FIT2086 Lecture 3 Notes, Part II
Parameter Estimation and Maximum Likelihood

Dr. Daniel F. Schmidt*

August 25, 2020

1 Introduction

The first two lectures of this unit introduced us to the basic building blocks of data science: random variables and parametric probability distributions. Furthermore, we saw why these tools are useful and how they are used by data scientists to build robust models of reality that take into account the various sources of uncertainty that are unavoidable in empirical, data-based science. There was, however, an elephant in the room that we did not address. The parametric probability distributions described in Lecture 2 might be suitable for modelling real-world processes such as the number of hot days in a month or the weights of individuals with diabetes – but they all require that the parameters of the distributions are set to values that appropriately reflect reality.

In practice we do not know these values ahead of time – instead, the central quantity in the field of data science is the *observed data*. Whether this is the outcome of some controlled experiments, or data observed from users on the web, the data is all we have access to. As discussed in Lecture 1, the basic aim of data science as a discipline is to take this data and build models that (hopefully) accurately reflect the unknown process that generated the data. The process of going from data to a model is called inductive inference, or sometimes simply statistical inference. In Part II of this book we will examine the three key components of statistical inference:

1. Point estimation: learning a model from data by finding the parameters that best fit the data, and hopefully captures properties of the unknown data generating process (i.e., the population).
2. Interval estimation: quantifying the *accuracy* of the model that we have learned, and therefore the degree of confidence that we have in our model.
3. Hypothesis testing: testing whether a specified hypothesis about the data generating process is plausible.

All three tools are complementary in the sense that they provide different information about our models, and all three are necessary for sound data science practice. We will begin our exploration by starting with point estimation.

2 Motivation and some Notation

Consider the usual situation: we have observed some data $\mathbf{y} = (y_1, \ldots, y_n)$ from a population (with unknown characteristics) and we would like to fit a parametric distribution $p(\mathbf{y} \mid \boldsymbol{\theta})$ to the data to use as a *model* of the population. The process of finding good values for the parameters of a statistical model is variously called point estimation, parameter estimation or sometimes "learning a model".

Before we continue, it is useful to discuss two pieces of notation common in the data science literature. The first is the use of "θ" as a standard placeholder for a generic parameters of a model. This can sometimes be confusing to students as some models (for example the Bernoulli) use the symbol θ as a specific parameter. In the remainder of the book it will usually be clear from the context whether we are using θ to refer to a generic parameter of a some arbitrary model, or a specific parameter of a specific distribution such as the Bernoulli or binomial.

The second piece of notation standard in the data science discipline is that estimates of parameters are usually represented by putting a "hat" on top of the appropriate symbol. For example, we would use $\hat{\theta}$ to denote an estimate of a generic parameter θ. Using this notation we can then define the estimation problem: given some data, and parametric distribution $p(y \mid \theta)$ with some tunable parameter θ that we need to specify, how do we go about *formally* arriving at a reasonable estimate $\hat{\theta}$ for θ so that the distribution (at least somewhat) reflects the population that generated the data?

2.1 An Example Problem

To make things a little more concrete, and set the scene we will consider a specific example of estimation. Imagine that we have performed a small experiment and collected the following heights (measured in meters) on $n = 9$ random individuals from the Australian population:

$$\mathbf{y} = (1.75, 1.64, 1.81, 1.55, 1.51, 1.67, 1.83, 1.63, 1.72).$$

This is our data; as heights are continuous we might choose to use a normal distribution (see Lecture 2) as a statistical model to represent the population from which these people were sampled. Recall that a normal distribution has two free parameters: μ, which controls the mean of the distribution, and σ^2 which controls the variance of the distribution. We obviously do not know the average height and variance of heights in our population; if we did, there would be no need to gather data and perform inference. So the question is, how do we utilise the observed data $\mathbf{y}$ to try and guess values for μ and σ^2 that lead to our model (a normal distribution) being a reasonable representative of the unknown population?

2.2 A Heuristic Approach: Minimum Squared Error

Let us first concentrate on the mean μ. If we recall that the mean of a distribution represents in some sense the average, or typical value, of a realisation (or sample) from the distribution, then it would make sense that we should choose a value of μ that is in some sense close to the data points we have observed. The question is: how to measure "closeness" in a formal way? One mathematically convenient choice is to use the sum of the squared distances of each observed data point from our choice of μ, i.e.,

$$\text{SSE}(\mu) = \sum_{i=1}^{n} (y_i - \mu)^2.$$

This is sometimes called the "squared error" because it represents how much in error the mean is if used as a representative value for all the data points. The squared error is obviously always greater than zero (unless all the data points are identical!), and the smaller the squared-error the "closer" the

value of μ is to the values of the data points. To use this as a basis for estimating μ we could propose that we should use the value of μ that minimises $\text{SSE}(\mu)$ as our best guess at the population value of μ, i.e., use the value that is closest to our observed data. Intuitively, this appears to be a reasonable strategy as we expect that our sampled data will be close (in some sense) to the unknown population mean – and therefore an estimate $\hat{\mu}$ that is close to the values in the sample should also be close to the unknown population mean.

2.3 Finding the Minimum Squared Error Estimate

While heuristic in nature, we do now have a formal mathematical specification for how to choose an estimate for the mean μ of a normal distribution using a set of observations:

$$\hat{\mu} = \arg\min_{\mu} \left\{ \sum_{i=1}^{n} (y_i - \mu)^2 \right\}.$$

We can read "$\arg\min_x\{f(x)\}$" as "find the value of x that minimises the function $f(x)$", and as before, we note that we place a hat over μ to denote our estimate. So we can read the above formula as saying: find the value of μ that minimises the sum-of-squared distances between μ and the n values of $\mathbf{y}$, and return it into $\hat{\mu}$. Due to the choice of squared distance, the above minimisation problem is easy to solve.

Recall from basic calculus the general approach finding the value of the argument x that minimises (maximises) a function $f(x)$: first, we differentiate the function $f(x)$ with respect to x; then we set the derivative equal to zero and solve for x. This finds the turning points of the function, at least one of which will minimise the function (if such a minimum value exists). For the majority of problems we will be examining there will be a single turning point, and this turning point will always be a minima. Let us now apply this find a formula for our minimum squared-error estimate of μ:

1. First, we differentiate $\text{SSE}(\mu)$ with respect to μ:

$$\begin{aligned} \frac{d\,\text{SSE}(\mu)}{d\mu} &= \sum_{i=1}^{n} \frac{d}{d\mu}(y_i - \mu)^2, \\ &= -2\sum_{i=1}^{n}(y_i - \mu), \\ &= -2\sum_{i=1}^{n} y_i + 2n\mu. \end{aligned}$$

2. Then we set the derivative to zero, and solve for μ, yielding:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

   which is readily identified as the *sample mean*.

The fact that the choice of μ which minimises the squared-error is equal to the sample mean is interesting; if we recall that due to the weak law of large numbers the sample mean *converges* to the population mean (see Lecture 2) we have some confidence that our estimate is probably close to the population mean, at least for large numbers of samples. That is a good property and offers some justification for our heuristic approach of minimising the squared-error.

Applying this estimator to the example data given in Section 2.1 yields a guess for μ of $\hat{\mu} = 1.6789$. However, as we know, the normal distribution has two parameters: the mean μ and variance σ^2. While the minimum squared-error approach gives us a way to estimate the mean, it is not immediately obvious how we could extend this measure of closeness to include the variance σ^2. In light of the correspondence between the minimum squared-error estimate for μ and the sample mean, we might be tempted to suggest that the sample variance could be used as an estimate for σ^2; however, we now are straying further into the territory of ad-hoc heuristics. Furthermore, if we encounter a statistical model for which the parameters do not correspond to simple sample quantities like the mean and variance (consider the success probability of a geometric distribution), this approach immediately comes unstuck. Instead, what is required is a *general* approach. There are a number of general approaches to estimation in the data science literature. We will now examine one of the most important: the principle of maximum likelihood.

3 Maximum Likelihood

The principle of maximum likelihood is a extremely general, and widely used, strategy for parameter estimation of arbitrary parametric statistical models. We can use maximum likelihood to find the values of the parameters $\boldsymbol{\theta}$ of a model that "best fit" the data. In essence maximum likelihood is based on the same idea as the minimum squared-error estimate we examined in the previous Section. The difference is in the way we measure "closeness" of our parameters to the data. They key idea underlying maximum likelihood is to measure closeness, or more correctly, "goodness of fit", of a model to observed data by the probability that it assigns to that data, $p(\mathbf{y} \mid \boldsymbol{\theta})$, i.e., the *joint probability* of $y_1, \ldots, y_n$ if the model parameters were $\boldsymbol{\theta}$.

The principle of 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 $p(\mathbf{y} \mid \boldsymbol{\theta})$ 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* $\boldsymbol{\theta}$. Formally:

Definition 1. Given a paramatric model $p(y \mid \boldsymbol{\theta})$ with unknown paramters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ and observed data $\mathbf{y} = (y_1, \ldots, y_n)$, the maximum likelihood estimate of $\boldsymbol{\theta}$ is given by:

$$\hat{\boldsymbol{\theta}}_{\text{ML}}(\mathbf{y}) = \arg\max_{\boldsymbol{\theta}} \{p(\mathbf{y} \mid \boldsymbol{\theta})\}.$$

The principle of maximum likelihood says that we want to find the distribution that is *most compatible* with the data. Distributions for which the observed data would be extremely unlikely to have arisen are considered incompatible with the data; distributions for which the data has a high probability of arising are considered more compatible. It is more common to solve the alternate minimisation problem

$$\hat{\boldsymbol{\theta}}_{\text{ML}}(\mathbf{y}) = \arg\min_{\boldsymbol{\theta}} \{L(\mathbf{y} \mid \boldsymbol{\theta})\}, \tag{1}$$

where $L(\mathbf{y} \mid \boldsymbol{\theta}) = -\log p(\mathbf{y} \mid \boldsymbol{\theta})$ is called the negative log-likelihood. We do this as the negative logarithm of the likelihood is usually easier to work with mathematically. The logarithm function is monotonic, so that if $a > b$ then $\log a > \log b$. Therefore, taking the negative logarithm of the likelihood transforms the problem from a maximisation to a minimisation, and the value of $\boldsymbol{\theta}$ that maximises the likelihood is the same as the value of $\boldsymbol{\theta}$ that minimises the negative log-likelihood.

To solve the equation (1) we can use the usual process: differentiate the negative log-likelihood with respect to the model parameters, and then solve for those values of the parameters that set the derivative to zero (find the turning point), i.e., for a parametric probability distribution with a single variable parameter θ, we solve

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

for θ. This value would be the maximum likelihood estimate for θ, given we have observed the data $\mathbf{y}$. For many of the problems we will examine this equation can be solved directly to yield a formula for the maximum likelihood estimate. However, in general, this is often not possible. In this case we must resort to using numerical techniques that search over the possible values of $\boldsymbol{\theta}$ for the ones that set the gradient very close to zero.

3.1 Independently and identically distributed data

In Lecture 1 we examined the concept of independent and identically distributed random variables, and noted that they play an important role in data science. We will now examine an important consequence of independence in the context of maximum likelihood estimation. If $y_1, \ldots, y_n$ are independent and identically distributed (i.i.d.) random variables the likelihood simplifies significantly. In this case, we can factorise the joint probability of $\mathbf{y}$ as

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

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

$$L(\mathbf{y}\,|\,\boldsymbol{\theta}) = -\sum_{i=1}^{n} \log p(y_i\,|\,\boldsymbol{\theta}) \tag{2}$$

which is just the sum of the negative log-probabilities of each data point (a consequence of the likelihood being a product of probabilities). From (2) we see another advantage of working with the negative logarithm of the likelihood: we have transformed the product of probabilities into a sum of negative log-probabilities. In addition to the logarithms improving the numerical stability (because we are no longer working with very small numbers), the transformation from a product to a summation also makes mathematical manipulation substantially easier. The derivative of a sum is simply the sum of the derivatives; in contrast, the derivative of a product is much more complex.

3.2 Example 1: Maximum Likelihood Estimation of the Normal

We have introduced the principle of maximum likelihood as a general tool for estimation. Let us now see how it handles estimation of a parameter such as the variance σ^2 of a normal distribution, which the more heuristic approach of minimum squared-error examined in Section 2.3 struggled to solve. This also gives us a good look at the process of deriving maximum likelihood estimates for a relatively simple, but very important, distribution.

Recall the general process for finding ML estimates detailed above: first, we derive the likelihood of our data under our chosen model; second, we take the negative logarithm of the likelihood; third, we differentiate the negative logarithm with respect to the model parameters; and finally, we solve for the values of the parameters that set the derivative to zero. If we are fitting a normal distribution $N(\mu, \sigma^2)$ the values of the mean μ and variance σ^2 then the probability distribution for our model is:

$$p(y\,|\,\mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{1}{2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right). \tag{3}$$

Under this standard normal model, the observations are independent and identically distributed, and so the likelihood function is given by plugging (3) into (3.1) which yields:

$$\begin{aligned} p(\mathbf{y} \,|\, \mu, \sigma^2) &= \prod_{i=1}^{n} p(y_i \,|\, \mu, \sigma^2), \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{1}{2}} \exp\left(-\frac{(y_1-\mu)^2}{2\sigma^2}\right) \cdots \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{1}{2}} \exp\left(-\frac{(y_n-\mu)^2}{2\sigma^2}\right), \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} y-\mu)^2\right) \end{aligned} \tag{4}$$

where we use the property that $e^a e^b = e^{a+b}$ to move the product from outside the exponential function to a sum inside the exponential function. To find the negative log-likelihood we then take the negative logarithm of (4) which yields:

$$L(\mathbf{y} \,|\, \mu, \sigma^2) = \frac{n}{2}\log\left(2\pi\sigma^2\right) + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y-\mu)^2 \tag{5}$$

where we use the facts that $\log ab = \log a + \log b$, $\log a^b = b \log a$ and $\log 1/a = -\log a$. To minimise this for μ and σ we need to differentiate equation (5) with respect to μ and σ and find the values that set the (partial) derivatives to zero, i.e., we need to solve the simultaneous equations:

$$\begin{aligned} \partial L(\mathbf{y}|\mu,\sigma)/\partial\mu &= 0, \\ \partial L(\mathbf{y}|\mu,\sigma)/\partial\sigma &= 0. \end{aligned}$$

It turns out for this problem, this is actually quite straightforward. First we find the partial derivative with respect to μ:

$$\begin{aligned} \frac{\partial L(\mathbf{y} \,|\, \mu, \sigma)}{\partial \mu} &= -\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\mu) \\ &= -\frac{1}{\sigma^2}\sum_{i=1}^{n} y_i + \frac{n\mu}{\sigma^2} \end{aligned} \tag{6}$$

which we note is similar to the derivative of the sum-of-squared errors we used to derive the minimum squared-error estimator in Section 2.3. In fact, setting equation (6) to zero and solving for μ yields

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

which is again simply the sample mean. So in this case, applying the principle of maximum likelihood yields the same result as the minimum squared-error approach. However, maximum likelihood also gives us a clear recipe for estimating σ^2. To remove μ as an unknown we can plug $\hat{\mu}$ into $L(\mathbf{y}|\mu,\sigma^2)$, and then find the derivative with respect to σ^2:

$$\begin{aligned} \frac{\partial L(\mathbf{y} \,|\, \hat{\mu}, \sigma^2)}{\partial \sigma^2} &= \frac{\partial}{\partial\sigma^2}\frac{n}{2}\left[\log\sigma^2 + \log(2\pi)\right] + \sum_{i=1}^{n}(y_i-\hat{\mu})^2\frac{\partial}{\partial\sigma^2}\frac{1}{2\sigma^2}, \\ &= \frac{n}{2\sigma^2} - \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i-\hat{\mu})^2, \end{aligned} \tag{7}$$

where we use the facts that $\log(ab) = \log b + \log b$, $\frac{\partial}{\partial x} K f(z) f(x) = K f(z) \frac{\partial}{\partial x} f(x)$ and $\frac{\partial}{\partial x} f(z) = 0$. To find the estimate for σ^2 we set the derivative to zero and solve as before. After doing this, the estimates that minimise the negative log-likelihood are:

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

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

3.3 Example 2: Maximum Likelihood Estimation of the Poisson

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

$$p(y \,|\, \lambda) = \frac{\lambda^y \exp(-\lambda)}{y!}. \tag{8}$$

As the samples are assumed to be i.i.d. in the Poisson model, the likelihood function is given by plugging (8) into (3.1) which yields:

$$\begin{aligned} p(\mathbf{y} \,|\, \lambda) &= \prod_{i=1}^{n} p(y_i \,|\, \lambda), \\ &= \left(\frac{\lambda^{y_1} \exp(-\lambda)}{y_1!} \right) \cdot \left(\frac{\lambda^{y_2} \exp(-\lambda)}{y_2!} \right) \cdots \left(\frac{\lambda^{y_n} \exp(-\lambda)}{y_n!} \right), \\ &= \lambda^{y_1 + y_2 + \cdots + y_n} \exp(-n\lambda) \prod_{i=1}^{n} \frac{1}{y_i!}, \end{aligned} \tag{9}$$

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

$$L(\mathbf{y} \,|\, \lambda) = -\sum_{i=1}^{n} y_i \log \lambda + n\lambda + \sum_{i=1}^{n} \log y_i!. \tag{10}$$

To find the maximum likelihood estimator we differentiate (10) with respect to λ

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

Now we set this derivative to zero, and solve for λ:

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

so that the maximum likelihood estimator of λ (i.e., the value of λ that maximises the likelihood (9)) is

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

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

Example 1: ML example: Poisson distribution

Imagine that we have observed the following counts

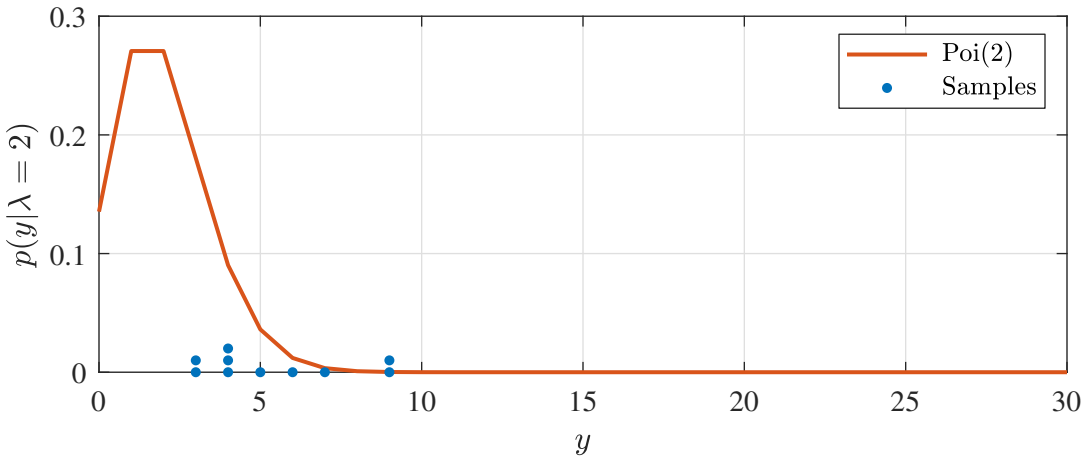
$$\mathbf{y} = (7, 9, 3, 5, 3, 4, 4, 9, 4, 6)$$

and want to use ML to fit a Poisson distribution to this data as a model of the (unknown) population from which the data was drawn. We have previously derived the ML estimate for λ and found that it is simply equal to the sample mean; therefore

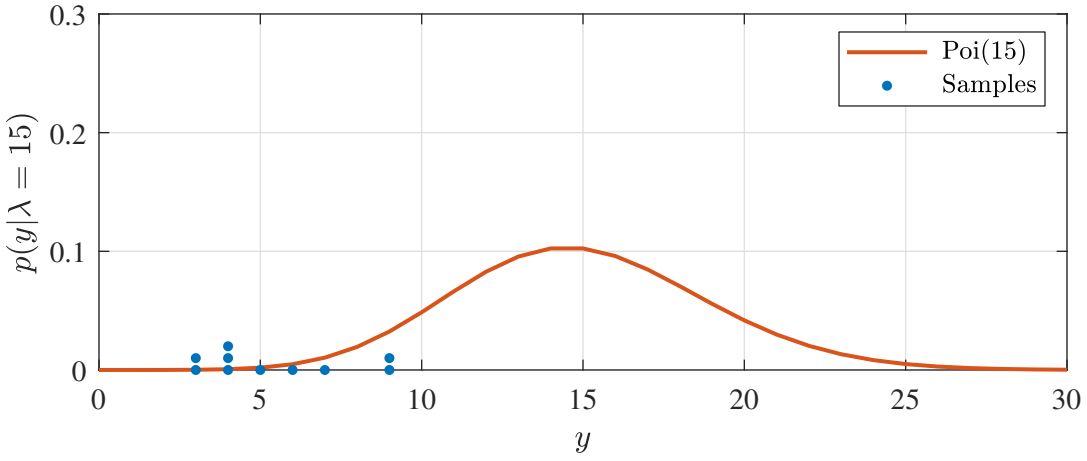
$$\hat{\lambda}_{\mathrm{ML}}(\mathbf{y}) = \left(\frac{1}{10}\right)(7 + 9 + 3 + 5 + 3 + 4 + 4 + 9 + 4 + 6) = 5.4.$$

You can think of the principle of maximum likelihood as saying "examine all the possible values of λ and choose the one that leads to the Poisson model that assigns the greatest probability to data sample we have observed". To understand a little more why the value of λ that maximises the likelihood for this sample is 5.4, consider instead the simpler problem of comparing three possible choices of λ, say $\lambda = 2$, $\lambda = 5.4$ and $\lambda = 15$ on the basis of their likelihoods. These three distributions are plotted in Figure 1, along with points representing the location of the $n = 10$ values in our sample.

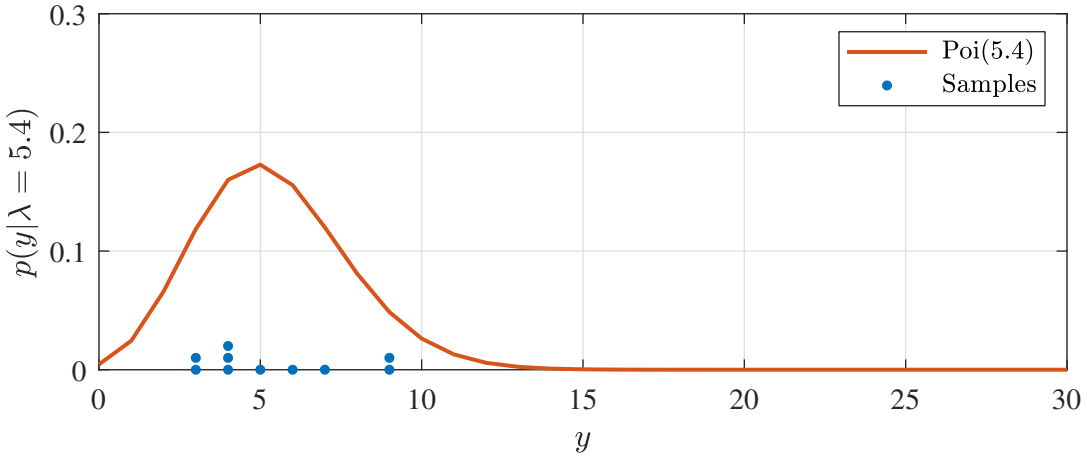
- If we examine Figure 1(a), when $\lambda = 2$, we see that it says values close to zero are quite likely, and counts greater than around 7 are *very unlikely*. This not really compatible with our observed data sample, as it has two values of $y_i = 9$.
- If we examined Figure 1(b), when $\lambda = 15$, we see that it says the values of counts from around 10 to 20 are highly likely, but values less than 5 are extremely unlikely. This is also not really compatible with our observed data sample, as it contains five values (3,3, 4,4,4) all smaller than five.
- If we examine Figure 1(c), when $\lambda = 5.4$ (i.e., the maximum likelihood value), we see that the Poisson distribution says values of counts anywhere from around 1 to 10 are quite likely, with the bulk of the probability concentrated around counts of 3 to 9. Such a probability distribution is highly compatible with our observed data. None of our observations are particularly unlikely under this model. This model seems the most compatible with our observations, and in fact, assigns the greatest probability to our data of the three models under consideration.



(a) $-\log p(\mathbf{y} \mid \lambda = 2) \approx 41.18$



(b) $-\log p(\mathbf{y} \mid \lambda = 15) \approx 62.38$



(c) $-\log p(\mathbf{y} \mid \lambda = 5.4) \approx 21.55$

Figure 1: Three possible Poisson distributions plotted along with an observed data sample of counts $\mathbf{y} = (7, 9, 3, 5, 3, 4, 4, 9, 4, 6)$.

We could compare the likelihoods of the three values of λ via their likelihood *ratios*. We see that

$$\frac{p(\mathbf{y} \mid \lambda = 5.4)}{p(\mathbf{y} \mid \lambda = 2)} \approx e^{19.63}$$

so that the observed data would be around e^{20} times ($\approx 10^8$) more likely to be generated by a Poisson distribution with $\lambda = 5.4$ than by one with a Poisson distribution of $\lambda = 2$; and

$$\frac{p(\mathbf{y} \mid \lambda = 5.4)}{p(\mathbf{y} \mid \lambda = 15)} \approx e^{40.83}$$

so that the observed data would be around e^{40} times ($\approx 10^{17}$!) more likely to be generated by a Poisson distribution with $\lambda = 5.4$ than by one with a Poisson distribution of $\lambda = 15$. This shows just how incompatible with the observed data the two choices $\lambda = 2$ and $\lambda = 15$ really are. Of course, for full maximum likelihood estimation we don't restrict our attention to only three competing values of λ; rather, we examine every value of λ and choose the one that maximises the likelihood on the observed data. □

3.4 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., $p(y \mid \hat{\theta}_{\text{ML}})$. 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

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

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.

4 Comparing Estimators

We have seen that the method of maximum likelihood offers us a general recipe for estimating the parameters of a probability distribution. However, unsurprisingly, this is only one of many types of estimators that statisticians have developed over the century or so that data science has been a discipline. As an example, consider the problem of estimating the variance σ^2 of a normal distribution. We saw that the ML estimator for the variance is

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

However, in many standard packages, and throughout the data science literature, it is common to see an alternative estimator for σ^2 being used: the so-called "unbiased estimator", which is given by

$$\hat{\sigma}^2_{\text{U}} = \left(\frac{1}{n-1}\right) \sum_{i=1}^{n} (y_i - \hat{\mu}_{\text{ML}})^2.$$

Sample	Mean $\hat{\mu}_{\text{ML}}$
$\{1.81, 2.01, 1.76\}$	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*.

When confronted with a choice such as this it is natural to ask: which one of these two estimators is better? To answer this question, we need to first determine how to define "better". This problem has been examined extensively by data scientists, and loosely speaking, the general consensus is that an estimator is better than another estimator if the estimate it produces is *on average closer to the population value* of the parameter. Of course, this opens the further questions: (i) how do we measure "on average", and (ii) how do we define "close"? We will begin by answering the first question by introducing the very important concept of sampling distributions.

4.1 Sampling distributions

Remember that in data science, we are always 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 from our population, we would invariably get a different set of observations. Their statistical properties would be *similar*, but the particular values would be different. If we are using a sample to estimate a parameter of a statistical model using, for example, maximum likelihood, each different sample would result in a different estimate of the parameter. For example. To make this idea more concrete, 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 mean height is $(1.8 + 1.38 + 1.81 + 2.01 + 1.76)/5 = 1.752m$. Now imagine that we draw a sample of size $n = 3$ from our population (i.e., we randomly pick three of the people in our population and measure their heights), and then *estimate* the population mean height using the mean of this sample. There are $\binom{5}{3} = 10$ 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 have 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 concept for our particular 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 estimates will vary from sample to sample would intuitively seem to

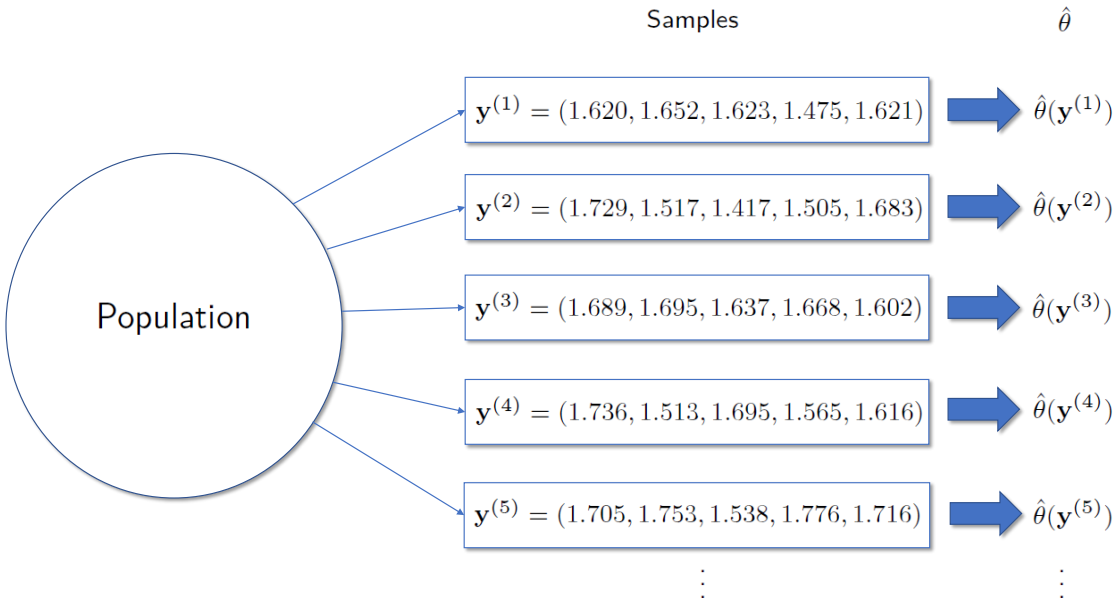


Figure 2: An (infinite) number of different random samples can be drawn from a population. Each sample would lead to a potentially different estimate $\hat{\theta}$ of a population parameter θ. The distribution of these estimates is called the sampling distribution of $\hat{\theta}$.

depend on both the size of the sample that we take (the bigger the sample, the more similar the statistical properties of the sample and the less variability) and also the natural variability present in the population – that is, how much the different values in the population vary from each other. By enumerating all of the possible values of the sample mean that we could find by sampling from our population we implicitly define a probability distribution over the sample mean. 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. Figure 2 demonstrates ideas behind the concept of repeated sampling when the population is infinitely large. The population is represented by the large circle; each of the lines represents the action of sampling from the population, each of the rectangles shows a specific sample (of size $n = 5$) from the infinitely many different possible samples, and these samples are mapped by the arrows onto some estimate of the unknown population parameter θ we are trying to estimate. How do we get a sampling distribution in the situation of an infinitely large population? As stated above, we cannot enumerate all the samples. Instead, we usually do the following: first, we make some assumptions about the population, such as assuming that it follows some specific distribution (for example, a normal distribution); we calculate what the sampling distribution *would be if the population happened to follow these assumptions.*

4.2 Sampling Distribution of the ML Estimate of the Normal Mean

The concept of repeated sampling – of the idea that the sample we have observed is just one of an infinite number of different possible samples we may have observed but did not – and sampling distributions is one of the trickiest concepts in data science. It tends to be a concept that students initially struggle with because of its abstract nature. A concrete example may help to solidify some of the ideas. The example is also useful as it concerns one of the most basic building blocks of data science: the ML estimate of the mean of a normal distribution.

Consider the following setting: we believe that our population follows a normal distribution, and we are interested in estimating the unknown mean of the normal distribution from a sample $Y = (Y_1, \ldots, Y_n)$ of size n using the maximum likelihood estimator. Recall that our "population follows a normal distribution" means that each of the measurements we record follows a normal distribution, i.e., $Y_i \sim N(\mu, \sigma^2)$. The measurements could be the heights of people, or their blood pressure, or the number of insects observed in various areas of a rainforest, and so on. We leave the mean and variance of the population, as well as the sample size, as variables without fixed values; this allows us to study the dependency of the sampling distribution of the maximum likelihood estimator on these quantities. To begin, we first recall from Section 3.2 that the maximum likelihood estimator of the mean of a normal distribution is given by

$$\hat{\mu}_{\mathrm{ML}}(Y) = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

which we identify as being equivalent to the sample mean. We know that by our assumption of a normal population model each of the measurements Y_i in our sample follows a normal distribution with mean μ and variance σ^2. We can now use two properties of the normal distribution discussed in Lecture 2 to derive the sampling distribution of $\hat{\mu}_{\mathrm{ML}}$ under the assumption that the population follows a normal distribution. Recall that:

1. if $Y_1 \sim N(\mu_1, \sigma_1^2)$ and $Y_2 \sim N(\mu_2, \sigma_2^2)$ then $Y_1 + Y_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$; and
2. if $Y \sim N(\mu, \sigma^2)$, then $Y/c \sim N(\mu/c, \sigma^2/c^2)$.

By using the first fact we can see that $\sum_{i=1}^{n} Y_i \sim N(n\mu, n\sigma^2)$, and combining this result with the second fact we can see that

$$\hat{\mu}_{\mathrm{ML}}(Y) \sim N(\mu, \sigma^2/n).$$

What does this mean? How do we interpret or understand this result? Essentially, what this result is telling us is that the population of measurements is distributed as per a normal distribution with mean μ and variance σ^2, and we drew a sample of n measurements randomly from the population and compute the sample mean, then the resulting number will behave as if it was itself drawn from a normal distribution with a mean equal to the population mean and a variance equal to the population variance divided by the sample size n. This has a number of implications. The first is that if we repeated the sampling process and drew many, many samples of size n randomly from our population, and for each of these samples computed the sample mean, and then produced a histogram of all of these sample means, the histogram would look like a normal distribution.

A second implication is that this result tells us that the chance of randomly drawing a sample from our population that leads to a sample mean more than $1.96\,\sigma/\sqrt{n}$ units away from the true, population mean μ, is only 5%. This result allows us to quantify how confident we are in our estimate $\hat{\mu}$: if the sample size is large, or the population variance (i.e., the variability of measurements in the population) is small, then $\sigma/\sqrt{n}$ will be small and our estimate will be likely to be close to the true population mean. On the flipside, if σ is large or the sample size is small, then our estimate may differ substantially from the population mean. This result is the basis of the method of confidence intervals that are used to quantify statistical uncertainty, and will be studied in more detail in Lecture 4.

4.3 Comparing Estimators

What can we do once we have obtained the sampling distribution? There are actually a number of uses for this information – the one we will examine in this chapter is in the evaluation of estimators. For a given estimator of a population parameter – say the population mean – we can make some assumptions about the population and then use the sampling distribution to determine how well our estimator would do at estimating the population parameter *on average*, if the population satisfied our assumptions. By varying or changing our population assumptions we can see how robust the estimator is to different types of populations, and get an insight into how well the estimator performs in across different situations. To make a judgement we need some metrics of performance of an estimator. While there exists an enormous literature on estimator performance metrics, we will restrict ourselves to the four most common metrics seen in the data science literature: (i) bias, (ii) variance, (iii) mean squared error and (iv) consistency.

4.3.1 Bias

The bias of an estimator measures the degree to which an estimator tends to over or under-estimate the value of the population parameter. Formally, let $Y = (Y_1, \ldots, Y_n)$ denote the sample from our population, and let $\hat{\theta}(Y)$ denote an estimator of a population parameter θ from the sample Y. Remember that the measurements in our sample are random variables, and that in this context we are using "θ" to denote a general model parameter – in practice this could be the mean μ of a normal distribution, or the success probability θ of a Bernoulli, or the rate parameter λ of a Poisson distribution. Given this setup, the bias is then defined as follows.

Definition 2. The bias of an estimator $\hat{\theta}(y_1, \ldots, y_n) \equiv \hat{\theta}$ is given by

$$b_\theta(\hat{\theta}) = \mathbb{E}\left[\hat{\theta}(Y)\right] - \theta \tag{11}$$

where the expectation is taken with respect to the (population) distribution of our sample, i.e.,

$$\begin{aligned} \mathbb{E}\left[\hat{\theta}(Y)\right] &= \int_{\hat{\theta}} \hat{\theta}\, p(\hat{\theta})\, d\hat{\theta} \\ &= \int \cdots \int \hat{\theta}(y_1, \ldots, y_n) p(y_1, \ldots, y_n \,|\, \theta) dy_1 \cdots dy_n, \end{aligned}$$

where $p(\hat{\theta})$ is the sampling distribution of $\hat{\theta}$ and $p(y_1, \ldots, y_n \,|\, \theta)$ is the population distribution of the data $y_1, \ldots, y_n$.

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*, weighted by how likely such a sample would be under our assumed 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 θ
2. If $b_\theta(\hat{\theta}) > 0$, then the estimator tends to *overestimate* (be greater, on average, than) the population parameter θ
3. If $b_\theta(\hat{\theta}) = 0$, then the estimator, on average, neither overestimates or underestimates the population parameter θ

We note that the bias is a function of the population parameter θ – 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 θ. In this case we say $\hat{\theta}$ is an unbiased estimator of the population parameter θ. When the bias is non-zero, there is exists a *systematic* (i.e., non-random) error in estimating the population parameter.

4.3.2 Variance

The second metric we will look at is estimator variance.

Definition 3. The variance of an estimator $\hat{\theta}$ is given by

$$\text{Var}_\theta(\hat{\theta}) = \mathbb{E}\left[\left(\hat{\theta}(Y) - \mathbb{E}\left[\hat{\theta}(Y)\right]\right)^2\right] = \mathbb{V}\left[\hat{\theta}(Y)\right] \tag{12}$$

where the expectation is again, taken with respect to the (population) distribution of $Y = (Y_1, \ldots, Y_n)$ (or equivalently, the sampling distribution of $\hat{\theta}$).

The variance of an estimator measures how much, on average, we expected our parameter estimate to vary from sample to sample, if we were able to repeatedly resample (i.e., draw new samples of size n) 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 previously 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 $p(\hat{\theta})$.

4.3.3 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 to answer this question is to measure how close, on average, the estimators are to the population parameter. To measure how close our estimator is we need to define what it means to be "close". A standard measure used in data science is to 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 of the estimator.

Definition 4. The mean-squared error (MSE) of an estimator $\hat{\theta}$ of population parameter θ is given by

$$\text{MSE}_\theta(\hat{\theta}) = \mathbb{E}\left[(\hat{\theta}(Y) - \theta)^2\right],$$

with the expectation again being taken with respect to the population distribution of $Y = (Y_1, \ldots, Y_n)$.

The MSE measures how far, on average, the estimator is from the population parameter in a squared-sense. 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*, weighted by how likely they are to be appear in a sample randomly drawn from population. Squared-error is obviously just

one measure of distance we can use; other measures, like absolute error could be equally plausible as measures of distance. The fact that squared-error is often used is primarily due to its nice mathematical properties. Perhaps the most important of these properties is the so called bias-variance decomposition of MSE. This useful property allows us to write the MSE as:

$$\text{MSE}_\theta(\hat{\theta}) = b_\theta^2(\hat{\theta}) + \text{Var}_\theta(\hat{\theta}) \tag{13}$$

so that the MSE of any estimator is simply the sum of the square of the bias of the estimator plus the variance of the estimator. This holds for any estimator – as long as we know the bias and variance, we can calculate the MSE directly from equation (13). This is even easier in the case of unbiased estimators, for which the MSE is just equal to the variance.

This decomposition is a fundamental relation in data science. The idea of balancing bias and variance of an estimator to reduce or minimise the mean-squared error plays a central role in much of data science model fitting and parameter estimation. For the moment we will put aside the idea that we can control the bias and variance explicitly, but we will return to these concepts in Chapter 9 when we study regularisation and penalised estimation.

4.3.4 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 utlise if we know the bias and variance of an estimator. In particular, if an estimator $\hat{\theta}$ satisfies

$$b_\theta(\hat{\theta}) \quad \to \quad 0, \tag{14}$$

$$\text{Var}_\theta(\hat{\theta}) \quad \to \quad 0, \tag{15}$$

as the sample size $n \to \infty$, then we can conclude that 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. This is intuitive: if the bias tends to zero as $n \to \infty$, this tells us that for large samples there is no systematic error in estimating θ, and if the variance tends to zero, this tells us that for large sample sizes we expect little variation from sample to sample in the estimates that our estimator produces. From the bias-variance decomposition formula (13) we see that these conditions imply that the mean-squared error tends to zero as $n \to \infty$.

4.4 Bias, Variance, MSE and Consistency of the Sample Mean

In Section 4.2 we examined an example derivation of a sampling distribution for a specific estimator (the ML estimator) in the case of the normal distribution. In this example we noted that, in general, derivation of the sampling distribution for an arbitrary estimator for an arbitrary population distribution is extremely difficult or potentially impossible. This would on the surface appear to render the computation of quantities such as bias and variance, which depend on the sampling distribution, as impractical. However, there are a several neat properties of these quantities that mean they are actually more straightforward to compute than might appear at first glance.

The first of these properties is that bias, variance, MSE and consistency (as we have presented it) all depend on the sampling distribution through its mean and variance. The specific way in which the

probability is distributed by $p(\hat{\theta})$ is not important – only its expected value, and expected squared-value appear in the formulae presented in Section 4.3. As long as we know these two quantities we can compute the estimator performance metrics previously discussed.

This fact becomes even more relevant when we consider the particular case of the *sample mean*. The case of the sample mean is important for two reasons: (i) for the particular case of the sample mean we can obtain very general results from very weak assumptions about the population; and (ii) the results 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 the data values (i.e., the observations in our sample) from the population Y_i

(A1) have a mean of $\mathbb{E}[Y_i] = \mu$;

(A2) have a variance of $\mathbb{V}[Y_i] = \sigma^2$;

(A3) and are independent.

We assume nothing further 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 given by

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

Let us first derive the bias of the sample mean. Referring to equation (11) we see that we need the expected value of the estimator (where the expectation is taken with respect to the population distribution of $Y_1, \ldots, Y_n$). To do this, we can re-write the sample mean as

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

and recalling 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 Assumption A1 that we made about our population, $\mathbb{E}[Y_i] = \mu$; substituting this into the above equation yields

$$\mathbb{E}\left[\bar{Y}\right] = \mu. \tag{16}$$

Using this expectation in the bias equation (11) yields the bias of the sample mean

$$b(\bar{Y}) = 0.$$

We see that this value is always zero, irrespective of the value of the population mean μ. Therefore, under the assumptions that data from the population has a mean of μ, we see that the sample mean is always 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 can write

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

We now utilise the fact that the Y_i are assumed to be independent (Assumption A3) so that the variance of a sum becomes the sum of the variances, and the fact that $\mathbb{V}[c\,X] = c^2\,\mathbb{V}[X]$ 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 Assumption A2 above, $\mathbb{V}[Y_i] = \sigma^2$; substituting this into the above equation gives us the variance of our estimator:

$$\text{Var}(\bar{Y}) = \frac{\sigma^2}{n}. \tag{17}$$

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

1. The larger the variance of the data from the population, σ^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 (13). As the estimator is unbiased (bias is always zero), the MSE is just equal to the variance (17), i.e., $\text{MSE}(\bar{Y}) = \sigma^2/n$. Finally, we can establish the consistency of the sample mean under Assumptions 1 through 3. We note that the bias is always zero, so it satisfies the first condition (equation (14)). The variance is σ^2/n, which goes to zero as $n \to \infty$, so it also satisfies the second condition (equation (15)).

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. The strength of this fact is that it means that we can apply these results to get the bias and variance of *any estimator* that is equivalent to the sample mean. In contrast, the explicit calculation of the sampling distribution (as per the example in Section 4.2) required detailed knowledge of the population distribution.

However, this generality is not free – the sampling distribution, if it can be calculated, provides far more nuanced information about the behaviour of the estimator than the results we calculated above. This was shown in Section 4.2 in which we used the sampling distribution to characterise the probability of the error of our estimator being greater in magnitude than a certain size. This type of statement is not possible if we make only assumptions about the mean and variance of our population. However, all is not lost – there exists a remarkable result called the central limit theorem that allows us to make (approximate) probability statements of this type even if all we assume about our population distribution is that it has a finite mean and variance! We will see study this in more detail in the next Chapter.

Example 2: ML Estimator of Poisson Rate Parameter

As a concrete example of the usefulness of the results developed for the sample mean let us examine the ML estimator for the Poisson rate parameter that we derived in Section 3.3:

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

We can readily identify this as being equivalent to the sample mean which means that we can apply the results from Section 4.4 to find the bias, variance and MSE of this estimator. This is useful as the exact sampling distribution for the ML estimator of the Poisson rate is nontrivial to derive.

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 λ. We observe a sample of counts of size n and we estimate λ 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 (16) we find that the expected value of $\hat{\lambda}_{\text{ML}}$ to be

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

This result, and equation (17) let us find both 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 λ is an unbiased estimator of the population rate parameter λ, with a variance of λ/n. This tells us that for larger values of the population rate λ, the estimator is expected to vary more from sample to sample than for small values of λ. As the bias is zero, the MSE of the ML estimator of λ is simply

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

As the bias is always zero, and $\lambda/n \to 0$ as $n \to \infty$, we see from our consistency conditions (equations (14) and (15)) that the ML estimator of the rate parameter is consistent. □

4.5 Going Beyond the Sample Mean

The results developed in Section 4.4 are only directly applicable in the particular case that our estimator of interest was equivalent to the sample mean. While many estimators do take this form, this restriction does somewhat limit the applicability of the results. We now briefly discuss how we can use some basic expectation tools to extend the results to estimators that take the form

$$\hat{\theta}(Y) = f(\bar{Y})$$

where $f(\cdot)$ is a twice-differentiable function. Essentially, if our estimator of interest happens to take the form of a *function of* the sample mean then we can apply the Taylor series expansion-based results from Lecture 2 to obtain *approximate* formulae for the bias and variance. We know that $\bar{Y}$ is a random variable with mean and variance

$$\mathbb{E}\left[\bar{Y}\right] = \mu \text{ and } \mathbb{V}\left[\bar{Y}\right] = \sigma^2/n$$

respectively, where μ and σ^2 are the mean and variance of a single observation Y_i from our population; if $f(\cdot)$ is twice differentiable we can apply the Taylor series approach from Lecture 2 to $f(\bar{Y})$ yielding:

$$\begin{aligned} b(\hat{\theta}) &\approx \left(\frac{\sigma^2}{2n}\right)\left[\left.\frac{d^2 f(x)}{dx^2}\right|_{x=\mu}\right], \\ \text{Var}(\hat{\theta}) &\approx \frac{\sigma^2}{n}\left[\left.\frac{df(x)}{dx}\right|_{x=\mu}\right]^2. \end{aligned}$$

Despite being only approximations, these formulae still give us insight into the behaviour of more general estimators. We note for example, that as long as the first and second derivatives of $f(x)$ are bounded near $x = \mu$ then both the bias and variance will tend to zero as $n \to \infty$. A corollary of this of course is that the estimator will be consistent if these conditions on $f(\cdot)$ are met.