13

STATISTICAL ESTIMATION THEORY

13.1 Parameter estimators

13.1.1 Overview

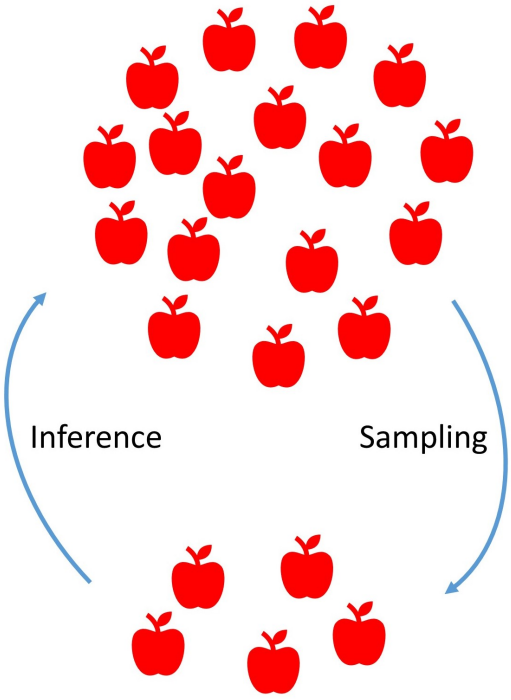


Figure 13.1.1: Statistical estimation and inference scheme.

Given observations of a random variable X, the goal of statistical estimation inference is to infer the population distribution of X from observed samples [Figure 13.1.1]. Direct inference on the population distribution using empirical distribution have several drawbacks. The data requirement usually goes up exponentially with the dimensionality. Further, empirical distributions lack the analytical tractability and convenience if when we need to develop further models based on the empirical distribution.

Instead, we often assume the distribution of X has some parametric form(e.g. Gaussian, Binomial,Poisson). More formally, we assume the distribution of X belongs to a family of distributions for X, $\mathcal{P} = \{P_\theta, \theta \in \Omega\}$, where Ω is the parameter space containing all possible values of θ. Note that a statistical model is a hypothesis, which might be correct or incorrect.

With the statistical model proposed, we estimate the model parameter θ from the data. Once estimated, we have a way to describe the distribution of X, which finishes the inference task.

There are two major components in statistical inference: **proposing statistical models** and **estimating model parameters**.

In proposing statistical models, we consider a number of factors including boundedness of the observed data, mathematically convenience, and tolerance of modeling error. For example, Gaussian distributions have well-established properties and modeling capacity.

In estimating model parameters, we will first design a statistic δ, which is a function of random samples $D = \{X_1, ..., X_N\}$, such that δ is closed to our target θ. In this way, we use a statistic to connect observations D to model parameter θ.

How to design a statistic? It turns out that not all the statistics are created equal. Some are biased and some are more efficient in terms of using observed samples to infer target parameters.

In this chapter, we present foundation and principles in designing good statistics with balanced trade-offs, inferring model parameters, and testing hypothesis regarding statistical models.

13.1.2 Statistic and Estimator

13.1.2.1 *Statistic*

Let $X_1, X_2, ..X_n$ denote random samples from a distribution. Let $T = T(X_1, X_2, ..., X_n)$ be a function of these samples. Then T is called a **statistic**. T is also a random variable.

In the statistical estimation tasks, statistics are design in a way such that their means are model parameters to be estimated. To start with, the most common, and perhaps simplest, statistics are the following. In the subsequent sections, we will discuss a wide variety of statistics specific to different statistical models tasks.

Definition 13.1.1 (common statistic). *Given a random sample $X_1, ..., X_n$ from X, we have following definitions:*

- *Sample mean:*

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

- *Sample variance:*
$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$
- *Sample standard deviation:*
$$S = \sqrt{S^2}.$$

Remark 13.1.1 (another equivalent form of sample variance). Note that $\sum_{i=1}^{n}(X_i - \overline{X})^2$ can also be written by $\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i - X_j)^2$. We have

$$\begin{aligned}\sum_{i=1}^{n}(X_i - \overline{X})^2 &= \sum_{i=1}^{n}(X_i^2 - 2\overline{X}X_i + \overline{X}^2)\\ &= \sum_{i=1}^{n} X_i^2 - 2\overline{X}n\overline{X} + n\overline{X}\\ &= \sum_{i=1}^{n} X_i^2 - n\overline{X}\overline{X},\end{aligned}$$

and

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i - X_j)^2 &= \sum_{i=1}^{n} 2nX_i^2 - \sum_{i=1}^{n}\sum_{j=1}^{n} 2X_iX_j\\ &= \sum_{i=1}^{n} 2nX_i^2 - \sum_{i=1}^{n} 2X_in\overline{X}\\ &= 2n(\sum_{i=1}^{n} X_i^2 - n\overline{X}\overline{X})\end{aligned}$$

13.1.2.2 *Estimator properties*

Given random samples $X_1, ..., X_n$ drawn from a distribution, we often design statistics or estimators $\hat{\theta}(X_1, ..., X_n)$ and use their mean $E[\hat{\theta}]$ as an estimate for parameters θ that describes the distribution.

In terms of estimation quality, the first important characteristic of $\hat{\theta}$, is the estimator's bias. The bias of an estimator $\hat{\theta}$ depicts on average how far $\hat{\theta}$ is from the real value of θ, which we elaborate in the following definition.

Definition 13.1.2 (unbiased estimator). *Let $X_1, X_2, ..X_n$ denote random samples from a distribution. Let $\hat{\theta} = \hat{\theta}(X_1, X_2, ..., X_n)$ be a statistic.*

- *The **bias of an estimator** $\hat{\theta}$ is*

$$Bias(\hat{\theta}) = E[\hat{\theta}] - \theta,$$

  where θ is the true value.
- *If $Bias(\hat{\theta}) = 0$, then estimator $\hat{\theta}$ is said to be **unbiased**, i.e., $E[\hat{\theta}] = \theta$.*

Example 13.1.1 (sample mean estimator is unbiased). The sample mean estimator

$$\hat{\theta} = \overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

is an unbiased estimator of $\theta = E[X_i]$.

Besides the mean, we also need the variance of the estimator as another metric to gauge the quality of an estimator. This is motivated by the following example: In estimating population mean, we can also choose $\hat{\theta} = X_1$, then $\hat{\theta}$, which is an unbiased estimator of θ as well.

However, $\overline{X}$ has a smaller variance than X_1 [a direct result from law of large numbers, Theorem 11.11.1, meaning that $\overline{X}$, as a random variable, has a higher probability of being closer to the true value.

Definition 13.1.3 (variance of an estimator). *The variance of an estimator is defined by*

$$Var[\hat{\theta}] = E[(\hat{\theta} - \theta)^2].$$

*If $\hat{\theta}$ is an **unbiased** estimator, then*

$$Var[\hat{\theta}] = E[\hat{\theta}^2] - \theta^2.$$

Example 13.1.2. Continue the mean estimation example. We assume $Var[X_1] = \sigma^2$. The variance of the estimator $\overline{X}$ is given by

$$Var[\overline{X}] = \sigma^2/n \leq Var[X_1].$$

Example 13.1.3. Suppose that $Y_1, \ldots, Y_n$ are a random sample from a Uniform $(0, \theta)$ distribution where $\theta > 0$ is a parameter. Consider the estimator

$$\hat{\theta}_1 = 2\overline{Y}.$$

We can show that $E[2\overline{Y}] = 2 \cdot \frac{\theta}{2} = \theta$, $Var[Y_i] = \theta^2/12$ for all i, and

$$Var[\overline{\theta}] = Var[2\overline{Y}] = \frac{4}{n} Var[Y_1] = \frac{\theta^2}{3n}.$$

The mean and variance metric further can be unified using mean squared error (MSE) in the following. In general, an estimators will smaller MSE are preferred.

Definition 13.1.4 (mean squared error of an estimator). *The mean squared error(MSE) of an estimator is defined by*

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2].$$

13.1.2.3 *Variance-bias decomposition*

In previous section, we introduce MSE to measure the average squared difference between the estimator $\hat{\theta}$ and the parameter θ. Apart from its simplicity, MSE also has critical connections with bias and variance.

Theorem 13.1.1 (variance bias decomposition). *The MSE of an estimator is related to its variance and bias via*

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] = Var[\hat{\theta}] + (Bias(\hat{\theta}))^2$$

where $Var[\hat{\theta}] = E[(\hat{\theta} - E[\hat{\theta}])^2]$. Particularly, if the estimator is unbiased (i.e. $Bias(\hat{\theta}) = 0$), we have

$$MSE(\hat{\theta}) = Var[\hat{\theta}]$$

Proof. Make $(\hat{\theta} - \theta) = (\hat{\theta} - E[\hat{\theta}] + E[\hat{\theta}] - \theta)$ and note that $\hat{\theta}$ is a random variable. Specifically,

$$\begin{aligned}
MSE(\hat{\theta}) &= E[(\hat{\theta} - \theta)^2] \\
&= E[((\hat{\theta} - E[\hat{\theta}] + E[\hat{\theta}] - \theta))^2] \\
&= E[((\hat{\theta} - E[\hat{\theta}])^2] + 2E[((\hat{\theta} - E[\hat{\theta}])(E[\hat{\theta}] - \theta))] + E[(E[\hat{\theta}] - \theta))^2] \\
&= Var[\hat{\theta}] + 2E[((\hat{\theta} - E[\hat{\theta}])](E[\hat{\theta}] - \theta)) + Bias[\hat{\theta}]^2 \\
&= Var[\hat{\theta}] + 0 + Bias[\hat{\theta}]^2
\end{aligned}$$

□

Remark 13.1.2 (bias can be useful). At first glance, it may seem that bias is always undesired. However, biased estimator might have smaller variance [Figure 13.1.2]. As a consequence, biased estimator can have smaller MSE than unbiased estimator. Also consider the following example.

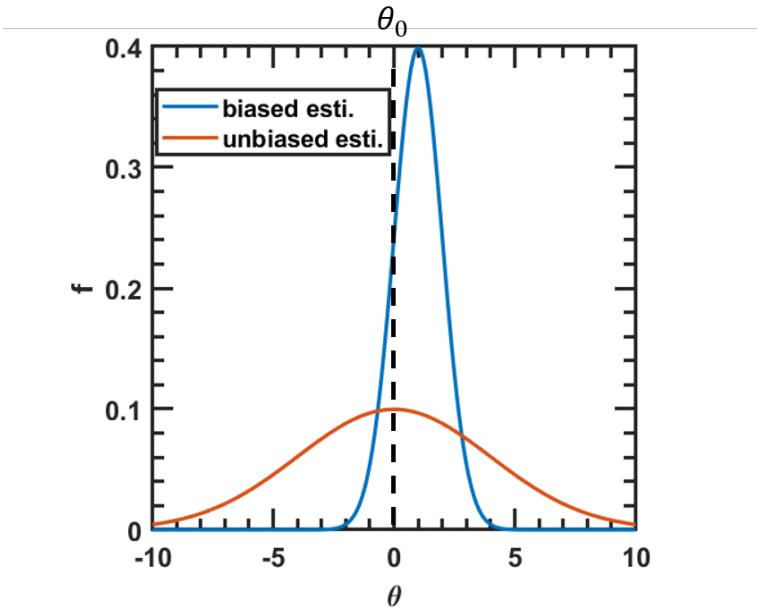


Figure 13.1.2: An example of biased estimator with smaller variance than unbiased estimator

Example 13.1.4. Consider a sample $X_1, X_2, ..., X_n$ of iid normal random variable with unknown mean and variance. Consider two variance estimator

$$S_1^2 = \frac{\sum_{i=1}^n (X_i - \overline{X})}{n-1}, S_2^2 = \frac{\sum_{i=1}^n (X_i - \overline{X})}{n}.$$

Then

- The MSE for S_1^2 is

$$\begin{aligned} MSE[S_1^2] &= Var[S_1^2] + [Bias]^2 \\ &= \frac{2\sigma^4}{n-1} + 0 = \frac{2\sigma^4}{n-1} \end{aligned}$$

  where we use the fact that

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1),$$

from Theorem 12.4.2 and

$$Var[\frac{(n-1)S^2}{\sigma^2}] = \frac{(n-1)^2}{\sigma^4} Var[S^2] = 2(n-1),$$

where use $Var[\chi^2(n)] = 2n$ in Lemma 12.1.31.

- The MSE for S_2^2 is

$$\begin{aligned} MSE[S_2^2] &= (Var[S_2^2] + [Bias])^2 \\ &= \frac{2(n-1)\sigma^4}{n} + (E[S_2^2] - \sigma^2)^2 \\ &= \frac{2(n-1)\sigma^4}{n} + (\frac{(n-1)\sigma^2}{n} - \sigma^2)^2 \\ &= \frac{2n-1}{n^2}\sigma^4 \end{aligned}$$

- $MSE[S_2^2] < MSE[S_1^2]$. That is, the maximum-likelihood estimator has smaller MSE than the unbiased estimator.

13.1.2.4 *Consistence*

We also like to investigate the behavior of estimators as the sample size n gets larger. Understanding of this behavior can be particularly useful in designing experiments involving large scale data in the big data era, We say an estimator is consistent if $\hat{\theta}$ converges to the real value θ when sample size approaches infinity. More precisely, we have the following definition.

Definition 13.1.5 (consistent estimator). *We say $\hat{\theta}$ is a **consistent** estimator of θ if $\hat{\theta}$ converges to θ in probability, i.e.,*

$$\lim_{n\to\infty} P(\left|\hat{\theta}(X_1, X_2, ..., X_n) - \theta\right| < \epsilon) = 1, \forall \epsilon > 0.$$

Example 13.1.5 (sample mean estimator is consistent). Let $X_1, X_2, X_3, \ldots, X_n$ be random samples with the same mean θ, and variance σ^2. We can use Chebyshev's inequality [Theorem 11.9.1] to write

$$P(|\bar{X} - \theta| \geq \epsilon) \leq \frac{Var[\bar{X}]}{\epsilon^2}$$
$$= \frac{\sigma^2}{n\epsilon^2}$$

which goes to 0 as $n \to \infty$.

A more convenient criterion to check if an estimator is consistent is via the following MSE criterion.

Theorem 13.1.2 (MSE criterion for consistent estimator). *An estimator $\hat{\theta}$ is consistent if*

$$\lim_{n\to\infty} MSE(\hat{\theta}(X_1, X_2, ..., X_n)) = 0,$$

*In particular, an **unbiased** estimator $\hat{\theta}$ is consistent if*

$$\lim_{n\to\infty} Var(\hat{\theta}(X_1, X_2, ..., X_n)) = 0,$$

Proof. Overall, we can use Theorem 11.10.3 (convergence in mean square implies convergence in probability). Specifically, we have

$$P\left(\left|\hat{\theta}_n - \theta\right| \geq \epsilon\right) = P\left(\left|\hat{\theta}_n - \theta\right|^2 \geq \epsilon^2\right)$$
$$\leq \frac{E\left[\hat{\theta}_n - \theta\right]^2}{\epsilon^2} \quad \text{(by Markov's inequality)}$$
$$= \frac{MSE\left(\hat{\theta}_n\right)}{\epsilon^2}$$

which goes to 0 as $n \to \infty$ and $MSE \to 0$ by the assumption. □

Remark 13.1.3 (consistence vs. bias)**.**

- A consistent estimator is at least **asymptotically unbiased**. However, some unbiased estimators can be inconsistent (i.e. the variance does not converge to 0), say X_1 as the mean estimator.
- If the sample size is large, consistent estimators are considered better than unbiased estimators because consistent estimators ensure that estimator variance goes to sufficiently smaller when sample size is large.

- Inconsistent estimators usually should be avoided, since increasing the number of samples will not necessarily reduce the variance.

13.1.2.5 *Efficiency*

We now introduce the concept of **efficiency** of an estimator. Essentially, we characterize an efficient estimator by the fact that **given a fixed number of samples**, a more efficient estimator has a lower MSE/variance. Further, the **relative efficiency** of two **unbiased** estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ is the ratio of their variance

$$\frac{Var[\hat{\theta}_1]}{Var[\hat{\theta}_2]}.$$

Example 13.1.6 (sample mean is the most efficient linear estimator of population mean). Consider the linear estimator

$$\widehat{\theta}_n = \sum_{i=1}^{n} a_i X_i$$

where $E[X_i] = \theta, Var[X_i] = \sigma^2$ for $1 \leq i \leq n$.

$$E\left[\widehat{\theta}_n\right] = \sum_{i=1}^{n} a_i E[X_i] = \theta \sum_{i=1}^{n} a_i,$$

so the estimator is unbiased provided

$$\sum_{i=1}^{n} a_i = 1.$$

For i.i.d. random variables,

$$Var\left[\widehat{\theta}_n\right] = \sum_{i=1}^{n} a_i^2 \sigma^2.$$

Now from constrained optimization theory, we know that minimum is achieved iff $a_i = \frac{1}{n}$, $1 \leq i \leq n$. The conclusion then is that, if $\widehat{\theta}_n$ is a linear unbiased estimator of the form $\sum_{i=1}^{n} a_i X_i$ and if $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$, then

$$Var\left[\overline{X}\right] \leq Var\left[\widehat{\theta}_n\right].$$

Therefore, among all linear unbiased estimators, $\overline{X}$ *is most efficient* estimator of the population mean.

Remark 13.1.4 (unbiasedness and efficiency). A biased estimator with a small variance may be more useful than an unbiased estimator with a large variance.

13.1.2.6 *Robust statistics*

Robust statistics are statistics that resilient to sample outliers, samples that are roughly significantly distinct from the majority of the samples. To pave the way for the discussion on robust statistics, we first introduce the concept of **breakdown point** for an estimator. The finite sample **breakdown point** of an estimator is the smallest fraction α of data points such that if $[n\alpha]$ points approach ∞, then the estimator approach ∞.

Given sample size n, the breakdown point for sample mean estimator using the arithmetic mean is $1/n$; that is one point can ruin the mean. To see this, we have

$$\begin{aligned}\overline{x}_n &= \frac{1}{n}\sum_{i=1}^{n} x_i \\ &= \frac{1}{n}(\sum_{i=1}^{n} x_i + x_n) \\ &= \frac{n-1}{n}(\overline{x}_{n-1}) + \frac{1}{n}x_n\end{aligned}$$

Suppose $x_1, ..., x_{n-1}$ are well-behaving points. When x_n, as one outlier, approaches to ∞, $\overline{x}_n$ will approach ∞. As a comparison, the sample median, as an estimate of a population median, can tolerate up to 50% bad values, i.e., its breakdown point is 0.5.

In the following, we introduce some robust estimators for mean and variance.

Definition 13.1.6 (α trimmed mean). *Let $k = n\alpha$ rounded to an integer(k is the number of observation removed from both ends for calculation). The* ***α-trimmed mean*** *is defined as*

$$\overline{X}_\alpha = \sum_{i=k+1}^{n-k} \frac{X_i}{n-2k}.$$

Definition 13.1.7 (median absolute deviation). *[1, p. 122] A robust estimator of standard deviation of iid random sample $X_1, X_2, ..., X_n$ is the* ***MAD(median absolute deviation)***

$$\hat{\sigma}^{MAD} = 1.4826 \times median\{|X_i - median(X_i)|\}.$$

Remark 13.1.5 (interpretation).

- For normally distributed data, $median\{|X_i - median(Y_i)|\}$ is the estimator of $\Phi^{-1}(0.75)\sigma = \sigma/1.4826$.
- For an iid normal random sample, as sample size $n \to \infty$, the MAD is the unbiased estimates of σ.

13.1.3 Method of moments

The method of moments is a straight forward approach to set up the equation to find estimators, although the quality of the found estimators can be of low quality and sovling equations can be troublesome in some cases.

To start with, let $X_1, X_2, ..., X_n$ be a sample from a population with pdf or pmf $f(x|\theta_1, ..., \theta_k)$. The first k moments of the random samples are given by

$$\begin{aligned} m_1 &= \frac{1}{n}\sum_{i=1}^{n} X_i \\ m_2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 \\ &\cdots \\ m_k &= \frac{1}{n}\sum_{i=1}^{n} X_i^k \end{aligned}$$

From the assumed pdf or pmf $f_X(x;\theta)$ of a random sample, we can derive the theoretical moments as a function of parameter θ. By equating theoretical moments and sample moment, we can solve θ. We elaborate this approach in the following.

Methodology 13.1.1 (method of moments for parameter estimation). *[2, p. 312] Let $X_1, X_2, ..., X_n$ be a random sample of X from a population with pdf or pmf $f(x|\theta_1, ..., \theta_k)$.*

Define $\mu_i = E[X^i], i = 1, 2, ..., k$. *The method of moments is aimed at solving* $\theta_1, \theta_2, ..., \theta_k$ *from the k equations*

$$\begin{aligned} m_1 &= \mu_1(\theta_1, ..., \theta_k) \\ m_2 &= \mu_2(\theta_1, ..., \theta_k) \\ &\cdots \\ m_k &= \mu_k(\theta_1, ..., \theta_k) \end{aligned}$$

where $m_i, i = 1, ..., k$ *are the k sample moments, and* $\mu_i, i = 1, .., k$ *are theoretical samples given by*

$$\mu_i E[X^i] = \int x^i f(x) dx.$$

Example 13.1.7 (estimating normal distribution parameter via method of moments). Suppose $X_1, X_2, ..., X_n$ are iid random samples with distribution $N(\mu, \sigma^2)$. It follows that

- Theoretical moments are $\mu_1 = \mu, \mu^2 + \sigma^2 = \mu_2$.
- The moment of method estimators for (μ, σ) are

$$\hat{\mu} = m_1, \hat{\sigma}^2 = m_2 - m_1^2.$$

Example 13.1.8 (estimating t distribution parameter via method of moments). Suppose $X_1, X_2, ..., X_n$ are iid random variable with $t_v(\mu, \sigma^2), v > 2$. It follows that

- Theoretical moments are $m_1 = \mu, \mu^2 + \sigma^2 \frac{v}{v-2} = m_2$.
- The moment of method estimators for (μ, σ) are

$$\hat{\mu} = m_1, \hat{\sigma}^2 = (m_2 - m_1^2) \frac{v-2}{v}.$$

13.1.4 Maximum likelihood estimation

13.1.4.1 *Basic concepts*

The method of moments is quite dated. Nowadays, the arguably most popular point estimation method is maximum likelihood estimation (MLE). The strengths of MLE include its simplicity, generality and efficiency [subsection 13.2.4]. Its dominance is further promoted by recent progress of numerical software and technology (e.g., automatic

differentiation), which make MLE a viable computational tool for large-scale, complex statistical modeling problems.

In general, Given observations $\boldsymbol{x} = X_1, X_2, ..., X_n$ random samples, MLE is comprised of two steps

- Find the likelihood function $L(\boldsymbol{x}|\theta)$ based on observed samples and assumed parametric distribution. The likelihood function is a function of the parameter θ.
- Find the optimal θ that maximizes the likelihood function L. θ^* is the MLE.

Definition 13.1.8 (likelihood function and MLE). *Assuming a statistical model parameterized by a fixed and unknown θ, the likelihood $L(\boldsymbol{x}|\theta)$ is the probability of the observations $\boldsymbol{x} = (x_1, x_2, ..., x_n)$ of iid random samples $X_1, X_2, ..., X_n$ as a function of θ. It can be written as*

$$L(\boldsymbol{x}|\theta) = \prod_{i=1}^{n} f(X = x_i|\theta)$$

And the corresponding log-likelihood function is defined by

$$\log L(\boldsymbol{x}|\theta) = \sum_{i=1}^{n} f(X = x_i|\theta).$$

A maximum likelihood estimator(MLE) of the parameter θ based on given observations $\boldsymbol{x}$ is

$$\hat{\theta} = \max_{\theta} \log L(\boldsymbol{x}|\theta),$$

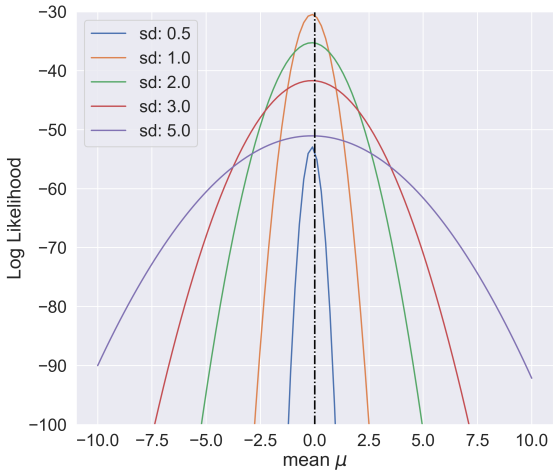
Or alternatively, $\hat{\theta}$ satisfies

$$s(\theta, \boldsymbol{x}) = \frac{\partial \log L}{\partial \theta} = 0,$$

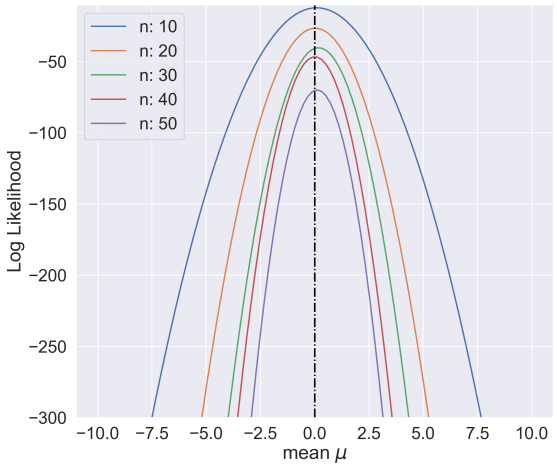
*where $s(\theta, \boldsymbol{x})$ is called **score function**.*

In chapter 12, we examine the log-likelihood functions for samples drawn from a normal distribution $N(0, 1)$. We assume the population distribution is normal governed by parameters μ and σ.

For a set of fixed samples, the log-likelihood function varies with μ and σ [Figure 13.1.3(a)]. Clearly, log-likelihoods achieve peak values when μ is **near** the true value. Log-likelihood is also a function of the sample size n [[Figure 13.1.3(b)]]. For large sample size, log-likelihood will be more sensitive to the parameters, that is, changing more rapidly when parameters change.



(a) Log-likelihood function as a function of μ and θ. Sample size 20.



(b) Log-likelihood function as a function of μ and sample size n. $\sigma = 1$.

Figure 13.1.3: Visualization of log-likelihood function for normal distributed samples.

13.1.4.2 *MLE examples*

Example 13.1.9 (Normal distribution MLE). The log-likelihood function for n iid observations $x_1, ..., x_n$ drawn from normal distribution is given by

$$\begin{aligned}\log L(\theta_1, \theta_2) &= \prod_{i=1}^{n} f(x_i \theta_1, \theta_2) \\ &= \theta_2^{-n/2}(2\pi)^{-n/2} \exp(-\frac{1}{2\theta_2}\sum_{i=1}^{n}(x_i - \theta_1)^2) \\ &= -\frac{n}{2}\log\theta_2 - \frac{n}{2}\log(2\pi) - \frac{\sum_{i=1}^{n}(x_i - \theta_1)^2}{2\theta_2}\end{aligned}$$

where $\theta_1 = \mu, \theta_2 = \sigma^2$. Setting derivatives to zeros, we have

$$\begin{aligned}\frac{\partial \log L(\theta_1, \theta_2)}{\partial \theta_1} &= \frac{\sum_{i=1}^{n}(x_i - \theta_1)}{\theta_2} = 0 \\ \frac{\partial \log L(\theta_1, \theta_2)}{\partial \theta_2} &= -\frac{n}{2\theta_2} + \frac{\sum_{i=1}^{n}(x_i - \theta_1)^2}{\theta_2^2} = 0\end{aligned}$$

which produces

$$\hat{\mu} = \hat{\theta}_1 = \frac{\sum_{i=1}^{n} x_i}{n} = \overline{x}, \hat{\sigma}^2 = \hat{\theta}_2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x}}{n}.$$

Example 13.1.10 (Bernoulli trial MLE). Consider a series of independent Bernoulli trials with success probability θ such that we have probability mass function given by

$$Pr(Y_i = y) = (1-\theta)^{1-y}\theta^y, y \in \{0, 1\}.$$

- The log-likelihood function based on n observations $Y = \{Y_1, ..., Y_N\}$ can be written by

  $$\log L(\theta; Y) = \sum_{i=1}^{n}((1-y_i)\log(1-\theta) + y_i \log\theta) = n((1-\overline{y})\log(1-\theta) + \overline{y}\log(\theta)),$$

  where $\overline{y}$ is the sample mean.
- The MLE is given by

  $$\hat{\theta} = \overline{y}.$$

Example 13.1.11 (exponential distribution MLE). Consider an exponential distribution with parameter α such that its pdf is given by

$$f(x;\alpha) = \alpha e^{-\alpha x}, x \geq 0.$$

The MLE for α from an iid random sample $X_1, ..., X_n$ is given by $\hat{\alpha} = 1/\overline{X}$ since

$$\log L(\alpha) = n \log \alpha - \alpha \sum_{i=1}^{n} X_i$$
$$\partial \log L(\alpha)/\partial \alpha = \frac{n}{\alpha} - \sum_{i=1}^{n} X_i$$
$$\partial \log L(\alpha)/\partial \alpha = 0 \implies \hat{\alpha} = 1/\overline{X}.$$

Example 13.1.12. Consider an exponential distribution with parameter α such that its pdf is given by

$$f(x;\alpha) = \alpha e^{-\alpha x}, x \geq 0.$$

- The MLE for α from an iid random sample $X_1, ..., X_n$ is given by $\hat{\alpha} = 1/\overline{X}$ since

$$\log L(\alpha) = n \log \alpha - \alpha \sum_{i=1}^{n} X_i$$
$$\partial \log L(\alpha)/\partial \alpha = \frac{n}{\alpha} - \sum_{i=1}^{n} X_i$$
$$\partial \log L(\alpha)/\partial \alpha = 0 \implies \hat{\alpha} = 1/\overline{X}.$$

- The Fisher information is given by $I(\alpha) = \frac{1}{\alpha^2}$ since

$$\log f(x;\alpha) = \log \alpha - \alpha x$$
$$\partial^2 \log L(\alpha)/\partial \alpha^2 = -\frac{1}{\alpha^2}$$
$$I(\alpha) = -E[\partial^2 \log L(\alpha)/\partial \alpha^2] = \frac{1}{\alpha^2}.$$

- The MLE $\hat{\alpha} = 1/\overline{X}$ is asymptotic normal and

$$\sqrt{n}(\hat{\alpha} - \alpha_0) \to N(0, \alpha_0^2).$$

Example 13.1.13. Let $X_1, X_2, X_3, \ldots, X_n$ be a random sample from a U niform $(0, \theta)$ distribution, where θ is unknown. Find the maximum likelihood estimator (MLE) of θ based on this random sample. Solution If $X_i \sim$Uniform$(0, \theta)$, then

$$f_X(x) = \begin{cases} \frac{1}{\theta} & 0 \leq x \leq \theta \\ 0 & \text{otherwise} \end{cases}$$

The likelihood function is given by

$$\begin{aligned} L(x_1, x_2, \cdots, x_n; \theta) &= f_{X_1 X_2 \cdots X_n}(x_1, x_2, \cdots, x_n; \theta) \\ &= f_{X_1}(x_1; \theta) f_{X_2}(x_2; \theta) \cdots f_{X_n}(x_n; \theta) \\ &= \begin{cases} \frac{1}{\theta^n} & 0 \leq x_1, x_2, \cdots, x_n \leq \theta \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

Note that $\frac{1}{\theta^n}$ is a decreasing function of θ. Thus, to maximize it, we need to choose the smallest possible value for θ. For $i = 1, 2, \ldots, n$, we need to have $\theta \geq x_i$. Thus, the smallest possible value for θ is

$$\hat{\theta}_{ML} = \max(x_1, x_2, \cdots, x_n)$$

Therefore, the MLE can be written as

$$\hat{\theta}_{ML} = \max(X_1, X_2, \cdots, X_n)$$

Note that this is one of those cases wherein $\hat{\theta}_{ML}$ cannot be obtained by setting the derivative of the likelihood function to zero. Here, the maximum is achieved at an endpoint of the acceptable interval.

13.1.4.3 *Bias and consistence of MLE*

An MLE could be biased. For example, the ML variance estimator for normal samples $X_1, X_2, ..., X_n$ is

$$S^2_{ML} = \frac{\sum_{i=1}^n (X_i - \overline{X})}{n},$$

which is biased, as opposed to the unbiased sample variance estimator

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

However, the ML variance estimator has a smaller variance, as showed in the following example.

Example 13.1.14 (ML variance estimator has a smaller variance). Consider samples $X_1, X_2, ..., X_n$ of iid normal random variable with unknown mean and variance. Consider two variance estimator

$$S_1^2 = \frac{\sum_{i=1}^n (X_i - \overline{X})}{n-1}, S_2^2 = \frac{\sum_{i=1}^n (X_i - \overline{X})}{n}.$$

Then

- The MSE for S_1^2 is

$$\begin{aligned} MSE[S_1^2] &= Var[S_1^2] + [Bias]^2 \\ &= \frac{2\sigma^4}{n-1} + 0 = \frac{2\sigma^4}{n-1} \end{aligned}$$

  where we use the fact that

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1),$$

  from Theorem 12.4.2 and

$$Var[\frac{(n-1)S^2}{\sigma^2}] = \frac{(n-1)^2}{\sigma^4} Var[S^2] = 2(n-1),$$

  where use $Var[\chi^2(n)] = 2n$ in Lemma 12.1.31.
- The MSE for S_2^2 is

$$\begin{aligned} MSE[S_2^2] &= (Var[S_2^2] + [Bias])^2 \\ &= \frac{2(n-1)\sigma^4}{n} + (E[S_2^2] - \sigma^2)^2 \\ &= \frac{2(n-1)\sigma^4}{n} + (\frac{(n-1)\sigma^2}{n} - \sigma^2)^2 \\ &= \frac{2n-1}{n^2}\sigma^4 \end{aligned}$$

- $MSE[S_2^2] < MSE[S_1^2]$. That is, the maximum-likelihood estimator has smaller MSE than the unbiased estimator.

Despite its possible bias for finite samples, one important property of MLEs is that as sample size $n \to \infty$, they would converge to the true value. We say that they are asymptotically unbiased, or consistent.

Theorem 13.1.3 (consistence of MLE). *Let $\hat{\theta}$ be the MLE of coefficient associated with distribution $f(x;\theta)$. Let θ_0 be the true value of the parameter. Let $I_1(\theta)$ be the Fisher information matrix associated with distribution $f(x;\theta)$. It follows that MLEs are consistent; that is*

$$\operatorname*{plim}_{n\to\infty} \hat{\theta} = \theta_0.$$

Proof. (sketch) Define the log-likelihood function associated with n samples

$$L_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log f(X_i|\theta).$$

Denote MLE $\hat{\theta}_n = \arg\max_\theta L_n(\theta)$. Consider the function $L(\theta) = \int (\log f(x;\theta) f(x;\theta_0) dx$, now we can show that the **true parameter θ_0 is the maximizer of** $L(\theta)$; that is, for any θ, we have

$$L(\theta) \le L(\theta_0).$$

$$\begin{aligned}
L(\theta) - L(\theta_0) &= E_{\theta_0}[\log f(X;\theta) - \log f(X;\theta_0)] \\
&= E_{\theta_0}[\log \frac{f(X;\theta)}{f(X;\theta_0)}] \\
&\le E_{\theta_0}[\frac{f(X;\theta)}{f(X;\theta_0)} - 1] \\
&= \int (\frac{f(X;\theta)}{f(X;\theta_0)} - 1) f(x;\theta_0) dx \\
&= \int f(x;\theta) dx - \int f(x;\theta_0) dx \\
&= 1 - 1 = 0
\end{aligned}$$

where we use inequality $\log x \le x - 1$.

From the law of large numbers, $L_n(\theta)$ converge to $L(\theta)$ in probability. Since MLE $\hat{\theta}$ is the maximizer for $L_n(\theta)$, $\hat{\theta}$ converges to θ_0 in probability. □

13.2 Information and efficiency

13.2.1 Fisher information

One central task of statistical estimation is to construct efficient estimators that can extract most information from finite number of samples. In previous section [subsubsection 13.1.2.5], efficiency of an unbiased estimator is characterized by its variance. In this section, we discuss Fisher information framework, which provides a lower bound on the variance of estimators given finite number of samples. The knowledge of lower bound enables us to assert if the estimator we found has the lowest possible variance.

To start with, we discuss Fisher information definitions and its important properties; in the subsequent sections, we examine how Fisher information is utilized to derive the lower bound on variances.

Fisher information requires some reguarlity conditions on the pdf for continuous random variables.

Assumption 13.1 (Fisher information regularity assumption). *For a pdf $f(x;\theta)$ of random variable X with parameter θ. We make the following regularity assumptions:*

- *The set $A = \{x|p(x;\theta) > 0\}$ does not depend on θ. For all $x \in A, \theta \in \Theta$, $\frac{\partial}{\partial\theta}\log p(x;\theta)$ exists and is finite. Here Θ is the parameter space.*
- *If T is any statistic of X such that $E[|T|] < \infty$ for all $\theta \in \Theta$, then integration and differentiation by θ can be interchanged in the following way:*

$$\frac{\partial}{\partial\theta}[\int T(x)f(x;\theta)dx] = \int T(x)\frac{\partial}{\partial\theta}f(x;\theta)dx,$$

whenever the right-hand side is finite.

The definition of Fisher information for discrete or continuous random variables is given by the following.

Definition 13.2.1 (Fisher information). *Given one dimensional parametric family of pdf or pmf $f(x;\theta)$, which is differentiable respect to θ, we define the Fisher information for $\theta \in \mathbb{R}$ as*

$$I(\theta) = E[(\frac{d}{d\theta}\log f(x;\theta))^2],$$

where the expectation is ***taken with respect to*** x*. In particular, if* $\theta \in \mathbb{R}^N$*, we have Fisher information matrix defined by*

$$I(\theta) = E[\frac{\partial \log(f(x;\theta))}{\partial \theta}(\frac{\partial \log(f(x;\theta))}{\partial \theta})^T].$$

The derivative $\frac{d}{d\theta}\log f(x|\theta)$ is known as the **score function**, which characterize the sensitivity of f with respect to θ at a particular θ. Then intuitively, the Fisher information measures the overall on average sensitivity.

How the sensitivity will affect the variance of the estimator? Suppose Fisher information is large, so the parametric distribution will change rapidly when parameters change and be quite different from the true distribution. The large difference can be particularly useful in designing numerical algorithms to search for the true parameters. Conversely, if the Fisher information is small and the distribution is insensitive to model parameters, it would be difficult to estimate the true parameters.

Theorem 13.2.1 (basic properties of Fisher information). *Let* $f(x;\theta)$ *be a pdf parameterized by* $\theta \in \mathbb{R}$ *with Assumption 13.1 holds, then*

- *Score function is zero mean:*

$$E[\frac{d}{d\theta}\log f(x;\theta)] = 0$$

- *Fisher information is the variance of the score function:*

$$I(\theta) = Var[\frac{d}{d\theta}\log f(x;\theta)].$$

- *Further assume* $f(x;\theta)$ *is twice differentiable and interchange between integration and differentiation is permitted. Then*

$$I(\theta) = E[(\frac{d}{d\theta}\log f(x;\theta))^2] = -E[\frac{\partial^2 \log(f(x;\theta))}{\partial \theta^2}].$$

- *For* $\theta \in \mathbb{R}^N$*,*

$$I(\theta) = E[\frac{\partial \log(f(x;\theta))}{\partial \theta}(\frac{\partial \log(f(x;\theta))}{\partial \theta})^T] = -E[\frac{\partial^2 \log(f(x;\theta))}{\partial \theta \partial \theta^T}],$$

for each entry in the matrix,

$$I(\theta)_{ij} = -E[\frac{\partial^2 \log(f(x;\theta))}{\partial\theta_i \partial\theta_j}].$$

Proof. (1) The equivalence of these two expressions can be showed as:

$$\begin{aligned} E[\frac{d}{d\theta}\log f(x;\theta)] &= \int \frac{1}{f(x;\theta)}\frac{d}{d\theta}f(x;\theta)f(x;\theta)dx \\ &= \int \frac{d}{d\theta}f(x;\theta)dx \\ &= \frac{d}{d\theta}\int f(x;\theta)dx \\ &= \frac{d}{d\theta}1 \\ &= 0 \end{aligned}$$

(2) Based on definition, we have

$$\begin{aligned} Var[\frac{d}{d\theta}\log f(x;\theta)] &= E[(\frac{d}{d\theta}\log f(x;\theta))^2] - E[\frac{d}{d\theta}\log f(x;\theta)]^2 \\ &= E[(\frac{d}{d\theta}\log f(x;\theta))^2] - 0 \\ &= I(\theta) \end{aligned}$$

(3)

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

Take expectation with respect to x on both sides and note that

$$E[\frac{1}{f(x;\theta)}\frac{\partial^2}{\partial\theta^2}f(x;\theta)] = \int \frac{\partial^2}{\partial\theta^2}f(x;\theta)dx = \frac{\partial^2}{\partial\theta^2}\int f(x;\theta)dx = \frac{\partial^2}{\partial^2\theta}1 = 0.$$

□

Since our ultimate goal is to estimate parameters from a set of random samples, it is beneficial to distinguish between Fisher information associated with the distribution of

a random variable and Fisher information associated with the joint distribution of a set of iid random samples. Let X denote a single random variable, and let X^n denote N iid random samples. Then based on iid assumption, we have

$$I_{X^n}(\theta) = nI_X(\theta),$$

where we use the fact that $f_{X^n} = [f_X]^n$. For simplicity, we might also use $I_1(\theta)$ and $I_n(\theta)$ to distinguish them in the following sections.

Example 13.2.1 (Fisher information for Bernoulli distribution). Let the pmf of Bernoulli distribution be $f(x;\theta) = \theta^x(1-\theta)^{1-x}, x \in \{0,1\}$. Then

$$I(\theta) = \frac{1}{\theta(1-\theta)}.$$

To see this, we have

$$\begin{aligned} I(\theta) &= -E[\frac{\partial^2 \log(f(x;\theta))}{\partial\theta^2}] \\ &= E[\frac{x}{\theta^2} + \frac{1-x}{(1-\theta)^2}] \\ &= \theta(1/\theta^2) + (1-\theta)(1/(1-\theta)^2) \\ &= \frac{1}{\theta(1-\theta)} \end{aligned}$$

Example 13.2.2 (Fish information matrix for univariate normal distribution). Let the pdf of normal distribution parameterized by

$$f(x;\theta) = (2\pi\theta_2)^{-1/2}\exp(-\frac{1}{2\theta_2}\sum_{i=1}^{n}(x-\theta_1)^2).$$

$$\begin{aligned} \frac{\partial^2 \log f}{\partial\theta_1^2} &= -\frac{1}{\theta_2} = -\frac{1}{\sigma^2} \\ \frac{\partial^2 \log f}{\partial\theta_2^2} &= \frac{1}{2\theta_2^2} - \frac{1}{\theta_2^3}(x-\theta_1)^2 \\ \frac{\partial^2 \ln f}{\partial\theta_1\partial\theta_2} &= -\frac{1}{\theta_2^2}(x_i-\theta_1) \end{aligned}$$

Finally, take expectation with respect to x and we have

$$I(\theta_1, \theta_2) = \begin{bmatrix} 1/\sigma^2 & 0 \\ 0 & 1/(2\sigma^4) \end{bmatrix}.$$

13.2.2 Cramer-Rao lower bound

13.2.2.1 *Preliminary: information inequality*

Theorem 13.2.2 (information inequality for a statistic). *Let $T(X)$ be any statistic such that $Var[T(X)] < \infty$ for all θ. Denote $E[T(X)]$ by $\phi(\theta)$. Suppose Assumption 13.1 holds and $0 < I(\theta) < \infty$. Then **for all** θ*

$$Var[T(X)] \geq \frac{[\phi'(\theta)]^2}{I(\theta)}.$$

Proof. Based on the Assumption 13.1, ,we have

$$\phi'(\theta) = \int T(x) \frac{\partial}{\partial \theta} f(x;\theta) dx = \int T(x) \frac{\partial \log f(x;\theta)}{\partial \theta} f(x;\theta) dx.$$

Therefore, we can view

$$\phi'(\theta) = E[T(X) \frac{\partial \log f(x;\theta)}{\partial \theta}] = Cov[T(X), \frac{\partial \log f(x;\theta)}{\partial \theta}],$$

since $E[\frac{\partial \log f(x;\theta)}{\partial \theta}] = 0 \implies E[T(X)]E[\frac{\partial \log f(x;\theta)}{\partial \theta}] = 0$ [Theorem 13.2.1].

Using Cauchy-Schwartz inequality [Theorem 11.9.4], we have

$$\left|\phi'(\theta)\right|^2 = Cov[T(X), \frac{\partial \log f(x;\theta)}{\partial \theta}]^2 \leq Var[T(X)] Var[\frac{\partial \log f(x;\theta)}{\partial \theta}].$$

At last, use the fact [Theorem 13.2.1] that $I(\theta) = Var[\frac{\partial \log f(x;\theta)}{\partial \theta}]$, we can get the final result. □

Corollary 13.2.2.1 (information lower bound for general estimators). *Let $T(X)$ be a (generally biased) estimator of θ such that*

$$\phi(\theta) \triangleq E[T(X)] = \theta + \underbrace{b(\theta)}_{bias}.$$

Then

- *the variance of $T(X)$ is*
$$Var[T(X)] \geq \frac{\left|1 + b'(\theta)\right|}{I(\theta)}.$$
- *the MSE of $T(X)$ is*
$$MSE[T(X)] \geq \frac{\left|1 + b'(\theta)\right|}{I(\theta)} + b(\theta)^2.$$

13.2.2.2 *Cramer-Rao lower bound: univariate case*

Theorem 13.2.3 (Cramer-Rao lower bound in univariate estimation). *Let $\hat{\theta}$ be an arbitrary univariate estimator as a function of iid random samples $X_1, ..., X_n$, whose distribution is parameterized by single parameter θ. Let θ_0 be the true value. Then the variance of the estimator $\hat{\theta}$ is bounded by*
$$Var(\hat{\theta}) \geq \frac{(\frac{d}{d\theta} E[\hat{\theta}])^2}{nI_1(\theta_0)},$$
where $I_1(\theta)$ is the Fish information associated with distribution $f(x;\theta)$ and the expectation is taken with respect to x. Particularly, if the estimator $\hat{\theta}$ is unbiased (that is $E[\hat{\theta}] = \theta$), we have
$$Var(\hat{\theta}) \geq \frac{1}{nI_1(\theta_0)}.$$

Proof. Note that the Fisher information $I(\theta)$ associated with the joint distribution of $(X_1, ..., X_n)$ can be expressed by $I(\theta) = nI_1(\theta)$, where $I_1(\theta)$ is the Fisher information associated with $f(x;\theta)$. This is because under iid assumption,
$$E[\log f(x_1, ..., x_n; \theta)] = nE[\log f(x;\theta)].$$
Then use the information inequality [Theorem 13.2.2], we have
$$Var(\hat{\theta}) \geq \frac{(\frac{d}{d\theta} E[\hat{\theta}])^2}{nI_1(\theta_0)}.$$
□

Example 13.2.3 (univariate estimation for normal distributions).

- Consider an unbiased mean estimator $\hat{\mu}$ and an unbiased variance estimator $\hat{\sigma}^2$ for normal distribution with unknown mean μ and variance σ^2. Because

the information matrix is given by Example 13.2.2, the mean estimator has a bounded variance given by

$$Var[\hat{\mu}] \geq \frac{1}{nI_1(\theta)} = \sigma^2/n.$$

- Consider a normal distribution with known mean μ. The variance estimator has a bounded variance given by

$$Var[\hat{\sigma}^2] \geq \frac{1}{nI_1(\theta)} = 2\sigma^4/n.$$

- It is clear that
  - Increasing sample size n will reduce the estimator variance.
  - Mean/variance estimators of random samples drawn from small-variance distributions have inherent smaller variances in their estimators.

13.2.2.3 *Cramer-Rao lower bound: multivariate case*

Theorem 13.2.4 (information inequality for statistic: multivariate case). *Let $T(X)$ be any statistic such that $Var[T(X)] < \infty$ for all θ. Denote $E[T(X)]$ by $\phi(\theta)$. Suppose Assumption 13.1 holds and $0 < I(\theta) < \infty$. Then* ***for all*** *θ*

$$Var[T(X)] \geq [\nabla_\theta \phi]^T [I(\theta)]^{-1} [\nabla_\theta \phi].$$

Proof. Based on the Assumption 13.1, ,we have

$$\phi'(\theta) = \int T(x) \frac{\partial}{\partial \theta} f(x;\theta) dx = \int T(x) \frac{\partial \log f(x;\theta)}{\partial \theta} dx.$$

Therefore, we can view

$$\phi'(\theta) = E[T(X) \frac{\partial \log f(x;\theta)}{\partial \theta}] = Cov[T(X), \frac{\partial \log f(x;\theta)}{\partial \theta}],$$

since $E[\frac{\partial \log f(x;\theta)}{\partial \theta}] = 0 \implies E[T(X)]E[\frac{\partial \log f(x;\theta)}{\partial \theta}] = 0.$

Using Cauchy-Schwartz inequality [Theorem 11.9.4], we have

$$\left|\phi'(\theta)\right|^2 = Cov[T(X), \frac{\partial \log f(x;\theta)}{\partial \theta}]^2 \leq Var[T(X)] Var[\frac{\partial \log f(x;\theta)}{\partial \theta}].$$

At last, use the fact [Theorem 13.2.1] that $I(\theta) = Var[\frac{\partial \log f(x;\theta)}{\partial \theta}]$, we can get the final result. □

Proof. Similar to Theorem 13.2.4, we can show that

$$\frac{\partial \phi(\theta)}{\partial \theta_j} = Cov(T(X), \frac{\partial \log f(x;\theta)}{\partial \theta_j}).$$

For constants $c_1, c_2, ..., c_p$, note that

$$Var[T(X) - \sum_{j=1}^{p} c_j \frac{\partial \log f(x;\theta)}{\partial \theta_j}] = Var[T(X)] + c^T I(\theta) c - 2c^T [\nabla_\theta \phi]$$
$$\geq 0$$

Particularly, the minimum is achieved at $c* = [I(\theta)]^{-1} \nabla_\theta \phi$. Then

$$Var[T(X) - \sum_{j=1}^{p} c_j \frac{\partial \log f(x;\theta)}{\partial \theta_j}] = Var[T(X)] - [\nabla_\theta \phi]^T [I(\theta)]^{-1} [\nabla_\theta \phi] \geq 0.$$

□

Theorem 13.2.5 (Cramer-Rao lower bound in multivariate estimation). *Let $\hat{\theta}$ be a p-dimension **unbiased** estimator as a function of iid random samples $X_1, ..., X_n$, whose distribution is parameterized by parameter vector $\theta \in \mathbb{R}^p, \theta = (\theta_1, ..., \theta_p)$. Let θ_0 be the true value.*

Then the variance matrix of the estimator $\hat{\theta}_i$ is bounded by

$$Var(\hat{\theta}) \geq [n[I_1(\theta_0)]^{-1},$$

where $I_1(\theta_0)$ is the Fish information maxtrix associated with distribution $f(x;\theta)$ and the expectation is taken with respect to x.

Proof. Note that the Fisher information $I(\theta)$ associated with the joint distribution of $(X_1, ..., X_n)$ can be expressed by $I(\theta) = nI_1(\theta)$, where $I_1(\theta)$ is the Fisher information associated with $f(x, \theta)$. This is because under iid assumption,

$$E[\log f(x_1, ..., x_n; \theta)] = nE[\log f(x;\theta)].$$

Then use the information inequality [Theorem 13.2.4], we have

$$Var(\alpha^T \hat{\theta}) = \alpha^T Var[\hat{\theta}] \alpha \geq [\nabla_\theta \alpha^T \hat{\theta}]^T [nI_1(\theta)]^{-1} [\nabla_\theta \alpha^T \hat{\theta}] = \alpha^T [nI_1(\theta)]^{-1} \alpha,$$

where $\alpha \in \mathbb{R}^p$ is an arbitrary vector. □

Example 13.2.4 (multivariate estimation for normal distributions).

- Consider an unbiased mean estimator $\hat{\mu}$ for normal distribution with known variance σ^2. The information matrix is given by (Example 13.2.2)

$$I(\theta_1, \theta_2) = \begin{bmatrix} 1/\sigma^2 & 0 \\ 0 & 1/(2\sigma^4) \end{bmatrix}.$$

  Therefore,

$$Var[\hat{\mu}] \geq \sigma^2/n, Var[\hat{\sigma}^2] \geq 2\sigma^4/n,$$

- It is clear that
  - Increasing sample size n will reduce the estimator variance.
  - Mean/variance estimators of random samples drawn from small-variance distributions have inherent smaller variances.

13.2.3 Efficient estimators

With the lower bound on the variance of an estimator, we can define **efficient estimators**. An efficient estimator is optimal in the sense of using information to reduce uncertainty.

Definition 13.2.2 (efficient estimator). *An unbiased estimator $\hat{\theta}$ if variance achieves equality in the Cramer Rao lower bound for all $\theta \in \Theta$.*

The Cramer-Rao lower-bound enables us to judge whether one estimator is efficient: the closer to lower bound, the more efficient. In practice, efficient estimators are usually difficult to find. Efficient estimator is also called a **uniformly minimum-variance unbiased estimator** (UMVUE).

Example 13.2.5. Let the pmf of Bernoulli distribution parameterized by $f(x;\theta) = \theta^x(1-\theta)^{1-x}, x \in \{0,1\}$. Then

$$\begin{aligned} I(\theta) &= -E[\frac{\partial^2 \log(f(x;\theta))}{\partial\theta^2}] \\ &= E[\frac{x}{\theta^2} + \frac{1-x}{(1-\theta)^2}] \\ &= \theta(1/\theta^2) + (1-\theta)(1/(1-\theta)^2) \\ &= \frac{1}{\theta(1-\theta)} \end{aligned}$$

Consider the estimator $\hat{\theta} = \overline{X}$, then

$$E[\hat{\theta}] = E[\overline{X}] = E[\sum_{i=1}^{n} X_i/n] = \theta$$

and

$$Var[\hat{\theta}] = E[\overline{X}^2] - E[\hat{\theta}]^2 = \theta/n + \theta^2(n-1)/n - \theta^2 = \theta(1-\theta)/n$$

Therefore, the variance of the estimator is achieving the lower bound and therefore efficient.

Is UMVUE always desirable? UMVUE restricts estimator to be unbiased. However, in practice, there are many biased estimators that have smaller MSE than UMVUE.

13.2.4 Asymptotic normality and efficiency of MLE

So far, we have addressed the asymptotic unbiasedness, or consistence, of MLEs [Theorem 13.1.3], with the new tool in Fisher information, we would like to examine the asymptotic efficiency of MLEs

Theorem 13.2.6 (asymptotic normality of MLE). *[3, p. 553][4, p. 478] Let $\hat{\theta}$ be the MLE of coefficient associated with distribution $f(x;\theta)$. Let θ_0 be the true value of the parameter. Let $I_1(\theta)$ be the Fisher information matrix associated with distribution $f(x;\theta)$. It follows that MLEs are asymptotic normal; that is, in distribution,*

$$\sqrt{n}(\hat{\theta} - \theta_0) \to MN(0, [I_1(\theta_0)]^{-1}),$$

as $n \to \infty$.

It is easy see that MLE is **asymptotically efficient** because its asymptotic variance reaches the Cramer-Rao lower bound.

13.3 Sufficiency and data reduction

13.3.1 Sufficient estimators

When we estimate a model parameter θ, **not all the information in the data are relevant** to the estimation procedure. For example, if we want to estimate the mean, then the order of the sample is irrelevant. A **sufficient statistic** for a model parameter θ represents the **summary of all information from the data that are useful** for estimation of θ.

Definition 13.3.1 (sufficient statistics). *Let X be a random sample of size n. A statistic $T(X)$ is a sufficient statistic for θ if the conditional distribution of the sample X given the value of $T(X)$ does not depend on θ; that is*

$$P(X|T,\theta) = P(X|T)$$

In otherwise, X and θ are conditional independent given T.

Remark 13.3.1 (sufficient statistic as a lossless data compression). A statistic is sufficient means that $T(X)$ itself can capture all the information useful in estimating θ; the sample X might contain more information than $T(X)$(since $T(X)$ is usually not 1-1), but this additional information does not provide additional usefulness in estimating θ.

Example 13.3.1 (**trivial sufficient statistic**). The statistic $T(X_1, ..., X_n) = (X_1, ..., X_n)$ is always sufficient for any estimation task.

Example 13.3.2. Suppose $X_1, X_2 \sim B(n, \theta)$ and consider

$$P\left(X_1 = x | X_1 + X_2 = r\right)$$

$$= \frac{P\left(X_1 = x, X_1 + X_2 = r\right)}{P\left(X_1 + X_2 = r\right)}$$

$$= \frac{P\left(X_1 = x, X_2 = r - x\right)}{P\left(X_1 + X_2 = r\right)}$$

$$= \frac{\binom{n}{x}\theta^x (1-\theta)^{n-x}\binom{n}{r-x}\theta^{r-x}(1-\theta)^{n-r+x}}{\binom{2n}{r}\theta^r (1-\theta)^{2n-r}}$$

$$= \frac{\binom{n}{x}\binom{n}{r-x}}{\binom{2n}{r}}.$$

This does not contain θ, so that $X_1 + X_2$ is a sufficient statistic for θ.

13.3.2 Factorization theorem

Theorem 13.3.1 (Neyman-Fisher Factorization theorem). *[2, p. 276] Let $f(\boldsymbol{x}|\theta)$ denote the joint pdf or pmf of a sample $\boldsymbol{X}$. A statistic $T(\boldsymbol{X})$ is a sufficient statistic for θ if and only if there exist functions $g(T(\boldsymbol{x})|\theta))$ such that for all sample points $\boldsymbol{x} \in \mathcal{X}$ and all parameter points $\theta \in \Theta$, we have*

$$f(\boldsymbol{x}|\theta) = g(T(\boldsymbol{x})|\theta)h(\boldsymbol{x})$$

Proof. (1) Assume $T(\boldsymbol{X})$ is sufficient, then we have $f(\boldsymbol{x}|T(\boldsymbol{x}), \theta) = f(\boldsymbol{x}|T(\boldsymbol{x}))$. Then we have

$$\begin{aligned} f(\boldsymbol{x}|\theta) = f(\boldsymbol{x}|\theta)f(T(\boldsymbol{x})|\boldsymbol{x}, \theta) &= f(\boldsymbol{x}, T(\boldsymbol{x})|\theta) = f(T(\boldsymbol{x})|\theta)f(\boldsymbol{x}|T(\boldsymbol{x}), \theta) \\ &= f(T(\boldsymbol{x})|\theta)f(\boldsymbol{x}|T(\boldsymbol{x}))(\text{use sufficiency}) \\ &= h(\boldsymbol{x})g(T(\boldsymbol{x})|\theta) \end{aligned}$$

(2) Assume the factorization holds. Let $T(\boldsymbol{x}) = a$.

Because $f(\boldsymbol{x};\theta) = g\left(T(\boldsymbol{x})|\theta\right) h(\mathbf{x})$, we have

$$P\left(T(\boldsymbol{X}) = a\right) = \int_{\boldsymbol{y}\in T^{-1}(a)} p(\boldsymbol{y})d\boldsymbol{y} = g(a|\theta) \int_{\boldsymbol{y}\in T^{-1}(a)} h(\boldsymbol{y})d\boldsymbol{y}.$$

Hence

$$P\left(\boldsymbol{X} = \boldsymbol{x} \mid T(\boldsymbol{X}) = a\right) = \frac{h(\boldsymbol{x})}{\int_{\boldsymbol{y}\in s^{-1}(a)} h(\boldsymbol{y})d\boldsymbol{y}}$$

and this does not depend upon θ.

□

Example 13.3.3. $X_1, X_2, \ldots, X_n \sim U(0,\theta)$, so that

$$f(\boldsymbol{x}|\theta) = \theta^{-n}, \quad 0 < x_1, \ldots, x_n < \theta.$$

Or equivalently that

$$f(\boldsymbol{x}|\theta) = \theta^{-n}, \quad \theta > x_{(n)} \triangleq \max_i X_i.$$

We can factorize this as $T(\mathbf{x}) = x_{(n)}$ and $h\left(\mathbf{x}\right) = 1$, so that $X_{(n)}$ is a sufficient statistic for θ.

Example 13.3.4. Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a Bernoulli distribution. Then

$$f(\mathbf{x};\theta) = \theta^{\sum_i x_i}(1-\theta)^{n-\sum_i x_i}.$$

We can factorize this as $T(\mathbf{x}) = \sum_i x_i$ and $h\left(\mathbf{x}\right) = 1$.

Example 13.3.5. Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a $N(\mu,\sigma^2)$ distribution, where $(\mu,\sigma^2)^T$ is a vector of unknown parameters. Then

$$f(\mathbf{x};\mu,\sigma^2) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_i (x_i-\mu)^2\right]$$

$$= \left(2\pi\sigma^2\right)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_i (x_i-\overline{x})^2 + n(\overline{x}-\mu)^2\right].$$

We can factorize this as $T(\mathbf{x}) = \left(\overline{x}, \sum_i (x_i - \overline{x})^2\right)^T$.

13.4 Bayesian estimation theory

13.4.1 Overview

In our previous statistical estimation approach, we have a general setup: we are given samples drawn from a population with unknown distribution; we assume a parametric distribution and estimate the distribution parameters from the samples via properly designed estimators. In this approach, the distribution parameters θ are assumed to be some unknown and non-random quantities. This approach is generally referred to as **frequentist** approach.

In this section, we introduce an alternative approach, known as **Bayesian** approach. In the Bayesian framework, we assume the distribution parameter θ is a random variable following a prior distribution. After observing some samples, we update the distribution of θ and yield the posterior distribution of θ. The update step is usually carried out using Bayes' Rule.

Advantages of this Bayesian approach include

- It provides a natural and principled way of combining prior information with data. Such prior can incorporate expert domain knowledge or act as a form of model regularization.
- It provides uncertainty measure of the estimated parameters. For example, Bayesian approach can offer answers like the true parameter has a probability of 0.9 of falling in an interval.

The downsides of Bayesian approach include

- Additional efforts to select a prior. Mistakenly specified priors can give misleading conclusions.
- Higher computational cost, particularly for models with a large number of parameters. Bayesian approach usually require the usage of computational intensive simulation methods, i.e., Monte Carlo, to perform calculation.

13.4.2 Basics

Let $X_1, \ldots, X_n$ be random samples from a distribution. A Bayesian statistical model is composed of a **data generation model**, $X_i \sim p(x|\theta)$, and a **prior distribution model** on the model parameters, $p(\theta), \theta \in \mathbb{R}^n$. Using Bayes' theorem, the posterior distribution of θ given the data is

$$\pi\left(\theta|X_1, \ldots, X_n\right) = \frac{p\left(X_1, \ldots, X_n|\theta\right)\pi(\theta)}{m\left(X_1, \ldots, X_n\right)}$$

where

$$m\left(X_1,\ldots,X_n\right)=\int p\left(X_1,\ldots,X_n\mid\theta\right)\pi(\theta)d\theta.$$

Because m does not depend on θ, we can write

$$p\left(\theta|X_1,\ldots,X_n\right)\propto L(\theta)\pi(\theta)$$

where $L(\theta)=p\left(X_1,\ldots,X_n\mid\theta\right)$ is the likelihood function [Definition 13.1.8]. The interpretation is that $p\left(\theta|X_1,\ldots,X_n\right)$ represents the update on the subjective beliefs about θ after observing $X_1,\ldots,X_n$.

With posterior distribution, two commonly used estimators are

- Posterior mean estimator

$$\bar{\theta}=\mathbb{E}\left(\theta\mid X_1,\ldots,X_n\right)=\int\theta\pi\left(\theta\mid X_1,\ldots,X_n\right)d\theta=\frac{\int\theta L(\theta)\pi(\theta)}{\int L(\theta)\pi(\theta)}.$$

- Maximum A posterior estimator is given as

$$\hat{\theta}_{MAP}(x)=\arg\max_{\theta}p(\theta|x)=\arg\max_{\theta}p(x|\theta)p(\theta).$$

Example 13.4.1. Let the data generation model be $Bernoulli(\theta)$ and the prior distribution on p be $\theta\sim Beta(\alpha,\beta)$. Let $X=X_1,\ldots,X_n$ be random samples. Then

$$\pi(\theta)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)},\Gamma(\alpha)=\int_0^\infty t^{\alpha-1}e^{-t}dt.$$

Set $Y=\sum_i X_i$. Then

$$\pi(p|X)\propto\underbrace{\theta^Y(1-\theta)^{n-Y}}_{\text{likelihood}}\times\underbrace{\theta^{\alpha-1}(1-\theta)^{\beta-1}}_{\text{prior}}\propto\theta^{Y+\alpha-1}(1-\theta)^{n-Y+\beta-1}.$$

Therefore, $p|X\sim Beta(Y+\alpha,n-Y+\beta)$. The Posterior mean estimator [Lemma 12.1.28] is

$$\hat{p}=\frac{Y+\alpha}{(Y+\alpha)+(n-Y+\beta)}=\frac{Y+\alpha}{\alpha+\beta+n}.$$

In the above example, the prior and posterior distribution belong to the same family. This is an example of a conjugate prior.

Definition 13.4.1 (conjugate prior). *$p(\theta)$ is a conjugate prior for $p(x|\theta)$ if*

$$p(\theta) \in \mathcal{P} \Rightarrow p(\theta|x) \in \mathcal{P}$$

where $\mathcal{P}$ is a family of pdf parameterized by θ. In other words, $p(\theta)$ and $p(x|\theta)$ are in the same family.

It is beneficial to use use conjugate prior. When prior and posterior distribution are in the same family, it is easy to interpret how the observations x changes the prior distribution. For a comprehensive account of conjugate priors, see [5][6].

Example 13.4.2. Let $X_1, \ldots, X_n \sim N\left(\mu, \sigma^2\right)$ with σ^2 known. Let $\mu \sim N\left(m, \tau^2\right)$. Then the posterior distribution is given by

$$p(\mu|X) \propto \exp(\frac{\sum_{i=1}^{N}(\mu - X_i)^2}{2\sigma^2}) \exp(\frac{(\mu - m)^2}{2\tau^2}).$$

The maximum a posterior estimator $\hat{\mu}$ is given by maximizing posterior estimator

$$\hat{\mu} = \frac{\frac{\sum_{i=1}^{N} X_i}{\sigma^2} + \frac{m}{\tau^2}}{\frac{N}{\sigma^2} + \frac{1}{\tau^2}}.$$

13.5 Bootstrap method

Let $\boldsymbol{X} = (X_1, X_2, ..., X_n)$ be a random sample of size n and let $\hat{\theta}(\boldsymbol{X})$ be a statistic of interest. One central task of statistical estimation is characterize the variance of $\hat{\theta}(\boldsymbol{X})$. In simple cases, we might be able to directly derive the distribution of the estimator. For example, let $\boldsymbol{X}$ be random samples of a normal distribution, the sample variance S^2 will have $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$.In complex cases where obtaining standard deviation, confidence interval or even distributions of $\hat{\theta}$ is difficult. The goal of bootstrap methods is to measure the standard deviation, confidence interval, or even distributions of $\hat{\theta}$ by numerical simulation method.

On a high level, a bootstrap method of estimating the variance of an estimator consisting of the following steps

- Draw B bootstrap samples, a bootstrap sample is a set of N sample drawn from the original samples with replacement.
- On each bootstrap sample i, evaluate the estimator $\hat{\theta}(\boldsymbol{X})_i$.
- Estimate the variance of $\hat{\theta}(\boldsymbol{X})$ from $\hat{\theta}(\boldsymbol{X})_i, i = 1, ..., B$.

The intuition of the working mechanism underlying bootstrap method is the fact that we can view the bootstrap sample as a new set of samples drawn from the empirical sample distribution (the joint distribution of $(X_1, ..., X_N)$).

Remark 13.5.1 (resampling property). Given a sample of size n, we re-draw n sample with replacement. Then from probability, we have:

- The probability that ith sample is not being resampled is $(1 - \frac{1}{n})$ at the first time.
- The probability that ith sample is not being resampled is $(1 - \frac{1}{n})^n$ at the new sample of size n.
- The probability that ith sample is not being resampled is e^{-1} at the new sample when $n \to \infty$.
- On average, about ne^{-1} of original samples will not show in the new sample as $n \to \infty$.

Now we can summarize the basic procedure in a bootstrap method.

Methodology 13.5.1 (general bootstrap estimation). *Let $\hat{\theta}$ be a statistic as a function of $(X_1, ..., X_N)$. Let $\hat{\theta}_i$ be the estimation evaluated at bootstrap sample i. Then*

- *The mean estimation*

$$m = \frac{1}{B} \sum_{i=1}^{B} \hat{\theta}_i \approx E[\hat{\theta}].$$

- *The variance estimation is*

$$s = \frac{1}{B}\sum_{i=1}^{B}(\hat{\theta}_i - m)^2 \approx Var[\hat{\theta}].$$

Clearly, there are two sources of error in the bootstrap estimate: The first arises from finite sample size N, and the second arises from finite B. In practice, we usually take as large as possible, say B = 10000 or $\sim N^2$. As $B \to \infty$, $Var[\hat{\theta}]$ will converge.

By properly modifying the variance estimation procedure, we arrive at the following confidence level estimation method.

Methodology 13.5.2 (bootstrap confidence level). *Let $\hat{\theta}$ be a statistic as a function of $(X_1, ..., X_N)$. Let $\hat{\theta}_i$ be the estimation evaluated at bootstrap sample i. Let $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_B$ be sorted. Denote $k_1 = (B \times \frac{\alpha}{2}), k_2 = (B \times (1 - \frac{\alpha}{2}))$. Then $[\hat{\theta}_{k_1}, \hat{\theta}_{k_2}]$ is the α confidence interval such that*

$$Pr(\hat{\theta}_{k_1} \leq \theta \leq \hat{\theta}_{k_2}) = 1 - \alpha.$$

Remark 13.5.2. One variation of bootstrap method is the jackknife where the standard error is estimated by from $N - 1$ leaving-one-out subsamples.

13.6 Hypothesis testing theory

13.6.1 Basics

In statistical modeling of observed data, one may put forward a hypothesis regarding the specification of statistical models, statistical relationship among data or between different group of data, etc.

For example, a principal of a school claims that the students in his school have an average height at 5 feet. Suppose measurement of heights of 100 students, we get average height of 5.5 feet and standard deviation of 0.5 feet. Is there sufficient evidence to conclude the principal's statement?

In a typical hypothesis testing, we usually propose two **hypotheses**: **null hypothesis** and **alternative hypothesis**, denoted as H_0 and H_1. In general, null hypothesis and the alternative hypothesis are **complementary** to each other. Our goal is to determine which one is statistically correct or incorrect.

The null hypothesis is usually a simple hypothesis **the contradiction** to what we would like to prove.The alternative hypothesis is usually a hypothesis what we would like to prove. Alternative hypothesis can be **two-sided or one-sided**.

Example 13.6.1. Consider a clinical trial of a new drug. Treatment data are collected to compare two treatments.
The *null hypothesis* is usually no difference between treatments.

Depending on the purpose the test, the *alternative hypothesis* might be:

- New drug is different from old drug. (*two-sided*)
- New drug is better than old drug.(*one-sided*),
- Old drug is better than new drug. (*two-sided*).

Example 13.6.2. Given observation sampled from a normal distribution. A null hypothesis regarding the mean μ can be $\mu = \mu_0$.

And the alternative hypothesis can be

- $H_1 : \mu > \mu_0$, which is an **upper-tailed one-sided hypothesis**.
- $H_1 : \mu < \mu_0$, which is a **lower-tailed one-sided hypothesis**.
- $H_1 : \mu \neq \mu_0$, which is a **two-sided hypothesis**.

Mathematically, a hypothesis can be viewed as a statement about a population parameter θ. Let θ denote a population parameter. The general format of the null and

alternative hypothesis is $H_0 : \theta \in \Theta_0$, and $H_1 : \theta \in \Theta_1$, where Θ_0 and Θ_1 are disjoint subsets of the parameter space Θ.

Example 13.6.3. A coin is tossed and we hypothesize that it is fair. Hence Θ_0 is the set $\left\{\frac{1}{2}\right\}$ containing just one element of the parameter space $\Theta = [0, 1]$

Given H_0 and H_1, we need to decide which hypothesis to reject or accept, or which partition in Θ the population coefficient θ lies in. From this perspective, we can view hypothesis testing as a decision making problem with uncertainty.

Usually, decision is based on $p(\theta|x)$, the posterior distribution can be calculation as $p(x|\theta \in H_i)$. Given the observation data we can calculate $p(x|H_0)$ and $p(x|H_1)$. We will determine which, H_0 or H_1, is more appropriate, by either comparing $p(x|H_0)$ and $p(x|H_1)$ or specifying the value range of a test statistics T to accept or reject H_0.

A basic framework of hypothesis testing using test statistic can be summarized in the following method.

Methodology 13.6.1 (hypothesis testing via test statistic). *Suppose we are given random samples drawn from a normal distribution with known various σ^2 and unknow mean μ. A typical hypothesis testing on the μ involves the following procedures.*

- *State the Null hypothesis. For example $H0 : \mu = \mu_0$.*
- *State the Alternate Hypothesis. For example, $H1 : \mu > \mu_0$ or $\mu < \mu_0$ or $\mu \neq \mu_0$.*
- *State the significance level α, say $\alpha = 0.05$.*
- *Select the appropriate test statistic. For example,*

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}.$$

- *Determine the rejection region area [Figure 13.6.1]. For test statistic Z and $H_1 : \mu > \mu_0$, one rejection region could be $R = \{Z : Z > \phi^{-1}(1-\alpha)\}$, where ϕ^{-1} is the inverse cdf of Z. Rejection regions for other cases can be determined similarly.*
- *Calculate the test statistic value and accept or reject H_0 based on Z. If $Z \in R$, we reject the null hypothesis.*

The **significance level** α is a probability threshold below which the **null hypothesis will be rejected under the assumption that H_0 is true**. Common values are 0.05 and 0.01.

There are different ways to specify a decision rule. In general, the decision rule depends on whether the test is an upper-tailed, lower-tailed, or two-tailed test[Figure 13.6.1]. In an upper-tailed or lower-tailed test the rejection region is around the upper or lower

tail whereH_0 will will get rejected when the test statistic is larger or smaller than a critical value, respectively. In a two-tailed test, the rejection region is at both tails where H_0 will get rejected if the test statistic is extreme, either larger than an upper critical value or smaller than a lower critical value.

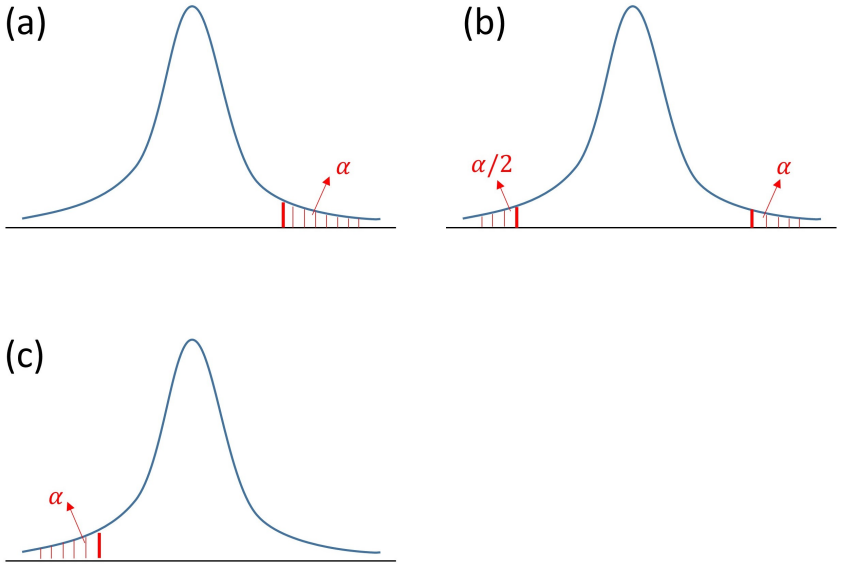


Figure 13.6.1: Demonstration for rejection regions for upper-tailed one-sided hypothesis (a), two-sided hypothesis(b), and lower-tailed one-sided hypothesis (c).

Definition 13.6.1 (p value). *The p value is the probability, assuming the null hypothesis is true, of observing the at least as extreme as (equal to or "more extreme" than) the observed test statistic in the alternative hypothesis direction.*
p value can also be interpreted as the smallest significant value that H_0 will be reject.

Methodology 13.6.2 (p value method). *Given a significance level α:*

- *If $p \leq \alpha$, then reject H_0.*
- *If $p > \alpha$, then accept H_0.*

Example 13.6.4. Consider a hypothesis test of n random samples from normal distribution $N(\mu, \sigma^2)$. Let $H_0 : \mu = \mu_0$. Let the test statistic be

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}.$$

- If $H_1 : \mu > 0$, the $p = 1 - \Phi^{-1}(Z)$. If $p \leq \alpha$, we reject H_0.
- If $H_1 : \mu < 0$, the $p = 1 - \Phi^{-1}(-Z)$. If $p \leq \alpha$, we reject H_0.
- If $H_1 : \mu \neq 0$, the $p = 2(1 - \Phi^{-1}(|Z|))$. If $p \leq \alpha$, we reject H_0.

where Φ is the cdf of a standard normal distribution.

13.6.2 Characterizing errors and power

We usually only test if H_0 is true or not and **do not test** the correctness H_1. For a binary hypothesis testing, there could be four types of results.

Definition 13.6.2 (Four types of results in binary hopothesis testing).

1. *Detection: H_0 true, decide H_0*
2. *False alarm/ **type I error:** H_0 true, we reject H_0, decide H_1.*
3. *Miss/ **type II error:** H_1 true, decide H_0(or H_0 is false, we do not reject H_0.)*
4. *Correctly rejection: H_1 true,decide H_1*

There are two types of possible error.

- A **Type I error** is the error of rejecting the null hypothesis H_0 when H_0 is true.
- A **Type II error** is the error of not rejecting the null hypothesis H_0 when H_0 is false.

We have following summary table.

	H_0 not rejected	H_0 rejected
H_0 true	no error	Type I error
H_0 false	Type II error	no error

We usually denote

$$\alpha = Pr(\text{Type I error}), \beta = Pr(\text{Type II error}).$$

Example 13.6.5 (type I, II error in hypothesis test of a normal distribution). Consider a hypothesis test of n random samples from normal distribution $N(0, \sigma^2)$. Let $H_0 : \mu = 0, H_1 : \mu > 0$. Let the test statistic be

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}.$$

Then

- To calculate type I error, we assume $H_0 : \mu = 0$ is correct, then

$$Pr(\text{type I error}) = Pr(\frac{\bar{X}}{\sigma/\sqrt{n}} > z_\alpha | H_0) = \alpha,$$

  where z_α is defined as $\Phi(z_\alpha) = 1 - \alpha$, $\Phi(z)$ is the cdf of a normal distribution.
- To calculate type II error, we assume $H_1 : \mu > 0$ is correct, then $Pr(\text{type II error}) = Pr(Z < z_\alpha | H_1)$ can be calculated in the following way:

$$\begin{aligned}\beta &= P(\frac{\bar{X}}{\sigma/\sqrt{n}} < z_\alpha) \\ &= P(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_\alpha - \frac{\mu}{\sigma/\sqrt{n}}) \\ &= \Phi(z_\alpha - \frac{\mu}{\sigma/\sqrt{n}})\end{aligned}$$

  There are several critical implications:
  - If μ increases, then β decreases.
  - If n increases, then β decreases.
  - If α increases, then z_α increases and β increases.

Remark 13.6.1 (interpretation on two types of errors).

- Under the null hypothesis H_0, significance level α is the probability measure (i.e., the size) of the rejection region where H_0 will be rejected. α bounds the type I error.
- Given a fixed size of samples, it is generally not possible to minimize both types of error.
- We usually consider type I error to be worse and try to minimize or bound type I error first and then minimize type II error.
- H_0 is usually conservative statement such that reject H_0 when it is true will have **significant bad consequence**.

13.6.3 Power of a statistical test

Hypothesis testing inherently involves two type of test errors, which usually can not be minimized together. In practice, we design hypothesis and choosing signficance level following the principle that

- **Minimize the probability of committing a Type I error.** That, is minimize $\alpha = P(TypeIError)$. Typically, $\alpha \leq 0.1$.
- **Maximize the power, or reduce the type II error** Note that $\beta = P(TypeIIError) = 1 - power$, typically $\beta \leq 0.2$.

In this section, we will give a close look at the statistical power.

Definition 13.6.3 (statistical power of a test). *The power of a hypothesis test is the probability of making the correct decision if the alternative hypothesis is true. That is, the power of a hypothesis test is the probability of rejecting the null hypothesis H_0 is incorrect(or when the alternative hypothesis H_1 is true):*

$$power = P(reject\ H_0|H_1).$$

Let's revisit the previous example. Consider a hypothesis test of n random samples from normal distribution $N(0, \sigma^2)$. Let $H_0 : \mu = 0, H_1 : \mu > 0$. Let the test statistic be

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}.$$

To calculate type II error, we assume $H_1 : \mu > 0$ is correct, then $Pr(\text{type II error}) = Pr(Z < z_\alpha|H_1)$ can be calculated in the following way:

$$\begin{aligned}\beta &= P(\frac{\bar{X}}{\sigma/\sqrt{n}} < z_\alpha) \\ &= P(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_\alpha - \frac{\mu}{\sigma/\sqrt{n}}) \\ &= \Phi(z_\alpha - \frac{\mu}{\sigma/\sqrt{n}})\end{aligned}$$

Since power equals $1 - \beta$, there are several critical implications:

- If μ increases, then β decreases, and power increases
- If n increases, then β decreases, and power increases
- If α increases, then z_α increases, β increases, power decreases

We have the following summary on factors affecting statistical power.

Note 13.6.1 (factors affecting statistical power). Statistical power may depend on a number of factors.

- the statistical significance criterion used in the test, i.e, α.
- the magnitude of the effect of interest in the population, i.e., μ.
- the sample size used to detect the effect, i.e., n.

Power analysis can be used to calculate the minimum sample size required so that one can be reasonably likely to detect an effect of a given size. For example: "how many times do I need to toss a coin to conclude it is unfair?

Example 13.6.6 (calculating required sample size). Following previous example, the power in a normal distribution mean test is given by

$$power = 1 - \Phi(z_\alpha - \frac{\mu}{\sigma/\sqrt{n}}),$$

If we need power to be greater than p_0, then via algebra, we can get

$$n \geq \frac{\sigma^2}{\mu^2}(z_\alpha - \Phi^{-1}(1-p_0)).$$

13.6.4 Common statistical tests

13.6.4.1 *Chi-square goodness-of-fit test*

Theorem 13.6.1 (Pearson's theorem). *Consider r boxes $B_1, ..., B_r$ and throw n balls $X_1, X_2, ..., X_n$ into these boxes independently of each other with probabilities*

$$P(X_1 \in B_1) = p_1, ..., P(X_r \in B_r) = p_r,$$

such that $p_1 + ... + p_r = 1$.

Let v_j be the number of balls in the jth box, i.e. $v_j = \sum_{i=1}^n \mathbf{1}_{X_i = B_j}$.

It follows that

- *The random variable*

$$\frac{v_j - np_j}{\sqrt{np_j}} \to N(0, 1-p_j) \text{ in distribution, as } n \to \infty$$

- *The random vector $Y = (Y_1, Y_2, ..., Y_r)$, $Y_j = \frac{v_j - np_j}{\sqrt{np_j}}$ will converge to $MN(0, \Sigma)$ in distribution, where*

$$\Sigma_{ii} = 1 - p_i, \Sigma_{ij} = -\sqrt{p_i p_j}.$$

- *The random variable*

$$\sum_{j=1}^r \frac{(v_j - np_j)^2}{np_j} \to \chi^2(r-1) \text{ in distribution, as } n \to \infty.$$

Proof. (1) Note that from Bernoulli distribution

$$E[\mathbf{1}(X_1 \in B_j)] = p_j, Var[\mathbf{1}(X_1 \in B_j)] = p_j(1-p_j).$$

By the central limit theorem

$$\frac{v_j - np_j}{\sqrt{np_j(1-p_j)}} \to N(0,1) \text{ in dist } \implies \frac{v_j - np_j}{\sqrt{np_j}} \to N(0,1-p_j) \text{ in dist.}$$

(2)

$$\begin{aligned} E[\frac{v_i - np_i}{\sqrt{np_i}} \frac{v_j - np_j}{\sqrt{np_j}}] &= \frac{1}{n\sqrt{p_i p_j}}(E[v_i v_j] - n^2 p_i p_j) \\ E[v_i v_j] &= E[\sum_{l=1}^{n} \mathbf{1}(X_l \in B_i) \sum_{k=1}^{n} \mathbf{1}(X_k \in B_i)] \\ &= E[\sum_{l=1}^{n} \sum_{k=1, k\neq l}^{n} \mathbf{1}(X_l \in B_i)\mathbf{1}(X_k \in B_i)] \\ &= 2E[\sum_{l=1}^{n} \sum_{k>1}^{n} \mathbf{1}(X_l \in B_i)\mathbf{1}(X_k \in B_i)] \\ &= n(n-1)p_i p_j \\ E[\frac{v_i - np_i}{\sqrt{np_i}} \frac{v_j - np_j}{\sqrt{np_j}}] &= -\sqrt{p_i p_j}. \end{aligned}$$

(3) Note that

$$Y^T Y = Z^T(I - UU^T)Z, Z \in MN(0, I_r), U = (\sqrt{p_1}, \sqrt{p_2}, ..., \sqrt{p_r}),$$

where UU^T is an rank 1 orthogonal projector($U^T U = p_1 + p_2 + \cdots + p_r = 1$).

From the chi-square decomposition theorem [Theorem 12.4.1], we know that $Y^T Y \to \chi^2(r-1).in.dist.$ □

Theorem 13.6.2 (chi-square goodness-of-fit test). *Suppose that we observe an iid sample $X_1, X_2, ..., X_n$ of random variable that take a finite number of values $B_1, B_2, ..., B_r$ with unknown probabilities $p_1, p_2, ..., p_r$. Consider hypotheses*

$$H_0 : p_i = p_i^0, for\ i = 1, 2, ..., r$$
$$H_1 : for\ some\ i, p_i \neq p_i^0$$

and the test statistic

$$T = \sum_{i=1}^{r} \frac{(v_i - np_i^0)^2}{np_i^0},$$

where $v_j = \sum_{i=1}^{n} \mathbf{1}_{X_i = B_j}$.

It follows that

- *If H_0 is true, then $T \to \chi^2(r) - 1$ in dist.*
- *If H_1 is true, then $T \to \infty$, as $n \to \infty$.*
- *The decision rule is reject H_0 if $T > c$ where $c = \inf\{z : F(z) \geq 0.99\}$.*

Proof. (1) From Pearson's theorem [Theorem 13.6.1]. (2) If we write

$$\frac{(v_i - np_i^0)}{\sqrt{np_i^0}} = \sqrt{\frac{p_i}{p_i^0}} \frac{(v_i - np_i)}{\sqrt{np_i^0}} + \sqrt{n} \frac{(v_i - n(p_i - p_i^0)}{\sqrt{np_i^0}},$$

then the second quantity will diverge as $n \to \infty$. □

Note 13.6.2 (p value method for chi-square test)**.** The p-value for a chi-square test is defined as the **tail area above the calculated test statistic**.

For example, consider an experiment with test statistic result

$$T = \sum_{i=1}^{r} \frac{(v_i - np_i^0)^2}{np_i^0}.$$

Then

$$p - value = Pr(\chi^2(r-1) \geq T).$$

Given a significance level α:

- If $p \leq \alpha$, then reject H_0.
- If $p > \alpha$, then accept H_0.

13.6.4.2 *Chi-square test for statistical independence*

Lemma 13.6.1. *link*

Denote

$$p_i = \sum_{j=1}^{c} \frac{O_{ij}}{N}, q_j = \sum_{i=1}^{r} \frac{O_{ij}}{N}, E_{ij} = N p_i q_j$$

$$\sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{i,j} - E_{ij})^2}{E_{ij}} \sim \chi^2(p),$$

where $p = (r-1)(c-1)$.

The hypothesis is given by

- *H_0: U is independent of V;*

- *H_1: there exists an statistical relationship between U and V.*

13.6.4.3 *Kolmogorov-Smirnov goodness-of-fit test*

Definition 13.6.4 (Kolmogorov-Smirnov(KS) goodness-of-fit test). *The Kolmogorov-Smirnov goodness-of-fit test for a random sample of size N has the following elements:*

- *Hypothesis:*
  - *H_0: the data follow a specified **continuous** distribution with cdf $F(t)$.*
  - *H_1: the data do not follow the specified distribution.*
- *For **ascending ordered** sample $Y_1, Y_2, ..., Y_N$. KS test statistic is defined as*

$$D = \max_{1 \le i \le N} (F(Y_i) - \frac{i-1}{N}, \frac{i}{N} - F(Y_i)).$$

- *The significance level α and critical value K_α.*
- *If $D > K_\alpha$, reject H_0.*

Remark 13.6.2 (interpretation and usage).

- The KS test statistic is measuring the distance of proposed distribution F is the empirical cdf given by $(i-1)/N$ and i/N.
- KS test is used for continuous distribution test. For discrete distribution test, see chi-square goodness-of-fit test [Theorem 13.6.2].
- For the KS critical value table, see link.

13.7 Hypothesis testing on normal distributions

Common notations in this sections:

- sample mean $\overline{X}$
- sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - E[X])^2$

13.7.1 Normality test

Before we proceed to hypothesis testing related to normal distributions, we need to review typical methods used to determine if samples are drawn from a normal distribution. Most straight forwards methods are qualitative graphical methods in where we plot the histogram or QQ plots. In QQ plot, we plot the quantiles of the sample against the theoretical quantiles of a standard normal distribution. For sample truly drawn from a normal distribution, we expect the plot is a perfect straight line; Different deviation from a straight line will be observed when the sample distribution is has non-zero excess Kurtosis (heavy tails or not) or skewness [Figure 13.7.1].

Additional quantitative testing methods include Shapiro–Wilk test, Kolmogorov–Smirnov test, Jarque–Bera test, Pearson's chi-squared test Theorem 13.6.1, D'Agostino's K-squared test, etc.

13.7.2 Sample mean with known variance

Consider we have n samples $X_1, ..., X_n$ for a random variable with $N(\mu, \sigma^2)$ with σ^2 known. The hypothesis testing involving the mean can be obtained by using the fact that

$$Z = \frac{\overline{X} - \mu}{\sigma / n}$$

is $N(0, 1)$. We can summarize the test as:

Table 13.7.1: Test on mean with known variance σ^2

H_0	test statistic	H_1	critical(rejection) region
$\mu \leq \mu_0$	$Z = \frac{\overline{X}-\mu}{\sigma/n}$	$\mu > \mu_0$	$z \geq z_\alpha$
$\mu \geq \mu_0$	$Z = \frac{\overline{X}-\mu}{\sigma/n}$	$\mu < \mu_0$	$z \leq -z_\alpha$
$\mu = \mu_0$	$Z = \frac{\overline{X}-\mu}{\sigma/n}$	$\mu \neq \mu_0$	$\lvert z \rvert \geq z_\alpha$

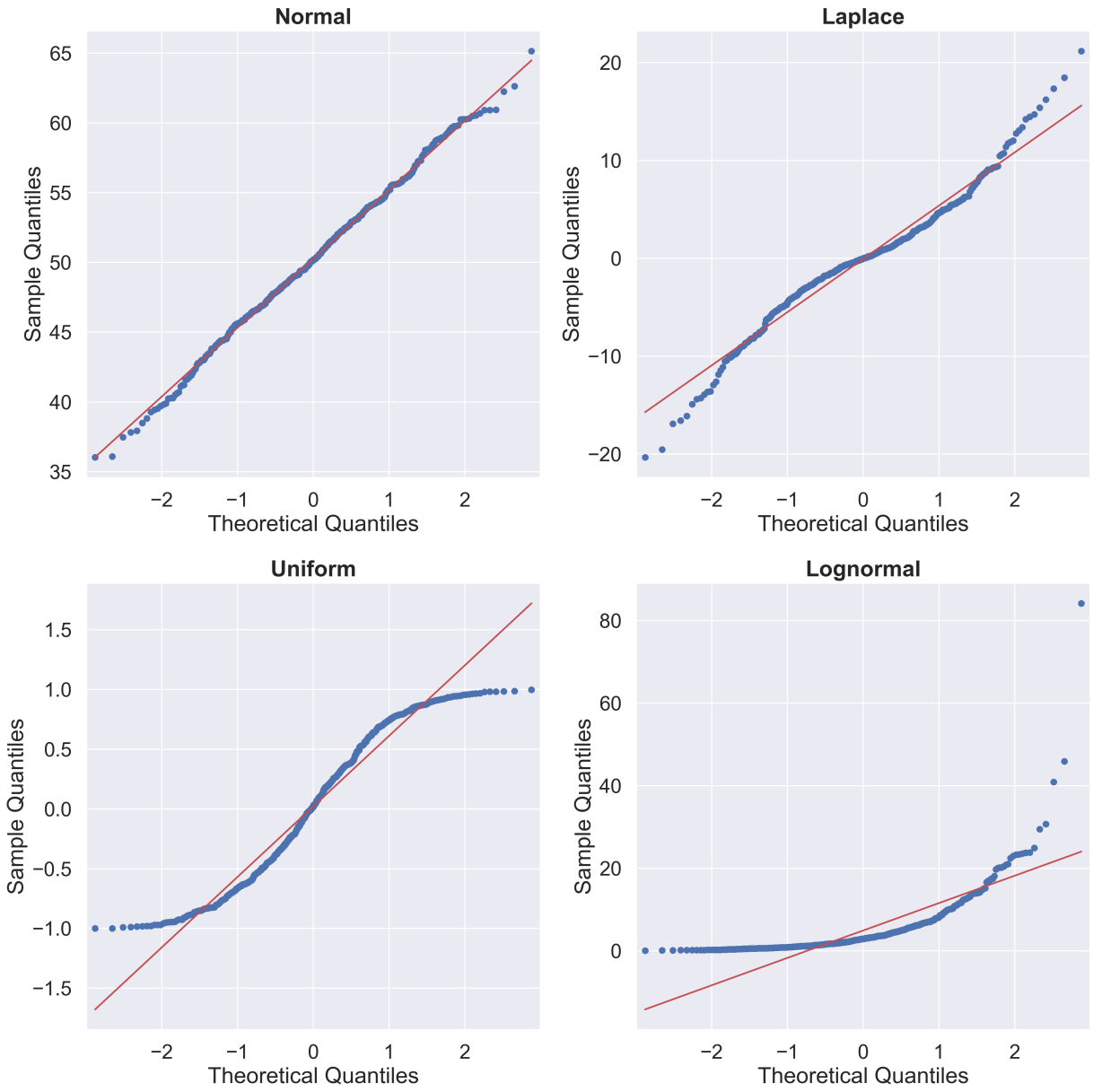


Figure 13.7.1: QQ plot with different sample distributions, including normal, Laplace ($b = 4$), Uniform $U([-1,1])$, Lognormal $LN(0,1)$. Red solid lines are the fitted linear lines.

13.7.3 Sample mean with unknown variance

Consider we have n samples $X_1, ..., X_n$ for a random variable with $N(\mu, \sigma^2)$ with σ^2 being unknown. The hypothesis testing involving the mean can be obtained by using the fact that

$$T = \frac{\overline{X} - \mu}{\sigma / n}$$

is $t(n-1)$. [Theorem 12.4.2] We can summarize the test as:

Table 13.7.2: Test on mean with unknown variance σ^2

H_0	test statistic	H_1	critical(rejection) region
$\mu \leq \mu_0$	$T = \frac{\overline{X}-\mu}{S/n}$	$\mu > \mu_0$	$t \geq t_\alpha(n-1)$
$\mu \geq \mu_0$	$T = \frac{\overline{X}-\mu}{S/n}$	$\mu < \mu_0$	$t \leq -t_\alpha(n-1)$
$\mu = \mu_0$	$T = \frac{\overline{X}-\mu}{S/n}$	$\mu \neq \mu_0$	$\lvert t \rvert \geq t_\alpha(n-1)$

13.7.4 Variance test

Consider we have n samples $X_1, ..., X_n$ for a random variable with $N(\mu, \sigma^2)$ with σ^2 being unknown. The hypothesis testing involving the mean can be obtained by using the fact that

$$T = \frac{\overline{X} - \mu}{\sigma / n}$$

is $t(n-1)$.[Theorem 12.4.2] We can summarize the test as:

Table 13.7.3: Test on variance

H_0	test statistic	H_1	critical(rejection) region
$\sigma^2 \leq \sigma_0^2$	$T = \frac{(n-1)S^2}{\sigma_0^2}$	$\sigma^2 \leq \sigma_0^2$	$t \geq \chi^2_\alpha$
$\sigma^2 \geq \sigma_0^2$	$T = \frac{(n-1)S^2}{\sigma_0^2}$	$\sigma^2 \leq \sigma_0^2$	$t \leq \chi^2_{1-\alpha}$
$\sigma^2 = \sigma_0^2$	$T = \frac{(n-1)S^2}{\sigma_0^2}$	$\sigma^2 \neq \sigma_0^2$	$t \leq \chi^2_{1-\alpha}$ or $t \geq \chi^2_\alpha$

13.7.5 Variance comparison test

Consider we have n samples $X_1, ..., X_n$ for a random variable with $N(\mu, \sigma^2)$ with σ^2 being unknown. The hypothesis testing involving the mean can be obtained by using the fact that

$$T = \frac{\overline{X} - \mu}{\sigma / n}$$

is $t(n-1)$. [Theorem 12.4.2] We can summarize the test as:

Table 13.7.4: Test on variance comparison between two samples

H_0	test statistic	H_1	critical(rejection) region
$\sigma_1^2 \leq \sigma_2^2$	$F = \frac{S_1^2}{S_2^2}$	$\sigma_1^2 \leq \sigma_2^2$	$t \geq \chi_\alpha^2$
$\sigma^2 \geq \sigma_0^2$	$F = \frac{S_1^2}{S_2^2}$	$\sigma_1^2 \leq \sigma_2^2$	$t \leq \chi_{1-\alpha}^2$
$\sigma_1^2 = \sigma_2^2$	$F = \frac{S_1^2}{S_2^2}$	$\sigma_1^2 \neq \sigma_2^2$	$F \leq \chi_{1-\alpha}^2$ or $F \geq \chi_\alpha^2$

13.7.6 Person correlation t test

Lemma 13.7.1 (Person correlation t test). *Let X and Y be random variable related by $Y = \beta X + \epsilon$, where $\beta \in \mathbb{R}$ and $\epsilon \sim N(\mu, \sigma)$. Let $\hat{\rho}$ be the correlation estimated from n samples of X and Y. Let the statistic*

$$T = \frac{\hat{\rho}}{\sqrt{1-\hat{\rho}^2}}\sqrt{n-2}$$

follows t distribution with degree of freedom $n-2$.

Proof. Note that if we construct the linear regression model on $Y \sim \beta X$, then [Theorem 15.1.11]

$$\hat{\rho}^2 = \frac{\hat{\beta}^2 S_{XX}}{S_{YY}}, 1 - \hat{\rho}^2 = \frac{SSE}{S_{YY}}.$$

Therefore

$$\frac{\hat{\rho}}{\sqrt{1-\hat{\rho}^2}}\sqrt{n-2} = \sqrt{n-2}\sqrt{\frac{S_{XX}}{SSE}}\hat{\beta}$$

is the t statistics follows $n-2$ degrees of freedom [Methodology 15.1.2]. □

13.7.7 Two sample tests

13.7.7.1 *Two-sample z test*

Bbasic setup of two-sample z test:

- $X_1, X_2, ..., X_m$ is a random sample from a distribution with mean μ_1 and variance σ_1^2.
- $Y_1, Y_2, ..., Y_n$ is a random sample from a distribution with mean μ_1 and variance σ_1^2.
- X and Y samples are independent of each other.

Lemma 13.7.2 (mean difference estimator). *[7, p. 363]Let $\overline{X}$ and $\overline{Y}$ denote the sample mean.*

- *$E[\overline{X} - \overline{Y}] = \mu_1 - \mu_2$, i.e., $\overline{X} - \overline{Y}$ is the unbiased estimator of $\mu_1 - \mu_2$.*
- $$Var[(\overline{X} - \overline{Y})] = \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}.$$

Proof. (1) Straight forwward. (2) Using independence, we have

$$Var[\overline{X} - \overline{Y}] = Var[\overline{X}] + Var[\overline{Y}] = \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}.$$

□

13.7.7.2 *Two-sample t test*

Basic setup two-sample t test:

- $X_1, X_2, ..., X_m$ is a random sample from a distribution with mean μ_1 and variance σ_1^2.
- $Y_1, Y_2, ..., Y_n$ is a random sample from a distribution with mean μ_2 and variance σ_2^2.
- X and Y samples are independent of each other.

Lemma 13.7.3 (mean difference estimator). *[7, p. 363] The standardized variable*

$$T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}}$$

has approximately a t distribution with degree of freedom v estimated to be (round to the nearest integer)

$$v = \frac{\left(\frac{S_1^2}{m} + \frac{S_2^2}{n}\right)^2}{\frac{\left(S_1^2/m\right)^2}{m-1} + \frac{\left(S_2^2/n\right)^2}{n-1}}$$

13.7.7.3 *Paired Data*

Basic setup for paired data

- The data consists of n independently selected pairs $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$, with $E[X_i] = \mu_1$ and $E[Y_i] = \mu_2$.
- Let $D_1 = X_1 - Y_1, D_2 = X_2 - Y_2, ..., D_n = X_n - Y_n$ so that $D_i's$ are the differences within pairs.
- The $D_i's$ are assumed to be normally distributed within mean value μ_D and variance σ_D^2.

Because we assume D_i are IID samples from normal distribution, we can apply the two-sample z test or two sample t test to test if D_i has zero mean.

Note 13.7.1 (caution!). X_i could be dependent on Y_i, but pairs are independent of each other. Here we assume D_i follows normal distribution, this is not necessarily true even if X and Y are normal distributions [Corollary 12.1.1.1].

13.7.8 Interval estimation for normal distribution

Definition 13.7.1 (confidence interval). *Let $X_1, X_2, ..X_n$ denote a random sample on a random variable X, where X has pdf $f(x;\theta)$. Let $\alpha (0 < \alpha < 1)$ be given. Let $L = L(X_1, X_2, ..X_n), U = U(X_1, X_2, ..X_n)$ be two statistics. We say that the interval (L, U) is a $(1-\alpha)$ confidence interval for θ if*

$$1 - \alpha = P_\theta(\theta \in (L, U))$$

Lemma 13.7.4 (confidence interval for mean of normal random sample). *Let X be a normal random variable $N(\mu, \sigma^2)$, Let $X_1, ..., X_n$ be the random sample, let S^2 and $\overline{X}$ be the sample variance and sample mean, then*

- *If σ is known, then the $(1-\alpha)$ confidence interval for μ is*

$$(\overline{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2}, \overline{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2})$$

- *If σ is unknown, then the $(1-\alpha)$ confidence interval for μ is*

$$(\overline{X} - \frac{S}{\sqrt{n}} t_{\alpha/2}(n-1), \overline{X} + \frac{S}{\sqrt{n}} t_{\alpha/2}(n-1))$$

where $z_{\alpha/2}, t_{\alpha/2)}(n-1)$ are the upper critical point of $\alpha/2$ for standard normal distribution and $t(n-1)$ distribution.

Proof. (1) Use the fact that

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

(2) Use the fact that

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

□

Remark 13.7.1 (knowing σ reduce uncertainty). Note that t distribution is wider(has big tails) than normal, which suggest larger confidence interval when σ is unknown.

Lemma 13.7.5 (Large sample confidence interval). *[8, p. 220] Let $X_1, ..., X_n$ be the random sample of a random variable with mean μ and variance σ^2. (Note that X is not necessarily normal). Then the $(1-\alpha)$ confidence interval for μ for large sample size is given as*

$$(\overline{X} - \frac{S}{\sqrt{n}} z_{\alpha/2}, \overline{X} + \frac{S}{\sqrt{n}} z_{\alpha/2})$$

Proof. When n is large, $S \approx \sigma$. Based on central limit theorem Theorem 11.11.3.

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

□

13.8 Notes on bibliography

For an advanced treatment on statistical estimation theory, see [2][8]. For likelihood based methods, see [9]. For large sample theory(asymptotic analysis), see [10].

For introductory level Bayesian statistics, see [11].

For a good treatment on statistical estimation theory, see [12].

For a tutorial on Fisher Information matrix, see [13]

For an introduction to robust statistics, see [14].

For an extensive discussion on statistical distribution, see [15][16].

BIBLIOGRAPHY

1. Ruppert, D. *Statistics and data analysis for financial engineering, 2ed* (Springer, 2015).
2. Casella, G. & Berger, R. L. *Statistical inference* (Duxbury Pacific Grove, CA, 2002).
3. Moon, T. K. S. & Wynn, C. *Mathematical methods and algorithms for signal processing* **621.39: 51 MON** (2000).
4. Greene, W. *Econometric Analysis* ISBN: 9780134461366 (Pearson, 2017).
5. Fink, D. A compendium of conjugate priors. *See http://www. people. cornell. edu/pages/df36/CONJINTRnew% 20TEX. pdf*, 46 (1997).
6. Wikipedia. *Conjugate prior — Wikipedia, The Free Encyclopedia* [Online; accessed 7-September-2016]. 2016.
7. Devore, J. L. *Probability and Statistics for Engineering and the Sciences* (Cengage learning, 2015).
8. Hogg, R. V., McKean, J. & Craig, A. T. Introduction to Mathematical Statistics, 7 ed (2012).
9. Pawitan, Y. *In all likelihood: statistical modelling and inference using likelihood* (Oxford University Press, 2001).
10. Lehmann, E. L. *Elements of large-sample theory* (Springer Science & Business Media, 1999).
11. Hoff, P. D. *A first course in Bayesian statistical methods* (Springer Science & Business Media, 2009).
12. Kay, S. M. Fundamentals of statistical signal processing, volume I: estimation theory (1993).
13. Ly, A., Marsman, M., Verhagen, J., Grasman, R. P. & Wagenmakers, E.-J. A tutorial on Fisher information. *Journal of Mathematical Psychology* **80,** 40–55 (2017).
14. Wilcox, R. R. *Fundamentals of modern statistical methods: Substantially improving power and accuracy* (Springer Science & Business Media, 2010).
15. Forbes, C., Evans, M., Hastings, N. & Peacock, B. *Statistical distributions* (John Wiley & Sons, 2011).
16. Krishnamoorthy, K. *Handbook of statistical distributions with applications* (CRC Press, 2016).