1.3 Estimating Central Tendency

As indicated in the previous section, statistics can be defined to estimate certain characteristics of a population distribution. One aspect of a population that is nearly always of interest is the central tendency of the population distribution. Several statistics that reflect this central tendency are formally defined in this section. We begin with the sample mean.

Sample mean

The **sample mean** is the most intuitive measure of central tendency. People naturally average data values in order to get a sense of the center of a probability distribution.

**Definition 1.3** Let $x_1, x_2, \ldots , x_n$ be experimental values associated with the random variables $X_1, X_2, \ldots , X_n$. The **sample mean** is

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$ 

As indicated previously, $\bar{X}$ is used in the abstract when there are no specific data values. When specific data values have been collected, the lower case version $\bar{x}$ is used to denote the sample mean. The sample mean is sometimes called the **sample arithmetic mean**.

**Example 1.15** Ten kindergarten children from ten different families are polled to find the number of children that are in their family. The resulting values, $x_1, x_2, \ldots , x_{10}$ are

$$3, 1, 5, 1, 3, 2, 1, 1, 3, 2.$$ 

Calculate the sample mean.

The sample mean is

$$\bar{x} = \frac{3+1+5+1+3+2+1+1+3+2}{10} = \frac{22}{10} = \frac{11}{5} = 2.2 \text{ children}.$$ 

This calculation is straightforward and can be conducted in R with the statements given below.

```r
x = c(3, 1, 5, 1, 3, 2, 1, 1, 3, 2)
mean(x)
```

Of course, polling ten different children would most likely result in a different sample mean. The experimental sample mean $\bar{x}$ given above is one instance from the sampling distribution of the random variable $\bar{X}$.

Another way of thinking about a sample mean is to consider it to be a special case of a **weighted average**, in which each of the data values is given a weight of $1/n$. If the data values constitute a random sample, then there is no reason to give more weight to one value over another. Returning to the kindergarten sibling data from the previous example, the sample mean could be written as

$$\bar{x} = \frac{3+1+5+1+3+2+1+1+3+2}{10} = \frac{1}{10} \cdot 4 + \frac{3}{10} + \frac{5}{10} + \frac{2}{10} + 3 \cdot \frac{1}{10} + 10 \cdot \frac{2}{10} + 3 \cdot \frac{1}{10} + 5 \cdot \frac{1}{10}.$$ 

This way of thinking emphasizes the fact that the sample mean is a weighted average, where the weights reflect the relative frequency of a particular data value. Compare the expression on the
far right with the formula for the population mean $E[X]$ for a discrete probability distribution from probability theory:

$$E[X] = \sum_{A} xf(x),$$

where $A$ is the support and $f(x)$ is the probability mass function. The weights 4/10, 2/10, 3/10, and 1/10 play the role of $f(x)$ from probability theory.

There is still another way to think about the sample mean. In order to develop this formulation, the notion of an empirical probability distribution must be defined.

**Definition 1.4** Let $x_1, x_2, \ldots, x_n$ be experimental values associated with the random variables $X_1, X_2, \ldots, X_n$. The *empirical probability distribution* associated with $x_1, x_2, \ldots, x_n$ is the discrete probability distribution defined by assigning probability $1/n$ to each $x_i$ value.

This empirical probability distribution can be expressed as either an *empirical probability mass function*, denoted by $\hat{f}(x)$, or an *empirical cumulative distribution function*, denoted by $\hat{F}(x)$, which are defined next. The empirical cumulative distribution function was introduced in the statistical graphics section as a way to avoid binning observations into cells when constructing a histogram.

**Definition 1.5** Let $x_1, x_2, \ldots, x_n$ be experimental values associated with the random variables $X_1, X_2, \ldots, X_n$. The *empirical probability mass function* associated with $x_1, x_2, \ldots, x_n$ is

$$\hat{f}(x) = \frac{\text{number of } x_i \text{ equal to } x}{n}. $$

The *empirical cumulative distribution function* associated with $x_1, x_2, \ldots, x_n$ is

$$\hat{F}(x) = \frac{\text{number of } x_i \text{ less than or equal to } x}{n}. $$

The empirical probability distribution, regardless of whether it is expressed in either of its equivalent forms as $\hat{f}(x)$ or $\hat{F}(x)$, is our best guess for the population probability distribution based on the data values $x_1, x_2, \ldots, x_n$.

Let’s return to the discussion of the sample mean. The empirical probability distribution associated with the data set has a population mean, which is typically called the “plug-in estimator of the population mean.” Using the formula for the population mean from probability, the formula for the plug-in estimator of the population mean is

$$\hat{\mu} = \sum_{A} x \hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} x_i,$$

where $A$ is the support of the population distribution. This is, once again, the formula for the sample mean.

So, regardless of whether you simply use the defining formula, think of the sample mean as a weighted average, or use the plug-in estimator of the population mean, the same value results for the sample mean.

Since the sample mean $\bar{X}$ is a random variable, we can calculate its sampling distribution. This sampling distribution depends on the population probability distribution from which the data values are drawn. The two examples that follow consider the sampling distribution of the sample mean for observations drawn from a discrete population and a continuous population.
Example 1.16 Let $X_1, X_2, \ldots, X_n$ be a random sample from a Poisson($\lambda$) distribution, where $\lambda$ is a positive unknown parameter. What is the sampling distribution of $\bar{X}$?

One could easily envision a real-world scenario in which averaging observations sampled from a Poisson population could occur, for example,

- averaging the number of customers that arrive to a drive-up window at a fast food restaurant during the lunch hour for five consecutive weekdays,
- averaging the number of potholes per mile on a particular stretch of highway, and
- averaging the number of web hits per day at a popular website during February.

The first step in finding the probability distribution of the sample mean $\bar{X}$ is to determine the support of the random variable $\bar{X}$. The Poisson population distribution has support on the nonnegative integers, so the numerator of $\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$ can also assume the values 0, 1, 2, $\ldots$. Therefore, the support of $\bar{X}$ is

$$A = \{ x \mid x = 0, 1, 2, n, \ldots \}.$$  

The next step is to determine the probabilities associated with each element in the support. The random variable $X_i$ has probability mass function

$$f_{X_i}(x) = \frac{\lambda^x e^{-\lambda}}{x!} \quad x = 0, 1, 2, \ldots$$

for $i = 1, 2, \ldots, n$. The numerator in the sample mean $\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$ consists of the sum of mutually independent and identically distributed Poisson($\lambda$) random variables because $X_1, X_2, \ldots, X_n$ is a random sample. Using a result from probability theory that can be proved by the moment generating function technique, the numerator $X_1 + X_2 + \cdots + X_n$ is Poisson($n\lambda$) with probability mass function

$$f_{X_1+X_2+\cdots+X_n}(x) = \frac{(n\lambda)^x e^{-n\lambda}}{x!} \quad x = 0, 1, 2, \ldots.$$  

Finally, dividing the numerator of $\bar{X}$ by $n$ gives the probability mass function

$$f_{\bar{X}}(x) = \frac{(n\lambda)^{nx} e^{-n\lambda}}{(nx)!} \quad x = 0, 1, 2, \ldots$$

by the transformation technique.

Now consider the sampling distribution of $\bar{X}$ for a particular sample size $n$ and a particular population mean $\lambda$, say $n = 5$ and $\lambda = 2$. Figure 1.22 is a graph of the probability mass function of the first seven support values of the population from which the data values are drawn, that is, a Poisson(2) distribution. The graph of $f_{\bar{X}}(x)$ continues to
Chapter 1. Random Sampling

Figure 1.22: Population probability mass function.

decline as \( x \) increases. Using the formula for \( f_{\bar{X}}(x) \), the probability mass function of \( \bar{X} \) is

\[
f_{\bar{X}}(x) = \frac{10^5 e^{-10}}{(5x)!} \quad x = 0, 1, 2, \ldots
\]

Figure 1.23 is a graph of the probability mass function of the sampling distribution of the statistic \( \bar{X} \) when \( n = 5 \) and \( \lambda = 2 \). The horizontal scales are identical, but the vertical scales differ on the two graphs. There are four observations that can be made concerning these two probability mass functions:

- Both the probability distribution of \( X_i \) and the probability distribution of \( \bar{X} \) have the same expected value: \( E[X_i] = E[\bar{X}] = 2 \) in this particular example. As will be seen subsequently, this result, stated more generally as \( E[\bar{X}] = \mu \), is true for any population distribution that has a finite population mean.
- The population variance of the sampling distribution of \( \bar{X} \) is less than the population variance of the population distribution. Although \( X_i \) and \( \bar{X} \) have the same

Figure 1.23: Probability mass function (sampling distribution) of the sample mean.
population mean, averaging the $n = 5$ observations decreases the population variance of $\bar{X}$ relative to $X_i$.

- The support of $\bar{X}$ is finer than the support of $X_i$. The data values can assume the values 0, 1, 2, ..., but the sample mean can assume the values 0, $1/5$, $2/5$, ...

- Even though the sample size is only $n = 5$, the central limit theorem is evident in the distribution of $\bar{X}$ as it has more of a bell shape than the population distribution. The limiting distribution of $\bar{X}$ in this example is normal.

The sampling distribution of $\bar{X}$ when the data values are drawn from a continuous population is determined in a similar fashion, as illustrated in the next example.

**Example 1.17** Let $X_1, X_2, \ldots, X_n$ be a random sample from a gamma($\lambda$, $\kappa$) distribution, where $\lambda$ and $\kappa$ are positive unknown scale and shape parameters. Find the sampling distribution of $\bar{X}$. Also find $P(\bar{X} < 2)$ for a sample size of $n = 4$ when $\lambda = 1$ and $\kappa = 3$.

The probability density function of $X_i$ sampled from a gamma($\lambda$, $\kappa$) population is

$$f_{X_i}(x) = \frac{\lambda^\kappa x^{\kappa-1} e^{-\lambda x}}{\Gamma(\kappa)} \quad x > 0$$

for $i = 1, 2, \ldots, n$. The corresponding moment generating function is

$$M_{X_i}(t) = \left(\frac{\lambda}{\lambda - t}\right)^\kappa \quad t < \lambda$$

for $i = 1, 2, \ldots, n$. Since the observations are a random sample, $X_1, X_2, \ldots, X_n$ are mutually independent and identically distributed random variables. Hence, the moment generating function of $\bar{X}$ is

$$M_{\bar{X}}(t) = E\left[e^{t\bar{X}}\right] = E\left[e^{t(X_1+X_2+\ldots+X_n)/n}\right] = M_{X_1+X_2+\ldots+X_n}(t/n) = M_{X_1}(t/n)M_{X_2}(t/n)\ldots M_{X_n}(t/n) = \left(\frac{\lambda}{\lambda - t/n}\right)^\kappa \left(\frac{\lambda}{\lambda - t/n}\right)^\kappa \ldots \left(\frac{\lambda}{\lambda - t/n}\right)^\kappa = \left(\frac{n\lambda}{n\lambda - t}\right)^{n\kappa} \quad t < n\lambda.$$

This moment generating function can be recognized as that of a gamma($n\lambda$, $n\kappa$) random variable.

Figure 1.24 is a plot of the population probability density function for $\lambda = 1$ and $\kappa = 3$. The $n = 4$ data values are sampled from this population probability distribution. Since $\bar{X} \sim$ gamma($n\lambda$, $n\kappa$), $\lambda = 1$, $\kappa = 3$, and $n = 4$,

$$\bar{X} \sim \text{gamma}(4, 12).$$

Figure 1.25 contains a plot of the probability density function of the sample mean $\bar{X} \sim \text{gamma}(4, 12)$. The horizontal scales in Figures 1.24 and 1.25 are identical, but the vertical scales differ on the two graphs. The same effect as in the previous example (when the random sampling was from a Poisson population) takes place here:
• The population probability distribution and the probability distribution of \( \bar{X} \) have the same central value, which in this case is \( E[X_i] = E[\bar{X}] = 3 \).
• The probability distribution of \( \bar{X} \) has a smaller population variance than the population probability distribution.
• The probability distribution of \( \bar{X} \) looks more bell-shaped than the population probability distribution because of the central limit theorem. The probability density function of \( \bar{X} \) is nearly symmetric. The limiting distribution of \( \bar{X} \) is normal.

The final part of the question is to determine the probability that the sample mean is less than 2 for sample size \( n = 4 \) and population parameters \( \lambda = 1 \) and \( \kappa = 3 \). One way to calculate this probability is to integrate the probability density function over the appropriate limits. Since \( \bar{X} \sim \text{gamma}(4, 12) \),

\[
P(\bar{X} < 2) = \int_{0}^{2} \frac{(n\lambda)^{\kappa}x^{n-1}e^{-n\lambda x}}{\Gamma(n\kappa)} dx = \int_{0}^{2} \frac{4^{12}1^{11}e^{-4x}}{\Gamma(12)} dx.
\]
Section 1.3. Estimating Central Tendency

This integral can be computed by hand using integration by parts repeatedly or can be calculated using a computer algebra system, giving the required probability as

\[
P(\bar{X} < 2) = 1 - \frac{41278941}{155925}e^{-8} \cong 0.1119.
\]

R can also be used to calculate the probability that the sample mean is less than 2. Using the `pgamma` function, which returns the cumulative distribution function of a gamma random variable, the single statement

`pgamma(2, 12, 4)`

also returns \( P(\bar{X} < 2) \cong 0.1119 \). Notice that R switches the order of the parameters as arguments relative to the convention \( \text{gamma}(\lambda, \kappa) \) used here.

Finally, to determine whether the derivation and associated numerical value are correct, a Monte Carlo simulation experiment can be conducted to estimate the probability that the sample mean is less than 2. The following R code generates one million sample means and prints the fraction of those sample means that are less than 2.

```r
nrep = 1000000
count = 0
for (i in 1:nrep) {
  xbar = mean(rgamma(4, 3, 1))
  if (xbar < 2) count = count + 1
}
print(count / nrep)
```

After a call to `set.seed(3)` to initialize the random number stream, five runs of this simulation yield the following estimates of \( P(\bar{X} < 2) \):

0.1118 0.1119 0.1120 0.1117 0.1125.

Since these values hover about the analytic value \( P(\bar{X} < 2) \cong 0.1119 \), the Monte Carlo simulation supports our analytic solution.

The two previous examples have shown that the probability distribution of the sample mean depends on the probability distribution associated with the population. Every population probability distribution that the data values are drawn from requires a separate derivation—some simple and others quite intricate—to determine the probability distribution of \( \bar{X} \). One piece of good news is that the expected value of \( \bar{X} \) and the population variance of \( \bar{X} \) can be computed with the same formulas for practically all probability distributions. Assuming that \( X_1, X_2, \ldots, X_n \) constitute a random sample from some population distribution (discrete or continuous) with finite population mean \( \mu \) and finite population variance \( \sigma^2 \), then the sample mean \( \bar{X} \) has expected value

\[
E[\bar{X}] = E\left[\frac{1}{n} \sum_{i=1}^{n} X_i\right] = \frac{1}{n} \sum_{i=1}^{n} E[X_i] = \frac{1}{n} (n\mu) = \mu
\]

and population variance

\[
V[\bar{X}] = V\left[\frac{1}{n} \sum_{i=1}^{n} X_i\right] = \frac{1}{n^2} \sum_{i=1}^{n} V[X_i] = \frac{1}{n^2} (n\sigma^2) = \frac{\sigma^2}{n}.
\]