Chapter 2: Theory of Estimation

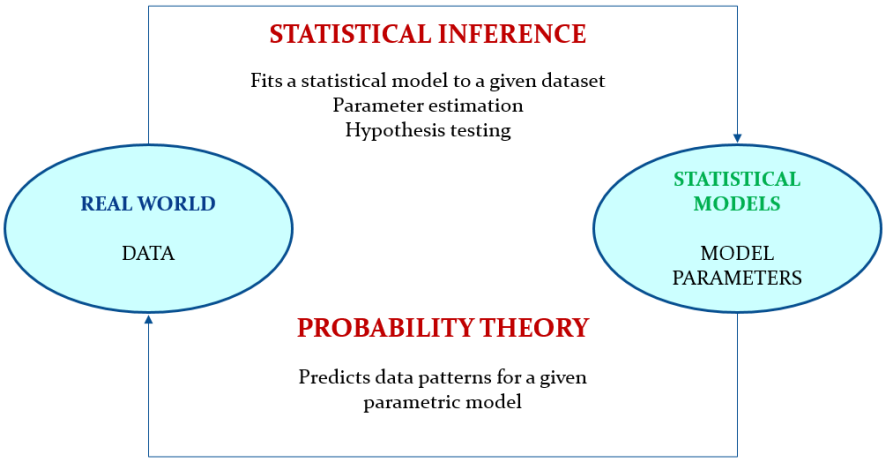

Figure 1: Interplay between Probability and Statistical Inference for Data Science

The word "inference" refers to drawing conclusions based on some evidence. Thus, *Statistical Inference* refers to drawing conclusions based on evidence obtained from the data[1].

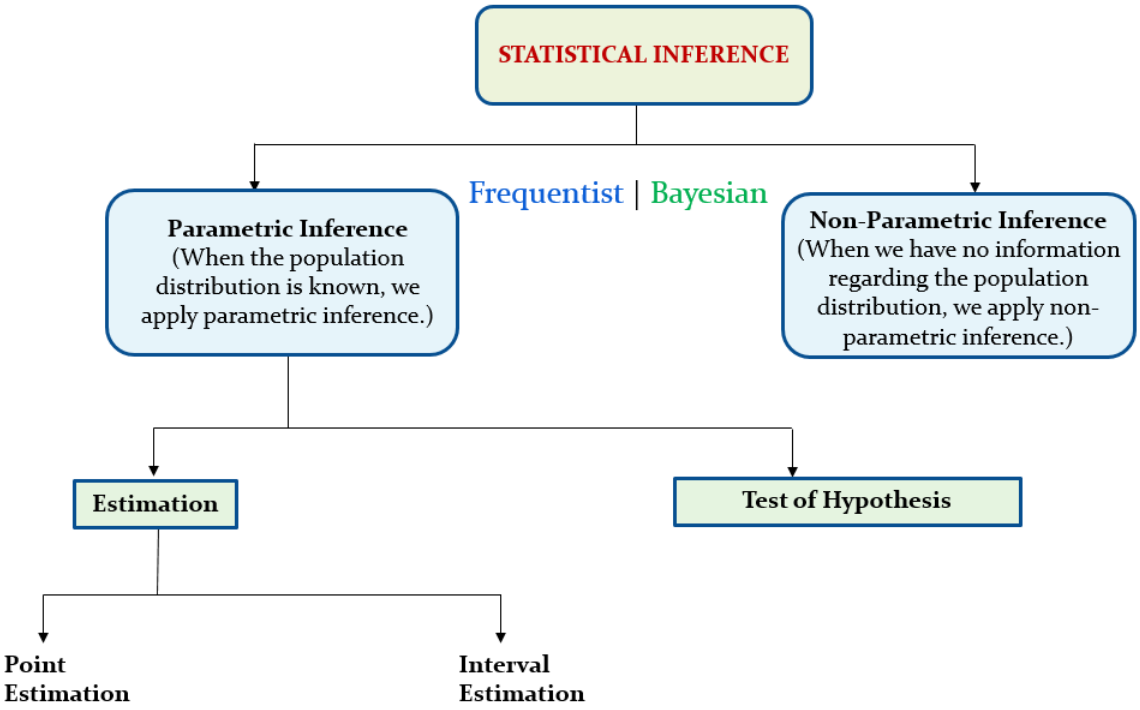

Figure 2: A loose Taxonomy of Statistical Inference

[1]Reference Book: Rice, John A. Mathematical statistics and data analysis. Cengage Learning, 2006.

The main challenges to do so are:

(i) How to summarise the information in the data using formal mathematical tools?

(ii) How to use these summaries to answer questions about the phenomenon of interest?

There is no unique way of answering these questions. There exist several schools of thought that perform statistical inference using different tools and starting from different philosophical views. These philosophical differences, as well as the different mathematical tools employed in these approaches, will be discussed in detail in this module. The two main schools of thought are the "Frequentist approach" and the "Bayesian approach". An appealing feature of this chapter is that both approaches are presented in parallel, giving the student a balanced perspective.
In many areas, researchers collect data as a means to obtain information about a phenomenon of interest or to collect information about a population. For example,

- The National Bureau of Statistics in UAE collects information about UAE residents, such as the age of those persons.
- UAE national cancer registry monitors the survival times of cancer patients diagnosed in the UAE in specific years (cohorts).
- Pharmaceutical companies conduct experiments (trials) to assess the effectiveness of new drugs on a group of people.
- The National Aeronautics and Space Administration (NASA) has a Data Portal (publicly available) with many data sets produced in their experiments and monitoring.
- Many companies monitor their financial performance by looking at the daily price of the saleable stocks of the company ("share price").
- Many others ...

We can understand the collected data as a sample of observations $x_1, \ldots, x_n$, where each $x_j, j = 1, \ldots, n$ can be either a scalar number or a vector. In statistical inference, this sample is interpreted as a realization of the random variables (vectors) $X_1, \ldots, X_n$. We will use bold capital letters to denote the vector of random variables $\mathbf{X} = (X_1, \ldots, X_n)^\top$, and bold lowercase letters to denote the sample of observations $\mathbf{x} = (x_1, \ldots, x_n)^\top$.

Data Reduction

To analyze, understand, and communicate the information contained in the sample $\mathbf{x}$, we need tools to summarise it. The process of summarising the data is known as data reduction or data summary. There exist many quantities that are used as summaries. These quantities are functions of the sample, and they are called "statistics". In mathematical terms, a statistic is any function of the sample $T(\mathbf{x})$, with $T : \mathbb{R}^n \to \mathbb{R}^m, 1 \leq m \leq n$. The statistic summarises the data in that, instead of reporting the entire sample, only the value of the statistic $T(\mathbf{x}) = t$ is reported. An example of a statistic (also known as "summary statistic") is the sample mean (or average) $T(\mathbf{x}) = \overline{\mathbf{x}} = \frac{x_1+\cdots+x_n}{n}$. For instance, if the sample $\mathbf{x}$ consists of the age of individuals in the UAE population, a way of summarising this sample is to report only the average age of the population, in which case $T(\mathbf{x})$ is the sample mean. In many cases, a statistical data analysis consists only of summarising a data set, using a choice of different summary statistics. This kind of analysis is known as "Descriptive Statistics" or "Descriptive Analysis". In fact, a descriptive analysis is usually the first step in statistical data analysis as it helps the statistician gain an understanding of the features of the data. Other summaries that are used in practice are the median (0.5 quantile) as well as other quantiles, the minimum of the sample, the maximum of the sample, etc. In fact, the set of summary statistics given by the minimum of the sample, the first quartile (0.25 quantile), the median, the third quartile (0.75 quantile), and the maximum is known as the "Five Number Summary". Visual tools (boxplots, violin plots, histograms, scatter plots, etc) are also used in applied statistics to understand other features of the data (covered in the Descriptive Statistics Course). Now, we will discuss two of the important concepts in the theory of estimation.

Point Estimation: In statistics, point estimation involves the use of sample data to calculate a simple value (known as a statistic) which serves as a "best estimate" of an unknown (fixed or random) population parameter.

Let $(X_1, X_2, \ldots.., X_n)$ be a random sample drawn from a population having distribution function $F_\theta, \theta \in \Theta$. Where the functional form of F is known except the parameter θ. If we are to guess a specific feature of the parent distribution, it can be explicitly written as a function of θ.

Suppose we are to guess $\gamma(\theta)$, a real-valued function of θ. The statistic $T(X_1, X_2, \ldots.., X_n)$ is said to be an estimator of $\gamma(\theta)$, if we guess $\gamma(\theta)$ by $T(X_1, X_2, \ldots, X_n)$ given $(X_1, X_2, \ldots\ldots, X_n) = (x_1, x_2, \ldots\ldots, x_n)$, $T(X_1, X_2, \ldots, X_n)$ is said to be an estimate of $\gamma(\theta)$.

Interval Estimation: In statistics, interval estimation is the use of sample data to calculate an interval of possible (probable) values, of an unknown population parameter, in contrast to point estimation, which is a single number (estimation by unique estimate).

An interval estimate of a real-values parameter θ is any pair of functions, $L(x_1, x_2, \ldots, x_n)$ and $U(x_1, x_2, \ldots, x_n)$, of a sample that satisfy $L(x) \leq U(x)$ for all $x \in X$. If $X = x$ is observed, the inference $L(x) \leq \theta \leq U(x)$ is made. The random interval $[L(X), U(X)]$ is called an interval estimator.

The most prevalent forms of interval estimation are Confidence intervals (a frequentist method) and credible intervals (a Bayesian method). Other common approaches to interval estimation, which are encompassed by statistical theory, are Tolerance and prediction intervals (used mainly in Regression Analysis).

- *Credible intervals* can readily deal with prior information, while confidence intervals cannot.
- *Confidence intervals* are more flexible and can be used practically in more situations than credible intervals: one area where credible intervals suffer in comparison is in dealing with non-parametric models.

1 Parametric models and methods of estimation

In this chapter, we discuss the question: How to estimate the parameter(s) of given probability distribution?

A **parametric model** is a family of probability distributions that can be described by a finite number of parameters[2]. We've already seen many examples of parametric models:

- The family of normal distributions $\mathcal{N}(\mu, \sigma^2)$, with parameters μ and σ^2.
- The family of Bernoulli distributions Bernoulli(p), with a single parameter p.
- The family of Gamma distributions Gamma(α, β), with parameters α and β.

We will denote a general parametric model by $\{f(x \mid \theta) : \theta \in \Omega\}$, where $\theta \in \mathbb{R}^k$ represents k **parameters**, $\Omega \subseteq \mathbb{R}^k$ is the **parameter space** to which the parameters must belong, and $f(x \mid \theta)$ is the PDF or PMF for the distribution having parameters

[2]The number of parameters is fixed and cannot grow with the sample size

θ. For example, in the $\mathcal{N}\left(\mu, \sigma^2\right)$ model above, $\theta = \left(\mu, \sigma^2\right)$, $\Omega = \mathbb{R} \times \mathbb{R}_+$ where $\mathbb{R}_+$ is the set of positive real numbers, and

$$f(x \mid \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Given data $X_1, \ldots, X_n$, the question of which parametric model we choose to fit the data usually depends on what the data values represent (number of occurrences over a period of time? aggregation of many small effects?) as well as a visual examination of the shape of the data histogram. This question is discussed in the context of several examples in Rice Sections 8.2-8.3.

Our main question of interest in this unit will be the following: After specifying an appropriate parametric model $\{f(x \mid \theta) : \theta \in \Omega\}$, and given observations

$$X_1, \ldots, X_n \overset{IID}{\sim} f(x \mid \theta).$$

How can we estimate the unknown parameter θ and quantify the uncertainty in our estimate?

1.1 Method of moments

If θ is a single number, then a simple idea to estimate θ is to find the value of θ for which the theoretical mean of $X \sim f(x \mid \theta)$ equals the observed sample mean $\bar{X} = \frac{1}{n}(X_1 + \ldots + X_n)$.

Example 1.1. *The **Poisson distribution** with parameter $\lambda > 0$ is a discrete distribution over the non-negative integers $\{0, 1, 2, 3, \ldots\}$ having* PMF

$$f(x \mid \lambda) = \frac{e^{-\lambda}\lambda^x}{x!}.$$

If $X \sim$ Poisson(λ), then it has mean $\mathbb{E}[X] = \lambda$. Hence for data $X_1, \ldots, X_n \overset{IID}{\sim}$ Poisson(λ), a simple estimate of λ is the sample mean $\hat{\lambda} = \bar{X}$.

Example 1.2. *The **exponential distribution** with parameter $\lambda > 0$ is a continuous distribution over $\mathbb{R}_+$ having* PDF

$$f(x \mid \lambda) = \lambda e^{-\lambda x}.$$

If $X \sim$ Exponential(λ), then $\mathbb{E}[X] = \frac{1}{\lambda}$. Hence for data $X_1, \ldots, X_n \overset{IID}{\sim}$ Exponential(λ), we estimate λ by the value $\hat{\lambda}$ which satisfies $\frac{1}{\hat{\lambda}} = \bar{X}$, i.e. $\hat{\lambda} = \frac{1}{\bar{X}}$.

More generally, for $X \sim f(x \mid \theta)$ where θ contains k unknown parameters, we may consider the first k **moments** of the distribution of X, which are the values

$$\begin{aligned} \mu_1 &= \mathbb{E}[X] \\ \mu_2 &= \mathbb{E}\left[X^2\right] \\ &\vdots \\ \mu_k &= \mathbb{E}\left[X^k\right], \end{aligned}$$

and compute these moments in terms of θ. To estimate θ from data $X_1, \ldots, X_n$, we solve for the value of θ for which these moments equal the observed sample moments

$$\begin{aligned} \hat{\mu}_1 &= \frac{1}{n}\left(X_1 + \ldots + X_n\right) \\ &\vdots \\ \hat{\mu}_k &= \frac{1}{n}\left(X_1^k + \ldots + X_n^k\right). \end{aligned}$$

(This yields k equations in k unknown parameters.) The resulting estimate of θ is called the **method of moments estimator**.

Example 1.3. *Let $X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$. If $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$, then $\mathbb{E}[X] = \mu$ and $\mathbb{E}\left[X^2\right] = \mu^2 + \sigma^2$. So the method of moments estimators $\hat{\mu}$ and $\hat{\sigma}^2$ for μ and σ^2 solve the equations*

$$\begin{aligned} \hat{\mu} &= \hat{\mu}_1, \\ \hat{\sigma}^2 + \hat{\mu}^2 &= \hat{\mu}_2. \end{aligned}$$

The first equation yields $\hat{\mu} = \hat{\mu}_1 = \bar{X}$, and the second yields

$$\hat{\sigma}^2 = \hat{\mu}_2 - \hat{\mu}_1^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \bar{X}^2 = \frac{1}{n}\left(\sum_{i=1}^n X_i^2 - 2\sum_{i=1}^n X_i\bar{X} + n\bar{X}^2\right) = \frac{1}{n}\sum_{i=1}^n \left(X_i - \bar{X}\right)^2.$$

Example 1.4. *Let $X_1, \ldots, X_n \overset{IID}{\sim} \text{Gamma}(\alpha, \beta)$. If $X \sim \text{Gamma}(\alpha, \beta)$, then $\mathbb{E}[X] = \frac{\alpha}{\beta}$ and $\mathbb{E}\left[X^2\right] = \frac{\alpha+\alpha^2}{\beta^2}$. So the method of moments estimators $\hat{\alpha}, \hat{\beta}$ solve the equations*

$$\begin{aligned} \frac{\hat{\alpha}}{\hat{\beta}} &= \hat{\mu}_1, \\ \frac{\hat{\alpha} + \hat{\alpha}^2}{\hat{\beta}^2} &= \hat{\mu}_2. \end{aligned}$$

Substituting the first equation into the second,

$$\left(\frac{1}{\hat{\alpha}} + 1\right) \hat{\mu}_1^2 = \hat{\mu}_2,$$

so

$$\hat{\alpha} = \frac{1}{\frac{\hat{\mu}_2}{\hat{\mu}_1^2} - 1} = \frac{\hat{\mu}_1^2}{\hat{\mu}_2 - \hat{\mu}_1^2} = \frac{\bar{X}^2}{\frac{1}{n}\sum_{i=1}^n \left(X_i - \bar{X}\right)^2}.$$

The first equation then yields

$$\hat{\beta} = \frac{\hat{\alpha}}{\hat{\mu}_1} = \frac{\bar{X}}{\frac{1}{n}\sum_{i=1}^n \left(X_i - \bar{X}\right)^2}.$$

1.2 Bias, variance, and mean-squared-error

Consider the case of a single parameter $\theta \in \mathbb{R}$. Any estimator $\hat{\theta} := \hat{\theta}\left(X_1, \ldots, X_n\right)$ is a statistic - it has variability due to the randomness of the data $X_1, \ldots, X_n$ from which it is computed. Supposing that $X_1, \ldots, X_n \overset{\text{IID}}{\sim} f(x \mid \theta)$ (so the parametric model is correct and the true parameter is θ), we can think about whether $\hat{\theta}$ is a "good" estimate of the true parameter θ in a variety of different ways:

- The **bias** of $\hat{\theta}$ is $\mathbb{E}_\theta[\hat{\theta}] - \theta$. Here and below, $\mathbb{E}_\theta$ denotes the expectation with respect to $X_1, \ldots, X_n \overset{\text{IID}}{\sim} f(x \mid \theta)$
- The **standard error** of $\hat{\theta}$ is its standard deviation $\sqrt{\text{Var}_\theta[\hat{\theta}]}$. Here and below, Var_θ denotes the variance with respect to $X_1, \ldots, X_n \overset{IID}{\sim} f(x \mid \theta)$
- The **mean-squared-error (MSE)** of $\hat{\theta}$ is $\mathbb{E}_\theta\left[(\hat{\theta} - \theta)^2\right]$.

The bias measures how close the average value of $\hat{\theta}$ is to the true parameter θ; the standard error measures how variable is $\hat{\theta}$ around this average value. An estimator with small bias need not be an accurate estimator, if it has large standard error, and conversely an estimator with small standard error need not be accurate if it has large bias. The mean-squared-error encompasses both bias and variance: For any random variable X and any constant $c \in \mathbb{R}$,

$$\begin{aligned}
\mathbb{E}\left[(X - c)^2\right] &= \mathbb{E}\left[(X - \mathbb{E}X + \mathbb{E}X - c)^2\right] \\
&= \mathbb{E}\left[(X - \mathbb{E}X)^2\right] + \mathbb{E}[2(X - \mathbb{E}X)(\mathbb{E}X - c)] + \mathbb{E}\left[(\mathbb{E}X - c)^2\right] \\
&= \text{Var}[X] + 2(\mathbb{E}X - c)\mathbb{E}[X - \mathbb{E}X] + (\mathbb{E}X - c)^2 \\
&= \text{Var}[X] + (\mathbb{E}X - c)^2,
\end{aligned}$$

where we used that $\mathbb{E}X - c$ is a constant and $\mathbb{E}[X - \mathbb{E}X] = 0$. Applying this to $X = \hat{\theta}$ and $c = \theta$

$$\mathbb{E}_\theta\left[(\hat{\theta} - \theta)^2\right] = \mathrm{Var}[\hat{\theta}] + \left(\mathbb{E}_\theta[\hat{\theta}] - \theta\right)^2$$

We obtain the **bias-variance decomposition** of mean-squared-error:

$$\text{Mean-squared-error} \;=\; \text{Variance} \;+\; \text{Bias}^2.$$

An important remark is that the bias, standard error, and MSE may depend on the true parameter θ and take different values for different θ. We say that $\hat{\theta}$ is **unbiased** for θ if $\mathbb{E}_\theta[\hat{\theta}] = \theta$ for all $\theta \in \Omega$

Example 1.5. *In the model* $X_1, \ldots, X_n \overset{IID}{\sim} \mathrm{Poisson}(\lambda)$*, the method-of-moments estimator of* λ *was* $\hat{\lambda} = \bar{X}$*. Then*

$$\mathbb{E}_\lambda[\hat{\lambda}] = \mathbb{E}_\lambda[\bar{X}] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_\lambda\left[X_i\right] = \lambda$$

where the last equality uses $\mathbb{E}[X] = \lambda$ *if* $X \sim \mathrm{Poisson}(\lambda)$*. So* $\mathbb{E}_\lambda[\hat{\lambda}] - \lambda = 0$ *for all* $\lambda > 0$*, and* $\hat{\lambda}$ *is an unbiased estimator of* λ*. Also,*

$$\mathrm{Var}_\lambda[\hat{\lambda}] = \mathrm{Var}_\lambda[\bar{X}] = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}_\lambda\left[X_i\right] = \frac{\lambda}{n}$$

where we have used that $X_1, \ldots, X_n$ *are independent and* $\mathrm{Var}[X] = \lambda$ *if* $X \sim \mathrm{Poisson}(\lambda)$*. Hence the standard error of* $\hat{\lambda}$ *is* $\sqrt{\lambda/n}$*, and the MSE is* λ/n*. Note that both of these depend on* λ*-they are larger when* λ *is larger.*

As we do not know λ, in practice to determine the variability of $\hat{\lambda}$ we may estimate the standard error by $\sqrt{\hat{\lambda}/n} = \sqrt{\bar{X}/n}$. For large n, this is justified by the fact that $\hat{\lambda}$ is unbiased with standard error of the order $1/\sqrt{n}$, so we expect $\hat{\lambda} - \lambda$ to be of this order. Hence the estimated standard error $\sqrt{\hat{\lambda}/n}$ should be very close to the true standard error $\sqrt{\lambda/n}$. (We expect the difference between $\sqrt{\lambda/n}$ and $\sqrt{\hat{\lambda}/n}$ to be of the smaller order $1/n$.)

Remark 1.1. ***A nonsense unbiased estimator.*** *Let* X *be a Poisson random variable with mean* $\lambda > 0$*. Recall that the pmf of* X *is given by*

$$p(x;\lambda) = P(X = x;\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}.$$

Suppose that we are interested in estimating the parameter $\theta = e^{-3\lambda}$ based on a sample of size one. Let $T(X) = (-2)^X$. Then, the expectation is

$$\begin{aligned} E(T) &= \sum_{x=0}^{\infty} (-2)^x \frac{\lambda^x e^{-\lambda}}{x!} \\ &= e^{-\lambda} \sum_{x=0}^{\infty} \frac{(-2\lambda)^x}{x!} \\ &= e^{-\lambda} e^{-2\lambda} = e^{-3\lambda}. \end{aligned}$$

Therefore, T is unbiased for $e^{-3\lambda}$. However, T is unreasonable in the sense that it may be negative (for X odd), even though it is an estimator of a strictly positive quantity. This suggests that one should not automatically assume that a unique unbiased estimator is necessarily good.

Example 1.6. *In the model $X_1, \ldots, X_n \overset{IID}{\sim}$ Exponential(λ), the method-of-moments estimator of λ was $\hat{\lambda} = 1/\bar{X}$. This estimator is biased: Recall Jensen's inequality, which says for any strictly convex function $g : \mathbb{R} \to \mathbb{R}$, $\mathbb{E}[g(X)] > g(\mathbb{E}[X])$. The function $x \mapsto 1/x$ is strictly convex, so*

$$\mathbb{E}_\lambda[\hat{\lambda}] = \mathbb{E}_\lambda[1/\bar{X}] > 1/\mathbb{E}_\lambda[\bar{X}] = 1/(1/\lambda) = \lambda,$$

where we used $\mathbb{E}_\lambda[\bar{X}] = \mathbb{E}_\lambda[X_1] = 1/\lambda$ when $X_1, \ldots, X_n \overset{IID}{\sim}$ Exponential(λ). So $\mathbb{E}_\lambda[\hat{\lambda}] - \lambda > 0$ for all $\lambda > 0$, meaning $\hat{\lambda}$ always has positive bias.

To compute exactly the bias, variance, and MSE of $\hat{\lambda}$, note that Exponential (λ) is the same distribution as Gamma$(1, \lambda)$. Then $\bar{X} = \frac{1}{n}(X_1 + \ldots + X_n) \sim$ Gamma$(n, n\lambda)$. (This may be shown by calculating the MGF of $\bar{X}$.) The distribution of $\hat{\lambda} = 1/\bar{X}$ is called the Inverse-Gamma $(n, n\lambda)$ distribution, which has mean $\frac{\lambda n}{n-1}$ and variance $\frac{\lambda^2 n^2}{(n-1)^2(n-2)}$ for $n \geq 3$. So the bias, variance, and MSE are given by

$$\begin{aligned} \text{Bias} &= \mathbb{E}_\lambda[\hat{\lambda}] - \lambda = \frac{\lambda n}{n-1} - \lambda = \frac{\lambda}{n-1}, \\ \text{Variance } &= \text{Var}_\lambda[\hat{\lambda}] = \frac{\lambda^2 n^2}{(n-1)^2(n-2)}, \\ \text{MSE} &= \frac{\lambda^2 n^2}{(n-1)^2(n-2)} + \left(\frac{\lambda}{n-1}\right)^2 = \frac{\lambda^2(n+2)}{(n-1)(n-2)}. \end{aligned}$$

Till now, we introduced the method of moments for estimating one or more parameters θ in a parametric model. In the next section, we discuss a different method called maximum likelihood estimation. The focus of the next section will be on how to compute this estimate; subsequent sections will study its statistical properties.

1.3 Maximum likelihood estimation

Consider data $X_1, \ldots, X_n \overset{\text{IID}}{\sim} f(x \mid \theta)$, for a parametric model $\{f(x \mid \theta) : \theta \in \Omega\}$. The "$x$ given θ" implies that given a particular value of θ, $f(\cdot \mid \theta)$ defines a density. The parameter θ can be a vector of parameters. Suppose we *simulate data* from F or collect some *real data* (say, waiting times in a popular restaurant), and we want to answer the following questions:

1. How to estimate θ?

2. How to construct confidence intervals around the estimator of θ?

A useful method of estimating θ is the method of maximum likelihood estimation. Given the observed data $\mathbf{X} = (X_1, \ldots, X_n)$, we define a function $L(\theta \mid \mathbf{X})$ (sometimes written as $\text{lik}(\theta)$)

$$L(\theta \mid \mathbf{X}) = \prod_{i=1}^{n} f\left(X_i \mid \theta\right) \quad \text{or} \quad \text{lik}(\theta) = f\left(X_1 \mid \theta\right) \times \ldots \times f\left(X_n \mid \theta\right)$$

of the parameter θ is called the **likelihood function**. If $f(x \mid \theta)$ is the PMF of a discrete distribution, then $\text{lik}(\theta)$ is simply the probability of observing the values $X_1, \ldots, X_n$ if the true parameter were θ. The **maximum likelihood estimator (MLE)** of θ is the value of $\theta \in \Omega$ that maximizes $\text{lik}(\theta)$ or $L(\theta \mid \mathbf{X})$. Intuitively, it is the value of θ that makes the observed data "most probable" or "most likely". The likelihood function measures "how likely is a particular value of θ given the data observed" and then finds the θ that maximizes this likelihood.

It is important to note that $L(\theta \mid \mathbf{X})$ is not a distribution over θ. The "most likely" value is the value that maximizes the likelihood

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta \in \Theta} L(\theta \mid \mathbf{X}).$$

Computing the MLE is an optimization problem. Maximizing $\text{lik}(\theta)$ or $L(\theta \mid \mathbf{X})$ is equivalent to maximizing its (natural) logarithm

$$l(\theta) = \log(\text{lik}(\theta)) = \sum_{i=1}^{n} \log f\left(X_i \mid \theta\right)$$

which in many examples is easier to work with as it involves a sum rather than a product. Let's work through several examples:

Example 1.7. *(Bernoulli). Let* $X_1, \ldots, X_n \overset{IID}{\sim} \text{Bernoulli}(p)$. *Then the likelihood is*

$$\begin{aligned} L(p \mid \mathbf{x}) &= \prod_{i=1}^{n} p\left(x_i \mid p\right) \\ &= \prod_{i=1}^{n} \left[p^{x_i}(1-p)^{1-x_i}\right] \\ &= p^{\sum x_i}(1-p)^{n-\sum x_i}. \end{aligned}$$

To obtain the MLE of p, *we will maximize the likelihood. Note that maximizing the likelihood is the same as maximizing the log of the likelihood, but the calculations are easier after taking a log. So we take a log:*

$$\begin{aligned} \Rightarrow l(p) := \log L(p \mid \mathbf{x}) &= \left(\sum_i^n x_i\right) \log p + \left(n - \sum_i^n x_i\right) \log(1-p) \\ \frac{dl(p)}{dp} &= \frac{\sum x_i}{p} - \frac{n - \sum x_i}{1-p} \overset{set}{=} 0 \\ \Rightarrow \hat{p} &= \frac{1}{n}\sum_{t=1}^{n} x_i. \end{aligned}$$

Verify for yourself that the second derivative is negative for this $\hat{p}$. *Thus,*

$$\hat{p}_{\text{MLE}} = \frac{1}{n}\sum_{t=1}^{n} x_i.$$

Example 1.8. *(Poisson). Let* $X_1, \ldots, X_n \overset{IID}{\sim} \text{Poisson}(\lambda)$. *Then*

$$\begin{aligned} l(\lambda) &= \sum_{i=1}^{n} \log \frac{\lambda^{X_i} e^{-\lambda}}{X_i!} \\ &= \sum_{i=1}^{n} \left(X_i \log \lambda - \lambda - \log\left(X_i!\right)\right) \\ &= \left(\log \lambda\right) \sum_{i=1}^{n} X_i - n\lambda - \sum_{i=1}^{n} \log\left(X_i!\right). \end{aligned}$$

This is differentiable in λ, *so we maximize* $l(\lambda)$ *by setting its first derivative equal to 0:*

$$0 = l'(\lambda) = \frac{1}{\lambda}\sum_{i=1}^{n} X_i - n.$$

Solving for λ yields the estimate $\hat{\lambda} = \bar{X}$. Since $l(\lambda) \to -\infty$ as $\lambda \to 0$ or $\lambda \to \infty$, and since $\hat{\lambda} = \bar{X}$ is the unique value for which $0 = l'(\lambda)$, this must be the maximum of l. In this example, $\hat{\lambda}$ is the same as the method-of-moments estimate.

Example 1.9. *(Normal). Let $X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$. Then*

$$\begin{aligned} l\left(\mu, \sigma^2\right) &= \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left(X_i-\mu\right)^2}{2\sigma^2}}\right) \\ &= \sum_{i=1}^{n} \left(-\frac{1}{2}\log\left(2\pi\sigma^2\right) - \frac{\left(X_i-\mu\right)^2}{2\sigma^2}\right) \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(X_i-\mu\right)^2. \end{aligned}$$

Considering σ^2 (rather than σ) as the parameter, we maximize $l(\lambda)$ by settings its partial derivatives with respect to μ and σ^2 equal to 0 :

$$\begin{aligned} 0 &= \frac{\partial l}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(X_i-\mu\right), \\ 0 &= \frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}\left(X_i-\mu\right)^2. \end{aligned}$$

Solving the first equation yields $\hat{\mu} = \bar{X}$, and substituting this into the second equation yields $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$. Since $l\left(\mu, \sigma^2\right) \to -\infty$ as $\mu \to -\infty$, $\mu \to \infty$, $\sigma^2 \to 0$, or $\sigma^2 \to \infty$, and as $\left(\hat{\mu}, \hat{\sigma}^2\right)$ is the unique value for which $0 = \frac{\partial l}{\partial \mu}$ and $0 = \frac{\partial l}{\partial \sigma^2}$, this must be the maximum of l. Again, the MLEs are the same as the method-of-moments estimates.

Example 1.10. *(Two parameter exponential). The density of a two parameter exponential distribution is*

$$f(x \mid \mu, \lambda) = \lambda e^{-\lambda(x-\mu)}; \quad x \geq \mu,\ \mu \in \mathbb{R},\ \lambda > 0.$$

We want to compute the MLEs of both λ and μ. The likelihood is

$$\begin{aligned} L(\lambda, \mu \mid \mathbf{x}) &= \prod_{t=1}^{n} f\left(x_i \mid \mu, \lambda\right) \\ &= \prod_{t=1}^{n} \lambda e^{-\lambda(x_i - \mu)} I\left(x_i \geq \mu\right) \\ &= \lambda^n \exp\left\{-\lambda\left(\sum_i x_i - n\mu\right)\right\} I\left(x_i \geq \mu\right), \quad \forall\, \mu. \end{aligned}$$

But if $X_1, \ldots, X_n \geq \mu \Rightarrow \min\{X_i\} \geq \mu$. So

$$L(\lambda, \mu \mid \mathbf{x}) = \lambda^n \exp\left\{-\lambda\left(\sum_i x_i - n\mu\right)\right\} I\left(\min_i\{x_i\} \geq \mu\right) \quad \forall\, \mu.$$

We will first try to maximize with respect to μ and then with respect to λ. Note that $L(\lambda, \mu)$ is an increasing function of μ within the restriction. So that the MLE of μ is the largest value in the support of μ where $\mu \leq \min\{X_i\}$. So

$$\hat{\mu}_{\text{MLE}} = \min_{1 \leq i \leq n}\{X_i\} = X_{(1)}.$$

Next, note that

$$\begin{aligned} L\left(X_{(1)}, \lambda \mid \mathbf{x}\right) &= \lambda^n \exp\left\{-\lambda\left(\sum_i X_i - nX_{(1)}\right)\right\} \\ \Rightarrow l\left(X_{(1)}, \lambda\right) := \log L\left(X_{(1)}, \lambda \mid \mathbf{x}\right) &= n\log\lambda - \lambda\left(\sum X_i - nX_{(1)}\right) \\ \Rightarrow \frac{dl}{d\lambda} &= \frac{n}{\lambda} - \left(\sum X_i - nX_{(1)}\right) \overset{set}{=} 0 \quad \textit{and} \\ \frac{d^2 l}{d\lambda^2} &= -\frac{n}{\lambda^2} < 0. \end{aligned}$$

So, the function is concave, thus there is a unique maximum. Set

$$\begin{aligned} \frac{dl}{d\lambda} &= 0 \\ \Rightarrow \frac{n}{\lambda} &= \sum_{t=1}^{n} X_i - nX_{(1)} \\ \Rightarrow \hat{\lambda}_{\text{MLE}} &= \frac{n}{\sum X_i - nX_{(1)}}. \end{aligned}$$

Example 1.11. ***The tank problem.***

"The enemy" has an unknown number N of tanks, which he has obligingly numbered $1, 2, \ldots, N$. Spies have reported sighting 8 tanks with numbers

$$\mathbf{x} = (137, 24, 86, 33, 92, 129, 17, 111)^\top.$$

Assume that sightings are independent and that each of the N tanks has a probability $1/N$ of being observed at each sighting. What is the MLE of N?

Let X_i be the serial number of tank i. Then, the pmf of these variables is:

$$P(x; N) = \begin{cases} \frac{1}{N}, & \text{for } x \leq N, \\ 0, & \text{for } x > N. \end{cases}$$

Given that each tank has the same probability of being observed, and that the largest sample value is $x_{(8)} = 137$, it follows that the likelihood function of N is

$$L(N \mid \mathbf{x}) = P(\text{ Event } N) = \begin{cases} \frac{1}{N^8}, & \text{for } N \geq 137, \\ 0, & \text{for } N < 137. \end{cases}$$

it is straightforward to see that the likelihood function is maximised at

$$\widehat{N} = \max_{i=1,\ldots,8} x_i = 137.$$

Q: *Would you trust this estimate?*

1.3.1 Why MLE?

One main reason of using MLE is that often (not always), the resulting estimators are consistent and asymptotically normal. That is, for a general likelihood $L(\theta \mid x)$ and n being the size of the data:

$$\hat{\theta}_{\text{MLE}} \xrightarrow{p} \theta \text{ as } n \to \infty$$

and under some additional conditions, we also have

$$\sqrt{n}\left(\hat{\theta}_{\text{MLE}} - \theta\right) \xrightarrow{d} N\left(0, \Sigma^*\right),$$

where Σ^* is an estimable matrix called the inverse Fisher information matrix. So if we use MLE estimation (and after verifying certain conditions), we know that we can construct confidence intervals around $\hat{\theta}_{\text{MLE}}$. The conditions required for consistency and asymptotic normality are important (to be discussed in Sec. 1.3.3).

1.3.2 No closed-form MLEs

In Inference, obtaining MLE estimates for a problem requires maximizing the likelihood. However, it is possible that no analytical form of the maxima is possible! This is a common challenge in many models and estimation problems, and requires sophisticated optimization tools. We will give examples in which we cannot get an analytical form of the MLE.

Example 1.12. *(Gamma). Let* $X_1, \ldots, X_n \overset{IID}{\sim} \text{Gamma}(\alpha, \beta)$. *Then*

$$\begin{aligned} l(\alpha, \beta) &= \sum_{i=1}^{n} \log \left(\frac{\beta^\alpha}{\Gamma(\alpha)} X_i^{\alpha-1} e^{-\beta X_i} \right) \\ &= \sum_{i=1}^{n} \left(\alpha \log \beta - \log \Gamma(\alpha) + (\alpha - 1) \log X_i - \beta X_i \right) \\ &= n\alpha \log \beta - n \log \Gamma(\alpha) + (\alpha - 1) \sum_{i=1}^{n} \log X_i - \beta \sum_{i=1}^{n} X_i. \end{aligned}$$

To maximize $l(\alpha, \beta)$, *we set its partial derivatives equal to 0:*

$$\begin{aligned} 0 = \frac{\partial l}{\partial \alpha} &= n \log \beta - \frac{n\Gamma'(\alpha)}{\Gamma(\alpha)} + \sum_{i=1}^{n} \log X_i; \\ 0 = \frac{\partial l}{\partial \beta} &= \frac{n\alpha}{\beta} - \sum_{i=1}^{n} X_i. \end{aligned}$$

The second equation implies that the MLEs $\hat{\alpha}$ *and* $\hat{\beta}$ *satisfy* $\hat{\beta} = \hat{\alpha}/\bar{X}$. *Substituting into the first equation and dividing by* $n, \hat{\alpha}$ *satisfies*

$$0 = \log \hat{\alpha} - \frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})} - \log \bar{X} + \frac{1}{n} \sum_{i=1}^{n} \log X_i. \tag{1}$$

The function $f(\alpha) = \log \alpha - \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$ *decreases from* ∞ *to 0 as* α *increases from 0 to* ∞, *and the value* $-\log \bar{X} + \frac{1}{n} \sum_{i=1}^{n} \log X_i$ *is always negative*

(by Jensen's inequality, $\mathbb{E}[g(X)] \geq g[\mathbb{E}(X)]$*)*

– hence (equation 1) always has a single unique root $\hat{\alpha}$, *which is the MLE for* α. *The MLE for* β *is then* $\hat{\beta} = \hat{\alpha}/\bar{X}$.

Unfortunately, there is no closed-form expression for this root $\hat{\alpha}$. *(In particular, the MLE* $\hat{\alpha}$ *is not the method-of-moments estimator for* α.*) We may compute the root*

numerically using the Newton-Raphson method: We start with an initial guess $\alpha^{(0)}$, which (for example) may be the method-of-moments estimator

$$\alpha^{(0)} = \frac{\bar{X}^2}{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}.$$

Having computed $\alpha^{(t)}$ for any $t = 0, 1, 2, \ldots$, we compute the next iteration $\alpha^{(t+1)}$ by approximating the equation 1 with a linear equation using a first-order Taylor expansion around $\hat{\alpha} = \alpha^{(t)}$, and set $\alpha^{(t+1)}$ as the value of $\hat{\alpha}$ that solves this linear equation. In detail, let $f(\alpha) = \log\alpha - \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$. A first-order Taylor expansion around $\hat{\alpha} = \alpha^{(t)}$ in (equation 1) yields the linear approximation

$$0 \approx f\left(\alpha^{(t)}\right) + \left(\hat{\alpha} - \alpha^{(t)}\right) f'\left(\alpha^{(t)}\right) - \log\bar{X} + \frac{1}{n}\sum_{i=1}^{n}\log X_i,$$

and we set $\alpha^{(t+1)}$ to be the value of $\hat{\alpha}$ solving this linear equation, i.e. [3]

$$\alpha^{(t+1)} = \alpha^{(t)} + \frac{-f\left(\alpha^{(t)}\right) + \log\bar{X} - \frac{1}{n}\sum_{i=1}^{n}\log X_i}{f'\left(\alpha^{(t)}\right)}.$$

The iterations $\alpha^{(0)}, \alpha^{(1)}, \alpha^{(2)}, \ldots$ converge to the MLE $\hat{\alpha}$.

Example 1.13. *Let $(X_1, \ldots, X_k) \sim \text{Multinomial}\,(n, (p_1, \ldots, p_k))$. (This is not quite the setting of n IID observations from a parametric model, as we have been considering, although you can think of $(X_1, \ldots, X_k)$ as a summary of n such observations $Y_1, \ldots, Y_n$ from the parametric model Multinomial$(1, (p_1, \ldots, p_k))$, where Y_i indicates which of k possible outcomes occurred for the i th observation.) The log-likelihood is given by*

$$l\,(p_1, \ldots, p_k) = \log\left(\binom{n}{X_1, \ldots, X_k} p_1^{X_1} \ldots p_k^{X_k}\right) = \log\binom{n}{X_1, \ldots, X_k} + \sum_{i=1}^{k} X_i \log p_i,$$

and the parameter space is

$$\Omega = \{(p_1, \ldots, p_k) : 0 \leq p_i \leq 1 \text{ for all } i \text{ and } p_1 + \ldots + p_k = 1\}.$$

*To maximize $l\,(p_1, \ldots, p_k)$ subject to the linear constraint $p_1 + \ldots + p_k = 1$, we may use the method of **Lagrange multipliers**: Consider the Lagrangian*

[3] If this update yields $\alpha^{(t+1)} \leq 0$, we may reset $\alpha^{(t+1)}$ to be a very small positive value.

$$L(p_1, \ldots, p_k, \lambda) = \log \binom{n}{X_1, \ldots, X_k} + \sum_{i=1}^{k} X_i \log p_i + \lambda (p_1 + \ldots + p_k - 1),$$

for a constant λ to be chosen later. Clearly, subject to $p_1 + \ldots + p_k = 1$, maximizing $l(p_1, \ldots, p_k)$ is the same as maximizing $L(p_1, \ldots, p_k, \lambda)$. Ignoring momentarily the constraint $p_1 + \ldots + p_k = 1$, the unconstrained maximizer of L is obtained by setting for each $i = 1, \ldots, k$

$$0 = \frac{\partial L}{\partial p_i} = \frac{X_i}{p_i} + \lambda,$$

which yields $\hat{p}_i = -X_i/\lambda$. For the specific choice of constant $\lambda = -n$, we obtain $\hat{p}_i = X_i/n$ and $\sum_{i=1}^{n} \hat{p}_i = \sum_{i=1}^{n} X_i/n = 1$, so the constraint is satisfied. As $\hat{p}_i = X_i/n$ is the unconstrained maximizer of $L(p_1, \ldots, p_k, -n)$, this implies that it must also be the constrained maximizer of $L(p_1, \ldots, p_k, -n)$, so it is the constrained maximizer of $l(p_1, \ldots, p_k)$. So the MLE is given by $\hat{p}_i = X_i/n$ for $i = 1, \ldots, k$.

Recall the Taylor expansion:

$$f(x) = f(a) + f'(a)(x-a) \overset{\text{set}}{=} 0 \Rightarrow f'(a)(x-a) = -f(a) \Rightarrow x - a = \frac{-f(a)}{f'(a)} \Rightarrow x = a - \frac{f(a)}{f'(a)}$$

1.3.3 Consistency and asymptotic normality of the MLE

We showed in the last section that given data $X_1, \ldots, X_n \overset{IID}{\sim} \text{Poisson}(\lambda)$, the maximum likelihood estimator for λ is simply $\hat{\lambda} = \bar{X}$. How accurate is $\hat{\lambda}$ for λ ? Recall from Section 1.1 the following computations:

$$\mathbb{E}_\lambda[\bar{X}] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[X_i] = \lambda$$

$$\text{Var}_\lambda[\bar{X}] = \frac{1}{n^2} \sum_{i=1}^{n} \text{Var}[X_i] = \frac{\lambda}{n}.$$

So $\hat{\lambda}$ is unbiased, with variance λ/n.

When n is large, asymptotic theory provides us with a more complete picture of the "accuracy" of $\hat{\lambda}$: By the Law of Large Numbers, $\bar{X}$ converges to λ in probability as $n \to \infty$. Furthermore, by the Central Limit Theorem,

$$\sqrt{n}(\bar{X} - \lambda) \to \mathcal{N}\left(0, \text{Var}\left[X_i\right]\right) = \mathcal{N}(0, \lambda)$$

in distribution as $n \to \infty$. So for large n, we expect $\hat{\lambda}$ to be close to λ, and the sampling distribution of $\hat{\lambda}$ is approximately $\mathcal{N}\left(\lambda, \frac{\lambda}{n}\right)$. This normal approximation is useful for many reasons - for example, it allows us to understand other measures of error (such as $\mathbb{E}[|\hat{\lambda} - \lambda|]$ or $\mathbb{P}[|\hat{\lambda} - \lambda| > 0.01]$), and will allow us to obtain a confidence interval for $\hat{\lambda}$. In a parametric model, we say that an estimator $\hat{\theta}$ based on $X_1, \ldots, X_n$ is **consistent** if $\hat{\theta} \to \theta$ in probability as $n \to \infty$. We say that it is **asymptotically normal** if $\sqrt{n}(\hat{\theta} - \theta)$ converges in distribution to a normal distribution (or a multivariate normal distribution, if θ has more than 1 parameter). So $\hat{\lambda}$ above is consistent and asymptotically normal.

The goal of this section is to explain why, rather than being a curiosity of this Poisson example, consistency and asymptotic normality of the MLE hold quite generally for many "typical" parametric models, and there is a general formula for its asymptotic variance. The following is one statement of such a result:

Theorem 1.1. *Let $\{f(x \mid \theta) : \theta \in \Omega\}$ be a parametric model, where $\theta \in \mathbb{R}$ is a single parameter. Let $X_1, \ldots, X_n \overset{IID}{\sim} f(x \mid \theta_0)$ for $\theta_0 \in \Omega$, and let $\hat{\theta}$ be the MLE based on $X_1, \ldots, X_n$. Suppose certain regularity conditions hold, including:*[a]

- *All PDFs/PMFs $f(x \mid \theta)$ in the model have the same support,*
- *θ_0 is an interior point (i.e., not on the boundary) of Ω,*
- *The log-likelihood $l(\theta)$ is differentiable in θ, and*
- *$\hat{\theta}$ is the unique value of $\theta \in \Omega$ that solves the equation $0 = l'(\theta)$.*

Then $\hat{\theta}$ is consistent and asymptotically normal, with $\sqrt{n}\left(\hat{\theta} - \theta_0\right) \to \mathcal{N}\left(0, \frac{1}{I(\theta_0)}\right)$ in distribution. Here, $I(\theta)$ is defined by the two equivalent expressions

$$I(\theta) := \text{Var}_\theta[z(X, \theta)] = -\mathbb{E}_\theta\left[z'(X, \theta)\right],$$

where Var_θ and $\mathbb{E}_\theta$ denote variance and expectation with respect to $X \sim f(x \mid \theta)$, and

$$z(x, \theta) = \frac{\partial}{\partial \theta} \log f(x \mid \theta), \quad z'(x, \theta) = \frac{\partial^2}{\partial \theta^2} \log f(x \mid \theta).$$

[a] Some technical conditions in addition to the ones stated are required to make this theorem rigorously true; these additional conditions will hold for the examples we discuss, and we won't worry about them in this class.

Here $z(x, \theta)$ is called the **score function**, and $I(\theta)$ is called the **Fisher information**. Heuristically for large n, the above theorem tells us the following about the MLE $\hat{\theta}$:

- $\hat{\theta}$ is asymptotically unbiased. More precisely, the bias of $\hat{\theta}$ is less than order $1/\sqrt{n}$. (Otherwise $\sqrt{n}\left(\hat{\theta} - \theta_0\right)$ should not converge to a distribution with mean 0.)
- The variance of $\hat{\theta}$ is approximately $\frac{1}{nI(\theta_0)}$. In particular, the standard error is of order $1/\sqrt{n}$, and the variance (rather than the squared bias) is the main contributing factor to the mean-squared-error of $\hat{\theta}$.
- If the true parameter is θ_0, the sampling distribution of $\hat{\theta}$ is approximately $\mathcal{N}\left(\theta_0, \frac{1}{nI(\theta_0)}\right)$.

Example 1.14. *Let's verify that this theorem is correct for the above Poisson example. There,*

$$\log f(x \mid \lambda) = \log \frac{\lambda^x e^{-\lambda}}{x!} = x \log \lambda - \lambda - \log(x!),$$

so the score function and its derivative are given by

$$z(x, \lambda) = \frac{\partial}{\partial \lambda} \log f(x \mid \lambda) = \frac{x}{\lambda} - 1, \quad z'(x, \lambda) = \frac{\partial^2}{\partial \lambda^2} \log f(x \mid \lambda) = -\frac{x}{\lambda^2}.$$

We may compute the Fisher information as

$$I(\lambda) = -\mathbb{E}_\lambda \left[z'(X, \lambda)\right] = \mathbb{E}_\lambda \left[\frac{X}{\lambda^2}\right] = \frac{1}{\lambda},$$

so $\sqrt{n}(\hat{\lambda} - \lambda) \to \mathcal{N}(0, \lambda)$ in distribution. This is the same result as what we obtained using a direct application of the CLT.

Proof Sketch of Theorem 1.1

Proof. We'll sketch heuristically the proof of Theorem 1.1, assuming $f(x \mid \theta)$ is the PDF of a continuous distribution. (The discrete case is analogous with integrals replaced by sums.)
To see why the MLE $\hat{\theta}$ is consistent, note that $\hat{\theta}$ is the value of θ which maximizes

$$\frac{1}{n} l(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log f\left(X_i \mid \theta\right).$$

Suppose the true parameter is θ_0, i.e. $X_1, \ldots, X_n \overset{IID}{\sim} f\left(x \mid \theta_0\right)$. Then for any $\theta \in \Omega$ (not necessarily θ_0), the Law of Large Numbers implies the convergence in probability

$$\frac{1}{n}\sum_{i=1}^{n}\log f\left(X_i \mid \theta\right) \to \mathbb{E}_{\theta_0}[\log f(X \mid \theta)].$$

Under suitable regularity conditions, this implies that the value of θ maximizing the left side, which is $\hat{\theta}$, converges in probability to the value of θ maximizing the right side, which we claim is θ_0. Indeed, for any $\theta \in \Omega$,

$$\mathbb{E}_{\theta_0}[\log f(X \mid \theta)] - \mathbb{E}_{\theta_0}\left[\log f\left(X \mid \theta_0\right)\right] = \mathbb{E}_{\theta_0}\left[\log \frac{f(X \mid \theta)}{f\left(X \mid \theta_0\right)}\right].$$

Noting that $x \mapsto \log x$ is concave, Jensen's inequality implies $\mathbb{E}[\log X] \leq \log \mathbb{E}[X]$ for any positive random variable X, so

$$\begin{aligned}\mathbb{E}_{\theta_0}\left[\log \frac{f(X \mid \theta)}{f\left(X \mid \theta_0\right)}\right] &\leq \log \mathbb{E}_{\theta_0}\left[\frac{f(X \mid \theta)}{f\left(X \mid \theta_0\right)}\right] \\ &= \log \int \frac{f(x \mid \theta)}{f\left(x \mid \theta_0\right)} f\left(x \mid \theta_0\right) dx = \log \int f(x \mid \theta) dx = 0.\end{aligned}$$

So $\theta \mapsto \mathbb{E}_{\theta_0}[\log f(X \mid \theta)]$ is maximized at $\theta = \theta_0$, which establishes consistency of $\hat{\theta}$. To show asymptotic normality, we first compute the mean and variance of the score:

Lemma 1.1. *(Properties of the score). For $\theta \in \Omega$,*

$$\mathbb{E}_\theta[z(X, \theta)] = 0, \quad \operatorname{Var}_\theta[z(X, \theta)] = -\mathbb{E}\left[z'(X, \theta)\right].$$

Proof. By the chain rule of differentiation,

$$z(x, \theta) f(x \mid \theta) = \left(\frac{\partial}{\partial \theta} \log f(x \mid \theta)\right) f(x \mid \theta) = \frac{\frac{\partial}{\partial \theta} f(x \mid \theta)}{f(x \mid \theta)} f(x \mid \theta) = \frac{\partial}{\partial \theta} f(x \mid \theta). \quad (2)$$

Then, since $\int f(x \mid \theta) dx = 1$,

$$\mathbb{E}_\theta[z(X, \theta)] = \int z(x, \theta) f(x \mid \theta) dx = \int \frac{\partial}{\partial \theta} f(x \mid \theta) dx = \frac{\partial}{\partial \theta} \int f(x \mid \theta) dx = 0.$$

Next, we differentiate this identity with respect to θ:

$$
\begin{aligned}
0 &= \frac{\partial}{\partial\theta}\mathbb{E}_\theta[z(X,\theta)] \\
&= \frac{\partial}{\partial\theta}\int z(x,\theta)f(x \mid \theta)dx \\
&= \int \left(z'(x,\theta)f(x \mid \theta) + z(x,\theta)\left(\frac{\partial}{\partial\theta}f(x \mid \theta)\right)\right)dx \\
&= \int \left(z'(x,\theta)f(x \mid \theta) + z(x,\theta)^2 f(x \mid \theta)\right)dx \\
&= \mathbb{E}_\theta\left[z'(X,\theta)\right] + \mathbb{E}_\theta\left[z(X,\theta)^2\right] \\
&= \mathbb{E}_\theta\left[z'(X,\theta)\right] + \mathrm{Var}_\theta[z(X,\theta)],
\end{aligned}
$$

where the fourth line above applies equation 2 and the last line uses $\mathbb{E}_\theta[z(X,\theta)] = 0$. □

Since $\hat{\theta}$ maximizes $l(\theta)$, we must have $0 = l'(\hat{\theta})$. Consistency of $\hat{\theta}$ ensures that (when n is large) $\hat{\theta}$ is close to θ_0 with high probability. This allows us to apply a first-order Taylor expansion to the equation $0 = l'(\hat{\theta})$ around $\hat{\theta} = \theta_0$:

$$
0 \approx l'\left(\theta_0\right) + \left(\hat{\theta} - \theta_0\right) l''\left(\theta_0\right),
$$

so

$$
\sqrt{n}\left(\hat{\theta} - \theta_0\right) \approx -\sqrt{n}\frac{l'\left(\theta_0\right)}{l''\left(\theta_0\right)} = -\frac{\frac{1}{\sqrt{n}}l'\left(\theta_0\right)}{\frac{1}{n}l''\left(\theta_0\right)}. \tag{3}
$$

For the denominator, by the Law of Large Numbers,

$$
\frac{1}{n}l''\left(\theta_0\right) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta^2}\left[\log f\left(X_i \mid \theta\right)\right]_{\theta=\theta_0} = \frac{1}{n}\sum_{i=1}^{n} z'\left(X_i,\theta_0\right) \to \mathbb{E}_{\theta_0}\left[z'\left(X,\theta_0\right)\right] = -I\left(\theta_0\right)
$$

in probability. For the numerator, recall by Lemma 1.1 that $z\left(X,\theta_0\right)$ has mean 0 and variance $I\left(\theta_0\right)$ when $X \sim f\left(x \mid \theta_0\right)$. Then by the Central Limit Theorem,

$$
\frac{1}{\sqrt{n}}l'\left(\theta_0\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\left[\log f\left(X_i \mid \theta\right)\right]_{\theta=\theta_0} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} z\left(X_i,\theta_0\right) \to \mathcal{N}\left(0, I\left(\theta_0\right)\right)
$$

in distribution. Applying these conclusions, the Continuous Mapping Theorem, and Slutsky's Lemma[4] to (equation 3)

$$
\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \to \frac{1}{I\left(\theta_0\right)}\mathcal{N}\left(0, I\left(\theta_0\right)\right) = \mathcal{N}\left(0, I\left(\theta_0\right)^{-1}\right)
$$

as desired. □

[4]Slutsky's Lemma says: If $X_n \to c$ in probability and $Y_n \to Y$ in distribution, then $X_n Y_n \to cY$ in distribution.

1.3.4 Applications of MLE in Regression Analysis

We start with the simple linear regression setup: Let $Y_1, Y_2, \ldots, Y_n$ be observations known as the response. Let $x_i = (x_{i1}, \ldots x_{ip})^T \in \mathbb{R}^p$ be the i th corresponding vector of covariates for the i th observation. Let $\beta \in \mathbb{R}^p$ be the regression coefficient so that for $\sigma^2 > 0$,

$$Y_i = x_i^T \beta + \epsilon_i \quad \text{where } \epsilon_i \sim N\left(0, \sigma^2\right).$$

Let $X = \left(x_1^T, x_2^T, \ldots, x_n^T\right)^T$. In vector form we have,

$$Y = X\beta + \epsilon \sim N_n\left(X\beta, \sigma^2 I_n\right).$$

The linear regression model is built to estimate β, which measures the linear effect of X on Y. We will show below how to use MLE in the computation of regression coefficients in linear regression model.

Example 1.15. *(MLE for Linear Regression).*

In order to understand the linear relationship between X and β, we will need to estimate β. We have

$$\begin{aligned} L\left(\beta, \sigma^2 \mid y\right) &= \prod_{t=1}^{n} f\left(y_i \mid X, \beta, \sigma^2\right) \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{1}{2}\frac{(Y - X\beta)^T(Y - X\beta)}{\sigma^2}\right\} \\ \Rightarrow l\left(\beta, \sigma^2\right) := \log L\left(\beta, \sigma^2 \mid y\right) &= -\frac{1}{2}\log(2\pi) - \frac{n}{2}\log\left(\sigma^2\right) - \frac{1}{2}\frac{(Y - X\beta)^T(Y - X\beta)}{\sigma^2}. \end{aligned}$$

Note that

$$\begin{aligned} (y - X\beta)^T(y - X\beta) &= \left(y^T - \beta^T X^T\right)(y - X\beta) \\ &= y^T y - y^T X\beta - \beta^T X^T y + \beta^T X^T X\beta \\ &= y^T y - 2\beta^T X^T y + \beta^T X^T X\beta. \end{aligned}$$

Using this we have

$$\begin{aligned} \frac{dl}{d\beta} &= -\frac{1}{2\sigma^2}\left[-2X^T y + 2X^T X\beta\right] = \frac{X^T y - X^T X\beta}{2\sigma^2} \overset{set}{=} 0 \\ \frac{dl}{d\sigma^2} &= -\frac{n}{2\sigma^2} + \frac{(y - X\beta)^T(y - X\beta)}{2\sigma^4} \overset{set}{=} 0. \end{aligned}$$

The first equation leads to $\hat{\beta}_{\text{MLE}}$ satisfying

$$X^T y - X^T X\hat{\beta}_{\text{MLE}} = 0 \Rightarrow \hat{\beta}_{\text{MLE}} = \left(X^T X\right)^{-1} X^T y, \quad \textit{if } \left(X^T X\right)^{-1} \textit{ exists.}$$

And $\hat{\sigma}^2_{MLE}$ *is*

$$\hat{\sigma}^2_{MLE} = \frac{\left(y - X\hat{\beta}_{\text{MLE}}\right)^T \left(y - X\hat{\beta}_{\text{MLE}}\right)}{n}.$$

Remark 1.2. *It can be easily verified that the second derivative is negative, and this is indeed the maximum. The next question is: What if* $\left(X^TX\right)^{-1}$ *does not exist? For example, if* $p > n$*, then the number of observations is less than the number of parameters, and since* X *is* $n \times p$*,* $\left(X^TX\right)$ *is* $p \times p$ *of rank* $n < p$*. So* X^TX *is not full rank and cannot be inverted. In this case, the MLE does not exist and other estimators need to be constructed. This is one of the motivations of penalized regression, which we will discuss next.*

Note that in the Linear regression setup, the MLE for β satisfies the following:

$$\hat{\beta}_{\text{MLE}} = \arg\min_{\beta} (y - X\beta)^T (y - X\beta).$$

Suppose X is such that $\left(X^TX\right)$ is not invertible, then the MLE does not exist, and we don't know how to estimate β. In such cases, we may use penalized likelihood that penalizes the coefficients β so that some of the βs are pushed towards zero. The corresponding Xs to those small βs are essentially not important, removing singularity from X^TX. The penalized likelihood is

$$\tilde{Q}(\beta) = L(\beta \mid y) + \tilde{P}(\beta),$$

where $P(\beta)$ is called the penalization function. Since the optimization of $L(\beta \mid y)$ only depends on $(y - X\beta)^T(y - X\beta)$ term, a penalized (negative) log-likelihood is used and the final penalized (negative) log-likelihood is

$$Q(\beta) = -\log L(\beta \mid y) + P(\beta).$$

There are many ways of penalizing β and each method yields a different estimator. A popular one is the ridge penalty.

Example 1.16. *(MLE for Ridge Regression).*

The ridge penalization term is $\lambda\beta^T\beta/2$ *for* $\lambda > 0$ *for*

$$Q(\beta) = \frac{(y - X\beta)^T(y - X\beta)}{2} + \frac{\lambda}{2}\beta^T\beta.$$

We will minimize $Q(\beta)$ *over the space of* β *and since we are adding a arbitrary term that depends on the size of* β*, smaller sizes of* β *will be preferred. Small sizes of* β

means X are less important, and this will eventually nullify the singularity in X^TX. The larger λ is, the more "penalization" there is for large values of β.

$$\hat{\beta} = \arg\min_{\beta} \left\{ \frac{(y - X\beta)^T (y - X\beta)}{2} + \frac{\lambda}{2} \beta^T \beta \right\}.$$

To carry out the minimization, we take the derivative:

$$\begin{aligned} \frac{dQ(\beta)}{d\beta} &= \frac{1}{2}\left(-2X^Ty + 2X^TX\beta\right) + \lambda\beta \overset{set}{=} 0 \\ &\Rightarrow \left(X^TX + \lambda I_p\right)\hat{\beta} - X^Ty = 0 \\ &\Rightarrow \hat{\beta}_{ridge} = \left(X^TX + \lambda I_p\right)^{-1} X^Ty \end{aligned}$$

(verify second derivative is positive for yourself). Note that $\left(X^TX + \lambda I_p\right)$ is always positive definite for $\lambda > 0$ since for any $a \in \mathbb{R}^p \neq 0$

$$a^T\left(X^TX + \lambda I_p\right)a = a^TX^TXa + \lambda a^Ta \geq 0.$$

Thus, the final ridge solution always exists even if X^TX is not invertible.

2 Fisher information and the Cramer-Rao bound

2.1 Fisher information for one or more parameters

For a parametric model $\{f(x \mid \theta) : \theta \in \Omega\}$ where $\theta \in \mathbb{R}$ is a single parameter, we showed in Section 1.3 that the MLE $\hat{\theta}_n$ based on $X_1, \ldots, X_n \overset{IID}{\sim} f(x \mid \theta)$ is, under certain regularity conditions, asymptotically normal:

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \to \mathcal{N}\left(0, \frac{1}{I(\theta)}\right)$$

in distribution as $n \to \infty$, where

$$I(\theta) := \mathrm{Var}_\theta\left[\frac{\partial}{\partial\theta}\log f(X \mid \theta)\right] = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(X \mid \theta)\right]$$

is the **Fisher information**. As an application of this result, let us study the sampling distribution of the MLE in a one-parameter Gamma model:

Example 2.1. *Let $X_1, \ldots, X_n \overset{IID}{\sim} \mathrm{Gamma}(\alpha, 1)$. (For this example, we are assuming that we know $\beta = 1$ and only need to estimate α.) Then*

$$\log f(x \mid \alpha) = \log\frac{1}{\Gamma(\alpha)}x^{\alpha-1}e^{-x} = -\log\Gamma(\alpha) + (\alpha - 1)\log x - x$$

The log-likelihood of all observations is then

$$\begin{aligned} l(\alpha) &= \sum_{i=1}^{n} \left(-\log\Gamma(\alpha) + (\alpha-1)\log X_i - X_i\right) \\ &= -n\log\Gamma(\alpha) + (\alpha-1)\sum_{i=1}^{n}\log X_i - \sum_{i=1}^{n} X_i. \end{aligned}$$

Introducing the digamma function $\psi(\alpha) = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$*, the MLE* $\hat{\alpha}$ *is obtained by (numerically) solving*

$$0 = l'(\alpha) = -n\psi(\alpha) + \sum_{i=1}^{n}\log X_i.$$

What is the sampling distribution of $\hat{\alpha}$*? We compute*

$$\frac{\partial^2}{\partial\alpha^2}\log f(x \mid \alpha) = -\psi'(\alpha).$$

As this does not depend on x*, the Fisher information is* $I(\alpha) = -\mathbb{E}_\alpha\left[-\psi'(\alpha)\right] = \psi'(\alpha)$*. Then for large* n*,* $\hat{\alpha}$ *is distributed approximately as* $\mathcal{N}\left(\alpha, \frac{1}{n\psi'(\alpha)}\right)$*.*

Asymptotic normality of the MLE extends naturally to the setting of multiple parameters:

Theorem 2.1. *Let* $\{f(x \mid \theta) : \theta \in \Omega\}$ *be a parametric model, where* $\theta \in \mathbb{R}^k$ *has* k *parameters. Let* $X_1, \ldots, X_n \overset{IID}{\sim} f(x \mid \theta)$ *for* $\theta \in \Omega$*, and let* $\hat{\theta}_n$ *be the MLE based on* $X_1, \ldots, X_n$*. Define the* ***Fisher information matrix*** $I(\theta) \in \mathbb{R}^{k\times k}$ *as the matrix whose* (i, j) *entry is given by the equivalent expressions*

$$I(\theta)_{ij} = \mathrm{Cov}_\theta\left[\frac{\partial}{\partial\theta_i}\log f(X \mid \theta), \frac{\partial}{\partial\theta_j}\log f(X \mid \theta)\right] = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(X \mid \theta)\right]. \tag{4}$$

Then under the same conditions as Theorem 1.1,

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \to \mathcal{N}\left(0, I(\theta)^{-1}\right),$$

where $I(\theta)^{-1}$ *is the* $k \times k$ *matrix inverse of* $I(\theta)$ *(and the distribution on the right is the multivariate normal distribution having this covariance).*

(For $k = 1$, this definition of $I(\theta)$ is exactly the same as our previous definition, and $I(\theta)^{-1}$ is just $\frac{1}{I(\theta)}$. The proof of the above result is analogous to the $k = 1$ case from previous section, employing a multivariate Taylor expansion of the equation $0 = \nabla l(\hat{\theta})$ around $\hat{\theta} = \theta_0$.)

Example 2.2. *Consider now the full Gamma model,* $X_1, \ldots, X_n \overset{IID}{\sim} \text{Gamma}(\alpha, \beta)$. *Numerical computation of the MLEs* $\hat{\alpha}$ *and* $\hat{\beta}$ *in this model was discussed in Section 1.3. To approximate their sampling distributions, note*

$$\log f(x \mid \alpha, \beta) = \log \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} = \alpha \log \beta - \log \Gamma(\alpha) + (\alpha - 1) \log x - \beta x,$$

so

$$\frac{\partial^2}{\partial \alpha^2} \log f(x \mid \alpha, \beta) = -\psi'(\alpha), \quad \frac{\partial^2}{\partial \alpha \partial \beta} \log f(x \mid \alpha, \beta) = \frac{1}{\beta}, \quad \frac{\partial^2}{\partial \beta^2} \log f(x \mid \alpha, \beta) = -\frac{\alpha}{\beta^2}.$$

These partial derivatives again do not depend on x*, so the Fisher information matrix is*

$$I(\alpha, \beta) = \begin{pmatrix} \psi'(\alpha) & -\frac{1}{\beta} \\ -\frac{1}{\beta} & \frac{\alpha}{\beta^2} \end{pmatrix},$$

and its inverse is

$$I(\alpha, \beta)^{-1} = \frac{1}{\psi'(\alpha)\frac{\alpha}{\beta^2} - \frac{1}{\beta^2}} \begin{pmatrix} \frac{\alpha}{\beta^2} & \frac{1}{\beta} \\ \frac{1}{\beta} & \psi'(\alpha) \end{pmatrix}.$$

$(\hat{\alpha}, \hat{\beta})$ *is approximately distributed as the bivariate normal distribution* $\mathcal{N}\left((\alpha, \beta), \frac{1}{n} I(\alpha, \beta)^{-1}\right)$. *In particular, the marginal distribution of* $\hat{\alpha}$ *is approximately*

$$\mathcal{N}\left(\alpha, \frac{1}{n\left(\psi'(\alpha)\frac{\alpha}{\beta^2} - \frac{1}{\beta^2}\right)} \frac{\alpha}{\beta^2}\right).$$

Suppose, in this example, that in fact the true parameter $\beta = 1$. Then the variance of $\hat{\alpha}$ reduces to $\frac{1}{n(\psi'(\alpha) - 1/\alpha)}$, which is not the variance $\frac{1}{n\psi'(\alpha)}$ obtained in example 2.1 – the variance here is larger. The difference is that in this example, we do not assume that we know $\beta = 1$ and instead are estimating β by its MLE $\hat{\beta}$. As a result, the MLEs of α in these two examples are not the same, and here our uncertainty about β is also increasing the variability of our estimate of α.

More generally, for any 2×2 Fisher information matrix

$$I = \begin{pmatrix} a & b \\ b & c \end{pmatrix},$$

the first definition of equation 4 implies that $a, c \geq 0$. The upper-left element of I^{-1} is $\frac{1}{a-b^2/c}$, which is always at least $\frac{1}{a}$. This implies, for any model with a single parameter θ_1 that is contained inside a larger model with parameters (θ_1, θ_2), that the variability of the MLE for θ_1 in the larger model is always at least that of the MLE for θ_1 in the smaller model; they are equal when the off-diagonal entry b is equal to 0. The same observation is true for any number of parameters $k \geq 2$ in the larger model.

This is a simple example of a trade-off between model complexity and accuracy of estimation, which is fundamental to many areas of statistics and machine learning: a complex model with more parameters might better capture the true distribution of data, but these parameters will also be more difficult to estimate than those in a simpler model.

2.2 The Cramer-Rao lower bound

Let's return to the setting of a single parameter $\theta \in \mathbb{R}$. Why is the Fisher information $I(\theta)$ called "information", and why should we choose to estimate θ by the MLE $\hat{\theta}$?

If $X_1, \ldots, X_n \overset{IID}{\sim} f(x \mid \theta_0)$ for a true parameter θ_0, and $l(\theta) = \sum_{i=1}^{n} \log f(X_i \mid \theta)$ is the log-likelihood function, then

$$I(\theta_0) = -\mathbb{E}_{\theta_0}\left[\frac{\partial^2}{\partial \theta^2}[\log f(X \mid \theta)]_{\theta=\theta_0}\right] = -\frac{1}{n}\mathbb{E}_{\theta_0}\left[l''(\theta_0)\right].$$

$I(\theta_0)$ measures the expected curvature of the log-likelihood function $l(\theta)$ around the true parameter $\theta = \theta_0$. If $l(\theta)$ is sharply curved around θ_0 - in other words, $I(\theta_0)$ is large - then a small change in θ can lead to a large decrease in the log-likelihood $l(\theta)$, and hence the data provides a lot of "information" that the true value of θ is close to θ_0. Conversely, if $I(\theta_0)$ is small, then a small change in θ does not affect $l(\theta)$ by much, and the data provides less information about θ. In this (heuristic) sense, $I(\theta_0)$ quantifies the amount of information that each observation X_i contains about the unknown parameter.

The Fisher information $I(\theta)$ is an intrinsic property of the model $\{f(x \mid \theta) : \theta \in \Omega\}$, not of any specific estimator. (We've shown that it is related to the variance of the MLE, but its definition does not involve the MLE.) There are various information-theoretic results stating that $I(\theta)$ describes a fundamental limit to how accurate any estimator of θ based on $X_1, \ldots, X_n$ can be. We'll prove one such result, called the **Cramer-Rao lower bound (CRLB)**:

Theorem 2.2. *Consider a parametric model $\{f(x \mid \theta) : \theta \in \Omega\}$ (satisfying certain mild regularity assumptions) where $\theta \in \mathbb{R}$ is a single parameter. Let T be any unbiased estimator of θ based on data $X_1, \ldots, X_n \overset{IID}{\sim} f(x \mid \theta)$. Then*

$$\mathrm{Var}_\theta[T] \geq \frac{1}{nI(\theta)}.$$

Proof. Recall the score function

$$z(x, \theta) = \frac{\partial}{\partial \theta} \log f(x \mid \theta) = \frac{\frac{\partial}{\partial \theta} f(x \mid \theta)}{f(x \mid \theta)},$$

and let $Z := Z(X_1, \ldots, X_n, \theta) = \sum_{i=1}^n z(X_i, \theta)$. By the definition of correlation and the fact that the correlation of two random variables is always between -1 and 1,

$$\mathrm{Cov}_\theta[Z, T]^2 \leq \mathrm{Var}_\theta[Z] \times \mathrm{Var}_\theta[T].$$

The random variables $z(X_1, \theta), \ldots, z(X_n, \theta)$ are IID, and by Lemma 1.1, they have mean 0 and variance $I(\theta)$. Then

$$\mathrm{Var}_\theta[Z] = n \,\mathrm{Var}_\theta[z(X_1, \theta)] = nI(\theta).$$

Since T is unbiased,

$$\theta = \mathbb{E}_\theta[T] = \int_{\mathbb{R}^n} T(x_1, \ldots, x_n) f(x_1 \mid \theta) \times \ldots \times f(x_n \mid \theta) \, dx_1 \ldots dx_n.$$

Differentiating both sides with respect to θ and applying the product rule of differentiation,

$$\begin{aligned} 1 = \int_{\mathbb{R}^n} T(x_1, \ldots, x_n) \Bigg(& \frac{\partial}{\partial \theta} f(x_1 \mid \theta) \times f(x_2 \mid \theta) \times \ldots \times f(x_n \mid \theta) \\ & + f(x_1 \mid \theta) \times \frac{\partial}{\partial \theta} f(x_2 \mid \theta) \times \ldots \times f(x_n \mid \theta) + \ldots \\ & + f(x_1 \mid \theta) \times f(x_2 \mid \theta) \times \ldots \times \frac{\partial}{\partial \theta} f(x_n \mid \theta) \Bigg) \, dx_1 \ldots dx_n \\ = \int_{\mathbb{R}^n} & T(x_1, \ldots, x_n) Z(x_1, \ldots, x_n, \theta) f(x_1 \mid \theta) \times \ldots \times f(x_n \mid \theta) \, dx_1 \ldots dx_n \\ = \mathbb{E}_\theta[TZ]. & \end{aligned}$$

Since $\mathbb{E}_\theta[Z] = 0$, this implies $\mathrm{Cov}_\theta[T, Z] = \mathbb{E}_\theta[TZ] = 1$, so $\mathrm{Var}_\theta[T] \geq \frac{1}{nI(\theta)}$ as desired. □

For two unbiased estimators of θ, the ratio of their variances is called their **relative efficiency**. An unbiased estimator is **efficient** if its variance equals the lower bound $\frac{1}{nI(\theta)}$. Since the MLE achieves this lower bound asymptotically, we say it is **asymptotically efficient**.

The Cramer-Rao bound ensures that no unbiased estimator can achieve asymptotically lower variance than the MLE. Stronger results, which we will not prove here, in fact, show that no estimator, biased or unbiased, can asymptotically achieve lower mean-squared-error than $\frac{1}{nI(\theta)}$, except possibly on a small set of special values $\theta \in \Omega$.[5] In particular, when the method-of-moments estimator differs from the MLE, we expect it to have higher mean-squared-error than the MLE for large n, which explains why the MLE is usually the preferred estimator in simple parametric models.

3 MLE under model misspecification

The eminent statistician George Box once said, *"All models are wrong, but some are useful."*

When we fit a parametric model to a set of data $X_1, \ldots, X_n$, we are usually not certain that the model is correct (for example, that the data truly have a normal or Gamma distribution). Rather, we think of the model as an approximation to what might be the true distribution of data. It is natural to ask, then, whether the MLE estimate $\hat{\theta}$ in a parametric model is at all meaningful, if the model itself is incorrect. The goal of this section is to explore this question and to discuss how the properties of $\hat{\theta}$ change under model misspecification.

3.1 MLE and the KL-divergence

Consider a parametric model $\{f(x \mid \theta) : \theta \in \Omega\}$. We'll assume throughout this section that $f(x \mid \theta)$ is the PDF of a continuous distribution, and $\theta \in \mathbb{R}$ is a single parameter.

Thus far, we have been measuring the error of an estimator $\hat{\theta}$ by its distance to the true parameter θ, via the bias, variance, and MSE. If $X_1, \ldots, X_n \overset{IID}{\sim} g$ for a PDF g that is not in the model, then there is no true parameter value θ associated to g. We will instead think about a measure of "distance" between two general PDFs:

[5] For example, the constant estimator $\hat{\theta} = c$ for fixed $c \in \Omega$ achieves 0 mean-squared-error if the true parameter happened to be the special value c, but at all other parameter values is worse than the MLE for sufficiently large n.

Definition 3.1. *For two PDFs f and g, the* ***Kullback-Leibler (KL) divergence*** *from f to g is*

$$D_{\mathrm{KL}}(g\|f) = \int g(x) \log \frac{g(x)}{f(x)} dx.$$

Equivalently, if $X \sim g$, then

$$D_{\mathrm{KL}}(g\|f) = \mathbb{E}\left[\log \frac{g(X)}{f(X)}\right].$$

D_{KL} has many information-theoretic interpretations and applications. For our purposes, we'll just note the following properties: If $f = g$, then $\log(g(x)/f(x)) \equiv 0$, so $D_{\mathrm{KL}}(g\|f) = 0$. By Jensen's inequality, since $x \mapsto -\log x$ is convex,

$$D_{\mathrm{KL}}(g\|f) = \mathbb{E}\left[-\log \frac{f(X)}{g(X)}\right] \geq -\log \mathbb{E}\left[\frac{f(X)}{g(X)}\right] = -\log \int \frac{f(x)}{g(x)} g(x) dx = 0.$$

Furthermore, since $x \mapsto -\log x$ is strictly convex, the inequality above can only be an equality if $f(X)/g(X)$ is a constant random-variable, so $f = g$. Thus, like an ordinary distance measure, $D_{\mathrm{KL}}(g\|f) \geq 0$ always, and $D_{\mathrm{KL}}(g\|f) = 0$ if and only if $f = g$.

Example 3.1. *To get an intuition for what the KL-divergence is measuring, let f and g be the PDFs of the distributions $\mathcal{N}\left(\mu_0, \sigma^2\right)$ and $\mathcal{N}\left(\mu_1, \sigma^2\right)$. Then*

$$\begin{aligned}
\log \frac{g(x)}{f(x)} &= \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu_1)^2}{2\sigma^2}} \Big/ \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu_0)^2}{2\sigma^2}}\right) \\
&= -\frac{(x-\mu_1)^2}{2\sigma^2} + \frac{(x-\mu_0)^2}{2\sigma^2} \\
&= \frac{2(\mu_1-\mu_0)x - \left(\mu_1^2 - \mu_0^2\right)}{2\sigma^2}.
\end{aligned}$$

So letting $X \sim g$,

$$\begin{aligned}
D_{\mathrm{KL}}(g\|f) = \mathbb{E}\left[\log \frac{g(X)}{f(X)}\right] &= \frac{1}{2\sigma^2}\left(2(\mu_1-\mu_0)\mathbb{E}[X] - \left(\mu_1^2 - \mu_0^2\right)\right) \\
&= \frac{1}{2\sigma^2}\left(2(\mu_1-\mu_0)\mu_1 - \left(\mu_1^2 - \mu_0^2\right)\right) = \frac{(\mu_1-\mu_0)^2}{2\sigma^2}.
\end{aligned}$$

Thus $D_{\mathrm{KL}}(g\|f)$ is proportional to the square of the mean difference normalized by the standard deviation σ. In this example we happen to have $D_{\mathrm{KL}}(f\|g) = D_{\mathrm{KL}}(g\|f)$, but

in general this is not true – for two arbitrary PDFs f and g, we may have $D_{\mathrm{KL}}(f\|g) \neq D_{\mathrm{KL}}(g\|f)$.

Suppose $X_1, \ldots, X_n \overset{IID}{\sim} g$, and consider a parametric model $\{f(x \mid \theta) : \theta \in \Omega\}$ which may or may not contain the true PDF g. The MLE $\hat{\theta}$ is the value of θ that maximizes

$$\frac{1}{n} l(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log f\left(X_i \mid \theta\right),$$

and this quantity by the Law of Large Numbers converges in probability to

$$\mathbb{E}_g[\log f(X \mid \theta)],$$

where $\mathbb{E}_g$ denotes expectation with respect to $X \sim g$. In Section 2, we showed that when $g(x) = f\left(x \mid \theta_0\right)$ (meaning g belongs to the parametric model, and the true parameter is θ_0), then $\mathbb{E}_g[\log f(X \mid \theta)]$ is maximized at $\theta = \theta_0$ - this explained consistency of the MLE. More generally, when g does not necessarily belong to the parametric model, we may write

$$\mathbb{E}_g[\log f(X \mid \theta)] = \mathbb{E}_g[\log g(X)] - \mathbb{E}_g\left[\log \frac{g(X)}{f(X \mid \theta)}\right] = \mathbb{E}_g[\log g(X)] - D_{\mathrm{KL}}(g\|f(x \mid \theta)).$$

The term $\mathbb{E}_g[\log g(X)]$ does not depend on θ, so the value of θ maximizing $\mathbb{E}_g[\log f(X \mid \theta)]$ is the value of θ that minimizes $D_{\mathrm{KL}}(g\|f(x \mid \theta)\ )$. This (heuristically) shows the following result:[6]

Theorem 3.1. *Let $X_1, \ldots, X_n \overset{IID}{\sim} g$ and suppose $D_{\mathrm{KL}}(g\|f(x \mid \theta))$ has a unique minimum at $\theta = \theta^*$. Then, under suitable regularity conditions on $\{f(x \mid \theta) : \theta \in \Omega\}$ and on g, the MLE $\hat{\theta}$ converges to θ^* in probability as $n \to \infty$.*

The density $f\left(x \mid \theta^*\right)$ may be interpreted as the "KL-projection" of g onto the parametric model $\{f(x \mid \theta) : \theta \in \Omega\}$. In other words, the MLE is estimating the distribution in our model that is closest, with respect to KL-divergence, to g.

3.2 The sandwich estimator of variance

When $X_1, \ldots, X_n \overset{\mathrm{IID}}{\sim} g$, how close is the MLE $\hat{\theta}$ to this KL-projection θ^*? Analogous to our proof in Section 2, we may answer this question by performing a Taylor expansion of the identity $0 = l'(\hat{\theta})$ around the point $\hat{\theta} = \theta^*$. This yields

[6] For a rigorous statement of necessary regularity conditions, see for example White, H. (1982). "Maximum likelihood estimation of misspecified models." Econometrica, `https://www.jstor.org/stable/1912526`.

$$0 \approx l'(\theta^*) + \left(\hat{\theta} - \theta^*\right) l''(\theta^*),$$

so

$$\sqrt{n}\left(\hat{\theta} - \theta^*\right) \approx -\frac{\frac{1}{\sqrt{n}} l'(\theta^*)}{\frac{1}{n} l''(\theta^*)}. \tag{5}$$

Recall the score function

$$z(x, \theta) = \frac{\partial}{\partial \theta} \log f(x \mid \theta).$$

The Law of Large Numbers applied to the denominator of equation 5 gives

$$\frac{1}{n} l''(\theta^*) = \frac{1}{n} \sum_{i=1}^{n} z'(X_i, \theta^*) \to \mathbb{E}_g\left[z'(X, \theta^*)\right]$$

in probability, while the Central Limit Theorem applied to the numerator of equation 5 gives

$$\frac{1}{\sqrt{n}} l'(\theta^*) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} z(X_i, \theta^*) \to \mathcal{N}\left(0, \mathrm{Var}_g\left[z(X, \theta^*)\right]\right)$$

in distribution. The quantity $z(X, \theta^*)$ has mean 0 when $X \sim g$ because θ^* maximizes $\mathbb{E}_g[\log f(X \mid \theta)]$, so differentiating with respect to θ yields

$$0 = \mathbb{E}_g\left[\frac{\partial}{\partial \theta}[\log f(X \mid \theta)]_{\theta=\theta^*}\right] = \mathbb{E}_g\left[z(X, \theta^*)\right].$$

Hence by Slutsky's lemma,

$$\sqrt{n}\left(\hat{\theta} - \theta^*\right) \to \mathcal{N}\left(0, \frac{\mathrm{Var}_g\left[z(X, \theta^*)\right]}{\mathbb{E}_g\left[z'(X, \theta^*)\right]^2}\right).$$

These are the same formulas as in Section 2 (with θ^* in place of θ_0), except expectations and variances are taken with respect to $X \sim g$ rather than $X \sim f(x \mid \theta^*)$. If $g(x) = f(x \mid \theta^*)$, meaning the model is correct, then $\mathrm{Var}_g\left[z(X, \theta^*)\right] = -\mathbb{E}_g\left[z'(X, \theta^*)\right] = I(\theta^*)$, and we recover our theorem from Section 2.However, when $g(x) \neq f(x \mid \theta^*)$, in general

$$\mathrm{Var}_g\left[z(X, \theta^*)\right] \neq -\mathbb{E}_g\left[z'(X, \theta^*)\right],$$

so we cannot simplify the variance of the above normal limit any further. We may instead estimate the individual quantities $\mathrm{Var}_g\left[z(X, \theta^*)\right]$ and $\mathbb{E}_g\left[z'(X, \theta^*)\right]$ using the sample variance of $z\left(X_i, \hat{\theta}\right)$ and the sample mean of $z'\left(X_i, \hat{\theta}\right)$ – this yields the **sandwich estimator** for the variance of the MLE.

Example 3.2. *Suppose we fit the model Exponential* (λ) *to data* $X_1, \ldots, X_n$ *by computing the MLE. The log-likelihood is*

$$l(\lambda) = \sum_{i=1}^{n} \log \lambda e^{-\lambda X_i} = n \log \lambda - \lambda \sum_{i=1}^{n} X_i$$

so the MLE solves the equation $0 = l'(\lambda) = n/\lambda - \sum_{i=1}^{n} X_i$*. This yields the MLE* $\hat{\lambda} = 1/\bar{X}$ *(which is the same as the method-of-moments estimator from Section 1.1).*

We may compute the sandwich estimate of the variance of $\hat{\lambda}$ *as follows: In the exponential model,*

$$z(x, \lambda) = \frac{\partial}{\partial \lambda} \log f(x \mid \lambda) = \frac{1}{\lambda} - x, \quad z'(x, \lambda) = \frac{\partial^2}{\partial \lambda^2} \log f(x \mid \lambda) = -\frac{1}{\lambda^2}.$$

Let $\bar{Z} = \frac{1}{n} \sum_{i=1}^{n} z\left(X_i, \hat{\lambda}\right) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{\hat{\lambda}} - X_i\right) = \frac{1}{\hat{\lambda}} - \bar{X}$ *be the sample mean of* $z\left(X_1, \hat{\lambda}\right), \ldots,$ $z\left(X_n, \hat{\lambda}\right)$*. We estimate* $\mathrm{Var}_g[z(X, \lambda)]$ *by the sample variance of* $z\left(X_1, \hat{\lambda}\right), \ldots, z\left(X_n, \hat{\lambda}\right)$*:*

$$\begin{aligned} \frac{1}{n-1} \sum_{i=1}^{n} \left(z\left(X_i, \hat{\lambda}\right) - \bar{Z}\right)^2 &= \frac{1}{n-1} \sum_{i=1}^{n} \left(\left(\frac{1}{\hat{\lambda}} - X_i\right) - \left(\frac{1}{\hat{\lambda}} - \bar{X}\right)\right)^2 \\ &= \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 = S_X^2. \end{aligned}$$

We estimate $\mathbb{E}_g\left[z'(X, \lambda)\right]$ *by the sample mean of* $z'\left(X_1, \hat{\lambda}\right), \ldots, z'\left(X_n, \hat{\lambda}\right)$*:*

$$\frac{1}{n} \sum_{i=1}^{n} z'\left(X_i, \hat{\lambda}\right) = \frac{1}{n} \sum_{i=1}^{n} -\frac{1}{\hat{\lambda}^2} = -\frac{1}{\hat{\lambda}^2}.$$

So the sandwich estimate of $\mathrm{Var}_g[z(X, \lambda)] / \mathbb{E}_g\left[z'(X, \lambda)\right]^2$ *is* $S_X^2 \hat{\lambda}^4 = S_X^2 / \bar{X}^4$*, and we may estimate the standard error of* $\hat{\lambda}$ *by* $S_X / \left(\bar{X}^2 \sqrt{n}\right)$*.*

4 Plugin estimators and the delta method

4.1 Estimating a function of θ

In the setting of a parametric model, we have been discussing how to estimate the parameter θ. We showed how to compute the MLE $\hat{\theta}$, derived its variance and sampling distribution for large n, and showed that no unbiased estimator can achieve variance

much smaller than that of the MLE for large n (the Cramer-Rao lower bound).

In many examples, the quantity we are interested in is not θ itself, but some value $g(\theta)$. The obvious way to estimate $g(\theta)$ is to use $g(\hat{\theta})$, where $\hat{\theta}$ is an estimate (say, the MLE) of θ. This is called the **plugin estimate** of $g(\theta)$, because we are just "plugging in" $\hat{\theta}$ for θ.

Example 4.1. *(Odds). You play a game with a friend, where you flip a biased coin. If the coin lands heads, you give your friend* \$1*. If the coin lands tails, your friend gives you* \$$x$*. What is the value of x that makes this a fair game?*

If the coin lands heads with probability p, then your expected winnings is $-p+(1-p)x$. The game is fair when $-p+(1-p)x = 0$, i.e. when $x = p/(1-p)$. This value $p/(1-p)$ is the odds of getting heads to getting tails. To estimate the odds from n coin flips

$$X_1, \ldots, X_n \overset{\text{IID}}{\sim} \text{Bernoulli}(p),$$

we may first estimate p by $\hat{p} = \bar{X}$. (This is both the method of moments estimator and the MLE.) Then the plugin estimate of $p/(1-p)$ is simply $\bar{X}/(1-\bar{X})$.

The odds falls in the interval $(0, \infty)$ and is not symmetric about $p = 1/2$. We oftentimes think instead in terms of the log-odds, $\log \frac{p}{1-p}$ - this can be any real number and is symmetric about $p = 1/2$. The plugin estimate for the log-odds is $\log \frac{\bar{X}}{1-\bar{X}}$.

Example 4.2. *(The Pareto mean). The Pareto (x_0, θ) distribution for $x_0 > 0$ and $\theta > 1$ is a continuous distribution over the interval $[x_0, \infty)$, given by the PDF*

$$f(x \mid x_0, \theta) = \begin{cases} \theta x_0^{\theta} x^{-\theta-1} & x \geq x_0 \\ 0 & x < x_0. \end{cases}$$

It is commonly used in economics as a model for the distribution of income. x_0 represents the minimum possible income; let's assume that x_0 is known and equal to 1 . We then have a one-parameter model with PDFs $f(x \mid \theta) = \theta x^{-\theta-1}$ supported on $[1, \infty)$.

The mean of the Pareto distribution is

$$\mathbb{E}_\theta[X] = \int_1^\infty x \cdot \theta x^{-\theta-1} dx = \theta \frac{x^{-\theta+1}}{-\theta+1}\Big|_1^\infty = \frac{\theta}{\theta-1},$$

so we might estimate the mean income by $\hat{\theta}/(\hat{\theta}-1)$ where $\hat{\theta}$ is the MLE. To compute $\hat{\theta}$ from observations $X_1, \ldots, X_n$, the log-likelihood is

$$l(\theta) = \sum_{i=1}^{n} \log\left(\theta X_i^{-\theta-1}\right) = \sum_{i=1}^{n} \left(\log\theta - (\theta+1)\log X_i\right) = n\log\theta - (\theta+1)\sum_{i=1}^{n}\log X_i.$$

Solving the equation

$$0 = l'(\theta) = \frac{n}{\theta} - \sum_{i=1}^{n}\log X_i$$

yields the MLE, $\hat{\theta} = n/\sum_{i=1}^{n}\log X_i$.

4.2 The delta method

We would like to be able to quantify our uncertainty about $g(\hat{\theta})$ using what we know about the uncertainty of $\hat{\theta}$ itself. When n is large, this may be done using a first-order Taylor approximation of g, formalized as the **delta method**:

Theorem 4.1. *(Delta method). If a function $g : \mathbb{R} \to \mathbb{R}$ is differentiable at θ_0 with $g'(\theta_0) \neq 0$, and if*

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \to \mathcal{N}\left(0, v\left(\theta_0\right)\right)$$

in distribution as $n \to \infty$ for some variance $v(\theta_0)$, then

$$\sqrt{n}\left(g(\hat{\theta}) - g\left(\theta_0\right)\right) \to \mathcal{N}\left(0, \left(g'\left(\theta_0\right)\right)^2 v\left(\theta_0\right)\right)$$

in distribution as $n \to \infty$.

Proof sketch. We perform a Taylor expansion of $g(\hat{\theta})$ around $\hat{\theta} = \theta_0$:

$$g(\hat{\theta}) \approx g\left(\theta_0\right) + \left(\hat{\theta} - \theta_0\right) g'\left(\theta_0\right).$$

Rearranging yields

$$\sqrt{n}\left(g(\hat{\theta}) - g\left(\theta_0\right)\right) \approx \sqrt{n}\left(\hat{\theta} - \theta_0\right) g'\left(\theta_0\right),$$

and multiplying a mean-zero normal variable by a constant c scales its variance by c^2. □

Example 4.3. *(Log-odds). Let $X_1, \ldots, X_n \overset{IID}{\sim}$ Bernoulli(p), and recall the plugin estimate of the log-odds $\log\frac{p}{1-p}$ given by $\log\frac{\bar{X}}{1-\bar{X}}$. By the Central Limit Theorem,*

$$\sqrt{n}(\bar{X} - p) \to \mathcal{N}(0, p(1-p))$$

in distribution, where $p(1-p)$ *is the variance of a* Bernoulli(p) *random variable. The function* $g(p) = \log \frac{p}{1-p} = \log p - \log(1-p)$ *has derivative*

$$g'(p) = \frac{1}{p} + \frac{1}{1-p} = \frac{1}{p(1-p)},$$

so by the delta method,

$$\sqrt{n}\left(\log \frac{\bar{X}}{1-\bar{X}} - \log \frac{p}{1-p}\right) \to \mathcal{N}\left(0, \frac{1}{p(1-p)}\right).$$

In other words, our estimate of the log-odds of heads to tails is approximately normally distributed around the true log-odds $\log \frac{p}{1-p}$*, with variance* $\frac{1}{np(1-p)}$*.*

Suppose we toss this biased coin $n = 100$ *times and observe 60 heads, i.e.* $\bar{X} = 0.6$*. We would estimate the log-odds by* $\log \frac{\bar{X}}{1-\bar{X}} \approx 0.41$*, and we may estimate our standard error by* $\sqrt{\frac{1}{n\bar{X}(1-\bar{X})}} \approx 0.20$*.*

Example 4.4. *(The Pareto mean). Let* $X_1, \ldots, X_n \overset{IID}{\sim}$ *Pareto* $(1, \theta)$*, and recall that the MLE for* θ *is* $\hat{\theta} = n / \sum_{i=1}^{n} \log X_i$*. We may use the maximum-likelihood theory developed in Section 1.3 to understand the distribution of* $\hat{\theta}$*: We compute (for* $x \geq 1$*) the following:*

$$\begin{aligned}
\log f(x \mid \theta) &= \log\left(\theta x^{-\theta-1}\right) = \log\theta - (\theta+1)\log x \\
\frac{\partial}{\partial\theta} \log f(x \mid \theta) &= \frac{1}{\theta} - \log x \\
\frac{\partial^2}{\partial\theta^2} \log f(x \mid \theta) &= -\frac{1}{\theta^2}.
\end{aligned}$$

Then the Fisher information is given by $I(\theta) = 1/\theta^2$*, so*

$$\sqrt{n}(\hat{\theta} - \theta) \to \mathcal{N}\left(0, \theta^2\right)$$

in distribution as $n \to \infty$*. For the function* $g(\theta) = \theta/(\theta-1)$*, we have*

$$g'(\theta) = \frac{1}{\theta-1} - \frac{\theta}{(\theta-1)^2} = -\frac{1}{(\theta-1)^2}.$$

So the delta method implies

$$\sqrt{n}\left(\frac{\hat{\theta}}{\hat{\theta}-1} - \frac{\theta}{\theta-1}\right) \to \mathcal{N}\left(0, \frac{\theta^2}{(\theta-1)^4}\right).$$

Say, for a data set with $n = 1000$ income values, we obtain the MLE $\hat{\theta} = 1.5$. We might then estimate the mean income as $\hat{\theta}/(\hat{\theta}-1) = 3$, and estimate our standard error by $\sqrt{\frac{\hat{\theta}^2}{n(\hat{\theta}-1)^4}} \approx 0.19$.

What if we decided to just estimate the mean income by the sample mean, $\bar{X}$? Since $\mathbb{E}[X_i] = \theta/(\theta-1)$, the Central Limit Theorem implies

$$\sqrt{n}\left(\bar{X} - \frac{\theta}{\theta-1}\right) \to \mathcal{N}\left(0, \mathrm{Var}[X_i]\right)$$

in distribution. For $\theta > 2$, we may compute

$$\mathbb{E}\left[X_i^2\right] = \int_1^\infty x^2 \cdot \theta x^{-\theta-1} dx = \theta \frac{x^{-\theta+2}}{-\theta+2}\Big|_1^\infty = \frac{\theta}{\theta-2},$$

so

$$\mathrm{Var}[X_i] = \mathbb{E}\left[X_i^2\right] - \left(\mathbb{E}[X_i]\right)^2 = \frac{\theta}{\theta-2} - \left(\frac{\theta}{\theta-1}\right)^2 = \frac{\theta}{(\theta-1)^2(\theta-2)}.$$

(If $\theta \leq 2$, the variance of X_i is actually infinite.) For any θ, this variance is greater than $\theta^2/(\theta-1)^4$.

Thus, if the Pareto model for income is correct, then our previous estimate $\hat{\theta}/(\hat{\theta}-1)$ is more accurate for the mean income than is the sample mean $\bar{X}$. Intuitively, this is because the Pareto distribution is heavy-tailed, and the sample mean $\bar{X}$ is heavily influenced by rare but extremely large data values. On the other hand, $\hat{\theta}$ is estimating the shape of the Pareto distribution and estimating the mean by its relationship to this shape in the Pareto model. The formula for $\hat{\theta}$ involves the values $\log X_i$ rather than X_i, so $\hat{\theta}$ is not as heavily influenced by extremely large data values. Of course, the estimate $\hat{\theta}/(\hat{\theta}-1)$ relies strongly on the correctness of the Pareto model, whereas $\bar{X}$ would be a valid estimate of the mean even if the Pareto model doesn't hold true.

That the plugin estimate $g(\hat{\theta})$ performs better than $\bar{X}$ in the previous example is not a coincidence - it is in certain senses the best we can do for estimating $g(\theta)$. For example, we have the following more general version of the Cramer-Rao lower bound:

Theorem 4.2. *For a parametric model $\{f(x \mid \theta) : \theta \in \Omega\}$ (satisfying certain mild regularity assumptions) where θ is a single parameter, let g be any function differentiable on all of Ω, and let T be any unbiased estimator of $g(\theta)$ based on data $X_1, \ldots, X_n \overset{IID}{\sim} f(x \mid \theta)$. Then*

$$\mathrm{Var}_\theta[T] \geq \frac{g'(\theta)^2}{nI(\theta)}.$$

Hint. The proof is identical to that of Theorem 2.2, except with the equation $\theta = \mathbb{E}_\theta[T]$ replaced by $g(\theta) = \mathbb{E}_\theta[T]$. (Differentiating this equation yields $g'(\theta) = \mathbb{E}_\theta[TZ] = \mathrm{Cov}_\theta[T, Z]$ as in Theorem 2.2) An estimator T for $g(\theta)$ that achieves this variance $g'(\theta)^2/(nI(\theta))$ is **efficient**. The plugin estimate $g(\hat{\theta})$ where $\hat{\theta}$ is the MLE achieves this variance asymptotically, so we say it is **asymptotically efficient**. This theorem ensures that no unbiased estimator of $g(\theta)$ can achieve variance much smaller than $g(\hat{\theta})$ when n is large, and in particular applies to the estimator $T = \bar{X}$ of the previous example. □

5 Confidence intervals (CI)

The estimation of a parameter by a single value is referred to as a point estimation. In a wide variety of inference problems, one is not interested in point estimation or testing of hypothesis of the parameter. Rather one wishes to establish a level or an upper bound or both for the parameter. An alternative procedure is to give an interval within which the parameter may be supposed to lie with a certain probability or confidence, which is called **Interval Estimation**.

We have seen how to understand the variability of an estimate $\hat{\theta}$ for a parameter θ, or of $g(\hat{\theta})$ for a quantity $g(\theta)$, in terms of its sampling distribution and its standard error. This understanding may be used to construct a confidence interval for θ or $g(\theta)$.

5.1 Exact confidence intervals

In a parametric model, let $g(\theta)$ be any quantity of interest (which might be the parameter θ itself). Informally, a **confidence interval** for $g(\theta)$ is a random interval calculated from the data that contains this value $g(\theta)$ with a specified probability. For example, a 90% confidence interval contains $g(\theta)$ with probability 0.90, and a 95% confidence interval contains $g(\theta)$ with probability 0.95. (If we construct 100 different 90% confidence intervals for θ using 100 independent sets of data, then we would expect about 90 of them to contain θ.)

WARNING! A common misconfusion is to interpret CIs as intervals with probability $(1-\alpha)$ of containing the true value θ, for a particular sample x. This interpretation is incorrect. The interpretation of CIs has to be made in terms of repeated sampling as discussed above (blue colored text).

More formally, what this means is the following: Let $X_1, \ldots, X_n$ be a sample of data. By *random interval*, we mean an interval whose lower and upper endpoints $L(X_1, \ldots, X_n)$ and $U(X_1, \ldots, X_n)$ are functions of the data $X_1, \ldots, X_n$. (Hence the interval is random in the same sense that the data itself is random - a different realization of the data leads to a different interval.) The interval $[L(X_1, \ldots, X_n), U(X_1, \ldots, X_n)]$ is a $100(1-\alpha)\%$ confidence interval for $g(\theta)$ if, for all $\theta \in \Omega$,

$$\mathbb{P}_\theta \left[L(X_1, \ldots, X_n) \leq g(\theta) \leq U(X_1, \ldots, X_n) \right] = 1 - \alpha;$$

where $\mathbb{P}_\theta$ denotes probability under $X_1, \ldots, X_n \overset{IID}{\sim} f(x \mid \theta)$.

A confidence interval for $g(\theta)$ is commonly constructed from an estimate of $g(\theta)$ and an estimate of the associated standard error:

Example 5.1. *Consider data $X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}(\mu, \sigma^2)$, where both μ and σ^2 are unknown. To construct a confidence interval for μ, consider the estimate $\bar{X}$. As $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$, the standard error of $\bar{X}$ is $\sigma/\sqrt{n}$, which we may estimate by $S/\sqrt{n}$ where*

$$S = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2}.$$

Recall from Chapter 2 Section 3 that when $X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}(\mu, \sigma^2)$, the quantity

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{\sqrt{n}(\bar{X} - \mu)}{S}$$

has a t-distribution with $n-1$ degrees of freedom. (In Chapter 2 Section 3 we assumed $\mu = 0$, but the distribution of this quantity doesn't depend on μ.) Letting $t_{n-1}(\alpha/2)$ be the upper-$\alpha/2$ point of the t_{n-1} distribution and noting that $-t_{n-1}(\alpha/2)$ is then the lower-$\alpha/2$ point by symmetry, this means

$$\mathbb{P}_{\mu,\sigma^2} \left[-t_{n-1}(\alpha/2) \leq \frac{\sqrt{n}(\bar{X} - \mu)}{S} \leq t_{n-1}(\alpha/2) \right] = 1 - \alpha.$$

The upper inequality above may be rearranged as

$$\bar{X} - \frac{S}{\sqrt{n}} t_{n-1}(\alpha/2) \leq \mu,$$

and the lower inequality may be rearranged as

$$\mu \leq \bar{X} + \frac{S}{\sqrt{n}} t_{n-1}(\alpha/2).$$

Hence

$$\mathbb{P}_{\mu,\sigma^2}\left[\bar{X} - \frac{S}{\sqrt{n}} t_{n-1}(\alpha/2) \leq \mu \leq \bar{X} + \frac{S}{\sqrt{n}} t_{n-1}(\alpha/2)\right] = 1 - \alpha,$$

so $\left[\bar{X} - \frac{S}{\sqrt{n}} t_{n-1}(\alpha/2), \bar{X} + \frac{S}{\sqrt{n}} t_{n-1}(\alpha/2)\right]$ *is a* $100(1-\alpha)\%$ *confidence interval for* μ. *We'll use the notation* $\bar{X} \pm \frac{S}{\sqrt{n}} t_{n-1}(\alpha/2)$ *as shorthand for this interval.*

Definition 5.1.

(i) An interval $I(\underset{\sim}{x})$ *which is a subset of* $\Omega \subseteq \mathbb{R}$ *is said to constitute a confidence interval with confidence coefficient* $(1-\alpha)$, *if* $P[I(\underset{\sim}{x}) \ni \theta] = 1 - \alpha \quad \forall\ \theta \in \Omega$, *i.e., the random interval* $I(\underset{\sim}{x})$ *covers the true parameter with probability* $= 1 - \alpha$.

(ii) A subset $S(\underset{\sim}{x})$ *of* $\Omega \subseteq \mathbb{R}^k$ *is said to constitute a confidence set at confidence* $(1-\alpha)$ *if* $P[S(\underset{\sim}{x}) \in \theta] \geqslant 1 - \alpha \quad \forall\ \theta \in \Omega$.

5.2 Methods of finding Confidence interval

Let θ be a parameter & T be a statistic based on a random sample of size n from a population. Most often it is possible to find a function $\psi(T, \theta)$ whose distribution is independent of θ. Then

$$P\left[\psi_{1-\alpha/2} < \psi(T,\theta) < \psi_{\alpha/2}\right] = 1 - \alpha,$$

where, ψ_α is independent of θ, as distribution of $\psi(T, \theta)$ is independent of θ.

Now, $\psi_{1-\alpha/2} < \psi(T,\theta) < \psi_{\alpha/2}$ can often be put in the form $\theta_1(T) \leqslant \theta \leqslant \theta_2(T)$.

Then $P\left[\theta_1(T) \leqslant \theta \leqslant \theta_2(T)\right] = 1-\alpha$ and the observed value of the interval $[\theta_1(T),\ \theta_2(T)]$ will be the confidence interval for θ with confidence coefficient $(1-\alpha)$.

Example 5.2. *Let $X_1, \ldots, X_n$ be a random sample from $\mathrm{N}\left(\mu, \sigma^2\right)$; μ and σ both are unknown. Find the confidence interval with confidence coefficient $(1-\alpha)$, for*

(i) μ; *(ii) σ^2;* *(iii) $\left(\mu, \sigma^2\right)$.*

Solutions.

(i) For confidence interval of μ, we select the statistic $T = \bar{X}$.
Then, $\psi(T, \mu) = \frac{\sqrt{n}(\bar{X}-\mu)}{s} \sim t_{n-1}$, which is independent of μ.

$$\text{Now,} \quad 1-\alpha = P\left[-t_{\alpha/2,n-1} < \frac{\sqrt{n}(\bar{X}-\mu)}{s} < t_{\alpha/2,n-1}\right]$$
$$= P\left[\bar{X} - \frac{s}{\sqrt{n}} t_{\alpha/2,n-1} \leqslant \mu \leqslant \bar{X} + \frac{s}{\sqrt{n}} t_{\alpha/2,n-1}\right].$$

Hence, $\left(\bar{X} - \frac{s}{\sqrt{n}} t_{\alpha/2,n-1},\ \bar{X} + \frac{s}{\sqrt{n}} t_{\alpha/2,n-1}\right)$ is an observed confidence interval for μ with confidence coefficient $(1-\alpha)$.

(ii) For confidence interval of σ^2, we select the statistic $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$.
Then, $\psi\left(s^2, \sigma^2\right) = (n-1)\frac{s^2}{\sigma^2} \sim \chi^2_{n-1}$, the distribution is independent of σ^2.
Now, $1-\alpha = P\left[\chi^2_{1-\alpha/2,n-1} \leqslant (n-1)\frac{s^2}{\sigma^2} \leqslant \chi^2_{\alpha/2,n-1}\right] = P\left[\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}} \leqslant \sigma^2 \leqslant \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}\right]$.
Hence, $\left(\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}{\chi^2_{\alpha/2,n-1}},\ \frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}{\chi^2_{1-\alpha/2,n-1}}\right)$ is an observed confidence interval for σ^2 with confidence coefficient $(1-\alpha)$.

(iii) $P\left[\bar{X} - \frac{s}{\sqrt{n}} t_{\alpha_1/2,n-1} \leqslant \mu \leqslant \bar{X} + \frac{s}{\sqrt{n}} t_{\alpha_1/2,n-1}\right] = 1-\alpha_1$
and $P\left[\frac{(n-1)s^2}{\chi^2_{\alpha_2/2,n-1}} \leqslant \sigma^2 \leqslant \frac{(n-1)s^2}{\chi^2_{1-\alpha_2/2,n-1}}\right] = 1-\alpha_2$.

Note that (**Boolsen Probability**): $P(A \cap B) \geqslant P(A) + P(B) - 1$.

$$\therefore P\left[\bar{X} - \frac{s}{\sqrt{n}} t_{\alpha_1/2,n-1} \leqslant \mu \leqslant \bar{X} + \frac{s}{\sqrt{n}} t_{\alpha_1/2,n-1};\ \frac{(n-1)s^2}{\chi^2_{\alpha_2/2,n-1}} \leqslant \sigma^2 \leqslant \frac{(n-1)s^2}{\chi^2_{1-\alpha_2/2,n-1}}\right]$$
$$\geqslant (1-\alpha_1) + (1-\alpha_2) - 1 = 1-\alpha, \text{ where } \alpha = \alpha_1 + \alpha_2.$$

Hence, $S(\underset{\sim}{X}) = \left(\bar{X} - \frac{s}{\sqrt{n}} t_{\alpha_1/2,n-1},\ \bar{X} + \frac{s}{\sqrt{n}} t_{\alpha_1/2,n-1}\right) \times \left(\frac{(n-1)s^2}{\chi_{\alpha_2/2,n-1}},\ \frac{(n-1)s^2}{\chi_{1-\alpha_2/2,n-1}}\right)$ is an observed confidence interval for (μ, σ^2) with confidence coefficient $(1-\alpha)$.

5.3 Asymptotic confidence intervals

In the previous example, we were able to construct an exact confidence interval because we knew the exact distribution of $\sqrt{n}(\bar{X}-\mu)/S$, which is t_{n-1} (and which does not

depend on μ and σ^2). Suppose that we had forgotten this fact. If n is large, we could have still reasoned as follows: By the Central Limit Theorem, as $n \to \infty$,

$$\sqrt{n}(\bar{X} - \mu) \to \mathcal{N}\left(0, \sigma^2\right)$$

in distribution. By our addendum at the end of Section 6.4 of Chapter 2, $S^2 \to \sigma^2$ in probability (meaning, the sample variance S^2 is consistent for σ^2). Then, applying the Continuous Mapping Theorem and Slutsky's Lemma,

$$\frac{\sqrt{n}(\bar{X} - \mu)}{S} = \frac{\sigma}{S} \times \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \to \mathcal{N}(0, 1)$$

in distribution, so

$$\mathbb{P}_{\mu,\sigma^2}\left[-z_{\alpha/2} \leq \frac{\sqrt{n}(\bar{X} - \mu)}{S} \leq z_{\alpha/2}\right] \to 1 - \alpha$$

as $n \to \infty$. Rearranging the inequalities above in the same way as the previous example yields a $100(1-\alpha)\%$ **asymptotic confidence interval** $\bar{X} \pm \frac{S}{\sqrt{n}} z_{\alpha/2}$ for μ. We expect this interval to be accurate (meaning its coverage of μ is close to $100(1-\alpha)\%$) for large n - indeed, for large n, $z_{\alpha/2} \approx t_{n-1}(\alpha/2)$ because the t_{n-1} distribution is very close to the standard normal distribution so that this interval is almost the same as the exact interval of the previous example.

This method may be applied to construct an approximate confidence interval from any asymptotically normal estimator, as we will see in the following examples.

Example 5.3. *Let* $X_1, \ldots, X_n \overset{IID}{\sim} \text{Poisson}(\lambda)$. *To construct an asymptotic confidence interval for* λ, *let's start with the estimator* $\hat{\lambda} = \bar{X}$. *By the Central Limit Theorem,*

$$\sqrt{n}(\hat{\lambda} - \lambda) \to \mathcal{N}(0, \lambda).$$

We don't know the variance λ *of this limiting normal distribution, but we can estimate it by* $\hat{\lambda}$. *By the Law of Large Numbers,* $\hat{\lambda} \to \lambda$ *in probability as* $n \to \infty$, *i.e.* $\hat{\lambda}$ *is consistent for* λ. *Then by the Continuous Mapping Theorem and Slutsky's Lemma,*

$$\frac{\sqrt{n}(\hat{\lambda} - \lambda)}{\sqrt{\hat{\lambda}}} = \frac{\sqrt{\lambda}}{\sqrt{\hat{\lambda}}} \times \frac{\sqrt{n}(\hat{\lambda} - \lambda)}{\sqrt{\lambda}} \to \mathcal{N}(0, 1),$$

so

$$\mathbb{P}_{\lambda}\left[-z_{\alpha/2} \leq \frac{\sqrt{n}(\hat{\lambda} - \lambda)}{\sqrt{\hat{\lambda}}} \leq z_{\alpha/2}\right] \to 1 - \alpha$$

Rearranging these inequalities yields the asymptotic $100(1-\alpha)\%$ *confidence interval* $\hat{\lambda} \pm \sqrt{\frac{\hat{\lambda}}{n}} z_{\alpha/2}$.
For various values of λ *and* n*, the table below shows the simulated true probabilities that the* 90% *and* 95% *confidence intervals constructed in this way cover* λ*:*

	Desired coverage: 90%			Desired coverage: 95%		
	$\lambda = 0.1$	$\lambda = 1$	$\lambda = 5$	$\lambda = 0.1$	$\lambda = 1$	$\lambda = 5$
$n = 10$	0.63	0.91	0.90	0.63	0.93	0.95
$n = 30$	0.79	0.89	0.90	0.80	0.93	0.95
$n = 100$	0.91	0.90	0.90	0.93	0.94	0.95

(Meaning, we simulated $X_1, \ldots, X_n \overset{IID}{\sim} \text{Poisson}(\lambda)$*, computed the confidence interval, checked whether it contained* λ*, and repeated this* $B = 1,000,000$ *times. The table reports the fraction of simulations for which the interval covered* λ*.) We observe that coverage is closer to the desired levels for larger values of* n*, as well as for larger values of* λ*. For small* n *and/or small* λ*, the normal approximation to the distribution of* $\hat{\lambda}$ *is inaccurate, and the simulations show that we underestimate the variability of* $\hat{\lambda}$*.*

Example 5.4. *More generally, let* $\{f(x \mid \theta) : \theta \in \Omega\}$ *be any parametric model satisfying the regularity conditions of Theorem 1.1, where* θ *is a single parameter. To obtain a confidence interval for* θ*, consider the MLE* $\hat{\theta}$*, which satisfies*

$$\sqrt{n}(\hat{\theta} - \theta) \to \mathcal{N}\left(0, I(\theta)^{-1}\right)$$

as $n \to \infty$*. We may estimate* $I(\theta)$ *by the plugin estimator* $I(\hat{\theta})$*. If* $I(\theta)$ *is continuous in* θ *and* $\hat{\theta}$ *is consistent for* θ*, then the Continuous Mapping Theorem implies* $I(\hat{\theta}) \to I(\theta)$ *in probability, and hence*

$$\sqrt{nI(\hat{\theta})}(\hat{\theta} - \theta) = \frac{\sqrt{I(\hat{\theta})}}{\sqrt{I(\theta)}} \times \sqrt{nI(\theta)}(\hat{\theta} - \theta) \to \mathcal{N}(0, 1).$$

So,

$$\mathbb{P}_\theta[-z_{\alpha/2} \leq \sqrt{nI(\hat{\theta})}(\hat{\theta} - \theta) \leq z_{\alpha/2}] \to 1 - \alpha,$$

and rearranging yields the asymptotic $100(1-\alpha)\%$ *confidence interval* $\hat{\theta} \pm \frac{1}{\sqrt{nI(\hat{\theta})}} z_{\alpha/2}$*. This is oftentimes called the* ***Wald interval*** *for* θ*.*

Example 5.5. *Let* $X_1, \ldots, X_n \overset{IID}{\sim} \text{Bernoulli}(p)$*. Suppose we wish to construct a confidence interval for the log-odds* $g(p) = \log \frac{p}{1-p}$*. In Section 4, we showed using the delta method that*

$$\sqrt{n}(g(\hat{p}) - g(p)) \to \mathcal{N}\left(0, \frac{1}{p(1-p)}\right),$$

where $\hat{p} = \bar{X}$*. Since* $\hat{p} \to p$ *in probability, by the Continuous Mapping Theorem and Slutsky's Lemma,*

$$\sqrt{n\hat{p}(1-\hat{p})}(g(\hat{p}) - g(p)) = \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{p(1-p)}}\sqrt{np(1-p)}(g(\hat{p}) - g(p)) \to \mathcal{N}(0,1),$$

so

$$\mathbb{P}_p[-z_{\alpha/2} \leq \sqrt{n\hat{p}(1-\hat{p})}(g(\hat{p}) - g(p)) \leq z_{\alpha/2}] \to 1 - \alpha.$$

An asymptotic $100(1-\alpha)\%$ *confidence interval for the log-odds* $g(p) = \log \frac{p}{1-p}$ *is then*

$$[L(\hat{p}), U(\hat{p})] := \left[\log \frac{\hat{p}}{1-\hat{p}} - \sqrt{\frac{1}{n\hat{p}(1-\hat{p})}} z_{\alpha/2}, \log \frac{\hat{p}}{1-\hat{p}} + \sqrt{\frac{1}{n\hat{p}(1-\hat{p})}} z_{\alpha/2}\right].$$

If we wish to obtain a confidence interval for the odds $\frac{p}{1-p}$ rather than the log-odds, note that $\mathbb{P}\left[L(\hat{p}) \leq \log \frac{p}{1-p} \leq U(\hat{p})\right] = \mathbb{P}\left[e^{L(\hat{p})} \leq \frac{p}{1-p} \leq e^{U(\hat{p})}\right]$, so that $\left[e^{L(\hat{p})}, e^{U(\hat{p})}\right]$ is a confidence interval for the odds. This interval is *not* symmetric around the estimate $\frac{\hat{p}}{1-\hat{p}}$, and is different from what we would have obtained if we instead applied the delta method directly to $g(p) = \frac{p}{1-p}$. The interval $\left[e^{L(\hat{p})}, e^{U(\hat{p})}\right]$ for the odds is typically used in practice the distribution of $\log \frac{\hat{p}}{1-\hat{p}}$ is less skewed than that of $\frac{\hat{p}}{1-\hat{p}}$ for small to moderate n, so the normal approximation and resulting confidence interval are more accurate if we consider odds on the log scale.

Let us caution that in the construction of these asymptotic confidence intervals, a number of different approximations are being made:

- The true distribution of $\sqrt{n}(\hat{\theta} - \theta)$ is being approximated by a normal distribution.
- The true variance of this normal distribution, say $I(\theta)^{-1}$, is being approximated by a plugin estimate $I(\hat{\theta})^{-1}$.
- In the case where we are interested in $g(\theta)$ and g is a nonlinear function, the value $g(\hat{\theta})$ is being approximated by the Taylor expansion $g(\theta) + (\hat{\theta} - \theta)g'(\theta)$. (This is what is done in the delta method.)

These approximations are all valid in the limit $n \to \infty$, but their accuracy is not guaranteed for the finite sample size n of any given problem. Coverage of asymptotic confidence intervals should be checked by simulation, as Example 5.3 illustrates that they might be severely overconfident for small n.

6 Bayesian analysis

Our treatment of parameter estimation thus far has assumed that θ is an unknown but non-random quantity - it is some fixed parameter describing the true distribution of data, and our goal was to determine this parameter. This is called the Frequentist Paradigm of statistical inference. In this and the next section, we will describe an alternative Bayesian Paradigm, in which θ itself is modeled as a random variable. The Bayesian paradigm naturally incorporates our prior belief about the unknown parameter θ and updates this belief based on observed data.

6.1 Prior and posterior distributions

Recall that if X, Y are two random variables having joint PDF or PMF $f_{X,Y}(x, y)$, then the **marginal distribution** of X is given by the PDF

$$f_X(x) = \int f_{X,Y}(x, y)dy$$

in the continuous case and by the PMF

$$f_X(x) = \sum_y f_{X,Y}(x, y)$$

in the discrete case; this describes the probability distribution of X alone. The **conditional distribution** of Y given $X = x$ is defined by the PDF or PMF

$$f_{Y|X}(y \mid x) = \frac{f_{X,Y}(x, y)}{f_X(x)},$$

and represents the probability distribution of Y if it is known that $X = x$. (This is a PDF or PMF as a function of y, for any fixed x.) Defining similarly the marginal distribution $f_Y(y)$ of Y and the conditional distribution $f_{X|Y}(x \mid y)$ of X given $Y = y$, the joint PDF $f_{X,Y}(x, y)$ factors in two ways as

$$f_{X,Y}(x, y) = f_{Y|X}(y \mid x)f_X(x) = f_{X|Y}(x \mid y)f_Y(y).$$

In Bayesian analysis, before data is observed, the unknown parameter is modeled as a random variable Θ having a probability distribution $f_\Theta(\theta)$, called the **prior distribution**. This distribution represents our prior belief about the value of this parameter.

Conditional on $\Theta = \theta$, the observed data X is assumed to have distribution $f_{X|\Theta}(x \mid \theta)$, where $f_{X|\Theta}(x \mid \theta)$ defines a parametric model with parameter θ, as in our previous chapters.[7] The joint distribution of Θ and X is then the product

$$f_{X,\Theta}(x, \theta) = f_{X|\Theta}(x \mid \theta) f_\Theta(\theta),$$

and the marginal distribution of X (in the continuous case) is

$$f_X(x) = \int f_{X,\Theta}(x, \theta) d\theta = \int f_{X|\Theta}(x \mid \theta) f_\Theta(\theta) d\theta.$$

The conditional distribution of Θ given $X = x$ is

$$f_{\Theta|X}(\theta \mid x) = \frac{f_{X,\Theta}(x, \theta)}{f_X(x)} = \frac{f_{X|\Theta}(x \mid \theta) f_\Theta(\theta)}{\int f_{X|\Theta}(x \mid \theta') f_\Theta(\theta') d\theta'}. \tag{6}$$

This is called the **posterior distribution** of Θ : It represents our knowledge about the parameter Θ after having observed the data X. We often summarize the preceding equation simply as

$$\begin{gathered} f_{\Theta|X}(\theta \mid x) \propto f_{X|\Theta}(x \mid \theta) f_\Theta(\theta) \\ \text{Posterior density } \propto \text{ Likelihood } \times \text{ Prior density} \end{gathered} \tag{7}$$

where the symbol $\propto$ hides the proportionality factor $f_X(x) = \int f_{X|\Theta}(x \mid \theta') f_\Theta(\theta') d\theta'$ which does not depend on θ.

Example 6.1. *Let $P \in (0, 1)$ be the probability of heads for a biased coin, and let $X_1, \ldots, X_n$ be the outcomes of n tosses of this coin. If we do not have any prior information about P, we might choose for its prior distribution Uniform $(0, 1)$, having* PDF $f_P(p) = 1$ *for all $p \in (0, 1)$. Given $P = p$, we model $X_1, \ldots, X_n \overset{IID}{\sim}$ Bernoulli(p). Then the joint distribution of $P, X_1, \ldots, X_n$ is given by*

$$\begin{aligned} f_{X,P}(x_1, \ldots, x_n, p) &= f_{X|P}(x_1, \ldots, x_n \mid p) f_P(p) \\ &= \prod_{i=1}^n p^{x_i} (1-p)^{1-x_i} \times 1 = p^{\sum_{i=1}^n x_i} (1-p)^{n - \sum_{i=1}^n x_i}. \end{aligned}$$

Let $s = x_1 + \ldots + x_n$. The marginal distribution of $X_1, \ldots, X_n$ is obtained by integrating $f_{X,P}(x_1, \ldots, x_n, p)$ over p:

$$f_X(x_1, \ldots, x_n) = \int_0^1 p^s (1-p)^{n-s} dp = B(s+1, n-s+1)$$

where $B(x, y)$ is the Beta function

[7] For notational simplicity, we are considering here a single data value X, but this extends naturally to the case where $\mathbf{X} = (X_1, \ldots, X_n)$ is a data vector and $f_{X|\Theta}(\mathbf{x} \mid \theta)$ is the joint distribution of $\mathbf{X}$ given θ.

$$B(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}.$$

Hence the posterior distribution of P given $X_1 = x_1, \ldots, X_n = x_n$ has PDF

$$f_{P|X}\left(p \mid x_1, \ldots, x_n\right) = \frac{f_{X,P}\left(x_1, \ldots, x_n, p\right)}{f_X\left(x_1, \ldots, x_n\right)} = \frac{1}{B(s+1, n-s+1)} p^s (1-p)^{n-s}.$$

This is the PDF of the $\text{Beta}(s+1, n-s+1)$ *distribution*[8]*, so the posterior distribution of P given $X_1 = x_1, \ldots, X_n = x_n$ is* $\text{Beta}(s+1, n-s+1)$*, where $s = x_1 + \ldots + x_n$.*

We computed explicitly the marginal distribution $f_X\left(x_1, \ldots, x_n\right)$ above, but this was not necessary to arrive at the answer. Indeed, equation (7) gives

$$f_{P|X}\left(p \mid x_1, \ldots, x_n\right) \propto f_{X|P}\left(x_1, \ldots, x_n \mid p\right) f_P(p) = p^s (1-p)^{n-s}.$$

This tells us that the PDF of the posterior distribution of P is proportional to $p^s(1-p)^{n-s}$, as a function of p. Then it must be the PDF of the $\text{Beta}(s+1, n-s+1)$ distribution, and the proportionality constant must be whatever constant is required to make this PDF integrate to 1 over $p \in (0,1)$. We will repeatedly use this trick to simplify our calculations of posterior distributions.

Example 6.2. *Suppose now we have a prior belief that P is close to $1/2$. There are various prior distributions that we can choose to encode this belief; it will turn out to be mathematically convenient to use the prior distribution* $\text{Beta}(\alpha, \alpha)$*, which has mean $1/2$ and variance $1/(8\alpha+4)$. The constant α may be chosen depending on how confident we are, a priori, that P is near $1/2$ - choosing $\alpha = 1$ reduces to the Uniform $(0,1)$ prior of the previous example, whereas choosing $\alpha > 1$ yields a prior distribution more concentrated around $1/2$.*

The prior distribution $\text{Beta}(\alpha, \alpha)$ *has* PDF $f_P(p) = \frac{1}{B(\alpha,\alpha)} p^{\alpha-1}(1-p)^{\alpha-1}$*. Then, applying equation (7), the posterior distribution of P given $X_1 = x_1, \ldots, X_n = x_n$ has* PDF

$$\begin{aligned} f_{P|X}\left(p \mid x_1, \ldots, x_n\right) &\propto f_{X|P}\left(x_1, \ldots, x_n \mid p\right) f_P(p) \\ &\propto p^s (1-p)^{n-s} \times p^{\alpha-1}(1-p)^{\alpha-1} = p^{s+\alpha-1}(1-p)^{n-s+\alpha-1}, \end{aligned}$$

where $s = x_1 + \ldots + x_n$ as before, and where the symbol $\propto$ hides any proportionality constants that do not depend on p. This is proportional to the PDF of the distribution Beta $(s+\alpha, n-s+\alpha)$, so this Beta distribution is the posterior distribution of P.

[8]The $\text{Beta}(\alpha, \beta)$ distribution is a continuous distribution on $(0,1)$ with PDF $f(x) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$.

In the previous example, the parametric form for the prior was (cleverly) chosen so that the posterior would be of the same form-they were both Beta distributions. This type of prior is called a **conjugate prior** for P in the Bernoulli model. Use of a conjugate prior is mostly for mathematical and computational convenience - in principle, any prior $f_P(p)$ on $(0,1)$ may be used. The resulting posterior distribution may be not be a simple named distribution with a closed-form PDF, but the PDF may be computed numerically from equation (6) by numerically evaluating the integral in the denominator of this equation.

Example 6.3. *Let $\Lambda \in (0,\infty)$ be the parameter of the Poisson model $X_1,\ldots,X_n \overset{IID}{\sim}$ Poisson(λ). As a prior distribution for Λ, let us take the Gamma distribution* Gamma(α,β). *The prior and likelihood are given by*

$$f_\Lambda(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}$$
$$f_{X|\Lambda}(x_1,\ldots,x_n \mid \lambda) = \prod_{i=1}^n \frac{\lambda^{x_i}e^{-\lambda}}{x_i!}.$$

Dropping proportionality constants that do not depend on λ, the posterior distribution of Λ given $X_1 = x_1,\ldots,X_n = x_n$ is then

$$\begin{aligned} f_{\Lambda|X}(\lambda \mid x_1,\ldots,x_n) &\propto f_{X|\Lambda}(x_1,\ldots,x_n \mid \lambda)\, f_\Lambda(\lambda) \\ &\propto \prod_{i=1}^n \left(\lambda^{x_i}e^{-\lambda}\right) \times \lambda^{\alpha-1}e^{-\beta\lambda} = \lambda^{s+\alpha-1}e^{-(n+\beta)\lambda}, \end{aligned}$$

where $s = x_1 + \ldots + x_n$. This is proportional to the PDF of the Gamma$(s+\alpha, n+\beta)$ *distribution, so the posterior distribution of Λ must be* Gamma$(s+\alpha, n+\beta)$.

As the prior and posterior are both Gamma distributions, the Gamma distribution is a **conjugate prior** for Λ in the Poisson model.

6.2 Point estimates and credible intervals

To the Bayesian statistician, the posterior distribution is the complete answer to the question: What is the value of θ? In many applications, though, we would still like to have a single estimate $\hat{\theta}$, as well as an interval describing our uncertainty about θ.

The **posterior mean** and **posterior mode** are the mean and mode of the posterior distribution of Θ; both of these are commonly used as a Bayesian estimate $\hat{\theta}$ for θ. A $100(1-\alpha)\%$ **Bayesian credible interval** is an interval I such that the posterior probability $\mathbb{P}[\Theta \in I \mid X] = 1-\alpha$, and is the Bayesian analogue to a frequentist

confidence interval. One common choice for I is simply the interval $\left[\theta^{(\alpha/2)}, \theta^{(1-\alpha/2)}\right]$ where $\theta^{(\alpha/2)}$ and $\theta^{(1-\alpha/2)}$ are the $\alpha/2$ and $1-\alpha/2$ quantiles of the posterior distribution of Θ. Note that the interpretation of a Bayesian credible interval is different from the interpretation of a frequentist confidence interval - in the Bayesian framework, the parameter Θ is modeled as random, and $1-\alpha$ is the probability that this random parameter Θ belongs to an interval that is fixed conditional on the observed data.

Example 6.4. *From Example 6.2, the posterior distribution of P is* Beta$(s+\alpha, n-s+\alpha)$. *The posterior mean is then $(s+\alpha)/(n+2\alpha)$, and the posterior mode is $(s+\alpha-1)/(n+2\alpha-2)$. Both of these may be taken as a point estimate $\hat{p}$ for p. The interval from the* 0.05 *to the* 0.95 *quantile of the* Beta$(s+\alpha, n-s+\alpha)$ *distribution forms a* 90% *Bayesian credible interval for p.*

Example 6.5. *From Example 6.3, the posterior distribution of Λ is* Gamma$(s+\alpha, n+\beta)$. *The posterior mean and mode are then $(s+\alpha)/(n+\beta)$ and $(s+\alpha-1)/(n+\beta)$, and either may be used as a point estimate $\hat{\lambda}$ for λ. The interval from the* 0.05 *to the* 0.95 *quantile of the* Gamma$(s+\alpha, n+\beta)$ *distribution forms a* 90% *Bayesian credible interval for λ.*

6.3 Conjugate priors and improper priors

Last section, we saw two examples of conjugate priors:

1. If $X_1, \ldots, X_n \overset{\text{IID}}{\sim}$ Poisson(λ), then a conjugate prior for λ is Gamma(α, β), and the corresponding posterior given $X_1 = x_1, \ldots, X_n = x_n$ is Gamma$(s+\alpha, n+\beta)$ where $s = x_1 + \ldots + x_n$. A Bayesian estimate of λ is the posterior mean

$$\hat{\lambda} = \frac{s+\alpha}{n+\beta} = \frac{n}{n+\beta} \cdot \frac{s}{n} + \frac{\beta}{n+\beta} \cdot \frac{\alpha}{\beta}.$$

2. If $X_1, \ldots, X_n \overset{IID}{\sim}$ Bernoulli(p), then a conjugate prior for p is Beta(α, β), and the corresponding posterior given $X_1 = x_1, \ldots, X_n = x_n$ is Beta$(s+\alpha, n-s+\beta)$.[9] A Bayesian estimate of p is the posterior mean

$$\hat{p} = \frac{s+\alpha}{n+\alpha+\beta} = \frac{n}{n+\alpha+\beta} \cdot \frac{s}{n} + \frac{\alpha+\beta}{n+\alpha+\beta} \cdot \frac{\alpha}{\alpha+\beta}.$$

In addition to being mathematically convenient, conjugate priors oftentimes have intuitive interpretations: In example 1 above, the posterior mean behaves as if we observed, a priori, β additional count observations that sum to α. β may be interpreted

[9] If we assume $\alpha = \beta$ then the prior is centered around 1/2, but the same calculation of the posterior distribution holds when $\alpha \neq \beta$.

as an effective prior sample size and α/β as a prior mean, and the posterior mean is a weighted average of the prior mean and the data mean. In example 2 above, the posterior mean behaves as if we observed, a priori, α additional heads and β additional tails. $\alpha + \beta$ is an effective prior sample size, $\alpha/(\alpha + \beta)$ is a prior mean, and the posterior mean is again a weighted average of the prior mean and the data mean. These interpretations may serve as a guide for choosing the prior parameters α and β.

Sometimes it is convenient to use the formalism of Bayesian inference, but with an "uninformative prior" that does not actually impose prior knowledge, so that the resulting analysis is more objective. In both examples above, the priors are "uninformative" for the posterior mean when α and β are small. We may take this idea to the limit by considering $\alpha = \beta = 0$. As the PDF of the Gamma distribution is proportional to $x^{\alpha-1}e^{-\beta x}$ on $(0, \infty)$, the "PDF" for $\alpha = \beta = 0$ may be considered to be

$$f(x) \propto x^{-1}.$$

Similarly, as the PDF of the Beta distribution is proportional to $x^{\alpha-1}(1-x)^{\beta-1}$ on $(0, 1)$, the "PDF" for $\alpha = \beta = 0$ may be considered to be

$$f(x) \propto x^{-1}(1-x)^{-1}.$$

These are *not* real probability distributions: There is no such distribution as Gamma$(0, 0)$, and $f(x) \propto x^{-1}$ does not actually describe a valid PDF on $(0, \infty)$, because $\int x^{-1}dx = \infty$ so that it is impossible to choose a normalizing constant to make this PDF integrate to 1. Similarly, there is no such distribution as Beta$(0, 0)$, and $f(x) \propto x^{-1}(1-x)^{-1}$ does not describe a valid PDF on $(0, 1)$. These types of priors are called **improper priors**.

Nonetheless, we may formally carry out Bayesian analysis using improper priors, and this oftentimes yields valid posterior distributions: In the Poisson example, we obtain the posterior PDF

$$\begin{aligned} f_{\Lambda|X}(\lambda \mid x_1, \ldots, x_n) &\propto f_{X|\Lambda}(x_1, \ldots, x_n \mid \lambda) f_\Lambda(\lambda) \\ &\propto \lambda^s e^{-n\lambda} \times \lambda^{-1} = \lambda^{s-1} e^{-n\lambda}, \end{aligned}$$

which is the PDF of Gamma(s, n). In the Bernoulli example, we obtain the posterior PDF

$$\begin{aligned} f_{P|X}(p \mid x_1, \ldots, x_n) &\propto f_{X|P}(x_1, \ldots, x_n \mid p) f_P(p) \\ &\propto p^s(1-p)^{n-s} \times p^{-1}(1-p)^{-1} = p^{s-1}(1-p)^{n-s-1}, \end{aligned}$$

which is the PDF of Beta$(s, n-s)$. These posterior distributions are real probability distributions (as long as $s > 0$ in the Poisson example and $s,\ n-s > 0$ in the Bernoulli example), and may be thought of as approximations to the posterior distributions that we would have obtained if we used proper priors with small but positive values of α and β.

6.4 Normal approximation for large n

For any fixed α, β in the above examples, as $n \to \infty$, the influence of the prior diminishes and the posterior mean becomes close to the MLE s/n. This is true more generally for parametric models satisfying mild regularity conditions, and in fact the posterior distribution is approximately a normal distribution centered at the MLE $\hat{\theta}$ with variance $\frac{1}{nI(\hat{\theta})}$ for large n, where $I(\theta)$ is the Fisher information. We sketch the argument for why this occurs:

Consider Bayesian inference applied with the prior $f_\Theta(\theta)$, for a parametric model $f_{X|\Theta}(x \mid \theta)$. Let $X_1, \ldots, X_n \overset{IID}{\sim} f_{X|\Theta}(x \mid \theta)$, and let

$$l(\theta) = \sum_{i=1}^{n} \log f_{X|\Theta}\left(x_i \mid \theta\right)$$

be the usual log-likelihood. Then the posterior distribution of Θ is given by

$$f_{\Theta|X}\left(\theta \mid x_1, \ldots, x_n\right) \propto f_{X|\Theta}\left(x_1, \ldots, x_n \mid \theta\right) f_\Theta(\theta) = \exp(l(\theta)) f_\Theta(\theta).$$

Applying a second-order Taylor expansion of $l(\theta)$ around the MLE of $\theta = \hat{\theta}$,

$$\begin{aligned} l(\theta) &\approx l(\hat{\theta}) + (\theta - \hat{\theta}) l'(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^2 l''(\hat{\theta}) \\ &\approx l(\hat{\theta}) - \frac{1}{2}(\theta - \hat{\theta})^2 \cdot nI(\hat{\theta}), \end{aligned}$$

where the second equality follows because $l'(\hat{\theta}) = 0$ if $\hat{\theta}$ is the MLE, and $l''(\hat{\theta}) \approx -nI(\hat{\theta})$ for large n. Since $\hat{\theta}$ is a function of the data $x_1, \ldots, x_n$ and doesn't depend on θ, we may absorb $\exp(l(\hat{\theta}))$ into the proportionality constant to obtain

$$f_{\Theta|X}\left(\theta \mid x_1, \ldots, x_n\right) \propto \exp\left(-\frac{1}{2}(\theta - \hat{\theta})^2 \cdot nI(\hat{\theta})\right) f_\Theta(\theta).$$

For large n, the value of $\exp\left(-\frac{1}{2}(\theta - \hat{\theta})^2 \cdot nI(\hat{\theta})\right)$ is small unless θ is within order $1/\sqrt{n}$ distance from $\hat{\theta}$. In this region of θ, the prior $f_\Theta(\theta)$ is approximately constant and equal to $f_\Theta(\hat{\theta})$. Absorbing this constant into the proportionality factor in $\propto$, we finally arrive at

$$f_{\Theta|X}\left(\theta \mid x_1, \ldots, x_n\right) \propto \exp\left(-\frac{1}{2}(\theta - \hat{\theta})^2 \cdot nI(\hat{\theta})\right).$$

This describes a normal distribution for Θ with mean $\hat{\theta}$ and variance $\frac{1}{nI(\hat{\theta})}$.

To summarize, the posterior mean of Θ is, for large n, approximately the MLE $\hat{\theta}$. Furthermore, a $100(1-\alpha)\%$ Bayesian credible interval is approximately given by $\hat{\theta} \pm z(\alpha/2)/\sqrt{nI(\hat{\theta})}$, which is exactly the $100(1-\alpha)\%$ Wald confidence interval for θ. In this sense, frequentist and Bayesian methods yield similar inferences for large n.

6.5 Prior distributions and average MSE

Last section we introduced the prior distribution for Θ as something that encodes our prior belief about its value. A different (but related) interpretation and motivation for the prior comes from the following considerations:

Let's return to the frequentist setting where we assume that there is a true parameter θ for a parametric model $\{f(x \mid \theta) : \theta \in \Omega\}$. Suppose we have two estimators for θ based on data $X_1, \ldots, X_n \sim f(x \mid \theta)$: $\hat{\theta}_1$ and $\hat{\theta}_2$. Which estimator is "better"? Without appealing to asymptotic (large n) arguments, one answer to this question is to compare their mean-squared-errors:

$$\begin{aligned}
\text{MSE}_1(\theta) = \mathbb{E}_\theta\left[\left(\hat{\theta}_1 - \theta\right)^2\right] &= \text{ Variance of } \hat{\theta}_1 + \left(\text{ Bias of } \hat{\theta}_1\right)^2; \\
\text{MSE}_2(\theta) = \mathbb{E}_\theta\left[\left(\hat{\theta}_2 - \theta\right)^2\right] &= \text{ Variance of } \hat{\theta}_2 + \left(\text{ Bias of } \hat{\theta}_2\right)^2.
\end{aligned}$$

The estimator with smaller MSE is "better". Unfortunately, the problem with this approach is that the MSEs might depend on the true parameter θ (hence why we have written MSE_1 and MSE_2 as functions of θ in the above), and neither may be uniformly better than the other. For example, suppose $X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}(\theta, 1)$. Let $\hat{\theta}_1 = \bar{X}$; this is unbiased with variance $\frac{1}{n}$, so its MSE is $\frac{1}{n}$. Let $\hat{\theta}_2 \equiv 0$ be the constant estimator that always estimates θ by 0. This has bias $-\theta$ and variance 0, so its MSE is θ^2. If the true parameter θ happens to be close to 0 - more specifically, if $|\theta|$ is less than $1/\sqrt{n}$ - then $\hat{\theta}_2$ is "better", and otherwise $\hat{\theta}_1$ is "better".

To resolve this ambiguity, we might consider a weighted average MSE,

$$\int \text{MSE}(\theta) w(\theta) d\theta,$$

where $w(\theta)$ is a weight function over the parameter space such that $\int_\Omega w(\theta)d\theta = 1$, and find the estimator that minimizes this weighted average. This weighted average MSE is called the **Bayes risk**. Writing the expectation in the definition of the MSE as an integral, and letting $\mathbf{x}$ denote the data and $f(\mathbf{x} \mid \theta)$ denote the PDF of the data, we may write the Bayes risk of an estimator $\hat{\theta}$ as

$$\int \left(\int (\hat{\theta}(\mathbf{x}) - \theta)^2 f(\mathbf{x} \mid \theta) d\mathbf{x} \right) w(\theta) d\theta.$$

Exchanging the order of integration, this is

$$\int \left(\int (\hat{\theta}(\mathbf{x}) - \theta)^2 f(\mathbf{x} \mid \theta) w(\theta) d\theta \right) d\mathbf{x}.$$

In order to minimize the Bayes risk, for each possible value $\mathbf{x}$ of the observed data, $\hat{\theta}(\mathbf{x})$ should be defined so as to minimize

$$\int (\hat{\theta}(\mathbf{x}) - \theta)^2 f(\mathbf{x} \mid \theta) w(\theta) d\theta.$$

Let us now interpret $w(\theta)$ as a prior $f_\Theta(\theta)$ for the parameter Θ, and $f(\mathbf{x} \mid \theta)$ as the likelihood $f_{X|\Theta}(\mathbf{x} \mid \theta)$ given $\Theta = \theta$. Then

$$\int (\hat{\theta}(\mathbf{x}) - \theta)^2 f(\mathbf{x} \mid \theta) w(\theta) d\theta = \int (\hat{\theta}(\mathbf{x}) - \theta)^2 f_{X,\Theta}(\mathbf{x}, \theta) d\theta = f_X(\mathbf{x}) \int (\hat{\theta}(\mathbf{x}) - \theta)^2 f_{\Theta|X}(\theta \mid \mathbf{x}) d\theta.$$

So given the observed data $\mathbf{x}$, $\hat{\theta}(\mathbf{x})$ should be defined to minimize

$$\int (\hat{\theta}(\mathbf{x}) - \theta)^2 f_{\Theta|X}(\theta \mid \mathbf{x}) d\theta = \mathbb{E}\left[(\hat{\theta}(\mathbf{x}) - \Theta)^2 \right],$$

where the expectation is with respect to the posterior distribution of Θ for the fixed and observed value of $\mathbf{x}$. For any random variable Y, $\mathbb{E}\left[(c - Y)^2\right]$ is minimized over c when $c = \mathbb{E}[Y]$ - hence the minimizer $\hat{\theta}(\mathbf{x})$ of the above is exactly the posterior mean of Θ. We have thus arrived at the following conclusion:

The posterior mean of Θ for the prior $f_\Theta(\theta)$ is the estimator that minimizes the average mean-squared-error

$$\int \mathrm{MSE}(\theta) f_\Theta(\theta) d\theta.$$

Thus a Bayesian prior may be interpreted as the weighting of parameter values for which we wish to minimize the weighted-average mean-squared-error.