Probability and Statistics for Engineers
Homework One
TMATH 390

Ben Foster*
Instructor: Julia Eaton

April 7, 2015

*Institute of Technology, University of Washington Tacoma

Contents

Problem 1

Collect data according to the following specifications. Most sources are allowed: web, books, papers, your own work, etc. but do not use the preloaded data that comes with R.

Specifications:

- Number of cases: 10 or more.
- 2 categorical or discrete variables.
- 2 continuous quantitative variables.
- These 4 variables must relate to a common problem; do not give me 4 unrelated variables.

Turn in a print out of the data. We will use this data in future problems.

Artist	Genre	Gender	Avg. length of song	Avg. $\frac{words}{song}$
Jimi Hendrix	Classic Rock	Male	7.35 min	189.4
Madonna	Pop	Female	5.22 min	120.2
Skrillex	Dubstep	Male	4.45 min	13.7
Eminem	Rap	Male	6.14 min	428.38
Aesop Rock	Rap	Male	7.78 min	932.2
Sia	Pop/Rock	Female	4.69 min	279.9
Lorde	Pop	Female	3.58 min	155.01
Beck	Alt Rock	Male	4.57 min	180.34
R. Kelly	R&B	Male	4.38 min	210.23
James Brown	R&B	Male	5.27	280.56

Table 1: Solo Artists

Problem 2

Temperature transducers of a certain type are shipped in batches of 50 units. A sample of 60 batches was selected, and the number of transducers in each batch not conforming to design specifications was determined, resulting in the following data:

2	1	2	4	0	1	3	2	0	5	3	3	1	3	2	4	7	0	2	3
0	4	2	1	3	1	1	3	4	1	2	3	2	2	8	4	5	1	3	1
5	0	2	3	2	1	0	6	4	2	1	6	0	3	3	3	6	1	2	3

(a) Determine the frequencies and relative frequencies for the observed values of x, which is the number of nonconforming transducers in a batch.
(b) What proportion of batches in the sample have at most five nonconforming transducers?

Answer to a There is a way in R to input data and get the frequencies of each. To do so by hand, you would count each time the specific number showed up in the data presented and keep a running count which would be the "frequency" of each. To find the "relative frequency" of each, you would divide each of the frequencies by the total number of data in the sample. Here is the code you would type in R to figure out the answers the easy way:

```
> data <- c(2,1,2,4,0,1,3,2,0,5,3,3,1,3,2,4,7,0,2,
3,0,4,2,1,3,1,1,3,4,1,2,3,2,2,8,4,5,1,3,1,5,0,2,3,
2,1,0,6,4,2,1,6,0,3,3,3,6,1,2,3)

> table(data)

> table(data) / length(data)
```

Here is the table of Frequency and Relative Frequency that came from R:

Value	Freq	Relative Freq.
0	7	0.11667
1	12	0.20
2	13	0.21667
3	14	0.23333
4	6	0.10
5	3	0.050
6	3	0.050
7	1	0.01667
8	1	0.01667

Answer to b To do this problem by hand, you would have to take the relative frequencies of all the values less than or equal to five and add them together. To do this in R, you just have to type the command `> sum(data <= 5) / length(data)` and that will give you the answer which is 0.91667.

Problem 3

A continuous variable x is said to have a *uniform* distribution if the density function is given by

$$f(x) = \begin{cases} \frac{1}{b-a}, & \text{if } a < x < b \\ 0, & otherwise. \end{cases}$$

The corresponding density "curve" has constant height over the interval from a to b. Suppose the time taken by a clerk to process a certain application form has a uniform distribution with $a = 4$ minutes and $b = 6$ minutes.

(a) Verify that the total area under the curve is indeed 1.
(b) In the long run, what proportion of forms will take between 4.5 min and 5.5 min to process?
(c) What value separates the slowest 50% of all processing times from the fastest 50% (the median of the distribution)?
(d) What value separates the fastest 10% of all processing times from the remaining 90%?

Answer to a In order to answer question a by hand, you would have to take the integral of the function from a to b with respect to x and the answer had to be equal to one. In this case, $a = 4$ and $b = 6$ which makes it a little easier to answer. In R, all you have to do is define the function and then integrate it. Here is the code to do that:

```
# Define the function as well as a and b
> fx <- function(x) \{ifelse(x > a & x < b,
1 / (b - a), 0)\}
> a = 4
> b = 6

# Integrate the function
> integrate(fx, a, b)
```

The answer you would get from this is 1. Thus verifying the total area under the curve.

Here's how to do it by hand:

$$\int_a^b \frac{1}{b-a} dx = \int_4^6 \frac{1}{6-4} dx = \frac{1}{2} \int_4^6 1 dx = \frac{1}{2} [6-4] = 1$$

Answer to b To answer this question using R, you would keep the same function that was defined as above and you would also use the integrate() function, however you would replace a with 4.5 and b with 5.5 like so: `> integrate(fx, 4.5, 5.5)`

Here's how to do it by hand:

$$\int_{4.5}^{5.5} \frac{1}{6-4} dx = \frac{1}{2} \int_{4.5}^{5.5} 1 dx = \frac{1}{2} \left[5.5 - 4.5\right] = 0.5$$

Answer to c Since it is easier to just answer this question by hand, then that is how I will do it. We have to find the value b such that $\int_4 b \frac{1}{6-4} dx = 0.5$.

$$\int_4^b \frac{1}{6-4} dx = \frac{1}{2} \int_4^b 1 dx = \frac{1}{2} \left[b - 4\right]$$

$$\frac{1}{2} \left[b - 4\right] = 0.5$$

$$b - 4 = 1$$

$$b = 5$$

Answer to d The process of figuring out the answer to this problem is very similar to the last one. All we have to do in this case is find the value b such that $\int_4 b \frac{1}{6-4} dx = 0.1$.

$$\int_4^b \frac{1}{6-4} dx = \frac{1}{2} \int_4^b 1 dx = \frac{1}{2} \left[b - 4\right]$$

$$\frac{1}{2} \left[b - 4\right] = 0.1$$

$$b - 4 = 0.2$$

$$b = 4.2$$

Problem 7

In the long run, what proportion of values selected from the standard normal distribution will satisfy each of the following conditions?

(a) Be at most 1.78?
(b) Be between 0.21 and 1.21?
(c) Be at least 2.00?

Overall note: The answers to these three problems can be figured out using the standard normal distribution table given to us or they can be solved using R.

Answer to a In order to figure out the answer to a, I decided to use R rather than try and figure out the answer from the handout. All I had to do was type in the command `> pnorm(1.78)`. The answer given by R is 0.962462.

Answer to b A similar process was used to find the answer to this problem. Since the `pnorm()` function finds the proportion of values to the left of the value given, then we needed to subtract the pnorm of 0.21 from 1.21. Here was the command I typed into R: `> pnorm(1.21) - pnorm(0.21)`. The answer given is 0.3036944.

Answer to c Since the standard normal distribution is symmetrical, then we can still use the pnorm function to find the proportion of values on the right side of the value given (at least a certain value). All we have to do is make the value negative like so: `> pnorm(-2.0)`. The given answer is 0.02275013.

Problem 8

Determine the following percentiles for the standard normal distribution:

(a) 91st percentile.
(b) 22nd percentile.
(c) 99.9th percentile.

Overall note: The answers to these three problems can be figured out using the standard normal distribution table given to us or they can be solved using R.

Answer to a In order to find the answer to this question, we would search in the handout given to us and find the value from that. Instead, we can use the `qnorm()` function in R. All we have to do is type in the command `> qnorm(0.91)` and we will get the answer 1.340755.

Answer to b The answer to this question is extremely similar. All we have to do is change the value given to us. This time, we type in the command `> qnorm(0.22)`. From this, we get the answer -0.7721932.

Answer to c The answer to this question is extremely similar. All we have to do is change the value given to us. This time, we type in the command `> qnorm(0.999)`. From this, we get the answer 3.090232.

Problem 9

The article "Characterization of Room Temperature Damping in Aluminum-Indium Alloys" suggests that $x = A1$ matrix grain size (μm) for an alloy consisting of 2% indium could be modeled with a normal distribution having $\mu = 96$ and $\sigma = 14$.

(a) What proportion of grain size exceed 100?
(b) What proportion of grain sizes are between 50 and 75?
(c) What interval (a, b) includes the central 90% of all grain sizes (so that 5% are below a and 5% are above b)?

Answer to a Since we are using a standard distribution, then we can alter this to become a standard normal distribution and use the handout by using the equation $z = \frac{x-\mu}{\sigma}$. We could also use the `pnorm()` function in R since it can handle different μ's and σ's. The command we input into R is: `> pnorm(96 - (100 - 96), 96, 14)`. This gives us an answer of 0.3875485. What was inside the pnorm function was a little different than what was expected because of the mu shift and because we are exceeding 100.

Answer to b Solving this problem is a little bit easier. It is very similar to question 7b. All we have to do here is type in the following command in R: `> pnorm(75, 96, 14) - pnorm(50, 96, 14)`. This yielded the answer: 0.06629858.

Answer to c The solution to this problem comes from finding which values along the standard distribution gives us the central 90% of all grain sizes. We do this by using the table or R to find what values cover the 5th and 95th percentile and that will be our range. For a we just have to input the command `> qnorm(0.05, 96, 14)` into R. For b we have to input the command `> qnorm(0.95, 96, 14)` into R. From inputting these commands, the answer is (72.97205, 119.028).

Problem 10

A transformation of data values by means of some mathematical function, such as x^2 or $\frac{1}{x}$, can often yield a set of numbers that has "nicer" statistical properties than the original data. In particular, it may be possible to find a function for which the histogram of transformed values is more symmetric (or even better, more like a bell-shaped curve) than the original data. For example, the article "Time Lapse Cinematographic Analysis of Beryllium Lung Ibroblast Interactions" (Envir. Research, 1983: 34-43) reported the results of experiments designed to study the behavior of certain individual cells that has been exposed to beryllium. An important characteristic of such an individual cell is its interdivision time (IDT). IDTs were determined for a number of cells both in exposed (treatment) and unexposed (control) conditions. The authors of the article used a logarithmic transformation. Consider the following representative IDT data:

28.1, 31.2, 13.7, 46.0, 25.8, 16.8, 34.8, 62.3, 28.0, 17.9, 19.5, 21.1, 31.9, 28.9, 60.1, 23.7, 18.6, 21.4, 26.6, 26.2, 32.0, 43.5, 17.4, 38.8, 30.6, 55.6, 25.5, 52.1, 21.0, 22.3, 15.5, 36.3, 19.1, 38.4, 72.8, 48.9, 21.4, 20.7, 57.3, 40.9.

Construct a histogram of this data based on classes with boundaries 10, 20, 30... Then calculate log10(x) for each observation, and construct a histogram of the transformed data using class boundaries 1.1, 1.2, 1.3... What is the effect of the transformation?

Answer We can easily create a histogram of both sets of data in R. All we have to do is input the following commands in R:

```
# Put the data into an array for future uses
> IDTdata <- c(28.1, 31.2, 13.7, 46.0, 25.8, 16.8,
34.8, 62.3, 28.0, 17.9, 19.5, 21.1, 31.9, 28.9, 60.1,
23.7, 18.6, 21.4, 26.6, 26.2, 32.0, 43.5, 17.4, 38.8,
30.6, 55.6, 25.5, 52.1, 21.0, 22.3, 15.5, 36.3, 19.1,
38.4, 72.8, 48.9, 21.4, 20.7, 57.3, 40.9)

# Create a histogram that displays the data.
> hist(IDTdata)

# Transform the data with $log_10$
> IDTlogData <- log10(IDTdata)

# Create a histogram that displays the new data
> hist(IDTlogData)
```

There is a difference between the two graphs. The logarithmic one has just one more bar representing data. This is mainly because of the data that the `hist()` function decided to show. If we had showed all data and chosen smaller breaks than R, then we should have seen basically the same histogram.

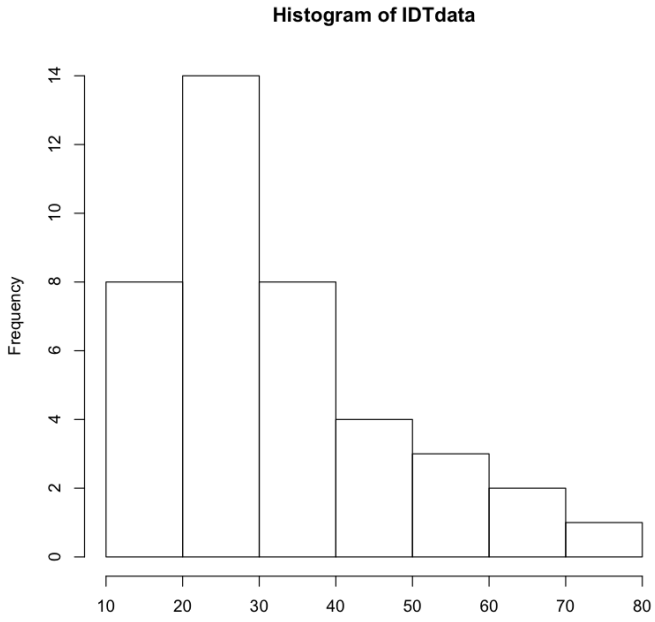


Figure 1: IDT Data Histogram

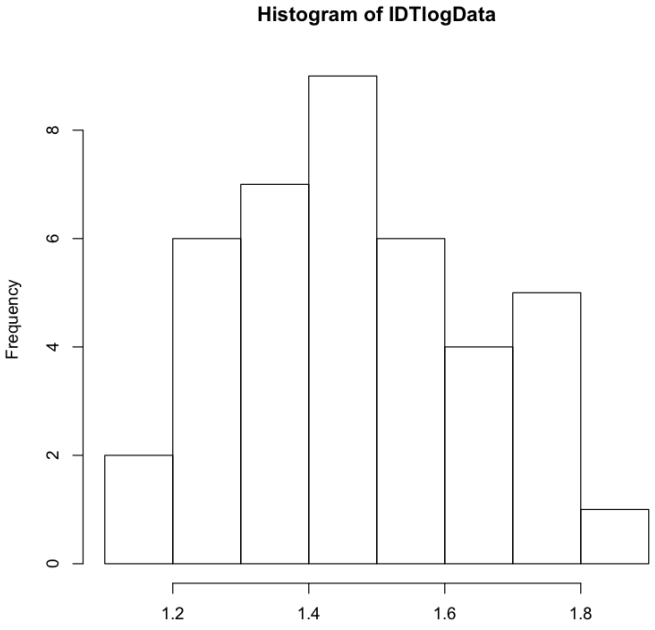


Figure 2: IDT Logarithmic Data Histogram

Problem 11

A mail-order computer business has six telephone lines. Let x denote the number of lines in use at a specified time. Suppose the mass function of x is given by

p(0) = 0.10, p(1) = 0.15, p(2) = 0.20, p(3) = 0.25,
p(4) = 0.20, p(5) = ????, p(6) =????.

(a) In the long run, what proportion of the time will at most three lines be in use?
(b) In the long run, what proportion of the time will at least five lines be in use?
(c) In the long run, what proportion of the time will between two and four lines, inclusive, be in use?
(d) In the long run, what proportion of the time will at least four lines NOT be in use?

Answer to a In order to solve this problem, we were already given all the data we need and all we have to do is sum up the given proportions for under and including 3 telephone lines. In this case, the answer is $P(x \leq 3) = p(0) + p(1) + p(2) + p(3) = 0.10 + 0.15 + 0.20 + 0.25 = 0.7$.

Answer to b The way to solve this problem in nearly the same as the previous way except we are summing together $p(5)$ and $p(6)$. We don't know what these values are, but it's easy to find out what the two of them summed together are. The answer is $P(x \geq 5) = 1 - P(x \leq 4) = 1 - (P(x \leq 3) + P(4)) = 1 - (0.70 + 0.20) = 0.10$.

Answer to c The way to solve this problem is almost exactly the same as problem a, except this time we are summing up $p(2)$, $p(3)$, and $p(4)$. The answer is $P(2 \leq x \leq 4) = p(2) + p(3) + p(4) = 0.20 + 0.25 + 0.20 = 0.65$.

Answer to d When answering this question, we have to think about which proportions we have to sum up so we have to think about the way that this problem is worded. When saying that at least four lines are not in use, then that means that, since there are only 6 lines, at most 2 lines are in use. The answer to this question is $P(x \leq 2) = p(0) + p(1) + p(2) = 0.10 + 0.15 + 0.20 = 0.45$.

Problem 12

Suppose the density function for x is given by the normal distribution with parameters μ, σ.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

(a) Compute the density function $f(z)$ for $z = \frac{x-\mu}{\sigma}$.

(Hint: Recall that $f(z)$ must satisfy $\int_{-\infty}^{\infty} f(z)dz = 1$. So start with $\int_{-\infty}^{\infty} f(x)dx = 1$ with $f(x)$ as above, then make the substitution so it looks like $\int_{-\infty}^{\infty} f(z)dz = 1$. The new integrand after the substitution is the density for the variable z.)

(b) From the form of $f(z)$, read off its μ and σ parameters.

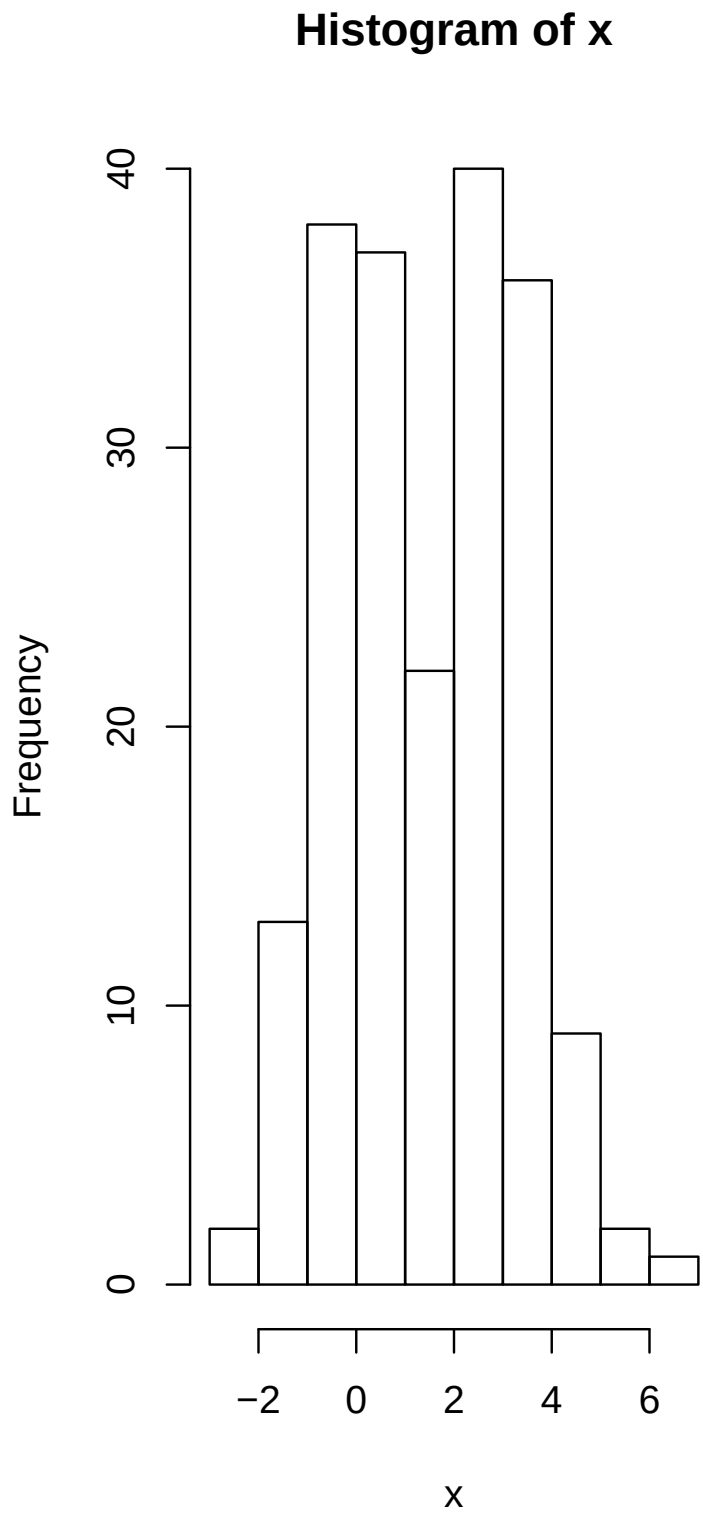
Answer to a Understanding this question was a little difficult at first. When we substitute z in for x in the standard distribution equation given above, then end up with the standard normal distribution where $\mu = 0$ and $\sigma = 1$. In this case, the answer is:

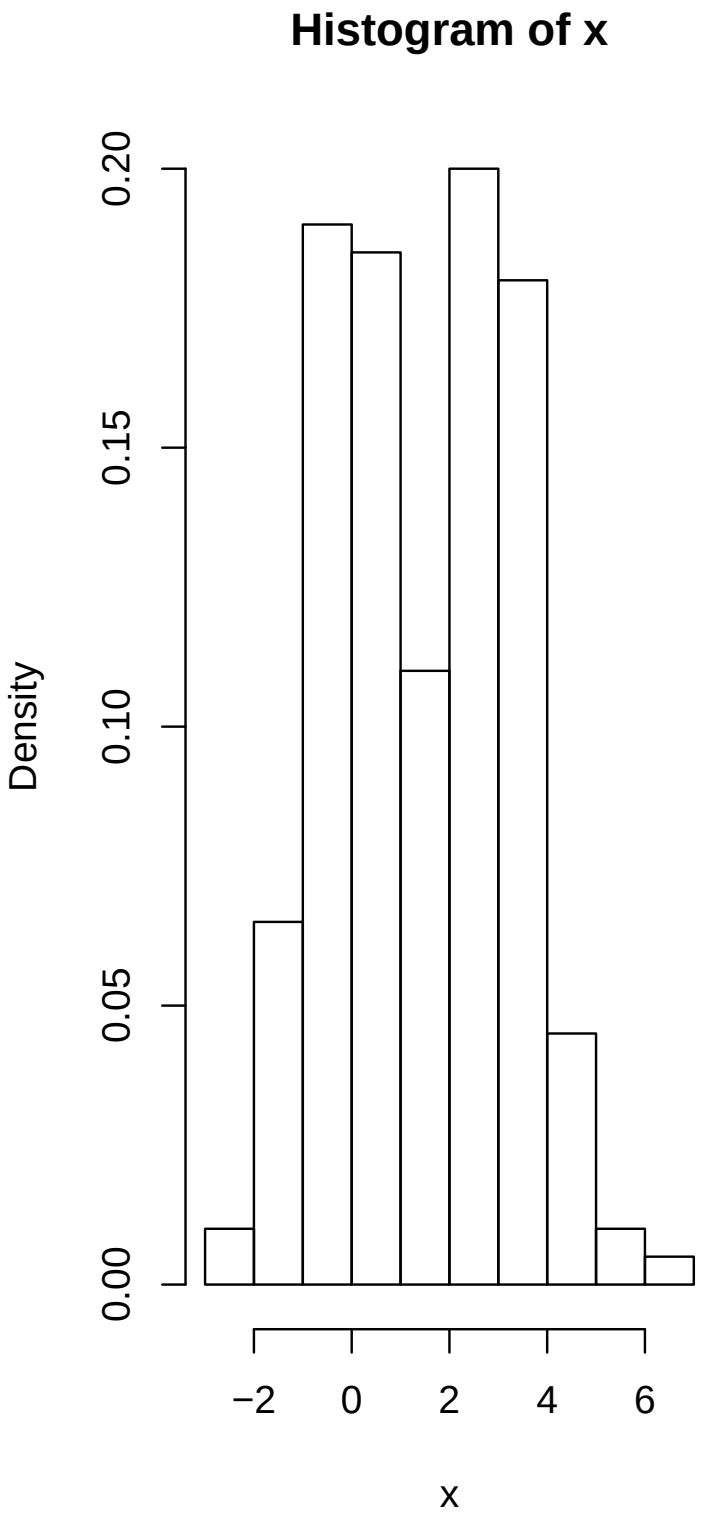
$$f(z) = \frac{1}{\sqrt{2\pi(1^2)}} e^{-\frac{1}{2}\left(\frac{z-0}{1}\right)^2}$$

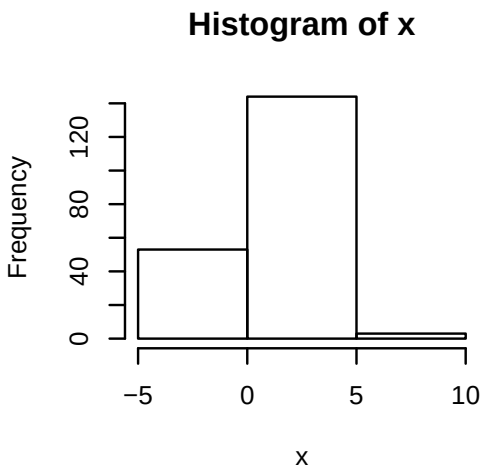
Answer to b According to the equation found in the previous part of this problem, $\mu = 0$ and $\sigma = 1$.

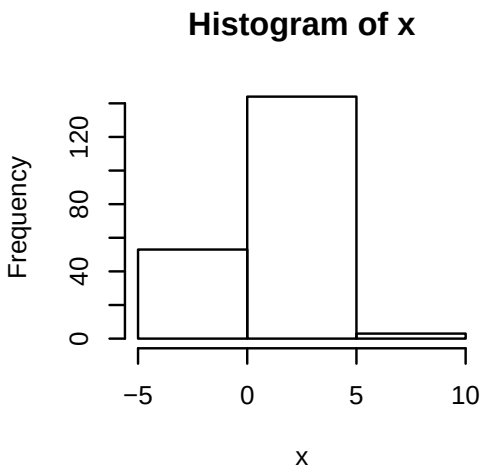
Lab 1 Printouts

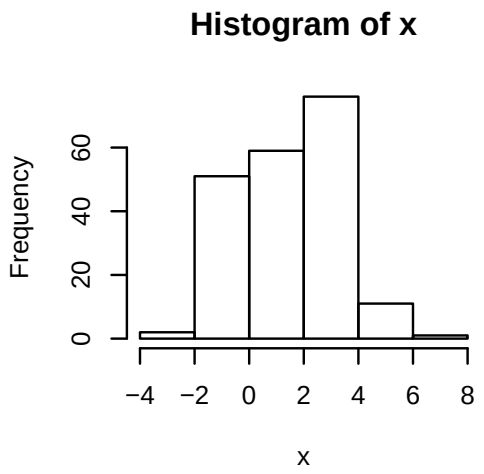
Attached are the two printouts required for Lab 1.

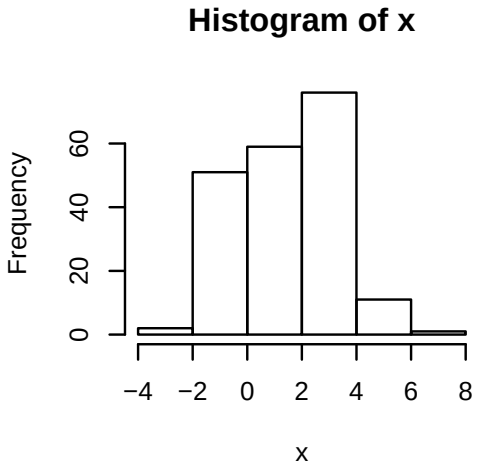

Histogram of x
Frequency
0
10
20
30
40
−2
0
2
4
6
x

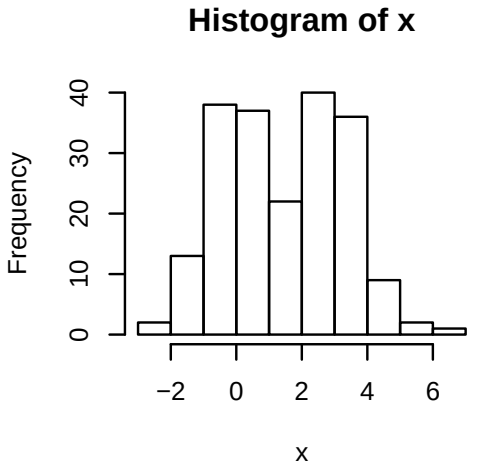

Histogram of x
Density
0.00
0.05
0.10
0.15
0.20
−2
0
2
4
6
x

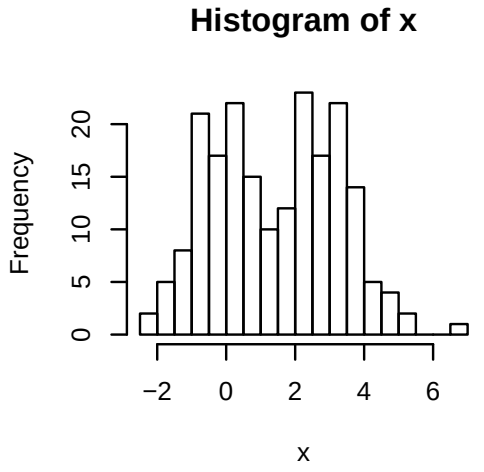

Histogram of x
Frequency
0
40
80
120
−5
0
5
10
x

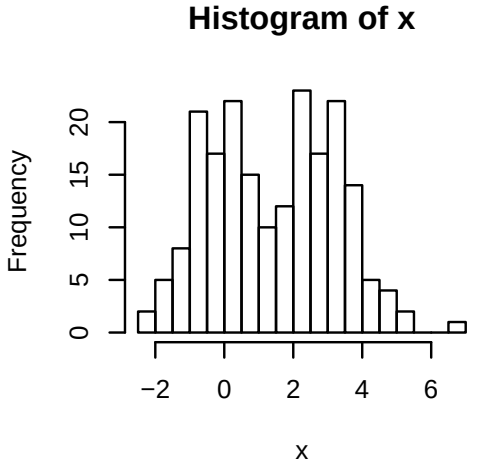

Histogram of x
Frequency
0
40
80
120
−5
0
5
10
x

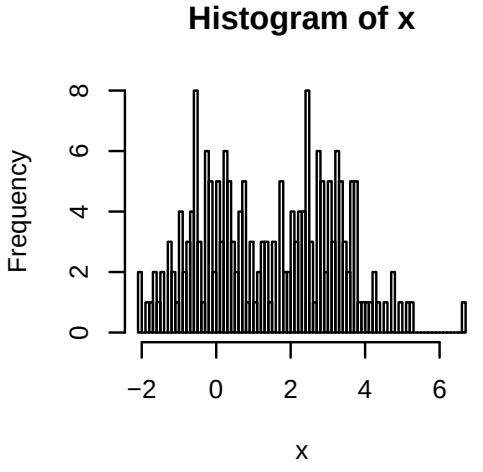

Histogram of x
Frequency
0
20
40
60
−4
−2
0
2
4
6
8
x

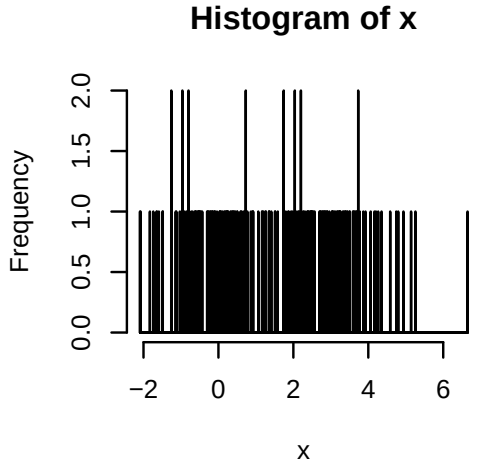

Histogram of x
Frequency
0
20
40
60
−4
−2
0
2
4
6
8
x


Histogram of x
Frequency
0
10
20
30
40
−2
0
2
4
6
x


Histogram of x
Frequency
0
5
10
15
20
−2
0
2
4
6
x


Histogram of x
Frequency
0
5
10
15
20
−2
0
2
4
6
x


Histogram of x
Frequency
0
2
4
6
8
−2
0
2
4
6
x


Histogram of x
Frequency
0.0
0.5
1.0
1.5
2.0
−2
0
2
4
6
x