FIT2086 Lecture 3 Summary
Parameter Estimation

Dr. Daniel F. Schmidt

August 8, 2018

# 1 Part I: Maximum likelihood

**Estimation**. We have data $\mathbf{y} = (y_1, \ldots, y_n)$ from a population and want to fit a parametric distribution $p(\mathbf{y} \,|\, \boldsymbol{\theta})$ to the data as a model of the population. The process of finding good values for the parameters is called parameter estimation. We usually denote an estimate by putting a "hat" on top of it, i.e., $\hat{\theta}$ denotes some estimate of a generic parameter $\theta$.

**Maximum likelihood (ML)**. A general strategy for parameter estimation. We can use maximum likelihood to find the values of the parameters $\boldsymbol{\theta}$ that "best fit" the data. Maximum likelihood says that the best fitting model to the data we have observed is the model that assigns the maximum probability to that data. Maximum likelihood finds the distribution that assigns the maximum probability to our observed data by searching over all the possible values of $\boldsymbol{\theta}$ and finding the one that maximises the likelihood function; the likelihood function is just the probability $p(\mathbf{y} \,|\, \boldsymbol{\theta})$ of the data $\mathbf{y}$ under parameter $\boldsymbol{\theta}$, with the twist that now the data $\mathbf{y}$ is fixed (as we have observed a single sample), and we are now varying the parameters $\boldsymbol{\theta}$. ML is then formally defined as

$$\hat{\boldsymbol{\theta}}_{\rm ML}(\mathbf{y}) = \mathop{\arg\max}_{\boldsymbol{\theta}} \left\{ p(\mathbf{y} \,|\, \boldsymbol{\theta}) \right\}.$$

Equivalently, we usually solve the alternate minimisation problem

$$\hat{\boldsymbol{\theta}}_{\rm ML}(\mathbf{y}) = \mathop{\arg\min}_{\boldsymbol{\theta}} \left\{ L(\mathbf{y} \,|\, \boldsymbol{\theta}) \right\}, \tag{1}$$

where $L(\mathbf{y} \,|\, \boldsymbol{\theta}) = -\log p(\mathbf{y} \,|\, \boldsymbol{\theta})$ is called the negative log-likelihood. We do this as the negative log of the likelihood is usually easier to work with and minimise mathematically. To solve the equation (1) we can differentiate with the negative log-likelihood with respect to the model parameters, and solve for those values of the parameters that set the derivative to zero (i.e., find the turning point), i.e., for a parametric probability distribution with a single variable parameter $\theta$, we solve

$$\frac{d\, L(\mathbf{y} \,|\, \theta)}{d\theta} = 0$$

for $\theta$; this value would be the maximum likelihood estimate for $\theta$, given we have observed the data $\mathbf{y}$.

**Independently and identically distributed data**. If $y_1, \dots, y_n$ are independent and identically distributed (i.i.d.) the likelihood simplifies significantly. In this case, we can write the joint probability of $\mathbf{y}$ as

$$p(\mathbf{y} \,|\, \boldsymbol{\theta}) = \prod_{i=1}^{n} p(y_i \,|\, \boldsymbol{\theta})$$

which is just the product of the marginal probabilities of each data point. The negative log-likelihood then becomes

$$L(\mathbf{y} \,|\, \boldsymbol{\theta}) = -\sum_{i=1}^{n} \log p(y_i \,|\, \boldsymbol{\theta})$$

which is just the sum of the negative log-probabilities of each data point (a consequence of the likelihood being a product of probabilities).

**Maximum likelihood estimation of the normal distribution**. If we are fitting a normal distribution $N(\mu, \sigma^2)$ the values of the mean $\mu$ and variance $\sigma^2$ that minimise the negative log-likelihood are

$$\begin{aligned} \hat{\mu}_{\text{ML}} &= \frac{1}{n} \sum_{i=1}^{n} y_i \\ \hat{\sigma}^2_{\text{ML}} &= \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{\mu}_{\text{ML}})^2 \end{aligned}$$

which are the sample mean (sometimes called $\bar{y}$) and the sample variance (see Lecture 1), respectively. The maximum likelihood estimate of $\sigma^2$ is the average squared deviation of the samples from the sample mean.

**Example: maximum likelihood estimation of the Poisson**. Let our data $\mathbf{y} = (y_1, \dots, y_n)$ be counts (non-negative integers). We can fit a Poisson model with rate parameter $\lambda$ to this data using maximum likelihood to find the "best" value of $\lambda$. Remember that the probability distribution for a Poisson model is:

$$p(y \,|\, \lambda) = \frac{\lambda^y \exp(-\lambda)}{y!}. \tag{2}$$

As the samples are assumed to be i.i.d. in the Poisson model, the likelihood function is given by plugging (2) into (1) which gives:

$$\begin{aligned} p(\mathbf{y} \,|\, \lambda) &= \prod_{i=1}^{n} p(y_i \,|\, \lambda) \\ &= \left( \frac{\lambda^{y_1} \exp(-\lambda)}{y_1!} \right) \cdot \left( \frac{\lambda^{y_2} \exp(-\lambda)}{y_2!} \right) \cdots \left( \frac{\lambda^{y_n} \exp(-\lambda)}{y_n!} \right) \\ &= \lambda^{y_1 + y_2 + \cdots + y_n} \exp(-n\lambda) \prod_{i=1}^{n} \frac{1}{y_i!} \end{aligned} \tag{3}$$

where we use $e^a e^b = e^{a+b}$ in the third step. The negative log-likelihood is then

$$L(\mathbf{y} \,|\, \lambda) = -\sum_{i=1}^{n} y_i \log \lambda + n\lambda + \sum_{i=1}^{n} \log y_i! \tag{4}$$

To find the maximum likelihood estimator we differentiate (4) with respect to $\lambda$

$$\begin{aligned}\frac{dL(\mathbf{y}\,|\,\lambda)}{d\lambda} &= -\sum_{i=1}^{n} y_i \frac{d}{d\lambda}\log\lambda + n \\ &= -\frac{\sum_{i=1}^{n} y_i}{\lambda} + n\end{aligned}$$

Now we set this derivative to zero, and solve for $\lambda$:

$$\begin{aligned} & -\frac{1}{\lambda}\sum_{i=1}^{n} y_i + n = 0 \\ \Rightarrow \quad & -\sum_{i=1}^{n} y_i + n\lambda = 0 \\ \Rightarrow \quad & n\lambda = \sum_{i=1}^{n} y_i \end{aligned}$$

so that the maximum likelihood estimator of $\lambda$ (i.e., the value of $\lambda$ that maximises the likelihood (3)) is

$$\hat{\lambda}_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

which is the sample mean of the data $\mathbf{y}$.

**Using ML to make predictions**. Once we have found maximum likelihood estimates, we have fitted a model to the data. We can use this model as a model of our population (from which the data was sampled) by "plugging" the estimated parameters into the probability density of our model, i.e., $p(y\,|\,\hat{\theta}_{\mathrm{ML}})$. This is called the plug-in distribution. For example, if our data was $\mathbf{y} = (3, 2, 6, 10)$, and we fitted a Poisson model to the data using maximum likelihood we would have $\hat{\lambda} = (3+2+6+10)/4 = 5.25$. Our estimated probability distribution for some future sample $y$ from population would then by

$$p(y\,|\,\lambda = 5.25) = \frac{5.25^{y}\exp(-5.25)}{y!}.$$

We could now use this to make predictions about the population. In packages such as R this is easily done by using the built-in functions for the various probability distributions and setting the distribution parameters to be the maximum likelihood estimates.

## 2 Part II: Comparing Estimators

**Sampling distributions**. Remember in data science, we are using a sample from a population. The data $\mathbf{y}$ we have observed is just one possible sample of $n$ datapoints from the population we could have observed – if we took another sample, we would invariably get a different set of observations. If we are using a sample to estimate a parameter of a statistical model, for example, using maximum likelihood, each sample would result in a different estimate of the parameter. For example. imagine our population consisted of five individuals with heights (measured in meters):

$$\mathbf{x} = (1.8, 1.38, 1.81, 2.01, 1.76)$$

In this case the population average (mean) height is $1.752m$. Imagine we draw a sample of size $n = 3$ from our population, and estimate the population mean height using the mean of the sample. There are

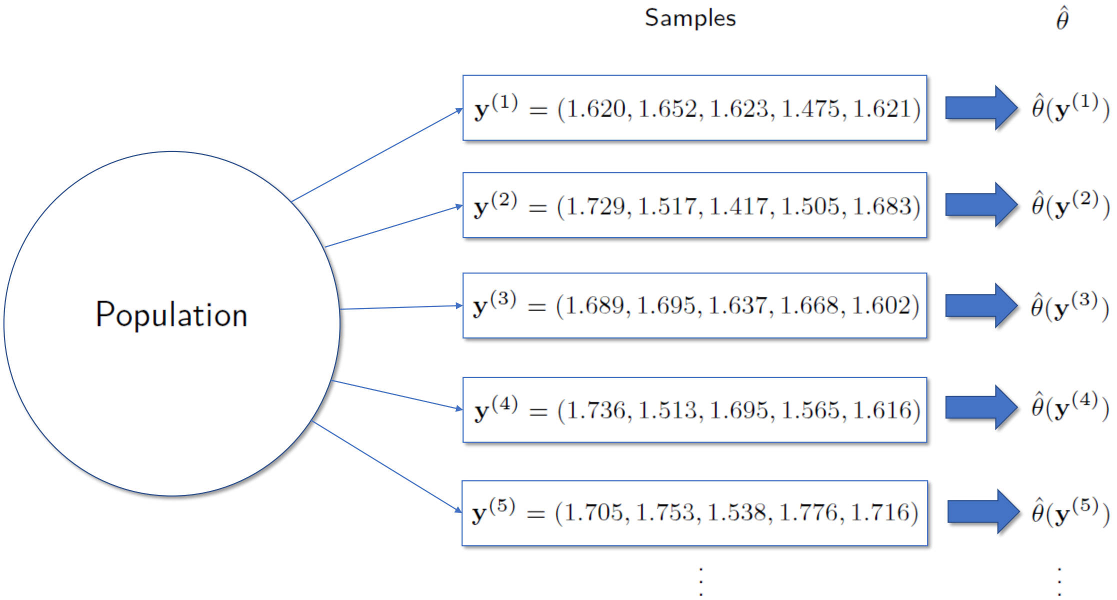
Samples
$\hat{\theta}$
Population
$\mathbf{y}^{(1)} = (1.620, 1.652, 1.623, 1.475, 1.621)$
$\mathbf{y}^{(2)} = (1.729, 1.517, 1.417, 1.505, 1.683)$
$\mathbf{y}^{(3)} = (1.689, 1.695, 1.637, 1.668, 1.602)$
$\mathbf{y}^{(4)} = (1.736, 1.513, 1.695, 1.565, 1.616)$
$\mathbf{y}^{(5)} = (1.705, 1.753, 1.538, 1.776, 1.716)$
$\hat{\theta}(\mathbf{y}^{(1)})$
$\hat{\theta}(\mathbf{y}^{(2)})$
$\hat{\theta}(\mathbf{y}^{(3)})$
$\hat{\theta}(\mathbf{y}^{(4)})$
$\hat{\theta}(\mathbf{y}^{(5)})$
$\vdots$
$\vdots$

Figure 1: An (infinite) number of different random samples can be drawn from a population. Each sample would lead to a potentially different estimate $\hat{\theta}$ of a population parameter $\theta$. The distribution of these estimates is called the sampling distribution of $\hat{\theta}$.

$\binom{5}{3} = 10$ possible samples we could draw from our population. Imagine our sample was $(1.8, 2.01, 1.38)$; then the sample mean would be 1.73. However, our sample, if randomly chosen, could quite as easily been $(1.38, 1.81, 1.76)$ for which the sample mean would be 1.65. Each of the ten possible samples we could have chosen from our population would result in a different sample mean, and therefore a different estimate of the population mean.

Table 1 demonstrates this for our population – each sample of size $n = 3$ is listed in the first column, and the sample mean for that sample is listed in the right hand column. The amount by which each of our estimates will vary from sample to sample would intuitively seem to depend on the size of the sample we take (the bigger the sample, the less variability) and also the natural variability in the population – that is, how much the different values in the population vary from each other. All the possible values of the sample mean that we could find by sampling from our population define a distribution. This is called the sampling distribution of the sample mean.

Of course, in reality, our population is very large – usually assumed to be infinitely large in comparison to the size of our sample, so we cannot find a sampling distribution by enumerating all of the possible samples we could see. Furthermore, the reason we sample from the population is that we do not have the resources to get data on the entire population – if we could we would not really need to estimate anything. So instead, we usually make some assumptions about the population, such as either assuming it follows some specific distribution, or weaker assumptions such as only assuming we know the values of the mean and variance of the population. Figure 1 demonstrates ideas behind the concept of repeated sampling.

What do we do when we have this information? There are actually a number of uses for this sampling distribution – the one we will look at today is in evaluating estimators. For a given estimator of a population parameter – say the population mean – we can make some assumptions about

| Sample | Mean $\hat{\mu}_{\mathrm{ML}}$ |
|---|---|
| $\{1.81, 2.01, 1.76\}$ | 1.8600 |
| $\{1.38, 2.01, 1.76\}$ | 1.7167 |
| $\{1.38, 1.81, 1.76\}$ | 1.6500 |
| $\{1.38, 1.81, 2.01\}$ | 1.7333 |
| $\{1.80, 2.01, 1.76\}$ | 1.8567 |
| $\{1.80, 1.81, 1.76\}$ | 1.7900 |
| $\{1.80, 1.81, 2.01\}$ | 1.8733 |
| $\{1.80, 1.38, 1.76\}$ | 1.6467 |
| $\{1.80, 1.38, 2.01\}$ | 1.7300 |
| $\{1.80, 1.38, 1.81\}$ | 1.6633 |

Table 1: This table demonstrates all possible samples of size $n = 3$ from the finite population $\{1.80, 1.38, 1.81, 2.01, 1.76\}$, along with the sample means for each of the 10 possible samples. The table clearly shows how taking a different sample of data from a population leads to a slightly different estimate. The distribution of this estimates is called the *sampling distribution.*

the population and then use the sampling distribution to determine how well our estimator would do at estimating the population parameter. By varying or changing our population assumptions we can see how robust the estimator is to different populations, and get an insight into how well the estimator performs in different situations. To do this we need some metrics of performance of an estimator. The ones we will look at are: (i) bias, (ii) variance, (iii) mean squared error and (iv) consistency.

**Bias**. Estimator bias is the degree to which an estimator tends to over or under-estimate the population parameter. Formally $\underset{\sim}{Y} = (Y_1, \ldots, Y_n)$ is our sample from the population, and $\hat{\theta}(\underset{\sim}{Y})$ is an estimator of a population parameter $\theta$. Remember, in this context we are using "$\theta$" to denote a general model parameter – in practice this could be the mean $\mu$ of a normal distribution, or the success probability $\theta$ of a Bernoulli, or the rate parameter $\lambda$ of a Poisson distribution. The bias is then defined as

$$b_\theta(\hat{\theta}) = \mathbb{E}\left[\hat{\theta}(\underset{\sim}{Y})\right] - \theta \tag{5}$$

where the expectation is taken with respect to the (population) distribution of our sample. That is, we are looking for the average difference between the estimator and the population parameter, where the average is taken over all the possible samples from our population. There are three distinct types of bias:

1. If $b_\theta(\hat{\theta}) < 0$, then the estimator tends to *underestimate* (be smaller, on average, than) the population parameter $\theta$

2. If $b_\theta(\hat{\theta}) > 0$, then the estimator tends to *overestimate* (be greater, on average, than) the population parameter $\theta$

3. If $b_\theta(\hat{\theta}) = 0$, then the estimator, on average, neither overestimates or underestimates the population parameter $\theta$

We note that the bias is a function of the population parameter $\theta$ – this implies that an estimator can be more or less biased for different values of the population parameter. This is an important point: when evaluating estimators we want to look to see how they perform under different value of the unknown population parameter. This lets us get an idea of how the estimator may behave on real data, and if there are certain populations for which it may perform better or worse.

A very special case is when $b_\theta(\hat{\theta}) = 0$ for all values of the population parameter $\theta$. In this case we say $\hat{\theta}$ is an unbiased estimator of the population parameter $\theta$.

**Variance**. The second metric we will look at is estimator variance. Given an estimator $\hat{\theta}(\underset{\sim}{Y})$ this is defined as

$$\mathrm{Var}_\theta(\hat{\theta}) = \mathbb{E}\left[\left(\hat{\theta}(\underset{\sim}{Y}) - \mathbb{E}\left[\hat{\theta}(\underset{\sim}{Y})\right]\right)^2\right] = \mathbb{V}\left[\hat{\theta}(\underset{\sim}{Y})\right] \tag{6}$$

where the expectation is again, taken with respect to the (population) distribution of $\underset{\sim}{Y}$. The variance measures how much, on average, we expected our parameter estimate to vary from sample to sample, if we were able to repeatedly resample from our population. The greater the variance, the more variation we expected to see in our estimate if we took a new sample from our population. The variance is once again a function of the population parameters, which is intuitive as we previousy discussed that the greater the variability in the values of the population, the greater the variability we expected in our estimator (in general). The estimator variance is equal to the variance of the sampling distribution.

**Mean squared error (MSE)**. When comparing two estimators, say $\hat{\theta}_1$ and $\hat{\theta}_2$, it is very possible that one estimator will have smaller bias but greater variance than the other. How do we decide which one is better? One way is to measure how close, on average, the estimates they produce are to the population parameter. To measure how close our estimator is, we could calculate the squared difference between our estimate and the population parameter. Averaging this over all the possible samples we could draw from our population gives us the mean-squared error (MSE) of the estimator:

$$\mathrm{MSE}_\theta(\hat{\theta}) = \mathbb{E}\left[(\hat{\theta}(\underset{\sim}{Y}) - \theta)^2\right].$$

The MSE measures how far, on average, the estimator is from the population parameter. The larger the value, the further on average the estimator is from the truth. Remember, by average, we mean averaged over all the possible samples from our population. Squared-error is just one measure of distance we can use; other measures, like absolute error could be just as plausible measures of distance. The squared-error is often used as it has nice mathematical properties. Perhaps the most important is the so called bias-variance decomposition of MSE. This means we can write the MSE as:

$$\mathrm{MSE}_\theta(\hat{\theta}) = b_\theta^2(\hat{\theta}) + \mathrm{Var}_\theta(\hat{\theta}) \tag{7}$$

so that the MSE is the sum of the squared bias plus the variance. This holds for any estimator – as long as we know the bias and variance, we can calculate the MSE directly from equation (7). This is even easier in the case of unbiased estimators, for which the MSE is simply equal to the variance.

**Consistency**. The final property of estimators we will examine is consistency. Loosely speaking, an estimator is considered consistent if for increasing sample size $n \to \infty$, the estimator gets closer and closer to the population value. It is essentially a guarantee that for large enough sample sizes, the estimator will basically estimate the population parameter without error. Clearly this is a very desirable property for an estimator to have. Proving that an estimator is consistent is not in general easy. However, there is a result we can use if we know the bias and variance of an estimator. If

$$b_\theta(\hat{\theta}) \to 0 \tag{8}$$

$$\mathrm{Var}_\theta(\hat{\theta}) \to 0 \tag{9}$$

as the sample size $n \to \infty$, then the estimator is consistent. In words, this says that if the bias and variance both go to zero for very large (asymptotic in) sample sizes $n$, then we can conclude that the estimator is consistent.

**Example: bias, variance, MSE and consistency of the sample mean**. We conclude by calculating the bias, variance and MSE for the sample mean. This example is important for two reasons:

1. It demonstrates that for the sample mean, we can get quite general results from quite weak assumptions about the population;

2. The result can be applied to any estimator that is equivalent to the sample mean.

We make the following assumptions about the population: we assume only that data values from the population $Y_i$ are independent, and have mean $\mathbb{E}[Y_i] = \mu$ and variance $\mathbb{V}[Y_i] = \sigma^2$. We assume nothing about the *distribution* of the values, beyond these three facts. Let $Y_1, \ldots, Y_n$ be a sample of size $n$ drawn from the population. The sample mean $\bar{Y}$ is then

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i.$$

Let us first derive the bias of the sample mean. Referring to equation (5) we see that we need the expected value of the estimator (where the expectation is taken with respect to the population distribution of $Y_1, \ldots, Y_n$). We can write

$$\mathbb{E}\left[\bar{Y}\right] = \mathbb{E}\left[\frac{Y_1 + Y_2 + \cdot + Y_n}{n}\right],$$

and remembering that the (i) the expectation of a sum is the sum of expectations, and (ii) that $\mathbb{E}[c\,Y_i] = c\,\mathbb{E}[Y_i]$, we can write this as

$$\mathbb{E}\left[\bar{Y}\right] = \frac{\mathbb{E}[Y_1]}{n} + \frac{\mathbb{E}[Y_2]}{n} + \cdots + \frac{\mathbb{E}[Y_n]}{n}$$

Finally, we note that by the assumptions we made about our population, $\mathbb{E}[Y_i] = \mu$; substituting this into the above equation yields

$$\mathbb{E}\left[\bar{Y}\right] = \mu. \qquad (10)$$

Using this expectation in the bias equation (5) gives us the bias of the sample mean

$$b_{\mu}(\bar{Y}) = 0.$$

We see that this value is always zero, irrespective of the value of the population mean. So, under the assumptions that data from the population has a mean of $\mu$ and a variance of $\sigma^2$, we see that the sample mean is an unbiased estimator of the population mean.

Now let us turn our attention to the variance of the estimator. Using an approach similar to the above, we write

$$\mathbb{V}\left[\bar{Y}\right] = \mathbb{V}\left[\frac{Y_1 + Y_2 + \cdots + Y_n}{n}\right].$$

Now we use the fact that the $Y_i$ are assumed to be independent so that the variance of a sum becomes the sum of the variances, and the fact that $\mathbb{V}[c\,X] = c^2\,\mathbb{V}[X]$ (see Lecture 2) to arrive at:

$$\mathbb{V}\left[\bar{Y}\right] = \frac{\mathbb{V}[Y_1]}{n^2} + \frac{\mathbb{V}[Y_2]}{n^2} + \cdots + \frac{\mathbb{V}[Y_n]}{n^2}.$$

Finally, we note that by our assumptions, $\mathbb{V}[Y_i] = \sigma^2$; substituting into the above equation gives us the variance of our estimator:

$$\mathrm{Var}_{\sigma^2}(\hat{\theta}) = \frac{\sigma^2}{n}. \qquad (11)$$

We see that the variance of the sample mean depends on two quantities:

1. The larger the variance of the data from the population, $\sigma^2$, the larger the estimator variance; this is intuitive as it says that the more variable the underlying population is, the harder it is to nail down the average value;

2. The larger the sample size $n$, the smaller the estimator variance. This is also intuitive as it says the more data we sample from the population, the better our estimate will become.

The MSE of the sample mean is easy to calculate given the bias and variance, using (7). As the estimator is unbiased (bias is always zero), the MSE is just equal to the variance (11), i.e., $\text{MSE}_{\sigma^2}(\bar{Y}) = \sigma^2/n$. Finally, we can establish the consistency of the sample mean. We note that the bias is always zero, so it satisfies condition 1 (equation (8)). The variance is $\sigma^2/n$, which goes to zero as $n \to \infty$, so it also satisfies condition 2 (equation (9)).

The importance of these results is that they were derived under very general assumptions – all we assumed was that the data from the population had some mean and variance. This means that we can apply them to get the bias and variance of any estimator that is equivalent to the sample mean. The next example demonstrates this when data comes from a Poisson population.

**Example 2: Maximum likelihood estimator of Poisson rate parameter**. Let us return to our Poisson rate estimator. In Part I of this document we derived the ML estimator for the Poisson rate parameter:

$$\hat{\lambda}_{\text{ML}} = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

which we can identify is equivalent to the sample mean. This means we can use the results from the previous section to find the bias, variance and MSE of this estimator. Let us assume that $Y_1, \ldots, Y_n \sim \text{Poi}(\lambda)$; that is, our population is a Poisson distribution with (unknown) rate parameter $\lambda$. We observe our sample of size $n$ and estimate $\lambda$ using the ML estimator $\hat{\lambda}_{\text{ML}}$. To find the bias and variance we first need to establish the values of the population mean and variance under our assumed population distribution. If $Y \sim \text{Poi}(\lambda)$, then we know (see Lecture 2) that $\mathbb{E}[Y_i] = \lambda$ and $\mathbb{V}[Y_i] = \lambda$. Using these assumptions and equation (10) we find that the expected value of $\hat{\lambda}_{\text{ML}}$ is

$$\mathbb{E}\left[\hat{\lambda}_{\text{ML}}\right] = \lambda.$$

This result, and equation (11) let us find the bias and variance:

$$\begin{aligned} b_\lambda(\hat{\lambda}_{\text{ML}}) &= 0, \\ \text{Var}_\lambda(\hat{\lambda}_{\text{ML}}) &= \frac{\lambda}{n}, \end{aligned}$$

so that the ML estimator of the Poisson rate parameter $\lambda$ is an unbiased estimator of the population rate parameter $\lambda$, with a variance of $\lambda/n$. So for larger values of the population rate parameter, the estimator is expected to vary more from sample to sample than for small values. As the bias is zero, the MSE of the ML estimator of $\lambda$ is just

$$\text{MSE}_\lambda(\hat{\lambda}_{\text{ML}}) = \frac{\lambda}{n}.$$

As the bias is always zero, and $\lambda/n \to 0$ as $n \to \infty$, we see from our conditions on consistency (equations (8 and 9)) that the ML estimator of the rate parameter is consistent.