MAXIMUM LIKELIHOOD FUNCTIONS

The maximum likelihood function is a method of estimating the most likely value of a parameter for a probability distribution given a sample of outcomes from that distribution. This handout will discuss in broad outlines the general method for constructing a maximum likelihood function and calculating the maximum likelihood estimate (MLE) from that function using calculus. Then we will go through a couple of specific worked examples.

In general terms, we consider the probability distribution $f(x,\lambda)$ and collect some samples of data that obey the distribution function. For each outcome, we measure the value of x, with the parameter λ still unknown. The maximum likelihood function is the product of these outcomes, i.e. $L(f) = \prod_{i=1}^{n} f(x_i,\lambda) = \prod_{i=1}^{n} f_i(\lambda)$. We will use this function to estimate the most likely value of the parameter λ. But, let's first construct the maximum likelihood function in a couple of specific examples.

Example 1. Construct the maximum likelihood function for the exponential distribution modeling the time between events in a Poisson process. We take several observations and obtain the following wait-times: $x_i = \{5,2,1,4,2,6,3,1,4,2\}$.

For the first observation, we obtained $x_1 = 5$. We substitution this into the exponential distribution $f(x,\lambda) = \lambda e^{-\lambda x}$ for x, obtaining $f_1(\lambda) = \lambda e^{-5\lambda}$. The second observation was $x_2 = 2$. So we substitution that into the exponential distribution for x, obtaining $f_2 = \lambda e^{-2\lambda}$. And so forth.

$$f_3(\lambda) = \lambda e^{-\lambda}, f_4(\lambda) = \lambda e^{-4\lambda}, f_5(\lambda) = \lambda e^{-2\lambda}, f_6(\lambda) = \lambda e^{-6\lambda}$$
$$f_7(\lambda) = \lambda e^{-3\lambda}, f_8(\lambda) = \lambda e^{-\lambda}, f_9(\lambda) = \lambda e^{-4\lambda}, f_{10}(\lambda) = \lambda e^{-2\lambda}$$

The maximum likelihood function is the product of these expressions: $L(f) = \prod_{i=1}^{10} f_i(\lambda) =$

$$L(f) = \lambda e^{-5\lambda}\,\lambda e^{-2\lambda}\,\lambda e^{-\lambda}\,\lambda e^{-4\lambda}\,\lambda e^{-2\lambda}\,\lambda e^{-6\lambda}\,\lambda e^{-3\lambda}\lambda e^{-\lambda}\lambda e^{-4\lambda}\,\lambda e^{-2\lambda}$$
$$L(f) = \lambda^{10} e^{-30\lambda}$$

Because this probability distribution contains exponentials, we convert a product to a sum in the exponent. In this case, the exponential distribution maximum likelihood function becomes

$$L(f) = \prod_{i=1}^{n} \lambda e^{-x_i\lambda} = \lambda^n e^{\lambda \sum_{i=1}^{n} x_i}$$

To build our maximum likelihood functions, it may be useful to review common probability distributions here.

Common Distributions

Discrete

$$binomial: b(x,n,p) = \binom{n}{x} p^x (1-p)^{n-x}, x = 0,1,2,\ldots$$

$$hypergeometric: h(x,n,M,N) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

$$negative\ binomial: nb(x,r,p) = \binom{x+r-1}{r-1} p^r (1-p)^x, x = 0,1,2,\ldots$$

$$Poisson: p(x,\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}, x = 0,1,2,\ldots$$

Continuous

$$uniform: f(x,A,B) = \frac{1}{B-A}, \qquad A \le x \le B$$

$$normal: f(x,\mu,\sigma) = \frac{1}{\sqrt{2\eth}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$exponential: f(x,\lambda) = \lambda e^{-\lambda x}, \qquad x \ge 0$$

$$gamma: f(x,\alpha,\beta) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\beta}}$$

$$Weibull: f(x,\alpha,\beta) = \frac{\alpha}{\beta^\alpha} x^{\alpha-1} e^{-\left(\frac{x}{\beta}\right)^\alpha}$$

$$lognormal: f(x,\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma x} e^{-\frac{(lnx-\mu)^2}{2\sigma^2}}, \qquad x \ge 0$$

$$beta: f(x,\alpha,\beta,A,B) = \frac{1}{B-A} \frac{\left(\Gamma(\alpha+\beta)\right)}{\Gamma(\alpha)\Gamma(\beta)} \left(\frac{x-A}{B-A}\right)^{\alpha-1} \left(\frac{B-x}{B-A}\right)^{\beta-1}, \qquad A \le x \le B$$

$$Student-T: f(x,\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

Some things to note about these distributions:

- If discrete distributions have continuous parameters, it is often possible to treat the maximum likelihood functions derived from these as though they were continuous functions. There are some notable exceptions to this, like the hypergeometric function.
- We can also neglect any constant multipliers in these functions, since it won't affect the outcome of where the best estimate is. So I will often state the $L(f)$ without those constants.
- It is possible to test for multiple parameters simultaneously.
- For the gamma function, if no β is specified, assume it is equal to 1. The χ^2 distribution is based on the gamma distribution with $\beta = 2$ and $\alpha = \frac{\nu}{2}$.

Example 2. Let's look at the binomial and negative binomial distributions. They behave pretty much the same once we have collected the samples.

Suppose that we have a Bernoulli random variable and we choose in advance to take 30 samples. Perhaps it is a weighted coin that we wish to test, so we flip it 30 times and record the outcome. After collecting the data, we find we have obtained the following sequence of heads and tails: HTTHHTTTHTHTTTTHTHTHHTTTTTHTTT. This is 10 heads and 20 tails.

But suppose instead that we took samples until we reached 20 tails and obtained the exact same sequence of heads and tails. If we are seek to calculate the probability for tails, then call this p, and the probability for heads, $1 - p$, then each time we obtain a head multiply by $(1 - p)$, and each time we get a tail, multiply by p. Since each data point collected was done one at a time, the coefficient in front of each product is $\binom{1}{1}$ or $\binom{1}{0}$ both of which are equal to 1. Thus, our maximum likelihood function is both cases is

$$L(f) = (1 - p)p^2(1 - p)^2p^3(1 - p)p(1 - p)p^4(1 - p)p(1 - p)p(1 - p)^2p^5(1 - p)p^3$$
$$L(f) = p^{20}(1 - p)^{10}$$

Since the functions themselves are just a product of such outcomes, the most they could differ by is the coefficient in front ($\binom{30}{20}p^{20}(1 - p)^{10}$ for the binomial or $\binom{29}{19}p^{20}(1 - p)^{10}$ for the negative binomial) which will not affect our future calculation for p. The difference here is only that the negative binomial demands that the last term is determined, since we stop counting at a particular success, so only the terms before that are actually free to be shuffled.

Example 3. Suppose that you have a collection of 1000 samples that are classified as Type A and Type B and want to use a small sample to estimate the number of Type A items in the entire collection. After collecting 20 samples, you obtain 8 samples of Type A. In the distribution, we know the values of n, N, x. So we substitute into the hypergeometric distribution. Because this distribution is derived from multiple products, we don't need to create the Π function here as we do for continuous distributions.

$$L(f) = \frac{\binom{M}{8}\binom{1000 - M}{20 - 8}}{\binom{1000}{20}} = \frac{\binom{M}{8}\binom{1000 - M}{12}}{\binom{1000}{20}}$$

Or, we can neglect the rather large constant in the denominator and use $L(f) = \binom{M}{8}\binom{1000 - M}{12}$.

Example 4. Let's consider a case of a normal distribution. Suppose that we wish to know the heights of male students in a particular department on campus. So 15 students are selected and their heights are measured. The value for x_i are obtained to be $\{69, 72, 75, 65, 66, 68, 70, 71, 73, 66, 68, 71, 69, 69, 63\}$ measured in inches. As we did in Example 1, each measurement is taken from a normal distribution, and we use this value for x in the function. $x_1 = 69$ gives us $= f_1(\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(69-\mu)^2}{2\sigma^2}}$. The second value $x_2 = 72$ gives us $f_2(\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(72-\mu)^2}{2\sigma^2}}$, and so forth. Thus the maximum likelihood function is

$$L(f) = \prod_{i=1}^{15} f_i(\mu, \sigma) =$$

$$\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(69-\mu)^2}{2\sigma^2}}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(72-\mu)^2}{2\sigma^2}}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(75-\mu)^2}{2\sigma^2}}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(65-\mu)^2}{2\sigma^2}}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(66-\mu)^2}{2\sigma^2}}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(68-\mu)^2}{2\sigma^2}}$$
$$\cdot\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(70-\mu)^2}{2\sigma^2}}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(71-\mu)^2}{2\sigma^2}}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(73-\mu)^2}{2\sigma^2}}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(66-\mu)^2}{2\sigma^2}}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(68-\mu)^2}{2\sigma^2}}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(71-\mu)^2}{2\sigma^2}}$$
$$\cdot\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(69-\mu)^2}{2\sigma^2}}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(69-\mu)^2}{2\sigma^2}}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(63-\mu)^2}{2\sigma^2}} =$$

$$L(f) = \frac{1}{(2\pi)^{\frac{15}{2}}\sigma^{15}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{15}(x_i-\mu)^2} =$$

$$\frac{1}{(2\pi)^{15/2}\sigma^{15}} e^{-\frac{1}{2\sigma^2}\left[3(69-\mu)^2+(72-\mu)^2+(75-\mu)^2+(65-\mu)^2+2(66-\mu)^2+2(68-\mu)^2+(70-\mu)^2+2(71-\mu)^2+(73-\mu)^2+(63-\mu)^2\right]}$$

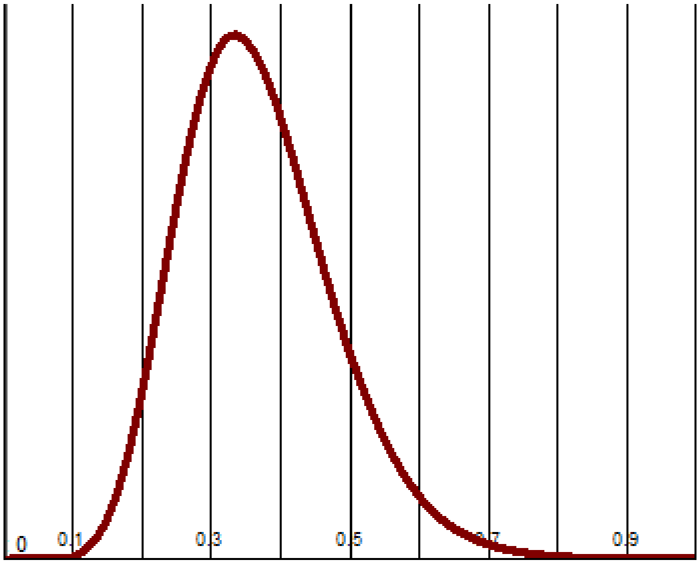
Practice Problems.

1. For each of the following situations, find the maximum likelihood function.
   a. You find a die at your friend's house and think that it's coming up 4 entirely too frequently to be fair. You suspect it is weighted. To test this, you roll the die 25 times and obtain the following sequence of rolls: $\{4, 5, 2, 3, 4, 4, 1, 6, 4, 2, 3, 5, 4, 2, 6, 1, 4, 4, 1, 4, 2, 3, 5, 1, 6\}$. Use this information to find the maximum likelihood function to estimate the value of p= probability of obtaining a 4.
   b. Suppose that you have 35 items in a collection, some of which are fake and some of which are genuine. You want to estimate the probability of fake items by taking a small sample. You test 8 items and find that 6 are genuine and 2 are fake. Use the hypergeometric distribution to obtain the maximum likelihood function to estimate the value of M.
   c. The number of customers that arrive at a certain drive-through between 2 and 3 p.m. each day can be modeled as a Poisson random variable. Suppose that you want to estimate the parameter for the Poisson distribution that applies to a new location, so you record the number of customers for that hour for a week. Your x_i values are $\{10, 15, 18, 22, 19, 16, 12\}$. Find the maximum likelihood function needed to estimate the parameter λ.
   d. Suppose that SAT scores are distributed normally, and you'd like to calculate the mean and standard deviation upon which they are based. You obtain a sample of 10 scores for the quantitative section given by {510, 580, 430, 710, 220, 620, 550, 490, 700, 330}. Find the maximum likelihood function for μ, σ.
   e. The wait-times for a Poisson process are modeled by the exponential distribution. Observations of the wait-times yield times of $\{22, 34, 16, 25, 29, 45, 32, 11, 27\}$. Use this information to find the maximum likelihood function for this situation.
   f. Body sizes are measured for a certain species of reptile and they find body lengths to be $\{8, 4, 12, 11, 9, 10, 9, 8, 7, 6, 9\}$ centimeters for a sample of specimens. Use the lognormal distribution to find the maximum likelihood function for this situation.

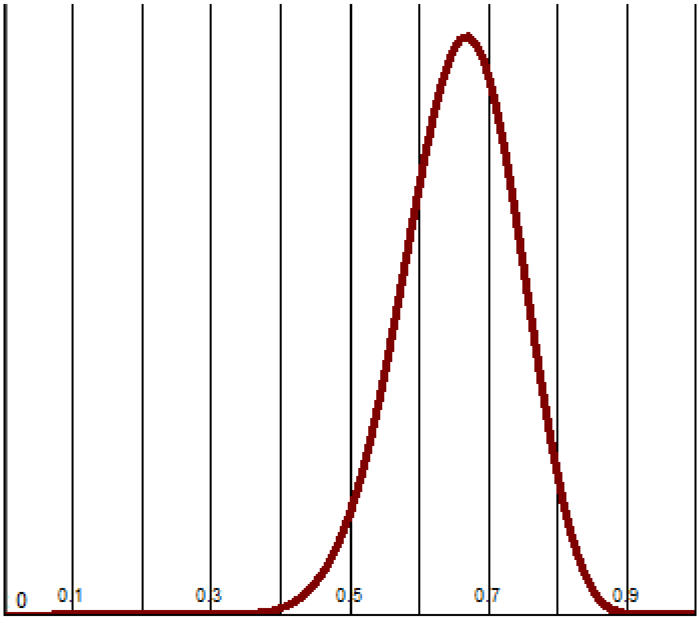
Once we obtain the maximum likelihood function $L(f)$, we then wish to use it to estimate the parameter(s). Since we are finding a maximum, we will take the derivative of continuous functions and set the derivative equal to zero (excluding any values that must be excluded, i.e. zeros are often disallowed, negative values, etc.) For functions like the hypergeometric, we can model them numerically with Excel to obtain the most likely values for the sought-after parameter.

Take the derivative of the general case is inconvenient at best. To take the derivative of many products, as in $L(f) = \prod_{i=1}^{n} f_i(\lambda)$, we would need to apply logarithmic differentiation to obtain the derivative. By taking the natural log of both sides: $\ln\big(L(f)\big) = \ln(\prod_{i=1}^{n} f_i(\lambda)) = \sum_{i=1}^{n} \ln\big(f_i(\lambda)\big)$. Taking the derivative of the sum is considerably easier in the most general case since we can do it term by term without needing to know the exact number of terms, and maximizing the log function yields the same results as maximizing the original function. However, when we are not trying to prove the most general case and have specific data, logarithmic differentiation is almost never needed. As long as algebra combines things nicely, we usually just have a simple product rule with maybe two terms to work with.

Example 5. In Example 1, we obtained the maximum likelihood function $L(f) = \lambda^{10}e^{-30\lambda}$. The derivative of this with respect to λ is $\frac{d}{d\lambda}\left[\lambda^{10}e^{-30\lambda}\right] = 10\lambda^9e^{-30\lambda} - 30\lambda^{10}e^{-30\lambda} = 0$. If we factor out the common terms, we obtain $10\lambda^9e^{-30\lambda}[1-3\lambda] = 0$. We can't use $\lambda = 0$, nor can the exponential piece be zero either, so we solve the remaining factor to obtain the maximum likelihood estimate (or MLE). $1 - 3\lambda = 0 \rightarrow \hat{\lambda} = \frac{1}{3}$. And this makes sense since in the exponential distribution, the mean is $\frac{1}{\lambda}$, and if we consider the data presented in Example 1, the mean does turn out to be 3. We can also see this is the appropriate value from looking at the graph of the function. The y-scale is extremely small. If trying to obtain this graph for yourself, you'll need to keep reducing the y-scale of the graph by 10s or 100s until it looks like more than a straight line.



Example 6. A similar procedure can be used to obtain the estimate for p in Example 2. The maximum likelihood function for that example is $L(f) = p^{20}(1-p)^{10}$. If we take the derivative with respect to p, we obtain $\frac{d}{dp}\left[p^{20}(1-p)^{10}\right] = 20p^{19}(1-p)^{10} - 10p^{20}(1-p)^9 = 0$. Factoring out the common terms we get $10p^{19}(1-p)^9[2(1-p)-p] = 0$. Since we can't use 0 or 1 as possible solutions, we solve for the remaining factor. $2(1-p) - p = 2 - 3p = 0$ or $\hat{p} = \frac{2}{3}$. We can see the peek on the attached graph as well.



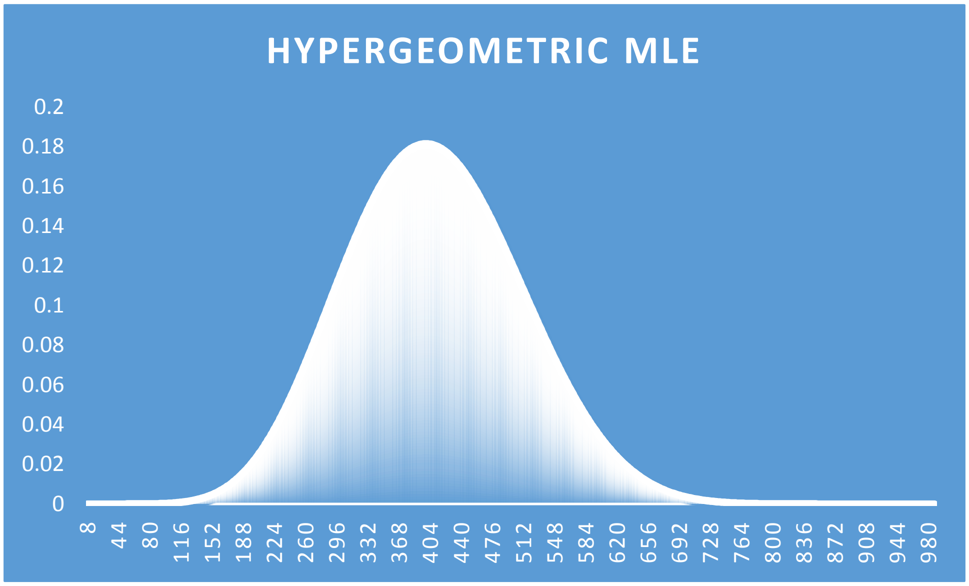
Example 7. In Example 3, we obtained the maximum likelihood function $L(f) = \binom{M}{8}\binom{1000-M}{12}$. Because this function is discrete (the combination formulas depend on factorials, which are not continuous, and which we cannot take the derivative of), we will need to do this one numerically. Note that initially M was restricted to values between 0 and 1000 (the population size). We've restricted the values now to a minimum of 8, and a maximum of 988 (since we know 12 of them at least are not of the correct type). Once we set up the formula in Excel, we can calculate all these values if we wish. If you find the values to be too large to work with, feel free to scale them, including using the original denominator.

The relevant section of the Excel calculation is this section of the table:

397	398	399	**400**	401	402	403	404	405
0.181464	0.181502	0.181523	**0.181529**	0.18152	0.181495	0.181455	0.1814	0.18133

The bolded value for $M = 400$ is where the most likely value of the function, so $\widehat{M} = 400$ for the described situation.

Shown here is also the distribution graphed, and you can see the peek on the graph right around 400.



Example 8. In Example 4, we obtained the maximum likelihood function for the normal distribution $L(f) = \frac{1}{(2\pi)^{\frac{15}{2}}\sigma^{15}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{15}(x_i-\mu)^2}$. In this case, it will be a bit easier to work with this slightly more general version of the formula. It will allow us to do just one chain rule on the sum, rather than on multiple terms. It will also have the added benefit of seeing how we obtain the formulas for μ and σ more generally. However, since there are two parameters, we will have to take two partial derivatives.

Thus $\frac{\partial L}{\partial \mu} = \frac{1}{(2\pi)^{\frac{15}{2}}\sigma^{15}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{15}(x_i-\mu)^2}\left(-\frac{2}{2\sigma^2}\sum_{i=1}^{15}(x_i-\mu)^1\right) = 0$, and $\frac{\partial L}{\partial \sigma} = \frac{-15}{(2\pi)^{\frac{15}{2}}\sigma^{16}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{15}(x_i-\mu)^2} + \frac{1}{(2\pi)^{\frac{15}{2}}\sigma^{15}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{15}(x_i-\mu)^2}\left(\frac{2}{2\sigma^3}\sum_{i=1}^{15}(x_i-\mu)^2\right) = 0$. In the derivative for μ, the initial constant can't be zero, and the exponential can't be zero, so that leaves us with $-\frac{2}{2\sigma^2}\sum_{i=1}^{15}(x_i-\mu)^1 = 0$, which essentially leaves us with $\sum_{i=1}^{15}(x_i-\mu) = 0$. But if we solve this we obtain $\sum_{i=1}^{15} x_i = \sum_{i=1}^{15}\mu = 15\mu$ or $\hat{\mu} = \frac{1}{15}\sum_{i=1}^{15} x_i$. Which is the formula for the mean we were expecting. In this case $\sum_{i=1}^{15} x_i = 967$, giving us $\hat{\mu} = \frac{967}{15} \approx 64.47$.

For the derivative with respect to σ, we can pull out the exponential piece and reduce a bit: $e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{15}(x_i-\mu)^2}\left(\frac{-15}{(2\pi)^{\frac{15}{2}}\sigma^{16}} + \frac{1}{(2\pi)^{\frac{15}{2}}\sigma^{15}}\cdot\frac{1}{\sigma^3}\sum_{i=1}^{15}(x_i-\mu)^2\right) =$

$$\frac{1}{(2\pi)^{\frac{15}{2}}\sigma^{16}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{15}(x_i-\mu)^2}\left(-15 + \frac{1}{\sigma^2}\sum_{i=1}^{15}(x_i-\mu)^2\right) = 0$$

This leaves us with $\left(-15 + \frac{1}{\sigma^2}\sum_{i=1}^{15}(x_i - \mu)^2\right) = 0$ or if we solve for $\sigma^2 = \frac{\sum_{i=1}^{15}(x_i-\mu)^2}{15}$. This is the same formula we use for the population (rather than the sample) standard deviation. Based on this, our best MLE for σ is $\hat{\sigma} \approx 3.0768$.

You'll notice this is not the same estimate obtained by other methods of estimation (such as the unbiased estimator method). When more than one method of estimation is available and one obtains different formulas, it's only through experience that statisticians have determined which one works best. As the sample size gets larger, the smaller the difference, and the less it matters which is used.

Practice Problems.

2. For each of the scenarios in Problem #1, calculate the maximum likelihood estimate (MLE) for each problem. If there are multiple parameters in the problem, find an estimate for each.
3. This method applies to any probability distribution, not just the standard one. For each of the problems below, find the maximum likelihood function for the distribution and included data set, and then use that to find the MLE for any parameters.
    a. Consider the probability distribution $f(x,\alpha) = \alpha^2 x e^{-\alpha x}, x \geq 0$. A collection of samples were obtained from this distribution and found to be $x_i = \{1, 1.14, 0.6, 0.5, 1.1, 0.2, 0.3, 0.2, 0.9\}$. Find the maximum likelihood function and use this to approximate the MLE for α.
    b. A collection of samples from the probability distribution $f(x,\alpha) = \frac{2\sqrt{\alpha}}{\pi(1+\alpha x^2)}, x \geq 0$ were obtained and found to be $x_i = \{0.1, 0.5, 0.6, 0.9, 1, 1.4, 1.7, 2.1, 3.2, 5.6\}$. Use this information to find the maximum likelihood function and the MLE for α.
    c. A certain class of objects is found to obey the probability distribution $f(x,\alpha,\beta) = \frac{4\alpha\sqrt{\beta}}{\pi(\alpha^2+\beta x^4)}, x \geq 0$. A collection of samples is obtained and found to be $x_i = \{1, 4, 5, 5, 7, 8, 10, 13\}$. Find the maximum likelihood function for this data and use it to estimate the values of α and β.