Solutions to Assignment #5

1. (Typographical Errors[1]) Typographical and spelling errors can be either "nonword errors" or "word errors." A nonword error is not a real word, as when "the" is typed as "teh." A word error is a real word, but not the right word, as when "lose" is typed as "loose." Spell–checking software will catch nonword errors but not word errors. Human proofreaders catch around 70% of word errors. You ask a fellow student to proofread and essay in which you have deliberately made 10 word errors.

   (a) If the student matches the usual 70% rate, what is the distribution of the number of errors caught? What is the distribution of the number of errors missed?

   Solution: Let X denote the number of word errors out of the 10 in the essay that are caught. Then, $X \sim B(n,p)$ where $n = 10$ and $p = 0.7$. The distribution for X is then given by

   $$p_X(k) = \binom{10}{k}(0.7)^k(0.3)^{10-k} \quad \text{for } k = 0, 1, 2, \ldots, 10.$$

   Similarly, if Y denotes the number of errors missed, then $Y \sim B(10, 0.3)$ and

   $$p_Y(k) = \binom{10}{k}(0.3)^k(0.7)^{10-k} \quad \text{for } k = 0, 1, 2, \ldots, 10.$$

   □

   (b) Missing 4 or more of the 10 errors seems a poor performance. What is the probability that a proofreader who catches 70% of the word errors misses 4 or more out of 10?

   Solution: We want $P(Y \geqslant 4)$. This is the same as $1 - P(Y \leqslant 3)$, or $1 - F_Y(3)$, where F_Y is the cumulative distribution function of Y. Using R to estimate $P(Y \geqslant 4)$ we obtain

   $$P(Y \geqslant 4) = 1 - \texttt{pbinom(3,10,0.3)} \approx 0.35$$

   or about 35%. □

[1] Adapted from Exercise 5.13 in Moore, McCabe abd Graig, *Introduction to the Practice of Statistics,* Sixth Edition, pp. 331–332

2. (Typographical Errors (continued)[2])

   (a) What is the mean number of errors caught? What is the mean number of errors missed?

   Solution: The mean number of errors caught is the expected value of X or $E(X) = 10 \cdot (0.7) = 7$.
   Similarly, the mean number of errors missed is the expected value of Y or $E(Y) = 10 \cdot (0.3) = 3$. □

   (b) What is the standard deviation, σ, of the number of errors caught?

   Solution: The variance of the number of errors caught is

   $$\text{Var}(X) = np(1-p) = 10(0.7)(0.3) = 2.1.$$

   Hence, the standard deviation of X is

   $$\sigma_X = \sqrt{Var(X)} \approx 1.45$$

   □

   (c) Suppose that a proof reader catches 90% of word errors, so that $p = 0.9$. What is the standard deviation, σ, in this case? What is σ is $p = 0.99$? What happens to the standard deviation as p approaches 1?

   Solution: $\sigma_X = \sqrt{np(1-p)}$, so that when $p = 0.9$, $\sigma_X = 0.3$; when $p = 0.99$, $\sigma_X \approx 0.0995$.

   $$\sigma_X \to 0 \quad \text{as} \quad p \to 1.$$

   □

3. Let X_i denote the number on the face that comes up in the i^{th} roll of a balanced die.

   (a) Give the expected value and variance for each X_i, for $i = 1, 2, 3, \ldots$

   Solution: Compute

   $$E(X_i) = 1p_{X_i}(1) + 2p_{X_i}(2) + \cdots + 6p_{X_i}(6) = 1\cdot\frac{1}{6} + 2\cdot\frac{1}{6} + \cdots + 6\frac{1}{6} = 3.5$$

[2] Adapted from Exercise 5.15 in Moore, McCabe abd Graig, *Introduction to the Practice of Statistics,* Sixth Edition, p. 332

Similarly,

$$\begin{aligned} E(X_i^2) &= 1^2 p_{X_i}(1) + 2^2 p_{X_i}(2) + \cdots + 6^2 p_{X_i}(6) \\ &= 1 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + \cdots + 36\frac{1}{6} \\ &= \frac{91}{6}. \end{aligned}$$

It then follows that the variance of each X_i is

$$\text{Var}(X_i) = E(X_i^2) - [E(X_i)]^2 = \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{35}{12}.$$

□

(b) n rolls of the die constitutes a random sample of size n:

$$X_1, X_2, X_3, \ldots, X_n.$$

Compute the sample mean

$$\overline{X}_n = \frac{X_1 + X_n + \cdots + X_n}{n},$$

and determine its expected value and variance.

Solution: $E(\overline{X}_n) = E(X_i) = 3.5$ for each i, and

$$\text{Var}(\overline{X}_n) = \frac{\text{Var}(X_i)}{n} = \frac{35}{12n}.$$

□

4. *(Continuation of Problem 3)*

(a) Use R to simulate rolling the die $n = 100$ times. This simulates collecting a random sample

$$X_1, X_2, X_3, \ldots, X_n$$

of size $n = 100$. Perform 1000 repetitions of the experiment to generate a simulation of the sampling distribution of $\overline{X}_n$. Plot a histogram of the simulations.

Solution: We can use the `sample()` function in R to generate samples of size 100 **with replacement** from a vector containing the numbers on the six faces of a die. The following R code generates 1000 of those samples, computes their means and stores them in a vector called `Xbar`.

```
faces <- 1:6
n <- 100
Nrep <- 1000    # number of repetitions
Xbar <-  mean(sample(faces,n,replace=T))
                # sets initial value of Xbar
L <- Nrep -1    # we need Nrep - 1 more repetitions
for (i in 1:L)  # Sets up a loop from 1 to L
{
Xbar <- c(Xbar, mean(sample(faces,n,replace=T)))
}
```

The histogram of `Xbar` is shown in Figure 1.

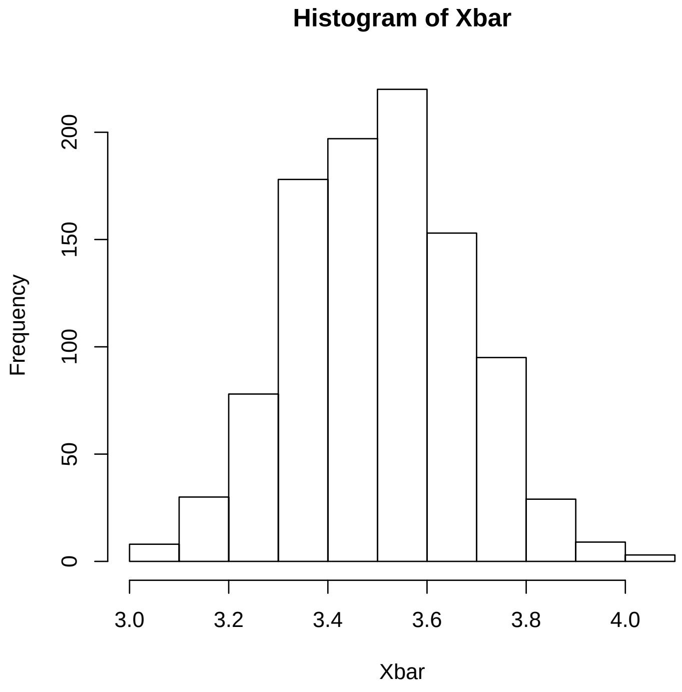

Figure 1: Histogram of `Xbar`

□

(b) Use the histogram generated in the previous part to obtain the probability distribution for the simulations of $\overline{X}_n$. In the same graph, plot the normal curve that approximates the distribution of $\overline{X}_n$ according to the Central Limit Theorem.

Solution: Typing `hist(Xbar,freq=FALSE)` in R yields the "density" histogram for `Xbar` shown in Figure 2. The density histogram in Figure 2 has the property that the area of the bars correspond to the probability that `Xbar` takes on values lying at the base of the bar. In this sense (i.e., in the sense of areas of the bars in the histogram), the picture in Figure 2 gives the probability distribution of `Xbar`.

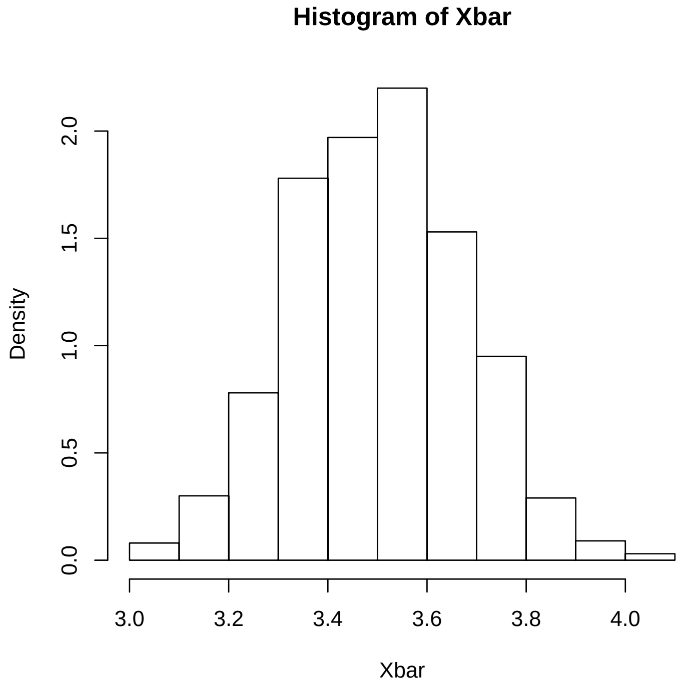

Figure 2: Density histogram of `Xbar`

To plot the normal approximation density given by the Central Limit Theorem, we need to plot the density function

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma}\, e^{-(x-\mu)^2/2\sigma^2} \quad \text{for } x \in \mathbf{R},$$

where $\mu = 3.5$ and $\sigma^2 = \dfrac{35}{1200} \approx 0.029167$. To superimpose the graph of f on the density plot in Figure 2, type the following commands in R:

```
x <- seq( 3, 4.1, length = 1000)
lines(x,dnorm(x,3.5,0.17078))
```

where 0.17078 is (approximately) the standard deviation of $\overline{X}_n$, or $\sqrt{0.029167}$. The graph is shown in Figure 3.

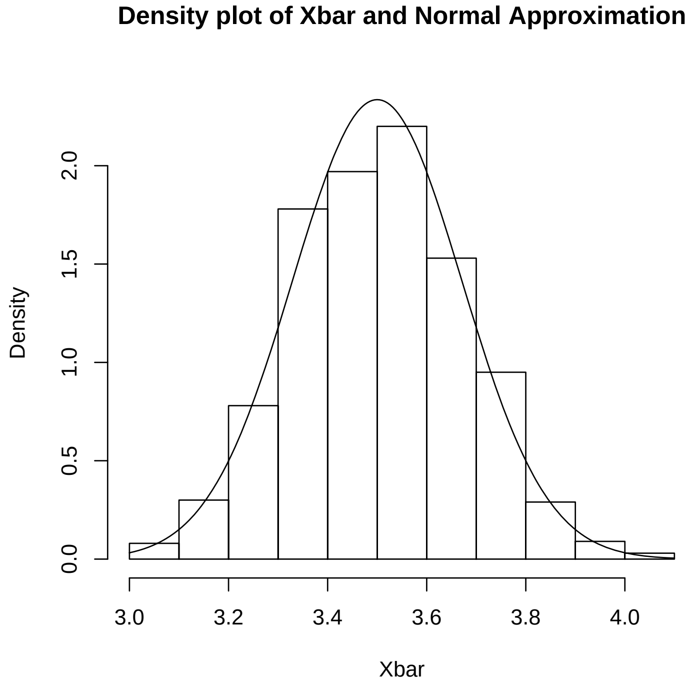

Figure 3: Density histogram of Xbar and Approximating Normal Density

□

5. (Fuel Efficiency[3]) Computers in some vehicles calculate various quantities related to performance. One of those quantities is fuel efficiency, or gas mileage, usually expressed as miles per gallon (mpg). For one vehicle equipped in this way, the mpg were recorded each time the gas tank was filled, and the com-

[3] Adapted from Exercise 6.26 in Moore, McCabe abd Graig, *Introduction to the Practice of Statistics,* Sixth Edition, p. 371

puter was then reset. The mpg values are contained in the MS Excel file `FuelEfficiencyData.xls`, which may be downloaded from

`http://pages.pomona.edu/~ajr04747`.

Assume that the standard deviation, σ, for the mpg random variable is known to be 3.5 mpg.

(a) Give the variance of the sample mean for a random sample of size n; that is, compute $\text{Var}(\overline{X}_n)$.

Solution: Compute

$$\text{Var}(\overline{X}_n) = \frac{\sigma^2}{n} = \frac{(3.5)^2}{20} \approx 0.6125.$$

□

(b) Give the 95% confidence interval for the true mean mileage of the vehicle.

Solution: The sample mean $\overline{X}_n$ for the $n = 20$ miles per gallon data is 43.2. This can be used as a point estimate and as the center of the 95% confidence interval

$$\left(\overline{X}_n - (1.96)\frac{\sigma}{\sqrt{n}}, \overline{X}_n + (1.96)\frac{\sigma}{\sqrt{n}}\right),$$

where

$$(1.96)\frac{\sigma}{\sqrt{n}} = (1.96)\frac{3.5}{\sqrt{20}} \approx 1.5.$$

Thus, the 95% confidence interval for the true mean mileage of the vehicle is $(41.7, 44.7)$. □