5.3 Problems

1. Computer Simulation of Distributions

(a) In Problem 4.4/8(c), it was formally proved that if \( Z \) follows a standard normal distribution – i.e., has density function

\[ \phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \]

– then its expected value is given by the mean \( \mu = 0 \). A practical understanding of this interpretation can be achieved via empirical computer simulation. For concreteness, suppose the random variable \( Z = \text{“Temperature (°C)”} \sim N(0°, 1°) \). Let us consider a single sample of \( n = 400 \) randomly generated z-value temperature measurements from a frozen lake, and calculate its mean temperature \( \bar{z} \), via the following R code.

```r
# Generate and display one random sample.
sample <- rnorm(400)
sort(sample)

# Compare density histogram of sample against population distribution Z \sim N(0, 1).
hist(sample, freq = F)
curve(dnorm(x), lwd = 2, col = "darkgreen", add = T)

# Calculate and display sample mean.
mean(sample)
```

Upon inspection, it should be apparent that there is some variation among these \( z \)-values.

```r
# Compare density histogram of sample against population distribution Z \sim N(0, 1).
hist(sample, freq = F)
curve(dnorm(x), lwd = 2, col = "darkgreen", add = T)

# Calculate and display sample mean.
mean(sample)
```

This sample mean \( \bar{z} \) should be fairly close to the actual expected value in the population, \( \mu = 0° \) (likewise, \( \text{sd(sample)} \) should be fairly close to \( \sigma = 1° \)), but it is only generated from a single sample. To obtain an even better estimate of \( \mu \), consider say, 500 samples, each containing \( n = 400 \) randomly generated \( z \)-values. Then average each sample to find its mean temperature, and obtain \( \{ \bar{z}_1, \bar{z}_2, \bar{z}_3, \ldots, \bar{z}_{500} \} \).

```r
# Generate and display 500 random sample means.
zbars <- NULL
for (s in 1:500) {sample <- rnorm(400)
  zbars <- c(zbars, mean(sample))}
sort(zbars)

# Compare density histogram of sample means against sampling distribution \( \bar{Z} \sim N(0, 0.05) \).
hist(zbars, freq = F)
curve(dnorm(x, 0, 0.05), lwd = 2, col = "darkgreen", add = T)

# Calculate and display mean of the sample means.
mean(zbars)
```

Upon inspection, it should be apparent that there is little variation among these \( \bar{z} \)-values.

```r
# Compare density histogram of sample means against sampling distribution \( \bar{Z} \sim N(0, 0.05) \).
hist(zbars, freq = F)
curve(dnorm(x, 0, 0.05), lwd = 2, col = "darkgreen", add = T)

# Calculate and display mean of the sample means.
mean(zbars)
```

This value should be extremely close to the mean \( \mu = 0° \), because there is much less variation about \( \mu \) in the sampling distribution, than in the population distribution. (In fact, via the Central Limit Theorem, the standard deviation is now only \( \sigma/\sqrt{n} = 1°/\sqrt{400} = 0.05° \). Check this value against \( \text{sd(zbars)} \).)
(b) Contrast the preceding example with the following. A random variable \( X \) is said to
follow a **standard Cauchy** (pronounced “ko-shee”) distribution if it has the density
function \( f_{\text{Cauchy}} (x) = \frac{1}{\pi} \frac{1}{1 + x^2} \), for \(-\infty < x < +\infty\), as illustrated.

- First, as in Problem 4.4/7, formally prove that this is indeed a valid density function.
- However, as in Problem 4.4/8, formally prove – using the appropriate “expected value” definition – that the mean \( \mu \) in fact does not exist!

Informally, there are too many outliers in both tails to allow convergence to a single mean value \( \mu \). To obtain a better appreciation of this subtle point, we once again rely on computer simulation.

```r
# Generate and display one random sample.
sample <- rcauchy(400)
sort(sample)
```

Upon inspection, it should be apparent that there is *much* variation among these \( x \)-values.

```r
# Compare density histogram of sample against population distribution \( X \sim \text{Cauchy} \).
hist(sample, freq = F)
curve(dcauchy(x), lwd = 2, col = "darkgreen", add = T)
```

```r
# Calculate and display sample mean.
mean(sample)
```

This sample mean \( \bar{x} \) is not necessarily close to an expected value \( \mu \) in the population, nor are the means \( \{ \bar{x}_1, \bar{x}_2, \bar{x}_3, ..., \bar{x}_{500} \} \) of even 500 random samples:

```r
# Generate and display 500 random sample means.
xbars <- NULL
for (s in 1:500) {
sample <- rcauchy(400)
    xbars <- c(xbars, mean(sample))
}
sort(xbars)
```

Upon inspection, it should be apparent that there is still *much* variation among these \( \bar{x} \)-values.

```r
# Compare density histogram of sample means against sampling distribution \( \bar{X} \sim \text{Cauchy} \).
hist(xbars, freq = F)
curve(dcauchy(x), lwd = 2, col = "darkgreen", add = T)
```

```r
# Calculate and display mean of the sample means.
mean(xbars)
```

Indeed, it can be shown that \( \bar{X} \) follows a Cauchy distribution as well, i.e., the Central Limit Theorem fails! *Gathering more data does not yield convergence to a mean \( \mu \).*
2. For which functions in Problem 4.4/9 does the Central Limit Theorem hold / fail?

3. Refer to Problem 4.4/16.

(a) Suppose that a random sample of \( n = 36 \) males is to be selected from this population, and the sample mean cholesterol level calculated. As in part (f), what is the probability that this sample mean value is between 202 and 238?

(b) How large a sample size \( n \) is necessary to guarantee that 80% of the sample mean values are within 5 mg/dL of the mean of their distribution? (Hint: First find the value of \( z \) that satisfies \( P(-z \leq Z \leq +z) = 0.8 \), then change back to \( X \), and solve for \( n \).)

4. Suppose that each of the four experiments in Problem 4.4/31 is to be performed \( n = 9 \) times, and the nine resulting distances averaged. Estimate the probability \( P(2 \leq \bar{X} \leq 4) \) for each of (a), (b), (c), and (d). [Note: Use the Central Limit Theorem, and the fact that, for (c), the mean and variance are \( \mu = 3 \) and \( \sigma^2 = 1.5 \), respectively.]

5. Bob suddenly remembers that today is Valentine’s Day, and rushes into a nearby florist to buy his girlfriend some (overpriced) flowers. There he finds a large urn containing a population of differently colored roses in roughly equal numbers, but with different prices: yellow roses cost $1 each, pink roses cost $2 each, and red roses cost $6 each. As he is in a hurry, he simply selects a dozen roses at random, and brings them up to the counter.

(a) Lowest and highest costs = ? How much money can Bob expect to pay, on average?

(b) What is the approximate probability that he will have to pay no more than $45? Assume the Central Limit Theorem holds.

(c) Simulate this in R: From many random samples (each with a dozen values) selected from a population of the prices listed above, calculate the proportion whose totals are no more than $45. How does this compare with your answer in (b)?

6. A geologist manages a large museum collection of minerals, whose mass (in grams) is known to be normally distributed, with some mean \( \mu \) and standard deviation \( \sigma \). She knows that 60% of the minerals have mass less than a certain amount \( m \), and needs to select a random sample of \( n = 16 \) specimens for an experiment. With what probability will their average mass be less than the same amount \( m \)? (Hint: Calculate the z-scores.)

7. Refer to Prob 4.4/5. Here, we sketch how formally applying the Central Limit Theorem to a binary variable yields the “normal approximation to the binomial distribution” (section 4.2). First, define the binary variable \( Y = \begin{cases} 1, \text{ with probability } \pi \\ 0, \text{ with probability } 1 - \pi \end{cases} \) and the discrete variable \( X = \#(Y = 1) \text{ in a random sample of } n \text{ Bernoulli trials} \sim \text{Bin}(n, \pi) \).

(a) Using the results of Problem 4.4/5 for \( \mu_Y \) and \( \sigma_Y^2 \), apply the Central Limit Theorem to the variable \( Y \).

(b) Why is it true that \( \bar{Y} = \frac{X}{n} \)? [Hint: Why is \( \#(Y = 1) \)” the same as \( \Sigma (Y = 1) ? \)]

Use this fact along with (a) to conclude that, indeed, \( X \approx \mathcal{N}(\mu_Y, \sigma_Y^2) \).

* Recall what the mean \( \mu_X \) and standard deviation \( \sigma_X \) of the Binomial distribution are.
8. Imagine performing the following experiment in principle. We are conducting a socioeconomic survey of an arbitrarily large population of households, each of which owns a certain number of cars $X = 0, 1, 2, \text{ or } 3$, as illustrated below. For simplicity, let us assume that the proportions of these four types of household are all equal (although this restriction can be relaxed).

Select $n = 1$ household at random from this population, and record its corresponding value $X = 0, 1, 2, \text{ or } 3$. By the “equal likelihood” assumption above, each of these four elementary outcomes has the same probability of being selected $(1/4)$, therefore the resulting uniform distribution of population values is given by:

<table>
<thead>
<tr>
<th>$x$</th>
<th>$f(x) = P(X = x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1/4</td>
</tr>
<tr>
<td>1</td>
<td>1/4</td>
</tr>
<tr>
<td>2</td>
<td>1/4</td>
</tr>
<tr>
<td>3</td>
<td>1/4</td>
</tr>
</tbody>
</table>

From this, construct the probability histogram of the “population distribution of $X$” values, on the graph paper below. Remember that the total area of a probability histogram $= 1$!

Next, draw an ordered random sample of $n = 2$ households, and compute the mean number of cars $\bar{X}$. (For example, if the first household has 2 cars, and the second household has 3 cars, then the mean for this sample is 2.5 cars.) There are $4^2 = 16$ possible samples of size $n = 2$; they are listed below. For each such sample, calculate and record its corresponding mean $\bar{X}$; the first two have been done for you. As above, construct the corresponding probability table and probability histogram of these “sampling distribution of $\bar{X}$” values, on the graph paper below. Remember that the total area of a probability histogram $= 1$; this fact must be reflected in your graph! Repeat this process for the $4^3 = 64$ samples of size $n = 3$, and answer the following questions.
1. Comparing these three distributions, what can generally be observed about their overall shapes, as the sample size $n$ increases?

2. Using the \textit{expected value} formula
\[
\mu = \sum_{x} x f(x),
\]
calculate the mean $\mu_x$ of the population distribution of $X$. Similarly, calculate the mean $\mu_{\bar{X}}$ of the sampling distribution of $\bar{X}$, for $n = 2$. Similarly, calculate the mean $\mu_{\bar{X}}$ of the sampling distribution of $\bar{X}$, for $n = 3$. Conclusions?

3. Using the \textit{expected value} formula
\[
\sigma^2 = \sum_{x} (x - \mu)^2 f(x),
\]
calculate the variance $\sigma_x^2$ of the population distribution of $X$. Similarly, calculate the variance $\sigma_{\bar{X}}^2$ of the sampling distribution of $\bar{X}$, for $n = 2$. Similarly, calculate the variance $\sigma_{\bar{X}}^2$ of the sampling distribution of $\bar{X}$, for $n = 3$. Conclusions?

4. Suppose now that we have some arbitrarily large study population, and a general random variable $X$ having an approximately symmetric distribution, with some mean $\mu_x$ and standard deviation $\sigma_x$. As you did above, imagine selecting all random samples of a moderately large, fixed size $n$ from this population, and calculate all of their sample means $\bar{X}$. Based partly on your observations in questions 1-3, answer the following.

(a) In general, how would the means $\bar{X}$ of most “typical” random samples be expected to behave, even if some of them do contain a few outliers, especially if the size $n$ of the samples is large? Why? Explain briefly and clearly.

(b) In general, how then would these two large collections – the set of all sample mean values $\bar{X}$, and the set of all the original population values $X$ – compare with each other, especially if the size $n$ of the samples is large? Why? Explain briefly and clearly.

(c) What effect would this have on the overall shape, mean $\mu_{\bar{X}}$, and standard deviation $\sigma_{\bar{X}}$, of the \textbf{sampling distribution} of $\bar{X}$, as compared with the shape, mean $\mu_x$ and standard deviation $\sigma_x$, of the \textbf{population distribution} of $X$? Why? Explain briefly and clearly.
<table>
<thead>
<tr>
<th>SAMPLES, ( n = 2 )</th>
<th>( \bar{X} )</th>
<th>SAMPLES, ( n = 3 )</th>
<th>( \bar{X} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Draw: 1st 2nd</td>
<td></td>
<td>Draw: 1st 2nd 3rd</td>
<td></td>
</tr>
<tr>
<td>1. 0 0</td>
<td>0.0</td>
<td>1. 0 0 0</td>
<td>0.00</td>
</tr>
<tr>
<td>2. 0 1</td>
<td>0.5</td>
<td>2. 0 0 1</td>
<td>0.33</td>
</tr>
<tr>
<td>3. 0 2</td>
<td></td>
<td>3. 0 0 2</td>
<td></td>
</tr>
<tr>
<td>4. 0 3</td>
<td></td>
<td>4. 0 0 3</td>
<td></td>
</tr>
<tr>
<td>5. 1 0</td>
<td></td>
<td>5. 0 1 0</td>
<td></td>
</tr>
<tr>
<td>6. 1 1</td>
<td></td>
<td>6. 0 1 1</td>
<td></td>
</tr>
<tr>
<td>7. 1 2</td>
<td></td>
<td>7. 0 1 2</td>
<td></td>
</tr>
<tr>
<td>8. 1 3</td>
<td></td>
<td>8. 0 1 3</td>
<td></td>
</tr>
<tr>
<td>9. 2 0</td>
<td></td>
<td>9. 0 2 0</td>
<td></td>
</tr>
<tr>
<td>10. 2 1</td>
<td></td>
<td>10. 0 2 1</td>
<td></td>
</tr>
<tr>
<td>11. 2 2</td>
<td></td>
<td>11. 0 2 2</td>
<td></td>
</tr>
<tr>
<td>12. 2 3</td>
<td></td>
<td>12. 0 2 3</td>
<td></td>
</tr>
<tr>
<td>13. 3 0</td>
<td></td>
<td>13. 0 3 0</td>
<td></td>
</tr>
<tr>
<td>14. 3 1</td>
<td></td>
<td>14. 0 3 1</td>
<td></td>
</tr>
<tr>
<td>15. 3 2</td>
<td></td>
<td>15. 0 3 2</td>
<td></td>
</tr>
<tr>
<td>16. 3 3</td>
<td></td>
<td>16. 0 3 3</td>
<td></td>
</tr>
</tbody>
</table>
Probability Histogram of Population Distribution of X
Probability Histogram of Sampling Distribution of $\bar{X}$, $n = 2$
Probability Histogram of Sampling Distribution of $\bar{X}$, $n = 3$