

Estimators and Parameters

Populations are characterized by numerical measures called *parameters*. In many statistical applications, we want to use information from a sample to estimate one or more population parameters. An *estimator* is a rule that tells us how to calculate the value of an estimate based on measurements contained in a sample. We will use the symbol \hat{q} to indicate the point estimator of a population parameter q . Of course, we want to use “good” estimators. One characteristic of a good estimator is that the mean of its sampling distribution equals the parameter value. Estimators with this characteristic are said to be *unbiased*. That is, \hat{q} is an unbiased estimator of q if $E(\hat{q}) = q$. Otherwise \hat{q} is said to be *biased*, and the bias of \hat{q} is defined as $B = E(\hat{q}) - q$.

There are other desirable characteristics of estimators; for example, we desire that the sampling distribution of the estimator have small spread. We define the *mean square error* of \hat{q} as $MSE = E[(\hat{q} - q)^2]$. This is the average value over all possible samples. The mean square error of an estimator provides information about its spread and is a function of both the variance and the bias, as the following theorem states.

Theorem: $MSE(\hat{q}) = V(\hat{q}) + B^2$

Proof:

$$\begin{aligned} MSE &= E[(\hat{q} - q)^2] \\ &= E[(\hat{q} - E(\hat{q}) + E(\hat{q}) - q)^2] = E\left[\{(\hat{q} - E(\hat{q})) + (E(\hat{q}) - q)\}^2\right] \\ &= E\left[\{(\hat{q} - E(\hat{q})) + B\}^2\right] = E[(\hat{q} - E(\hat{q}))^2] + E[2B \cdot (\hat{q} - E(\hat{q}))] + E[B^2] \\ &= V(\hat{q}) + 2B \cdot E[\hat{q} - E(\hat{q})] + B^2 = V(\hat{q}) + 2B \cdot 0 + B^2 \\ &= V(\hat{q}) + B^2 \end{aligned}$$

MSE can help us decide between several possible estimators for the same parameter. We prefer estimators with small mean square error. Often, however, there is a tradeoff between bias and variance.

Probability or Inference?

The following examples involving keys on a key ring are provided to illustrate the difference between probability and inference.

Probability Example: Suppose we have a ring with three keys, and we are trying to find the key that opens a particular lock. We try one key, and it does not work. We move it aside and try another key which does not work. As expected, since there are only three keys, the third key opens the lock. What is the probability that this would happen? Let F represent a failure and S

represent a success. Then $p(FFS) = \frac{2}{3} \cdot \frac{1}{2} \cdot 1 = \frac{1}{3}$. In this example involving probability, we begin by knowing what the world is and then ask questions about probabilities.

Inference Example: Now suppose we do not know the number of keys, N . We are trying to use data to estimate N . We will again assume that our data are FFS , as described above. So we know that N must be greater than or equal to 3. One strategy for determining the value of N is to consider all possible values of N and choose the value for which the observed result is most likely.

N	$p(FFS N)$
1	0
2	0
3	$\frac{2}{3} \cdot \frac{1}{2} \cdot 1 = \frac{1}{3}$
4	$\frac{3}{4} \cdot \frac{2}{3} \cdot \frac{1}{2} = \frac{1}{4}$
5	$\frac{4}{5} \cdot \frac{3}{4} \cdot \frac{1}{3} = \frac{1}{5}$
6	$\frac{5}{6} \cdot \frac{4}{5} \cdot \frac{1}{4} = \frac{1}{6}$

It should be obvious that for larger values of N , the probabilities will become smaller. Therefore, the value of N for which the observed result is most likely is $N = 3$.

We will now consider two commonly used techniques for deriving point estimators: method of moments and maximum likelihood. The previous example illustrates the method of maximum likelihood.

Method of Moments Technique for Deriving Point Estimators

Recall that we previously defined the k th moment of a random variable as

$$k\text{th moment} = m'_k = E(Y^k).$$

The corresponding k th sample moment is defined as

$$k\text{th sample moment} = m'_k = \frac{1}{n} \sum_{i=1}^n Y_i^k$$

The method of moments was developed by Karl Pearson in 1894. It is based on the assumption that sample moments should be good estimators of the corresponding population moments. The

method involves setting $E(Y^k) = \frac{1}{n} \sum_{i=1}^n Y_i^k$ and solving for the parameter.

Example 1

Let Y_1, Y_2, \dots, Y_n be independent identically distributed uniform random variables over the continuous interval from 0 to q where q is unknown. Use the method of moments to estimate the parameter q .

Solution:

We know that for a uniform distribution on $(0, q)$, the first moment $m'_1 = m = E(Y) = \frac{q}{2}$. We

also know, by definition, that $m'_1 = \frac{1}{n} \sum_{i=1}^n Y_i = \bar{Y}$. To get the method of moments estimator, set

$m'_1 = m$, that is set $\frac{q}{2} = \bar{Y}$, and solve for the unknown parameter q . Solving yields $q = 2\bar{Y}$; so the method of moments estimator is $\hat{q} = 2\bar{Y}$.

Example 2

Suppose Y_1, Y_2, \dots, Y_n denote a random sample from the exponential distribution with parameter b . Find the method of moments estimator for the unknown parameter b .

Solution:

$$f(y) = \frac{1}{b} e^{-y/b} \qquad m'_1 = m = E(Y) = b \qquad m'_1 = \frac{1}{n} \sum_{i=1}^n Y_i = \bar{Y}$$

Set the theoretical value of the first moment b equal to the sample value \bar{Y} . Thus, using the method of moments, $\hat{b} = \bar{Y}$

Example 3

Suppose Y_1, Y_2, \dots, Y_n denote a random sample from a Poisson distribution with mean λ . Find the method of moments estimator for λ .

Solution:

$$m'_1 = m = E(Y) = \lambda \qquad m'_1 = \frac{1}{n} \sum_{i=1}^n Y_i = \bar{Y}$$

Set λ equal to the sample value \bar{Y} to find that $\hat{\lambda} = \bar{Y}$ using the method of moments.

Example 4

If Y_1, Y_2, \dots, Y_n denote a random sample from the normal distribution with known mean $m = 0$ and unknown variance s^2 , find the method of moments estimator for s^2 .

Solution:

Since $\mathbf{s}^2 = E(Y^2) - [E(Y)]^2$, we have $\mathbf{m}'_2 = E(Y^2) = \mathbf{s}^2 + [E(Y)]^2 = \mathbf{s}^2 + \mathbf{m}^2 = \mathbf{s}^2$

$$m'_2 = \frac{1}{n} \sum_{i=1}^n Y_i^2$$

Set \mathbf{s}^2 equal $\frac{1}{n} \sum_{i=1}^n Y_i^2$ to find that $\hat{\mathbf{S}}^2 = \frac{1}{n} \sum_{i=1}^n Y_i^2$ by the method of moments.

We can use the method of moments to investigate more than one parameter at the same time. This is illustrated in the example that follows.

Example 5

If Y_1, Y_2, \dots, Y_n denote a random sample from the normal distribution with unknown mean \mathbf{m} and unknown variance \mathbf{s}^2 , find the method of moments estimators for \mathbf{m} and \mathbf{s}^2 .

Solution:

This solution is similar to Example 4, but requires us to set up a system of two equations in two unknowns.

$$\mathbf{m}'_1 = E(Y) = \mathbf{m} \quad \mathbf{m}'_1 = \frac{1}{n} \sum_{i=1}^n Y_i = \bar{Y} \quad \text{so } \hat{\mathbf{m}} = \bar{Y}$$

$$\mathbf{m}'_2 = E(Y^2) = V(Y) + [E(Y)]^2 = \mathbf{s}^2 + \mathbf{m}^2 \quad \mathbf{m}'_2 = \frac{1}{n} \sum_{i=1}^n Y_i^2$$

Set $\hat{\mathbf{S}}^2 + \hat{\mathbf{m}}^2 = \frac{1}{n} \sum_{i=1}^n Y_i^2$ and solve for the parameter $\hat{\mathbf{S}}^2$:

$$\hat{\mathbf{S}}^2 = \frac{1}{n} \sum_{i=1}^n Y_i^2 - \hat{\mathbf{m}}^2 = \frac{1}{n} \sum_{i=1}^n Y_i^2 - \bar{Y}^2 \quad (\text{using } \bar{Y} \text{ to estimate } \mathbf{m} \text{ since } \hat{\mathbf{m}} = \bar{Y})$$

$$= \frac{1}{n} \sum_{i=1}^n Y_i^2 - \bar{Y}^2 + \bar{Y}^2 - \bar{Y}^2 = \frac{1}{n} \sum_{i=1}^n Y_i^2 - 2\bar{Y}^2 + \bar{Y}^2$$

$$= \frac{1}{n} \sum_{i=1}^n Y_i^2 - 2\bar{Y} \frac{\sum_{i=1}^n Y_i}{n} + \frac{1}{n} n \bar{Y}^2 = \frac{1}{n} \left(\sum_{i=1}^n Y_i^2 - 2\bar{Y} \sum_{i=1}^n Y_i + n\bar{Y}^2 \right)$$

$$= \frac{1}{n} \left(\sum_{i=1}^n Y_i^2 - 2\bar{Y} \sum_{i=1}^n Y_i + \sum_{i=1}^n \bar{Y}^2 \right) = \frac{1}{n} \sum_{i=1}^n (Y_i^2 - 2Y_i\bar{Y} + \bar{Y}^2)$$

$$= \frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^2$$

So the method of moments estimators are $\hat{\mathbf{m}} = \bar{Y}$ and $\hat{\mathbf{S}}^2 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n}$.

Estimators derived by the method of moments are used less frequently than maximum likelihood estimators, which will be discussed in the section that follows. They are still useful, as they are generally easy to calculate. Often, however, the method of moments estimators are biased. We

have already seen that $\frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n-1}$ rather than $\frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n}$ is an unbiased estimator of σ^2 .

Maximum Likelihood

We have previously developed the method of moments for estimating population parameters. Sample moments were equated with corresponding population moments and the equations solved for the population parameters; these solutions provided estimators for the population parameters. The method of moments is intuitively pleasing and easy to apply, but may not lead to the "best" estimators, i.e. estimators with minimum variance and no bias.

We will now present a different approach to estimating parameters, the method of *maximum likelihood*. The method of maximum likelihood attempts to determine the "most likely" value of a parameter by calculating the conditional probability of getting the already acquired data, given different values of the parameter. Likelihood assumes the data to be fixed and treats the parameter as a variable. The process involves the construction of a *likelihood function* with domain all potential values of the parameter. The maximum likelihood estimator of the parameter is the function of the data that maximizes the likelihood. This method may be used to find estimates of multiple parameters from the same data: for example the mean and variance.

Definition of the likelihood

Let y_1, y_2, \dots, y_n be sample observations taken on corresponding random variables, Y_1, Y_2, \dots, Y_n whose distribution depends on a parameter \mathbf{q} . Then, if Y_1, Y_2, \dots, Y_n are discrete random variables, the *likelihood of the sample*, $L(y_1, y_2, \dots, y_n | \mathbf{q})$, is defined to be the joint probability of y_1, y_2, \dots, y_n . (Despite the complicated notation, L is a function of \mathbf{q} .) If Y_1, Y_2, \dots, Y_n are continuous random variables, the *likelihood* $L(y_1, y_2, \dots, y_n | \mathbf{q})$ is defined to be the joint density evaluated at y_1, y_2, \dots, y_n .

To find the maximum likelihood, we find the probability of the observed data given a specific value of \mathbf{q} and choose the value of \mathbf{q} that gives the largest probability.

If the set of random variables Y_1, Y_2, \dots, Y_n denotes a random sample from a discrete distribution with probability function $p(y | \mathbf{q})$, then

$$\begin{aligned} L(y_1, y_2, \dots, y_n | \mathbf{q}) &= P(Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n | \mathbf{q}) \\ &= p(y_1 | \mathbf{q}) \cdot p(y_2 | \mathbf{q}) \cdots p(y_n | \mathbf{q}) \end{aligned}$$

If the set of random variables Y_1, Y_2, \dots, Y_n denotes a random sample from a continuous distribution with density function $f(y | \mathbf{q})$, then

$$\begin{aligned} L(y_1, y_2, \dots, y_n | \mathbf{q}) &= f(y_1, y_2, \dots, y_n | \mathbf{q}) \\ &= f(y_1 | \mathbf{q}) \cdot f(y_2 | \mathbf{q}) \cdots f(y_n | \mathbf{q}) \end{aligned}$$

Notational convenience: To simplify notation, we will sometimes suppress the y_i 's and denote the likelihood by L , or $L(\mathbf{q})$.

Example 1

Suppose we have a box with 3 balls, \mathbf{q} of them red and $3 - \mathbf{q}$ of them white. We will sample, $n = 2$, without replacement and record the number that are red, y ; our problem is to estimate the number of red balls in the box. Suppose that the 2 balls we select are both red, that is, we observe $y = 2$. Notice that since we have already acquired two red balls, the only possible values of \mathbf{q} are two and three.

If $\mathbf{q} = 2$,

$$L(y = 2 | \mathbf{q} = 2) = P(y = 2 | \mathbf{q} = 2) = \frac{\binom{2}{2} \binom{1}{0}}{\binom{3}{2}} = \frac{1}{3}$$

If $\mathbf{q} = 3$,

$$L(y = 2 | \mathbf{q} = 3) = P(y = 2 | \mathbf{q} = 3) = \frac{\binom{3}{2}}{\binom{3}{2}} = 1$$

In this example, the value of $\mathbf{q} = 3$ makes the data most likely; of all possible values of the parameter, we pick the value that maximizes the likelihood of getting the results: $\hat{\mathbf{q}}_{MLE} = 3$.

Example 2

We will now reconsider the 3-ball problem in Example 1 under the assumption of different data. Suppose again the box with 3 balls, \mathbf{q} of them red and $3 - \mathbf{q}$ of them white. Again we sample, $n = 2$ balls, without replacement. This time let $y = 1$, i.e. one red ball and one white ball are drawn; our problem is once again to estimate the number of red balls in the box. This time we have acquired only one red ball, and possible values of \mathbf{q} are one and two.

If $\mathbf{q} = 1$,

$$L(y = 1 | \mathbf{q} = 1) = P(y = 1 | \mathbf{q} = 1) = \frac{\binom{1}{1} \binom{2}{1}}{\binom{3}{2}} = \frac{2}{3}$$

If $\mathbf{q} = 2$,

$$L(y = 1 | \mathbf{q} = 2) = P(y = 1 | \mathbf{q} = 2) = \frac{\binom{2}{1} \binom{1}{1}}{\binom{3}{2}} = \frac{2}{3}$$

In this example, the data are equally likely for both possible values of \mathbf{q} . (Maximum likelihood estimators are not necessarily unique.)

Example 3

Let Y_1, Y_2, \dots, Y_n be a random sample of observations from a uniform distribution with a probability density function $f(y_i | \mathbf{q}) = \frac{1}{\mathbf{q}}$ for $0 \leq y_i \leq \mathbf{q}$ and $i = 1, 2, \dots, n$. We want to find the maximum likelihood estimator of \mathbf{q} .

In the continuous case the likelihood is given by:

$$\begin{aligned} L(y_1, y_2, \dots, y_n | \mathbf{q}) &= L(\mathbf{q}) \\ &= f(y_1 | \mathbf{q}) \cdot f(y_2 | \mathbf{q}) \cdots f(y_n | \mathbf{q}) \quad (\text{Because of independence.}) \\ &= \prod_{i=1}^n \frac{1}{\mathbf{q}} \\ &= \begin{cases} \left(\frac{1}{\mathbf{q}}\right)^n, & 0 \leq y_i \leq \mathbf{q}, \text{ for } i = 1, 2, \dots, n \\ 0, & \text{elsewhere} \end{cases} \end{aligned}$$

Since all sample values must be between 0 and \mathbf{q} , we know that \mathbf{q} must be greater than or equal to the largest sample value. Of all these possible values of \mathbf{q} , the maximum value of $L(\mathbf{q})$ occurs where \mathbf{q} is as small as possible. Thus, $\hat{\mathbf{q}}_{MLE} = \text{maximum}\{y_1, y_2, \dots, y_n\}$. Note that in this case, the maximum likelihood estimator will generally underestimate the true value of \mathbf{q} and therefore will be biased low.

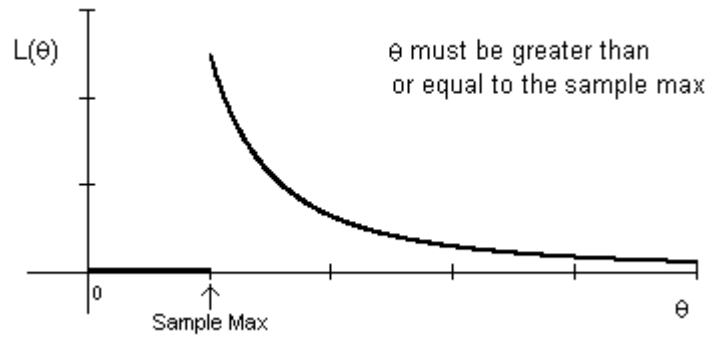


Figure 5: Maximum Likelihood Estimator of θ .

Example 4

We will now generalize the results of Example 3.

Often, we are interested in the relative magnitudes of observed random variables. Perhaps we want to know the maximum wind velocity or the largest rainfall. We will, therefore, order the observed random variables by their magnitudes. The resulting ordered variables are known as *order statistics*. There is a common notation for these order statistics. If Y_1, Y_2, \dots, Y_n denote continuous random variables, then the ordered random variables are denoted by $Y_{(1)}, Y_{(2)}, Y_{(3)}, \dots, Y_{(n)}$, where the subscripts denote the ordering. This means that $Y_{(1)} \leq Y_{(2)} \leq Y_{(3)} \leq \dots \leq Y_{(n)}$.

Let Y_1, Y_2, \dots, Y_n denote independent continuous random variables with distribution function $F(y)$ and density function $f(y)$. We can use the method of distribution functions to derive the density function of $Y_{(n)} = \max\{Y_1, Y_2, \dots, Y_n\}$.

$$\begin{aligned}
 Y_{(n)} &= \max\{Y_1, Y_2, \dots, Y_n\} \\
 \text{cdf}(y) = F_{(n)}(y) &= P(Y_{(n)} \leq y) \\
 &= P(Y_1 \leq y, Y_2 \leq y, \dots, Y_n \leq y) \\
 &= \prod_{i=1}^n P(Y_i \leq y) \text{ since } Y_i \text{'s are independent,} \\
 &= [F(y)]^n, \text{ since } Y_i \text{'s are identically distributed.}
 \end{aligned}$$

Letting $g_{(n)}(y)$ denote the density function of $Y_{(n)}$, we can find $g_{(n)}(y)$ by differentiation:

$$\begin{aligned}
 g_{(n)}(y) &= \frac{d}{dy} [F(y)]^n \\
 &= n[F(y)]^{n-1} \frac{dF}{dy} \\
 &= n[F(y)]^{n-1} f(y)
 \end{aligned}$$

To illustrate this general result, suppose that $Y_i \sim U[0,1]$. Then $f(y) = \begin{cases} 1, & \text{if } 0 \leq y \leq 1 \\ 0, & \text{elsewhere} \end{cases}$

Then the distribution function and density function for $Y_{(n)}$ are defined as follows:

$$\begin{aligned}
 F_{(n)}(y) &= \begin{cases} 0, & y < 0 \\ \int_0^y 1 \, dy, & 0 \leq y \leq 1 \\ 1, & y > 1 \end{cases} \\
 g_{(n)}(y) &= \begin{cases} n[y]^{n-1}, & \text{if } 0 \leq y \leq 1 \\ 0, & \text{elsewhere} \end{cases}
 \end{aligned}$$

We now find the expected value of this random variable, $Y_{(n)}$:

$$\begin{aligned}
 E(Y_{(n)}) &= \int_a^b y \cdot g_{(n)}(y) dy \\
 &= \int_0^1 y [ny^{n-1}] dy \\
 &= \int_0^1 ny^n dy \\
 &= n \left. \frac{y^{n+1}}{n+1} \right|_{y=0}^{y=1} \\
 &= \frac{n}{n+1}
 \end{aligned}$$

For example, if $n=1$, $E(Y_{(n)}) = E(Y_{(1)}) = \frac{1}{1+1} = \frac{1}{2}$. If $n=3$, $E(Y_{(n)}) = E(Y_{(3)}) = \frac{3}{3+1} = \frac{3}{4}$.

Similarly, it can be shown that for $Y_i \sim U[0, \mathbf{q}]$, $E(Y_{(n)}) = \frac{n}{n+1} \mathbf{q}$.

Example 5

It should be pointed out that though the theory was developed for a continuous uniform distribution, the results provide very good approximations for discrete uniform distributions.

Consider the "Petit Taxis" in Marrakech. These taxis are numbered sequentially $1, 2, \dots, \mathbf{q}$. Suppose we observe six taxis randomly and note that the largest taxi number in our sample is

$y_{(6)} = 435$. If the numbers are uniformly distributed over $[1, 2, \dots, \mathbf{q}]$, the maximum likelihood estimate for \mathbf{q} is $y_{(6)} = 435$. Since $E(Y_{(6)}) = \frac{6}{6+1}\mathbf{q}$, we set $\frac{6}{7}\hat{\mathbf{q}} = 435$, and $\hat{\mathbf{q}} = \frac{7}{6}(435) = 507.5$

Our "unbiased correction" to the maximum likelihood estimate of 435 estimates of the number of Petit Taxis at 508.

Example 6

A binomial experiment consisting of n trials results in observations, y_1, y_2, \dots, y_n , where $y_i = 1$ if the i th trial was a success, and $y_i = 0$ if the i th trial was otherwise. Then the likelihood function for these trials is defined as follows:

$$\begin{aligned} L(y_1, y_2, \dots, y_n | p) &= P(Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n | p) \\ &= P(Y_1 = y_1 | p) \cdot P(Y_2 = y_2 | p) \cdots P(Y_n = y_n | p) \\ &= p^{\sum y_i} (1-p)^{n-\sum y_i}. \end{aligned}$$

The order in which the particular y_i 's are drawn does not matter. The probability function of y successes, $y = \sum_{i=1}^n y_i$, is:

$$L(p) = \binom{n}{\sum_{i=1}^n y_i} p^{\sum y_i} (1-p)^{n-\sum y_i} = \binom{n}{y} p^y (1-p)^{n-y}.$$

Alternately, we could define the likelihood in terms of $Y =$ total number of successes. The results are the same.

Note that for purposes of finding the maximum likelihood estimator of p the factor $\binom{n}{y}$ is superfluous. For any given combination of y_i 's, the $\sum y_i$ will be equal to y and we could choose a different likelihood function that will reach its maximum at the same value of p . Remember that we are not finding the value of the likelihood functions, only the value of p at which the functions have the maximum value. We can search for the p that maximizes $L(p)$. Two special cases suggest themselves immediately. If y is zero, each of the y_i 's will be zero, and $L(p) = p^0(1-p)^n = (1-p)^n$. Therefore, $L(p)$ reaches a maximum at $p=0$. If p is one, each of the y_i 's will be one, $L(p) = p^n(1-p)^0 = p^n$, and $L(p)$ reaches a maximum at $p=1$. In each of these instances, there is only one possible distinct outcome of the experiment, so $\binom{n}{y} = 1$ in the likelihood function. For values of y strictly between 0 and n , we can find the maximum likelihood by differentiation:

$$\begin{aligned}
L(p) &= \binom{n}{y} p^y (1-p)^{n-y} \\
\frac{dL}{dp} &= \binom{n}{y} \left\{ p^y \frac{d}{dy} [(1-p)^{n-y}] + (1-p)^{n-y} \frac{d}{dp} p^y \right\} \\
&= \binom{n}{y} \left\{ p^y [(n-y)(1-p)^{n-y-1}(-1)] + [(1-p)^{n-y} y p^{y-1}] \right\} \\
&= \binom{n}{y} p^{y-1} (1-p)^{n-y-1} [-p(n-y) + (1-p)y] \\
&= \binom{n}{y} \left\{ p^{y-1} (1-p)^{n-y-1} [y - pn] \right\}
\end{aligned}$$

By inspection we see that the factors p^{y-1} and $(1-p)^{n-y-1}$ are greater than zero for all values of p . Thus the sign of the derivative will be determined by $[y - pn]$. The derivative is zero for $p = \frac{y}{n}$, positive for $p < \frac{y}{n}$, and negative for $p > \frac{y}{n}$. Thus, the maximum likelihood estimator for p is:

$$\hat{p}_{MLE} = \frac{y}{n}.$$

Example 7

While not a particular problem in Example 6, it sometimes happens that maximizing the likelihood function $L(y_1, y_2, \dots, y_n | \mathbf{q})$ requires tedious algebra due to differentiation of a plethora of products. Since the natural logarithm function is a monotonic strictly increasing function, both $L(y_1, y_2, \dots, y_n | \mathbf{q})$ and $\ln[L(y_1, y_2, \dots, y_n | \mathbf{q})]$ will have the same solution(s) for their relative extrema. For illustration, we will find \hat{p}_{MLE} for the function in Example 6 using logarithms. As before, $L(p) = \binom{n}{y} p^y (1-p)^{n-y}$, and...

$$\begin{aligned}
\ln[L(p)] &= \ln \binom{n}{y} + y \ln p + (n-y) \ln(1-p) \\
\frac{d}{dp} \{\ln[L(p)]\} &= 0 + y \frac{d}{dp} \ln p + (n-y) \frac{d}{dp} \ln(1-p) \\
&= y \left(\frac{1}{p} \right) + (n-y) \frac{1}{1-p} (-1)
\end{aligned}$$

To find the maximum likelihood estimator we maximize our likelihood function by setting the derivative equal to 0.

$$\begin{aligned}\frac{d}{dp}[\ln L(p)] &= 0 \\ \Rightarrow \frac{y}{p} - \frac{n-y}{1-p} &= 0 \\ \Rightarrow p &= \frac{y}{n}\end{aligned}$$

Checking the second derivative to determine whether we have a maximum or minimum value at $p = \frac{y}{n}$, we have

$$\frac{d}{dp}\left[\frac{y}{p} - \frac{n-y}{1-p}\right] = \frac{-y}{p^2} - \frac{(-1)(n-y)(-1)}{(1-p)^2} = -\left(\frac{y}{p} + \frac{n-y}{(1-p)^2}\right)$$

which is negative for all values of p and certainly for $p = \frac{y}{n}$. Thus, using this equivalent procedure, the maximum likelihood estimator for p is (once again):

$$\hat{p}_{MLE} = \frac{y}{n}.$$

Example 8

Let y_1, y_2, \dots, y_n be a random sample taken from a normal distribution with mean \mathbf{m} and variance \mathbf{s}^2 . We would like to find the maximum-likelihood estimators, $\hat{\mathbf{m}}_{MLE}$ and $\hat{\mathbf{s}}^2_{MLE}$ respectively. Since the random variables, Y_1, Y_2, \dots, Y_n , are continuous, L is the joint density function of the sample: $L(y_1, y_2, \dots, y_n | \mathbf{m}, \mathbf{s}^2) = L(\mathbf{m}, \mathbf{s}^2) = f(y_1, y_2, \dots, y_n | \mathbf{m}, \mathbf{s}^2)$. Since the sample is random and therefore independent,

$$\begin{aligned}L(\mathbf{m}, \mathbf{s}^2) &= f(y_1 | \mathbf{m}, \mathbf{s}^2) \cdot f(y_2 | \mathbf{m}, \mathbf{s}^2) \cdots f(y_n | \mathbf{m}, \mathbf{s}^2) \\ &= \left\{ \frac{1}{\mathbf{s}\sqrt{2\mathbf{p}}} \exp\left[-\frac{(y_1 - \mathbf{m})^2}{2\mathbf{s}^2}\right] \right\} \cdot \left\{ \frac{1}{\mathbf{s}\sqrt{2\mathbf{p}}} \exp\left[-\frac{(y_2 - \mathbf{m})^2}{2\mathbf{s}^2}\right] \right\} \cdots \left\{ \frac{1}{\mathbf{s}\sqrt{2\mathbf{p}}} \exp\left[-\frac{(y_n - \mathbf{m})^2}{2\mathbf{s}^2}\right] \right\} \\ &= \left(\frac{1}{\mathbf{s}\sqrt{2\mathbf{p}}}\right)^n \left\{ \exp\left[-\frac{(y_1 - \mathbf{m})^2}{2\mathbf{s}^2}\right] \right\} \cdot \left\{ \exp\left[-\frac{(y_2 - \mathbf{m})^2}{2\mathbf{s}^2}\right] \right\} \cdots \left\{ \exp\left[-\frac{(y_n - \mathbf{m})^2}{2\mathbf{s}^2}\right] \right\} \\ &= \left(\frac{1}{\mathbf{s}\sqrt{2\mathbf{p}}}\right)^n \exp\left[\sum_{i=1}^n \frac{-1}{2\mathbf{s}^2} (y_i - \mathbf{m})^2\right] \\ &= \left(\frac{1}{\mathbf{s}\sqrt{2\mathbf{p}}}\right)^n \exp\left[\frac{-1}{2\mathbf{s}^2} \sum_{i=1}^n (y_i - \mathbf{m})^2\right]\end{aligned}$$

Substituting $\sum_{i=1}^n (y_i - \mathbf{m})^2 = \sum_{i=1}^n (y_i - \bar{y})^2 + n(\bar{y} - \mathbf{m})^2$ (see page 23), we have

$$L(\mathbf{m}, \mathbf{s}^2) = \left(\frac{1}{\mathbf{s}\sqrt{2\mathbf{p}}} \right)^n \exp \left\{ -\frac{1}{2\mathbf{s}^2} \left[\sum_{i=1}^n (y_i - \bar{y})^2 + n(\bar{y} - \mathbf{m})^2 \right] \right\}$$

$$\ln L(\mathbf{m}, \mathbf{s}^2) = n \ln \left(\frac{1}{\mathbf{s}\sqrt{2\mathbf{p}}} \right) - \frac{1}{2\mathbf{s}^2} \left[\sum_{i=1}^n (y_i - \bar{y})^2 + n(\bar{y} - \mathbf{m})^2 \right]$$

We note here that the only occurrence of \mathbf{m} is in the term with a negative coefficient. Thus, for all values of \mathbf{s}^2 , setting $\mathbf{m} = \bar{y}$ will maximize the value of $\ln(L)$. That is,

$$\ln L(\bar{y}, \mathbf{s}^2) \geq \ln L(\mathbf{m}, \mathbf{s}^2) \quad \text{for all values of } \mathbf{s}^2.$$

More formally, we could take the partial derivatives of $\ln L(\mathbf{m}, \mathbf{s}^2)$ with respect to \mathbf{m} and \mathbf{s}^2 .

$$\begin{aligned} \frac{\partial}{\partial \mathbf{m}} \ln [L(\mathbf{m}, \mathbf{s})] &= \frac{\partial}{\partial \mathbf{m}} \left[n \ln \left(\frac{1}{\mathbf{s}\sqrt{2\mathbf{p}}} \right) - \frac{1}{2\mathbf{s}^2} \left[\sum_{i=1}^n (y_i - \bar{y})^2 + n(\bar{y} - \mathbf{m})^2 \right] \right] \\ &= -\frac{1}{2\mathbf{s}^2} 2n(\bar{y} - \mathbf{m})^1 (-1) = \frac{n}{\mathbf{s}^2} (\bar{y} - \mathbf{m}) \end{aligned}$$

Setting this partial derivative equal to zero yields $\hat{\mathbf{m}} = \bar{y}$.

Thus, $\hat{\mathbf{m}}_{MLE} = \bar{y}$ is the maximum likelihood estimator of the mean.

$$\begin{aligned} \frac{\partial \ln [L(\mathbf{m}, \mathbf{s}^2)]}{\partial \mathbf{s}^2} &= \frac{\partial}{\partial \mathbf{s}^2} \left\{ \left[-\frac{n}{2} \ln(2\mathbf{p}\mathbf{s}^2) \right] - \left[\frac{1}{2\mathbf{s}^2} \sum_{i=1}^n (y_i - \mathbf{m})^2 \right] \right\} \\ &= -\frac{n}{2} \frac{1}{\mathbf{s}^2} - \frac{1}{2} \cdot \sum_{i=1}^n (y_i - \mathbf{m})^2 \left[\frac{\partial}{\partial \mathbf{s}^2} \left(\frac{1}{\mathbf{s}^2} \right) \right] \\ &= -\frac{n}{2} \frac{1}{\mathbf{s}^2} - \frac{1}{2} \cdot \sum_{i=1}^n (y_i - \mathbf{m})^2 \left[-\frac{1}{(\mathbf{s}^2)^2} \right] \\ &= -\frac{n}{2} \frac{1}{\mathbf{s}^2} + \frac{1}{2\mathbf{s}^4} \sum_{i=1}^n (y_i - \mathbf{m})^2 \end{aligned}$$

Setting the derivative equal to 0 and solving, we have:

$$-\frac{n}{2\hat{\mathbf{S}}^2} + \frac{1}{2\hat{\mathbf{S}}^4} \sum_{i=1}^n (y_i - \hat{\mathbf{m}})^2 = 0$$

$$\frac{1}{2\hat{\mathbf{S}}^4} \sum_{i=1}^n (y_i - \hat{\mathbf{m}})^2 = \frac{n}{2\hat{\mathbf{S}}^2}$$

$$\hat{\mathbf{S}}^2 = \frac{\sum_{i=1}^n (y_i - \hat{\mathbf{m}})^2}{n}$$

Thus, $\hat{\mathbf{S}}^2_{MLE} = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n}$ is the maximum likelihood estimator for the variance. Notice that the maximum likelihood estimator for the variance is a biased estimator.

Sufficient Statistics

In general we would like to use a statistic that “summarizes” or “reduces” the data in a sample without losing any information about parameters of interest. Such statistics are said to have the property of *sufficiency*.

Definition: Let Y_1, Y_2, \dots, Y_n denote a random sample from a probability distribution with unknown parameter \mathbf{q} . Consider some statistic $U = g(Y_1, Y_2, \dots, Y_n)$ which is a function of the data. U is *sufficient* for \mathbf{q} if the conditional distribution of Y_1, Y_2, \dots, Y_n given U does not depend on \mathbf{q} .

Numerical Example

A certain game, which costs \$50 to play, consists of drawing a card from a non-standard deck of cards. If a red card is drawn, the player wins \$100. The parameter of interest is $\mathbf{q} = P(\text{red})$. There are two decks of cards available for this game. Dan holds these decks and each day chooses which deck to use. If he is in a good mood he uses the deck with $\mathbf{q} = 3/4$. When he is in a bad mood he uses the deck with $\mathbf{q} = 1/3$.

On a certain day Jon plays twice, not knowing which deck is being used, and gets y_1, y_2 where $y_i = 1$ if the card is red and $y_i = 0$ if the card is black. Jeff is interested in playing that day, and he would like to know how Jon’s draws turned out so that he can get some information about his chance of winning. (He obviously prefers to play when Dan uses the deck with $\mathbf{q} = 3/4$.) Jon reports that he got 1 red card. That is, he reports that the value of $u = y_1 + y_2$ is one but does not report the individual values y_1 and y_2 . We want to determine whether Jeff has lost information about \mathbf{q} by virtue of Jon reporting the sum $u = y_1 + y_2$ rather than the individual outcomes.

First we will look at the sample space for Y_1, Y_2 . This is displayed in the table below:

Sample Y_1, Y_2	$U = Y_1 + Y_2$	Probability if $\theta = 3/4$	Probability if $\theta = 1/3$
0,0	0	$1/4 \cdot 1/4 = 1/16$	$2/3 \cdot 2/3 = 4/9$
0,1	1	$1/4 \cdot 3/4 = 3/16$	$2/3 \cdot 1/3 = 2/9$
1,0	1	$3/4 \cdot 1/4 = 3/16$	$1/3 \cdot 2/3 = 2/9$
1,1	2	$3/4 \cdot 3/4 = 9/16$	$1/3 \cdot 1/3 = 1/9$

As we see in the table above, if $U = 0$ then $q = \frac{1}{3}$ is more likely than $q = \frac{3}{4}$. If $U = 2$, then $q = \frac{3}{4}$ is more likely than $q = \frac{1}{3}$.

This sample space will be useful in looking at the conditional distribution of Y_1, Y_2 given U for each value of θ . If the conditional distribution is the same for each value of θ we can conclude that U is sufficient for θ . This conditional distribution is provided in the table below:

Sample Y_1, Y_2	$P(Y_1, Y_2 U = 0)$		$P(Y_1, Y_2 U = 1)$		$P(Y_1, Y_2 U = 2)$	
	$\theta = 3/4$	$\theta = 1/3$	$\theta = 3/4$	$\theta = 1/3$	$\theta = 3/4$	$\theta = 1/3$
0,0	1	1	0	0	0	0
0,1	0	0	$1/2^*$	$1/2^*$	0	0
1,0	0	0	$1/2^*$	$1/2^*$	0	0
1,1	0	0	0	0	1	1

*Note from the previous table that 0,1 and 1,0 are equally likely outcomes when $q = 3/4$ and $U = 1$. These are also equally likely when $q = 1/3$.

Since the conditional distribution for Y_1, Y_2 given U is the same when $q = 3/4$ and $q = 1/3$, we can conclude that the conditional distribution does not depend on q . Thus, U is sufficient for q . All the relevant information about q is contained in reporting the value of U . That is, U provides all the information about q that the sample provides, and knowing the chronology of the values of Y_1, Y_2 is not relevant in determining $P(\text{red})$. The sample proportion (in this case $\hat{p} = \frac{1}{2}$) is also a sufficient statistic for p .

Instead of the sum, suppose Jon reports the value of $d = y_1 - y_2$, but he does not report the individual values of y_1 and y_2 . We want to determine in this situation whether Jeff loses information about q by virtue of Jon's reporting the difference rather than the individual values. To answer this question, we will again look at the sample space for Y_1, Y_2 . This table is displayed below:

Sample Y_1, Y_2	$D = Y_1 - Y_2$	Probability if $\theta = 3/4$	Probability if $\theta = 1/3$
0,0	0	$1/4 \cdot 1/4 = 1/16$	$2/3 \cdot 2/3 = 4/9$
0,1	-1	$1/4 \cdot 3/4 = 3/16$	$2/3 \cdot 1/3 = 2/9$
1,0	1	$3/4 \cdot 1/4 = 3/16$	$1/3 \cdot 2/3 = 2/9$
1,1	0	$3/4 \cdot 3/4 = 9/16$	$1/3 \cdot 1/3 = 1/9$

Now look at the conditional distribution of Y_1, Y_2 given D for each value of q . This conditional distribution is provided in the table below:

Sample Y_1, Y_2	$P(Y_1, Y_2 D = -1)$		$P(Y_1, Y_2 D = 0)$		$P(Y_1, Y_2 D = 1)$	
	$\theta = 3/4$	$\theta = 1/3$	$\theta = 3/4$	$\theta = 1/3$	$\theta = 3/4$	$\theta = 1/3$
0,0	0	0	1/10	4/5	0	0
0,1	1	1	0	0	0	0
1,0	0	0	0	0	1	1
1,1	0	0	9/10	1/5	0	0

Notice from the column for $P(Y_1, Y_2 | D = 0)$, it is clear that the conditional distribution is not the same for each value of q ; that is, when $D = 0$, the conditional distribution *does* depend on the value of q . Thus we conclude that $D = Y_1 - Y_2$ is *not* sufficient for q . If $D = 0$, we would like to know if the data were (0, 0), which makes $q = \frac{1}{3}$ look likely, or if they were (1, 1), which makes $q = \frac{3}{4}$ look likely.

Theoretical Example

Let Y_1, Y_2, \dots, Y_n denote independent and identically distributed Bernoulli variables such that $P(Y_i = 1) = q$ and $p(Y_i = 0) = 1 - q$. We want to show that $U = \sum Y_i$ is sufficient for q .

$$\begin{aligned}
 P(Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n | U = u) &= \frac{P(Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n \text{ and } U = u)}{P(U = u)} \\
 &= \frac{P(Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n) \cdot P(U = u | Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n)}{P(U = u)}
 \end{aligned}$$

Note that $P(Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n) = q^{y_1}(1-q)^{1-y_1} \cdot q^{y_2}(1-q)^{1-y_2} \dots q^{y_n}(1-q)^{1-y_n}$

$$\begin{aligned}
 &= q^{y_1+y_2+\dots+y_n}(1-q)^{n-(y_1+y_2+\dots+y_n)} \\
 &= q^u(1-q)^{n-u}
 \end{aligned}$$

Note also, $P(U = u | Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n) = 1$, if $m = \sum y_i$, and $P(U = u) = \binom{n}{u} \mathbf{q}^u (1-\mathbf{q})^{n-u}$ for $u = 0, 1, 2, \dots, n$. So,

$$P(Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n | U = u) = \frac{\mathbf{q}^u (1-\mathbf{q})^{n-u} \cdot 1}{\binom{n}{u} \mathbf{q}^u (1-\mathbf{q})^{n-u}} = \begin{cases} \frac{1}{\binom{n}{m}} & \text{if } m = \sum y_i \\ 0 & \text{otherwise} \end{cases}.$$

This probability does not depend on \mathbf{q} , so we conclude that $U = \sum Y_i$ is sufficient for \mathbf{q} .

Note: If we tried to show that $U = y_1$ is sufficient in this setting we would fail!

The following theorem provides a useful way to show sufficiency.

Factorization Theorem: Let U be a statistic based on the random sample Y_1, Y_2, \dots, Y_n . Then U is a sufficient statistic for the estimation of a parameter \mathbf{q} if and only if $L(y_1, y_2, \dots, y_n | \mathbf{q}) = g(u, \mathbf{q}) \cdot h(y_1, y_2, \dots, y_n)$ where g and h are non-negative functions, $g(u, \mathbf{q})$ is a function only of u and \mathbf{q} and $h(y_1, y_2, \dots, y_n)$ is not a function of \mathbf{q} .

We will not prove this theorem.

Example 1

Let Y_1, Y_2, \dots, Y_n denote independent and identically distributed Bernoulli random variables such that $P(Y_i = 1) = \mathbf{q}$ and $P(Y_i = 0) = 1 - \mathbf{q}$. Show that $U = \sum_{i=1}^n Y_i$ is sufficient for \mathbf{q} .

Solution:

$$\begin{aligned} L(y_1, y_2, \dots, y_n | \mathbf{q}) &= \mathbf{q}^{y_1} (1-\mathbf{q})^{1-y_1} \cdot \mathbf{q}^{y_2} (1-\mathbf{q})^{1-y_2} \cdot \dots \cdot \mathbf{q}^{y_n} (1-\mathbf{q})^{1-y_n} \\ &= \mathbf{q}^{\sum y_i} (1-\mathbf{q})^{n-\sum y_i} \\ &= \mathbf{q}^u (1-\mathbf{q})^{n-u} \cdot 1 \\ &= g(u, \mathbf{q}) \cdot h(y_1, y_2, \dots, y_n) \end{aligned}$$

$$\text{where } g(u, \mathbf{q}) = \mathbf{q}^u (1-\mathbf{q})^{n-u} \text{ and } h(y_1, y_2, \dots, y_n) = 1$$

$\therefore U = \sum_{i=1}^n Y_i$ is sufficient for \mathbf{q} by the factorization theorem.

Example 2

Let Y_1, Y_2, \dots, Y_n be a random sample in which Y_i possesses the probability density function

$$f(y_i | \mathbf{a}) = \frac{1}{\Gamma(\mathbf{a})} y_i^{\mathbf{a}-1} e^{-y_i} \text{ for } y_i > 0. \text{ Show that } U = \prod_{i=1}^n Y_i \text{ is a sufficient statistic for the}$$

estimation of \mathbf{a} .

Solution:

$$\begin{aligned}
 L(y_1, y_2, \dots, y_n | \mathbf{a}) &= \frac{1}{\Gamma(\mathbf{a})} y_1^{\mathbf{a}-1} e^{-y_1} \cdot \frac{1}{\Gamma(\mathbf{a})} y_2^{\mathbf{a}-1} e^{-y_2} \cdot \frac{1}{\Gamma(\mathbf{a})} y_3^{\mathbf{a}-1} e^{-y_3} \cdot \mathbf{L} \cdot \frac{1}{\Gamma(\mathbf{a})} y_n^{\mathbf{a}-1} e^{-y_n} \\
 &= \left(\frac{1}{\Gamma(\mathbf{a})} \right)^n (y_1 \cdot y_2 \cdot y_3 \cdot \mathbf{L} \cdot y_n)^{\mathbf{a}-1} e^{-\sum y_i} \\
 &= \left(\frac{1}{\Gamma(\mathbf{a})} \right)^n (u)^{\mathbf{a}-1} e^{-\sum y_i} \\
 &= g(u, \mathbf{a}) \cdot h(y_1, y_2, \mathbf{K} y_n), \\
 \text{where } g(u, \mathbf{a}) &= \left(\frac{1}{\Gamma(\mathbf{a})} \right)^n (u)^{\mathbf{a}-1} \text{ and } h(y_1, y_2, \mathbf{K} y_n) = e^{-\sum y_i}.
 \end{aligned}$$

Therefore, $U = \prod_{i=1}^n Y_i$ is sufficient for \mathbf{a} by the Factorization Theorem.