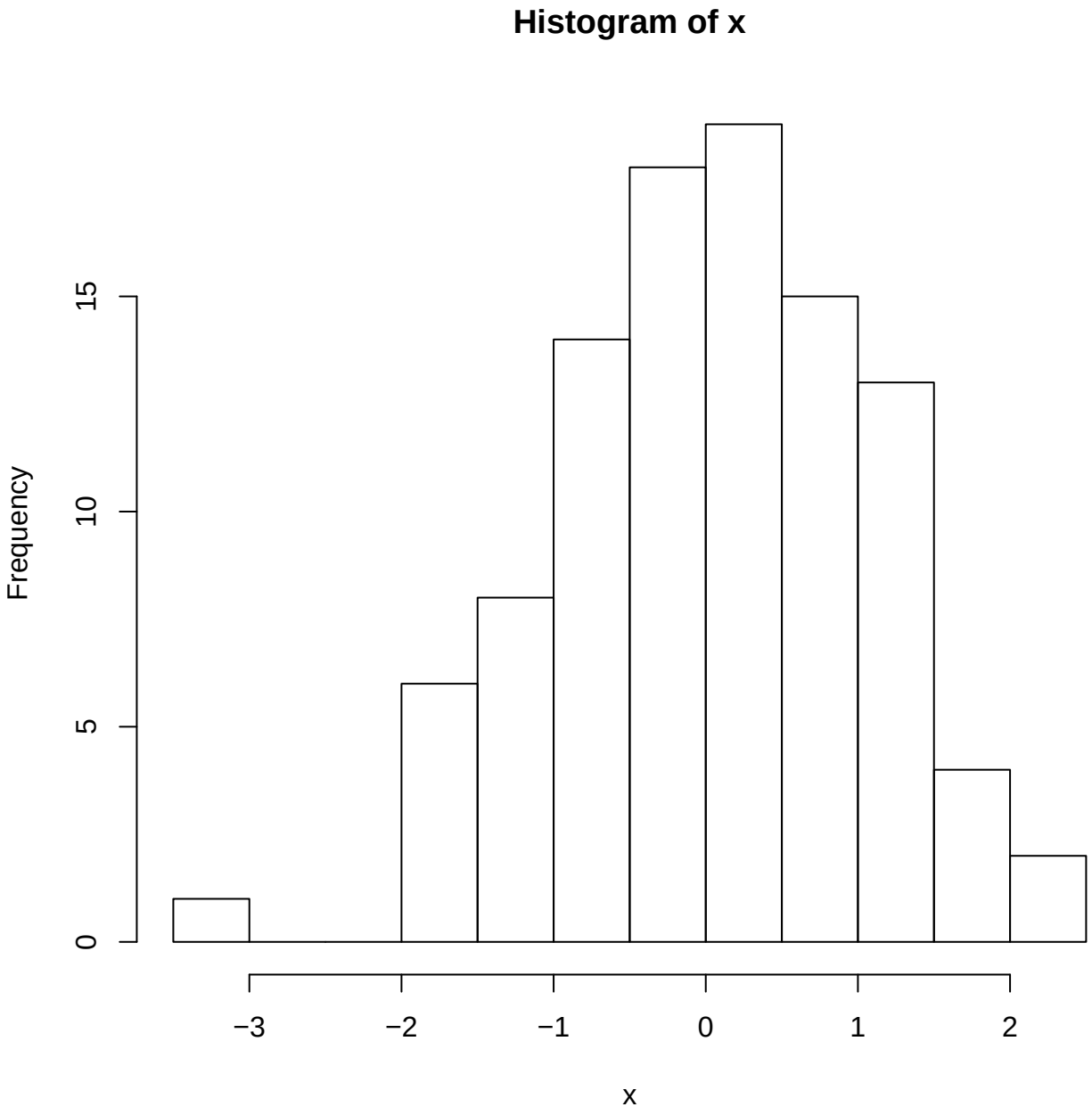

Figure 1: Histogram for Problem 1(a)

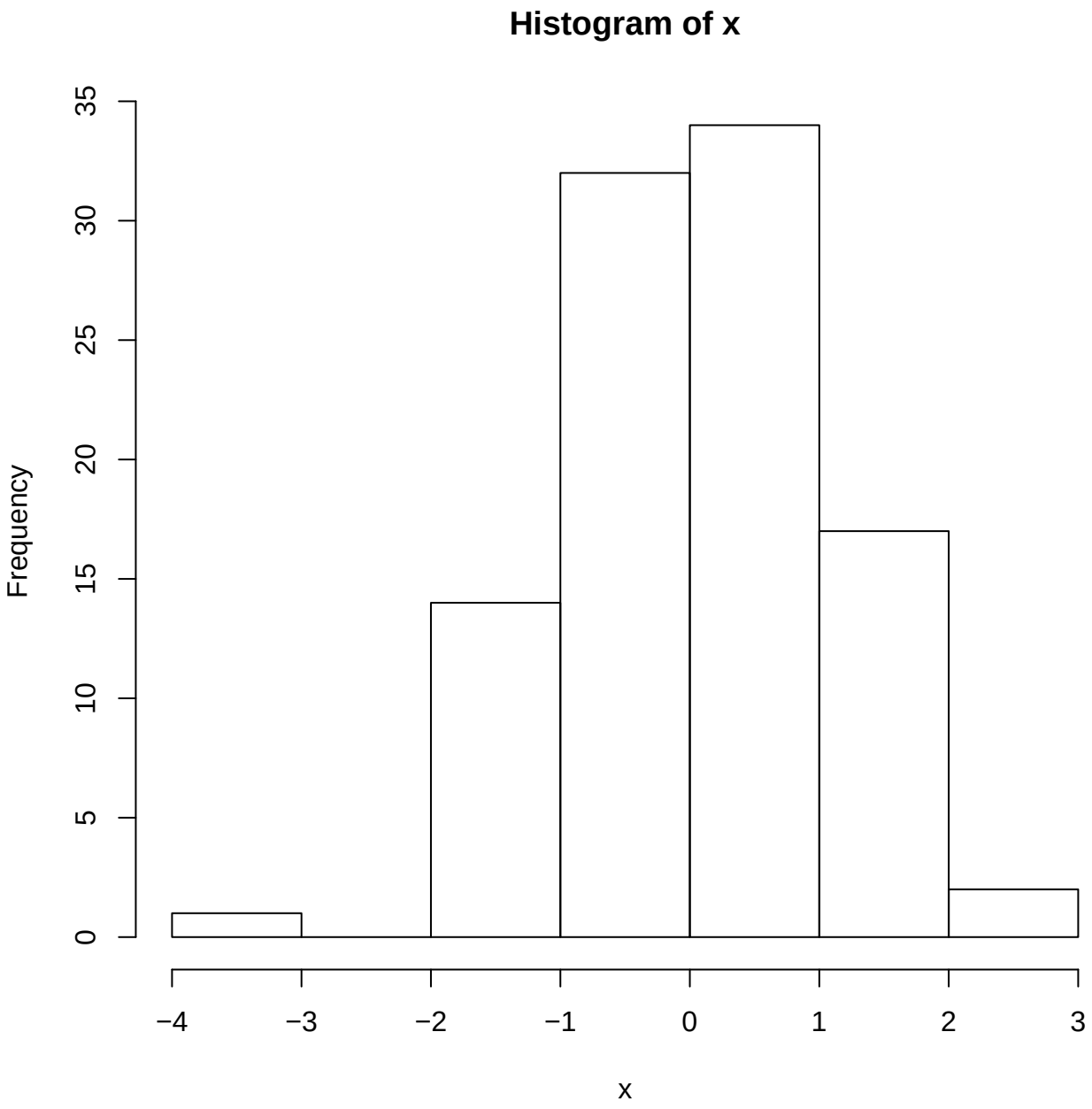

Figure 2: Histogram for Problem 1(b)

Homework 1, Ex.6a