Sample median

The sample mean is the "gold standard" in terms of estimating the central tendency of a population probability distribution and is used in a vast majority of applications in which central tendency is of interest. Occasions arise, however, when the *sample median* is a better measure of central tendency.

Definition 1.6 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 median* is

$$M = \begin{cases} X_{((n+1)/2)} & n \text{ odd} \\ (X_{(n/2)} + X_{(n/2+1)})/2 & n \text{ even} \end{cases}$$

where $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ are the order statistics (the data values sorted into ascending order).

If n is odd, the median is just the middle sorted value; if n is even, the median is the average of the two middle sorted values.

Economists frequently use the sample median, rather than the sample mean, when reporting statistics concerning certain economic measures, such as incomes or house prices. To see why this is the case, consider a small M.S. program in operations research that graduates just $n = 5$ students in one particular academic year. The students assume positions in industry and report the following annual salaries:

$$\$71{,}000 \qquad \$65{,}000 \qquad \$74{,}000 \qquad \$194{,}000 \qquad \$73{,}000.$$

Now which would be a more accurate way to report the salary data in a recruiting brochure for the new class of operations researchers: use the sample mean $\bar{x} = \$95{,}400$ or use the sample median $m = \$73{,}000$ as the measure of central tendency? The student who graduated and took a salary of \$194,000 might have joined a family business or had a lucrative overseas offering. The other four salaries are fairly tightly clustered around the sample median $m = \$73{,}000$. The one high salary is a rarity, so it can either be considered an outlier or it can be an observation from a very long right-hand tail of the population probability distribution. In either case, reporting the sample median is the appropriate statistic to go in the brochure for next year. It gives the students the most accurate assessment of what their salary will be when they finish the M.S. program.

Determining the sampling distribution of the sample median can vary from simple to very complex. The two examples that follow span the two extremes.

Example 1.20 Let $X_1, X_2, \ldots, X_9$ be a random sample from a $U(0, 1)$ distribution. Find the sampling distribution of the sample median.

Unlike the three examples associated with determining the sampling distribution of the sample mean, this time the population distribution does not have any unknown parameters. The probability density function of X_i drawn from a $U(0, 1)$ population is

$$f_{X_i}(x) = 1 \qquad 0 < x < 1,$$

for $i = 1, 2, \ldots, 9$. The corresponding cumulative distribution function on the support of X_i is

$$F_{X_i}(x) = x \qquad 0 < x < 1,$$

for $i = 1, 2, \ldots, 9$. The observations are mutually independent and identically distributed random variables because they constitute a random sample. So the distribution of the sample median M, which is $X_{(5)}$ because $n = 9$ is odd, can be found using the formula for the probability density function of the kth order statistic from a continuous population,

$$f_{X_{(k)}}(x) = \frac{n!}{(k-1)!(n-k)!}[F(x)]^{k-1} f(x)[1-F(x)]^{n-k},$$

for $a < x < b$ and $k = 1, 2, \ldots, n$, where a and b are the lower and upper limits of the support of the population probability distribution. Applying this formula to our sample of $n = 9$ observations from a $U(0, 1)$ population gives

$$f_M(x) = \frac{9!}{(5-1)!(9-5)!}x^{5-1} \cdot 1 \cdot (1-x)^{9-5} = 630x^4(1-x)^4 \qquad 0 < x < 1.$$

A Monte Carlo simulation experiment can be used to support our analytic work. The following R code generates 100 sample medians from 100 samples of size $n = 9$ drawn from a $U(0, 1)$ population distribution, plots a histogram, and overlays the histogram with the sampling distribution of the sample median derived above.

```
nrep = 100
medians = numeric(nrep)
for (i in 1:nrep) {
  x = runif(9)
  medians[i] = median(x)
}
hist(medians, probability = TRUE)
xx = seq(0, 1, by = 0.01)
yy = 630 * xx ^ 4 * (1 - xx) ^ 4
lines(xx, yy, type = "l")
```

Executing this code after a call to `set.seed(7)` yields the graph Figure 1.30. For both the analytic values represented by the curve and the sample values represented by the histogram, the effect of choosing the fifth largest of the nine values is to push the probability distribution away from the extremes at 0 and 1 toward the center of the distribution at 1/2. But are the histogram and the curve close enough to support our analytic work? The problem illustrated here is that we chose only `nrep = 100` replications

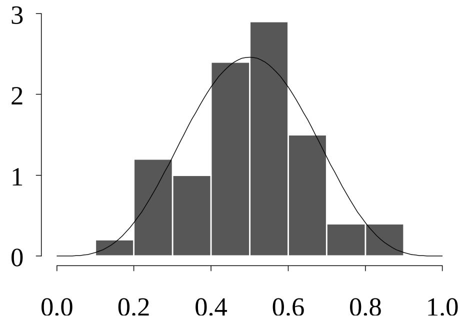

Figure 1.30: Sampling distribution of the sample median (100 replications).

of the experiment, resulting in a rather noisy histogram. Random sampling variability applies to Monte Carlo simulation as well as to collecting data. Figure 1.31 uses the same code, but this time with `nrep = 200000`. R chooses more cells for the histogram, which is much smoother this time. We now achieve a good match between the sampling distribution of M and its estimate via Monte Carlo simulation. This time our analytic work is supported by the simulation. The bell shape of the sampling distribution of M is not due to the central limit theorem, but rather due to the choice of the middle order statistic from a symmetric population probability distribution. Now that the sampling distribution of the sample median has been derived and supported by Monte Carlo simulation, it is often of value to know the expected value and variance of the statistic of interest. The APPL statements below calculate the probability density function of the sample median M and its expected value and its variance.

```
X := UniformRV(0, 1);
M := OrderStat(X, 9, 5);
Mean(M);
Variance(M);
```

The statements yield

$$E[M] = \frac{1}{2} \qquad \text{and} \qquad V[M] = \frac{1}{44}.$$

Notice that the expected value of the sample median equals the population median (this is $1/2$ by inspection because of the symmetry of the $U(0, 1)$ distribution). This is a good property for an estimator to have because the estimator is "on target" for estimating the population quantity. This property will be defined carefully in the next chapter, but for now $E[M] = E\left[X_{(5)}\right] = x_{0.5} = 1/2$ is stated in words as "the sample median is an unbiased estimator of the population median." The variance of M is an indication of how far the sample median might stray from its target. We would like the variance of M to be as small as possible. One way to decrease the variance of M is to increase the sample size n.

So the sample median seems like a reasonable estimator for the population median. But for this particular population distribution, the $U(0, 1)$ distribution, the population

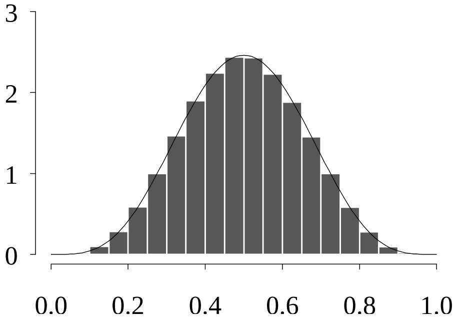

Figure 1.31: Sampling distribution of the sample median (200,000 replications).

median and the population mean are both equal to $1/2$. Would it be better to use the sample mean $\bar{X}$ to estimate the population median? One way to pin down the choice is to consider the variance of the estimates. From above, the variance of the sample median is

$$V[M] = \frac{1}{44},$$

but the variance of the sample mean $\bar{X}$ is

$$V\left[\bar{X}\right] = \sigma^2_{\bar{X}} = \frac{\sigma^2_X}{n} = \frac{1/12}{9} = \frac{1}{108}.$$

So the sample mean is more tightly clustered about $1/2$ than the sample median, and is therefore the preferred estimator of the population median.

The previous example had two factors which made the analytic work tractable: a particularly simple population distribution and an odd value for n. In the next example, we remove both of those advantages and see the extra work associated with an even n and a more complicated population distribution.

Example 1.21 Let $X_1, X_2, \ldots, X_6$ be a random sample from a population having probability density function

$$f(x) = 2x \qquad\qquad 0 < x < 1.$$

Find the sampling distribution of the sample median.

As in the previous example, the population distribution does not involve any parameters. In contrast to the previous example, there are two complicating factors at play in this question: the *even* sample size $n = 6$ and the *slightly* more complicated population distribution. The even sample size implies that two adjacent order statistics will be averaged in order to arrive at the sample median. As you will see, these two extra factors create lots of extra work in deriving the sampling distribution of the sample median. The problem is a good review, however, of the joint distribution of order statistics and the transformation technique. The random variable X_i has probability density function

$$f_{X_i}(x) = 2x \qquad\qquad 0 < x < 1,$$

for $i = 1, 2, \ldots, 6$. The associated cumulative distribution function of X_i on its support is

$$F_{X_i}(x) = \int_0^x 2w\,dw = \left[w^2\right]_0^x = x^2 \qquad\qquad 0 < x < 1,$$

for $i = 1, 2, \ldots, 6$. Since n is even, the sample median is calculated by averaging $X_{(3)}$ and $X_{(4)}$. Unfortunately, $X_{(3)}$ and $X_{(4)}$ are dependent random variables. So we begin the process of finding the probability density function of the sample median by finding the joint probability density function of $X_{(3)}$ and $X_{(4)}$. Using the same heuristic argument that gave us the probability density function of a single order statistic drawn from a continuous population, the joint probability density function of two order statistics $X_{(i)}$ and $X_{(j)}$, which is $f_{X_{(i)},X_{(j)}}\left(x_{(i)}, x_{(j)}\right)$ for $i < j$, is given by the expression

$$\frac{n!}{(i-1)!(j-i-1)!(n-j)!}\left[F\left(x_{(i)}\right)\right]^{i-1} f\left(x_{(i)}\right)\left[F\left(x_{(j)}\right)-F\left(x_{(i)}\right)\right]^{j-i-1} f\left(x_{(j)}\right)\left[1-F\left(x_{(j)}\right)\right]^{n-j}$$