1. ESTIMATION PROBLEMS

Consider the following examples:

(1) A coin has an unknown probability p of turning up head. We wish to determine the value of p. For this, we toss the coin 100 times and observe the outcomes $X_1, \ldots, X_{100}$. How to give a guess for the value of p based on the data?

(2) A factory manufacture light bulbs whose lifetimes may be assumed to be exponential random variables with a mean lifetime μ. We take a sample of 50 bulbs at random and measure their lifetimes $X_1, \ldots, X_{50}$. Based on this data, how can we present a reasonable guess for μ? We may want to do this so that the specifications can be printed on the product when sold.

(3) Can we guess the average height μ of students in MSO201A by taking a random sample of 100 students, and measuring their heights?

In such questions, there is an unknown parameter μ (there could be more than one unknown parameter too) whose value we are trying to guess based on the data. The data consists of i.i.d. random variables from a family of distributions. We assume that the family of distributions is known, and the only unknown is (are) the value of the parameter(s). Rather than present the ideas in abstract, let us see a few examples.

Example 1. Let $X_1, \ldots, X_n$ be i.i.d. random variables with Exponential density $f_\mu(x) = \frac{1}{\mu}e^{-x/\mu}$ for $x > 0$, where the value of $\mu > 0$ is unknown. How to *estimate* μ using the data $X = (X_1, \ldots, X_n)$?

This is the framework in which we would study the second example above, namely, the lifetime distribution of light bulbs. Observe that we have parametrized the exponential family of distributions differently from usual. We could equivalently have considered $g_\lambda(x) = \lambda e^{-\lambda x}$, but the interest is then in estimating $1/\lambda$ (which is the expected value) rather than λ. Here, are two methods.

Method of Moments (MME): We observe that $\mu = \mathbf{E}_\mu[X_1]$, the mean of the distribution (also called *population mean*). Hence, it seems reasonable to take the sample mean $\bar{X}_n$ as an estimate. On second thought, we realize that $\mathbf{E}_\mu[X_1^2] = 2\mu^2$ and hence $\mu = \sqrt{\frac{1}{2}\mathbf{E}_\mu[X_1^2]}$. Therefore, it also seems reasonable to take the corresponding sample quantity, $T_n := \sqrt{\frac{1}{2n}(X_1^2 + \cdots + X_n^2)}$ as an estimate for μ. One can go further and write μ in various ways as $\mu = \sqrt{\mathrm{Var}_\mu(X_1)}$, $\mu = \sqrt[3]{\frac{1}{6}\mathbf{E}_\mu[X_1^3]}$, etc. Each such expression motivates an estimate, just by substituting sample moments for population moments.

This idea is called estimating by the *method of moments* because we are equating the sample moments to population moments to obtain the estimate.

We can also use other features of the distribution, such as quantiles (we may call this the "method of quantiles"). In other words, obtain estimates by equating the sample quantiles to population quantiles. For example, the population median of X_1 is $\mu \log 2$ (Check!), hence a reasonable estimate for μ is $M_n/\log 2$, where M_n is a sample median. Alternately, the 25% quantile of Exponential$(1/\mu)$ distribution is $\mu \log(4/3)$, and hence another estimate for μ is $X_{([n/4])}/\log(4/3)$ and so on.

Maximum Likelihood Estimation (MLE): The joint density of $X_1, \ldots, X_n$ is

$$g_\mu(X_1, \ldots, X_n) = \mu^{-n} e^{-\frac{1}{\mu}(X_1 + \cdots + X_n)} \qquad \text{if all } X_i > 0,$$

since X_is are independent, the joint density is a product. This is called the *likelihood function*. In other words, define

$$L_X(\mu) := \mu^{-n} e^{-\frac{1}{\mu}\sum_{i=1}^n X_i}.$$

Two points: This is the joint density of $X_1, \ldots, X_n$, evaluated at the observed data. Further, we like to think of it as a function of μ with $X := (X_1, \ldots, X_n)$ being fixed (given or observed).

When μ is the actual value, then $L_X(\mu)$ is the "likelihood" of seeing the data that we have actually observed. The *maximum likelihood estimate* is that value of μ that maximizes the likelihood function. In our case, by differentiating and setting equal to zero we get,

$$0 = \frac{d}{d\mu} L_X(\mu) = -n\mu^{-n-1} e^{-\frac{1}{\mu}\sum_{i=1}^n X_i} + \mu^{-n} \left(\frac{1}{\mu^2} \sum_{i=1}^n X_i \right) e^{-\frac{1}{\mu}\sum_{i=1}^n X_i}$$

which is satisfied when $\mu = \frac{1}{n}\sum_{i=1}^n X_i = \bar{X}_n$. To distinguish this from the true value of μ which is unknown, it is customary to put a hat on the letter μ. We write $\hat{\mu}_{MLE} = \bar{X}_n$. We should really verify whether $L(\mu)$ is maximized, minimized or neither at this point. We leave it to you to do the checking (e.g., by looking at the second derivative).

Let us see the same methods at work in some more examples.

Example 2. Let $X_1, \ldots, X_n$ be i.i.d. Ber(p) random variables, where the value of p is unknown. How to *estimate* it using the data $X = (X_1, \ldots, X_n)$?

MME: We observe that $p = \mathbf{E}_p[X_1]$, the mean of the distribution (also called *population mean*). Hence, a method of moments estimator would be the sample mean $\bar{X}_n$. In this case, $\mathbf{E}_p[X_1^2] = p$ again, but we do not get any new estimate because $X_k^2 = X_k$ (as X_k is 0 or 1).

MLE: Now, we have a probability mass function instead of density. The joint pmf of $X_1, \ldots, X_n$ is $f_p(X_1, \ldots, X_n) = p^{\sum_{i=1}^n X_i}(1-p)^{n-\sum_{i=1}^n X_i}$ when each X_i is 0 or 1. The likelihood function is

$$L_X(p) := p^{\sum_{i=1}^n X_i}(1-p)^{n-\sum_{i=1}^n X_i} = p^{n\bar{X}_n}(1-p)^{n(1-\bar{X}_n)}.$$

We need to find the value of p that maximizes $L_X(p)$. Here is a trick that almost always simplifies calculations (try it in the previous example too!). Instead of maximizing $L_X(p)$, maximize $\ell_X(p) = \log L_X(p)$ (called the *log-likelihood function*). Since "log" is an increasing function, the maximizer will remain the same. In our case,

$$\ell_X(p) = \bar{X}_n \log p + n(1 - \bar{X}_n) \log(1 - p).$$

Differentiating and setting equal to 0, we get $\hat{p}_{MLE} = \bar{X}_n$. Again, the sample mean is the maximum likelihood estimate.

Example 3. Consider the two-parameter Laplace density $f_{\theta,\alpha}(x) = \frac{1}{2\alpha} e^{-\frac{|x-\theta|}{\alpha}}$ for all $x \in \mathbb{R}$. Check that $f_{\theta,\alpha}$ is indeed a density for all $\theta \in \mathbb{R}$ and $\alpha > 0$.

Now, suppose we have data $X_1, \ldots, X_n$ i.i.d. from $f_{\theta,\alpha}$, where we do not know the values of θ and α. How to estimate the parameters?

MME: We compute

$$\mathbf{E}_{\theta,\alpha}[X_1] = \frac{1}{2\alpha} \int_{-\infty}^{+\infty} t e^{-\frac{|t-\theta|}{\alpha}} dt = \frac{1}{2} \int_{-\infty}^{+\infty} (\alpha s + \theta) e^{-|s|} ds = \theta.$$

$$\mathbf{E}_{\theta,\alpha}[X_1^2] = \frac{1}{2\alpha} \int_{-\infty}^{+\infty} t^2 e^{-\frac{|t-\theta|}{\alpha}} dt = \frac{1}{2} \int_{-\infty}^{+\infty} (\alpha s + \theta)^2 e^{-|s|} ds = 2\alpha^2 + \theta^2.$$

Thus, the variance is $\mathrm{Var}_{\theta,\alpha}(X_1) = 2\alpha^2$. Based on this, we can take the method of moments estimate to be $\hat{\theta}_n = \bar{X}_n$ (sample mean) and $\hat{\alpha}_n = \frac{1}{\sqrt{2}} s_n$, where $s_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$. At this moment, the ideas of defining sample variance as s_n^2 may look strange, and it might be more natural to take $V_n := \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ as an estimate for the population variance. As we shall see later, s_n^2 has some desirable properties that V_n lacks. Whenever we say sample variance, we mean s_n^2 (unless stated otherwise).

MLE: The likelihood function of the data is

$$L_X(\theta, \alpha) = \prod_{k=1}^n \frac{1}{2\alpha} \exp\left\{ -\frac{|X_k - \theta|}{\alpha} \right\} = 2^{-n} \alpha^{-n} \exp\left\{ -\sum_{k=1}^n \frac{|X_k - \theta|}{\alpha} \right\}.$$

The log-likelihood function is

$$\ell_X(\theta, \alpha) = \log L(\theta, \alpha) = -n \log 2 - n \log \alpha - \frac{1}{\alpha} \sum_{k=1}^n |X_k - \theta|.$$

For a fixed α, we first maximize w.r.t. θ, which is equivalent to minimizing $\sum_{k=1}^n |X_k - \theta|$ w.r.t. θ.

We know that[1] for fixed $X_1, \ldots, X_n$, the value of $\sum_{k=1}^n |X_k - \theta|$ is minimized when $\hat{\theta} = M_n$, the median of $X_1, \ldots, X_n$ (strictly speaking the median may have several choices when n is even, all of them are equally good).

Alternative argument: The MLE $\hat{\theta}$ is by definition the maximizer of $\ell_X(\theta)$ for the given data X. This is equivalent to finding the minimizer of $\sum_{k=1}^n |X_k - \theta|$ for given X_k. The answer is $\hat{\theta} = M_n$.

Lemma 4. *Let $w_1 < w_2 < \cdots < w_n$ be real numbers. Then $f(t) = \sum_{k=1}^n |w_k - t|$ is minimized when t is equal to a median of $w_1, \ldots, w_n$.*

Proof. If $t \in [w_i, w_{i+1}]$, we can write $f(t) = \sum_{k=1}^{i}(t - w_k) + \sum_{k=i+1}^{n}(w_k - t)$. If $t, t+h$ ($h > 0$) are both in (w_i, w_{i+1}), then

$$\frac{f(t+h) - f(t)}{h} = \frac{hi - h(n-i)}{h} = 2i - n.$$

When $2i < n$ (respectively, $2i > n$), the derivative is negative (respectively, positive). Thus, on the interval $[w_i, w_{i+1}]$ we have

$$f(t) \text{ to be } \begin{cases} \text{decreasing} & \text{if } 2i < n, \\ \text{constant} & \text{if } 2i = n, \\ \text{increasing}, & \text{if } 2i > n. \end{cases}$$

So, $f(t)$ stays constant when $i = n/2$ and all values in $[w_{n/2}, w_{(n/2)+1}]$ are minimizing if n is even. Thus, f is minimized when t is a median.

Complete the proof by arguing that $w_{(n+1)/2}$ is the median when n is odd. ■

We fix $\hat{\theta} = M_n$, and then we maximize $\ell(\hat{\theta}, \alpha)$ over α by differentiation. We get $\hat{\alpha} = \frac{1}{n}\sum_{k=1}^n |X_k - M_n|$ (the sample mean absolute deviation about the median). Thus, the MLE of (θ, α) is $(\hat{\theta}, \hat{\alpha})$.

[1] If you do not know, here is an idea. Let $x_1 < x_2 < \cdots < x_n$ be n distinct real numbers and let $a \in \mathbb{R}$. Rewrite $\sum_{k=1}^n |x_k - a|$ as $(|x_1 - a| + |x_n - a|) + (|x_2 - a| + |x_{n-1} - a|) + \cdots$. By triangle inequality, we see that

$$|x_1 - a| + |x_n - a| \geq x_n - x_1, \quad |x_2 - a| + |x_{n-1} - a| \geq x_{n-1} - x_2, \quad |x_3 - a| + |x_{n-2} - a| \geq x_{n-2} - x_3, \ldots.$$

Further, the first inequality is an equality if and only if $x_1 \leq a \leq x_n$, the second inequality is an equality if and only if $x_2 \leq a \leq x_{n-1}$, etc. In particular, if a is a median, then all these inequalities become equalities and shows that a median minimizes the given sum.

Exercise 5. Find an estimate for the unknown parameters by the method of moments (MME) and the maximum likelihood method (MLE).

(1) $X_1, \ldots, X_n$ are i.i.d. $N(\mu, 1)$. Estimate μ. How does your estimate change if the distribution is $N(\mu, 2)$?

(2) $X_1, \ldots, X_n$ are i.i.d. $N(0, \sigma^2)$. Estimate σ^2. How does your estimate change if the distribution is $N(7, \sigma^2)$?

(3) $X_1, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$. Estimate μ and σ^2.

[**Note:** The first case is when σ^2 is known and μ is unknown. Then, the known value of σ^2 may be used to estimate μ. In the second case it is similar, now μ is known and σ^2 is not known. In the third case, both are unknown].

Hints for solution: Let $X_1, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$.

(i) The MME procedure to estimate the parameters μ and σ^2 is as follows. The first and second theoretical (population) moments for the normal distribution are

$$\mathbf{E}(X) = \mu \quad \text{and} \quad \mathbf{E}\left(X^2\right) = \mu^2 + \sigma^2.$$

The first and second sample moments are

$$m_1' = \bar{X} \text{ and } m_2' = \frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

Solving the equations, we get

$$\mu = \bar{X} \text{ and } \mu^2 + \sigma^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

We have the method of moment estimate

$$\hat{\mu} = \bar{X} \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2.$$

(ii) The MLE based on $X = (X_1, \ldots, X_n)$ can be applied to the likelihood function as follows:

$$L\left(\mu, \sigma^2 \mid X\right) = \frac{1}{\sqrt{(2\pi\sigma^2)^n}} \exp - \left\{\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(X_i - \mu\right)^2\right\}.$$

The log-likelihood is

$$\ell\left(\mu, \sigma^2 \mid X\right) = -\frac{n}{2}\left(\ln 2\pi + \ln \sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(X_i - \mu\right)^2.$$

Now,

$$\frac{\partial}{\partial \mu} \ell\left(\mu, \sigma^2 \mid X\right) = \frac{1}{\sigma^2} \sum_{i=1}^{n} \left(X_i - \mu\right) = \frac{1}{\sigma^2} n(\bar{X} - \mu) = 0.$$

Because the second partial derivative with respect to μ is negative, we have

$$\hat{\mu} = \bar{X}.$$

Now,

$$\frac{\partial}{\partial \sigma^2} \ell\left(\mu, \sigma^2 \mid X\right) = -\frac{n}{2\sigma^2} + \frac{1}{2\left(\sigma^2\right)^2} \sum_{i=1}^{n} \left(X_i - \mu\right)^2 = -\frac{n}{2\left(\sigma^2\right)^2} \left(\sigma^2 - \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \mu\right)^2\right) = 0.$$

Recalling that $\hat{\mu} = \bar{X}$, we obtain

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2.$$

Exercise: Compute the second partial derivative with respect to σ^2.

Exercise 6. $X_1, \ldots, X_n$ are i.i.d. Geo(p). Estimate $\mu = 1/p$.

Exercise 7. $X_1, \ldots, X_n$ are i.i.d. Pois(λ). Estimate λ.

Exercise 8. $X_1, \ldots, X_n$ are i.i.d. Beta(a, b). Estimate a, b.

Exercise 9. $X_1, \ldots, X_n$ are i.i.d. Uniform$[a, b]$. Estimate a, b.
This exercise is approachable by the same methods, but requires you to think a little.

2. Properties of estimates - Mean squared error

We have seen that there may be several competing estimates that can be used to estimate a parameter. How can one choose between these estimates? In this section, we present some properties that may be considered desirable in an estimator. However, having these properties does not lead to an unambiguous choice of one estimate as the best for a problem.

The setting: Let $X_1, \ldots, X_n$ be i.i.d. random variables with a common density f_θ. The parameter θ is unknown and the goal is to estimate it. Let T_n be an estimator for θ, this just means that T_n is a function of $X_1, \ldots, X_n$ (in words, if we have the data at hand, we should be able to compute the value of T_n).

Bias: Define the *bias* of the estimator as $\text{Bias}_{T_n}(\theta) := \mathbf{E}_\theta[T_n] - \theta$. If $\text{Bias}_{T_n}(\theta) = 0$ for all values of the parameter θ, then we say that T_n is *unbiased* for θ. Here, we write θ in the subscript of $\mathbf{E}_\theta$ to remind ourself that in computing the expectation we use the underlying density f_θ. However, we shall often omit the subscript for simplicity.

Mean-squared error: The *mean squared error* (MSE) of T_n is defined as $\text{MSE}_{T_n}(\theta) = \mathbf{E}_\theta[(T_n - \theta)^2]$. This is a function of θ. Smaller it is, better our estimate.

In computing the mean squared error, it is useful to observe the formula

$$\text{MSE}_{T_n}(\theta) = \text{Var}_{T_n}(\theta) + (\text{Bias}_{T_n}(\theta))^2 .$$

To prove this, consider a random variable Y with (population mean) $\mathbf{E}[Y] = \mu$ and observe that for any real number a, we have

$$\begin{aligned}\mathbf{E}[(Y-a)^2] &= \mathbf{E}[(Y-\mu+\mu-a)^2] = \mathbf{E}[(Y-\mu)^2] + (\mu-a)^2 + 2(\mu-a)\mathbf{E}[Y-\mu] \\ &= \mathbf{E}[(Y-\mu)^2] + (\mu-a)^2 = \text{Var}(Y) + (\mu-a)^2.\end{aligned}$$

Now, use this identity with T_n in place of Y and θ in place of a.

Remark 10. An analogy. Consider shooting with a rifle having a telescopic sight. A given target can be missed for two reasons. One, the marksman may be unskilled and shoot all over the place, sometimes a meter to the right of the target, sometimes a meter to the left, etc. In this case, the shots have a *large variance*. Another person may consistently hit a point 20cm to the right of the target. Perhaps the telescopic sight is not set right, and this caused the systematic error. This is *the bias*. Both bias and variance contribute to missing the target.

Example 11. Let $X_1, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$. Let $V_n = \frac{1}{n}\sum_{k=1}^n (X_k - \bar{X}_n)^2$ be an estimate for σ^2. By expanding the squares, we get

$$V_n = \bar{X}_n^2 + \frac{1}{n}\sum_{k=1}^n X_k^2 - \frac{2}{n}\bar{X}_n \sum_{k=1}^n X_k = \left(\frac{1}{n}\sum_{k=1}^n X_k^2\right) - \bar{X}_n^2.$$

It is given that $\mathbf{E}[X_k] = \mu$ and $\text{Var}(X_k) = \frac{\sigma^2}{n}$. Hence, $\mathbf{E}[X_k^2] = \mu^2 + \sigma^2$. We have seen before that $\text{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$ and $\mathbf{E}[\bar{X}_n] = \mu$. Hence, $\mathbf{E}[\bar{X}_n^2] = \mu^2 + \frac{\sigma^2}{n}$. Putting all this together, we get

$$\mathbf{E}\left[V_n\right] = \left(\frac{1}{n}\sum_{k=1}^n (\mu^2 + \sigma^2)\right) - \left(\mu^2 + \frac{\sigma^2}{n}\right) = \frac{n-1}{n}\sigma^2.$$

Thus, the bias of V_n is $\frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{1}{n}\sigma^2$.

Example 12. For the same setting as the previous example, suppose $W_n = \frac{1}{n}\sum_{k=1}^n (X_k - \mu)^2$. Then, it is easy to see that $\mathbf{E}[W_n] = \sigma^2$. Can we say that W_n is an unbiased estimate for σ^2? There is a hitch!

If the value of μ is unknown, then W_n is *not* an estimate (cannot compute it using $X_1, \ldots, X_n$!). However if μ is known, then it is an unbiased estimate. For example, if we knew that $\mu = 0$, then $W_n = \frac{1}{n}\sum_{k=1}^n X_k^2$ is an unbiased estimate for σ^2.

When μ is unknown, we define $s_n^2 = \frac{1}{n-1}\sum_{k=1}^n (X_k - \bar{X}_n)^2$. Clearly, $s_n^2 = \frac{n}{n-1}V_n$ and hence $\mathbf{E}[s_n^2] = \frac{n}{n-1}\mathbf{E}[V_n] = \sigma^2$. Thus, s_n^2 is an unbiased estimate for σ^2. Note that s_n^2 depends only on the data and hence it is an estimate, whether μ is known or unknown.

All the remarks in these two examples apply for *any distribution*, i.e.,

(1) The sample mean is unbiased for the population mean.

(2) The sample variance $s_n^2 = \frac{1}{n-1}\sum_{k=1}^n (X_k - \bar{X}_n)^2$ is unbiased for the population variance. But, $V_n = \frac{1}{n}\sum_{k=1}^n (X_k - \bar{X}_n)^2$ is not, in fact $\mathbf{E}[V_n] = \frac{n-1}{n}\sigma^2$.

It appears that s_n^2 is better, but the following remark says that one should be cautious in making such a statement.

Remark 13. In case of $N(\mu, \sigma^2)$ data, it turns out that although s_n^2 is unbiased and V_n is biased, the mean squared error of V_n is smaller! Further, V_n is the MLE of σ^2! Overall, unbiasedness is not so important as having smaller mean squared error, but for estimating variance (when the mean is not known), we always use s_n^2. Computation of the MSE is a bit tedious, so we skip it here.

Example 14. Let $X_1, \ldots, X_n$ be i.i.d. Ber(p). Then, $\bar{X}_n$ is an estimate for p. It is unbiased since $\mathbf{E}[\bar{X}_n] = p$. Hence, the MSE of $\bar{X}_n$ is just the variance which is equal to $p(1-p)/n$.

A puzzle: A coin C_1 has probability p of turning up head, while a coin C_2 has probability $2p$ of turning up head. All we know is that $0 < p < \frac{1}{2}$. Now, 20 tosses are given. You can choose all tosses from C_1, all tosses from C_2, or some tosses from each (the total is 20). If the objective is to estimate p, what do you do?

Solution: If we choose to have all $n = 20$ tosses from C_1, then we get $X_1, \ldots, X_n$ that are i.i.d. Ber(p). An estimate for p is $\bar{X}_n$ which is unbiased, and hence $\text{MSE}_{\bar{X}_n}(p) = \text{Var}(\bar{X}_n) = p(1-p)/n$. On the other hand, if we choose to have all 20 tosses from C_2, then we get $Y_1, \ldots, Y_n$ that are i.i.d. Ber$(2p)$. The estimate for p is now $\bar{Y}_n/2$ which is also unbiased and has $\text{MSE}_{\bar{Y}_n/2}(p) = \text{Var}(\bar{Y}_n/2) = 2p(1-2p)/4 = p(1-2p)/2$. It is not hard to see that for all $p < 1/2$, $\text{MSE}_{\bar{Y}_n/2}(p) < \text{MSE}_{\bar{X}_n}(p)$ and hence choosing C_2 is better (at least by mean-squared criterion)! It can be checked that if we choose to have k tosses from C_1 and the rest from C_2, the MSE of the corresponding estimate will be between the two MSEs found above. Hence, not better than $\bar{Y}_n/2$.

Another puzzle: A factory produces light bulbs having an exponential distribution with mean μ. Another factory produces light bulbs having an exponential distribution with mean 2μ. Your goal is to estimate μ. You are allowed to choose a total of 50 light bulbs (all from the first, all from the second, or some from each factory). What do you do?

Solution: If we pick all $n = 50$ bulbs from the first factory, we see $X_1, \ldots, X_n$ i.i.d. Exp$(1/\mu)$. The estimate for μ is $\bar{X}_n$ which has $\text{MSE}_{\bar{X}_n}(\mu) = \text{Var}(\bar{X}_n) = \mu^2/n$. If we choose all bulbs from factory 2 we get $Y_1, \ldots, Y_n$ i.i.d. Exp$(1/2\mu)$. The estimate for μ is $\bar{Y}_n/2$. But, $\text{MSE}_{\bar{Y}_n/2}(\mu) = \text{Var}(\bar{Y}_n/2) = (2\mu)^2/4n = \mu^2/n$. The two mean-squared errors are exactly the same!

Probabilistic thinking: Is there any calculation-free explanation why the answers to the two puzzles are as above? Yes, and it is illustrative of what may be called probabilistic thinking. Take the second puzzle. Why are the two estimates same by mean-squared error? Is one better by some other criterion?

Recall that if $X \sim \text{Exp}(1/\mu)$, then $Y = 2X \sim \text{Exp}(1/2\mu)$ and vice-versa. Therefore, if we have data from Exp$(1/\mu)$ distribution, then we can multiply all the numbers by 2 and convert it into data from Exp$(1/2\mu)$ distribution. Conversely, if we have data from Exp$(1/2\mu)$ distribution, then we can convert it into data from Exp$(1/\mu)$ distribution by dividing each number by 2. Hence, there should be no advantage in choosing either factory.

We leave it for you to think in analogous ways why in the first puzzle, C_2 is better than C_1.

3. PROPERTIES OF ESTIMATES - CONSISTENCY

Motivation: Consider the following example.

Question 15. Let $X_1, \ldots, X_n$ be i.i.d. random variables with Laplace density $f_\theta(x) = \frac{1}{2}e^{-|x-\theta|}$, where the value of $\theta \in \mathbb{R}$ is unknown. How to *estimate* it using the data $X = (X_1, \ldots, X_n)$?

Examples are $T_1(X) = X_1$, $T_2(X) = (X_1 + X_2)/2$ and $T_3(X) = \bar{X}_n$. One must notice that $\mathbf{E}_\theta[X_1] = \theta$. So, we have

$$\begin{aligned}
\text{MSE}_{T_1}(\theta) &= \text{Var}_\theta(X_1) = \frac{1}{2}\int_{-\infty}^{\infty} u^2 e^{-|u|} du = 2, \\
\text{MSE}_{T_2}(\theta) &= \frac{1}{4}(\text{Var}_\theta(X_1) + \text{Var}_\theta(X_2)) = 1, \text{ and} \\
\text{MSE}_{T_3}(\theta) &= \frac{2}{n}.
\end{aligned}$$

Which one should one choose?

MSE is a finite-sample criteria. In contrast, we might consider asymptotic properties which describe the behavior of a procedure as the sample size becomes infinite. In this module, we will look at one such property. The property of consistency seems to be quite a fundamental one, requiring that the estimator converges to the "correct" value as the sample size becomes infinite. It is such a fundamental property that the worth of an *inconsistent estimator* should be questioned (or, at least vigorously investigated). Consistency (as well as all asymptotic properties) concerns a sequence of estimators rather than a single estimator, although it is common to speak of a "consistent estimator". If we observe $X_1, X_2, \ldots$ according to a distribution f_θ, we can construct a sequence of estimators $W_n = W_n(X_1, \ldots, X_n)$ merely by performing the same estimation procedure for each sample size n. For example, $\bar{X}_1 = X_1, \bar{X}_2 = (X_1 + X_2)/2$, $\bar{X}_3 = (X_1 + X_2 + X_3)/3$, etc. We can now define a consistent sequence.

Definition 16. A sequence of estimators $W_n = W_n(X_1, \ldots, X_n)$ is a consistent sequence of estimators of the parameter θ if, for every $\epsilon > 0$

$$\lim_{n\to\infty} \mathbf{P}_\theta(|W_n - \theta| < \epsilon) = 1.$$

Informally, this statement says that as the sample size becomes infinite (and the sample information becomes better and better), the estimator will be arbitrarily close to the parameter with high probability, an eminently desirable property. Or, turning things around, we can say that the probability that a consistent sequence of estimators misses the true parameter is small.

Recall the definition of convergence in probability from Module 1 of Note 9. This definition essentially says that a consistent sequence of estimators converges in probability to the parameter

θ it is estimating. For each different value of θ, the probability structure associated with the sequence W_n is different. And the definition says that for each value of θ, the probability structure is such that the sequence converges in probability to the true θ. This is the usual difference between a probability definition and a statistics definition. The probability definition deals with one probability structure, but the statistics definition deals with an entire family of distributions.

Example 17. Let $X_1, X_2, \ldots$ be i.i.d. $N(\theta, 1)$, and consider the sequence

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Fix $\epsilon > 0$. Recall that $\bar{X}_n \sim N(\theta, 1/n)$. So,

$$\begin{aligned}
\mathbf{P}_\theta\left(\left|\bar{X}_n - \theta\right| < \epsilon\right) &= \int_{\theta-\epsilon}^{\theta+\epsilon} \left(\frac{n}{2\pi}\right)^{\frac{1}{2}} e^{-(n/2)(\bar{x}_n-\theta)^2} d\bar{x}_n && \text{(definition)} \\
&= \int_{-\epsilon}^{\epsilon} \left(\frac{n}{2\pi}\right)^{\frac{1}{2}} e^{-(n/2)y^2} dy && \text{(substitute } y = \bar{x}_n - \theta\text{)} \\
&= \int_{-\epsilon\sqrt{n}}^{\epsilon\sqrt{n}} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt && \text{(substitute } t = y\sqrt{n}\text{)} \\
&= \mathbf{P}(-\epsilon\sqrt{n} < Z < \epsilon\sqrt{n}) && (Z \sim N(0,1)) \\
&\to 1 \text{ as } n \to \infty.
\end{aligned}$$

In general, a detailed calculation is not necessary to establish consistency. Recall that for an estimator W_n, Chebyshev's inequality gives us

$$0 \leq \mathbf{P}_\theta\left(|W_n - \theta| \geq \epsilon\right) \leq \frac{\mathbf{E}_\theta\left[(W_n - \theta)^2\right]}{\epsilon^2}.$$

So, if

$$\lim_{n\to\infty} \mathbf{E}_\theta\left[(W_n - \theta)^2\right] = 0,$$

then the sequence of estimators is consistent. Again, recall the definition of convergence in quadratic mean from Module 1 of Note 9. Furthermore,

$$\mathbf{E}_\theta\left[(W_n - \theta)^2\right] = \operatorname{Var}_\theta W_n + \left[\operatorname{Bias}_\theta W_n\right]^2.$$

Putting this all together, we can state the following theorem.

Theorem 18. *If W_n is a sequence of estimators of a parameter θ satisfying*
i. $\lim_{n\to\infty} \operatorname{Var}_\theta W_n = 0$, *and*
ii. $\lim_{n\to\infty} \operatorname{Bias}_\theta W_n = 0$,
then W_n is a consistent sequence of estimators of θ.

Example 19. (Continuation of previous example) Since

$$\mathbf{E}_\theta[\bar{X}_n] = \theta \quad \text{and} \quad \text{Var}_\theta[\bar{X}_n] = \frac{1}{n},$$

the conditions of the theorem are satisfied and the sequence $\bar{X}_n$ is consistent. Furthermore, if there is i.i.d. sampling from any population (distribution) with $\text{mean}\,\theta$, then $\bar{X}_n$ is consistent for θ as long as the population (distribution) has a finite variance.

Theorem 20. *Let W_n be a consistent sequence of estimators of a parameter θ. Let a_n and b_n be sequences of constants satisfying*
i. $\lim_{n\to\infty} a_n = 1$,
ii. $\lim_{n\to\infty} b_n = 0$.
Then, the sequence $U_n = a_n W_n + b_n$ is a consistent sequence of estimators of θ.

Thus, for an estimand θ, typically, many consistent estimators exist. Also, it follows that a consistent estimator may *not* be unbiased.

Let us now consider the two estimators that we discussed earlier:

(i) Recall the MME for estimating the estimand θ. Let $m'_k = \frac{1}{n}\sum_{i=1}^n X_i^k$ and $\mathbf{E}_\theta\left(X_1^k\right) = h_k(\theta)$ for $k = 1, \ldots, p$. Assuming appropriate population moments exist, by WLLN we have

$$m'_k \xrightarrow{P} h_k(\theta) \text{ as } n \to \infty.$$

(ii) To have consistency of the MLE, the underlying density (likelihood function) must satisfy "regularity conditions" that we will not discuss here.

Let $X_1, X_2, \ldots, X_n$ be i.i.d. f_θ and $L(\theta|X) = \prod_{i=1}^n f_\theta\left(X_i\right)$ be the likelihood function. Let $\hat{\theta}$ denote the MLE of θ. Under some regularity conditions on f_θ, we can prove that

$$\hat{\theta} \xrightarrow{P} \theta \text{ as } n \to \infty.$$

Let $\tau(\theta)$ be a continuous function of θ. Then, we further have (recall continuous mapping theorem)

$$\tau(\hat{\theta}) \xrightarrow{P} \tau(\theta) \text{ as } n \to \infty.$$

That is, $\tau(\hat{\theta})$ is a consistent estimator of $\tau(\theta)$.

Remark 21. Although the MLE is consistent, the MLE:
1. may not be unique,
2. may be absurd,
3. may not be unbiased.