MATHEMATICAL STATISTICS
Lecture Notes for STA 200B

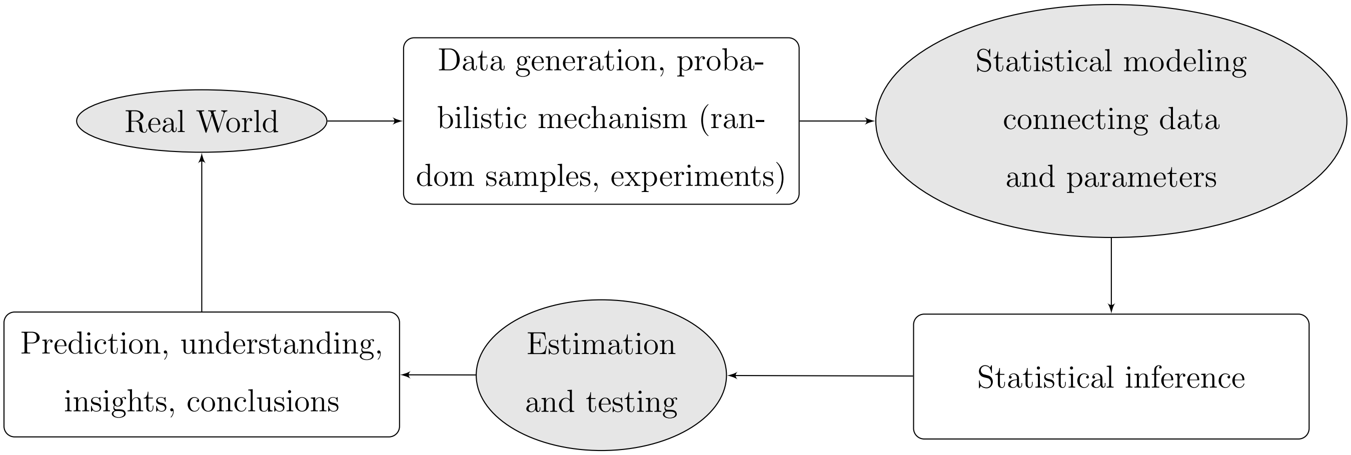
UC Davis

Winter 2020

1 Introduction

Statistics is the art and science of gathering, modeling and making inference from data.

For this class, the tools we will be using are mathematics and probability (the material of STA 200A and its prerequisites is required for STA 200B). Pioneers of modern statistics include K. Pearson, R.A. Fisher, J. Neyman. Below is a visual depiction of the general statistical framework.


Real World
Data generation, probabilistic mechanism (random samples, experiments)
Statistical modeling connecting data and parameters
Statistical inference
Estimation and testing
Prediction, understanding, insights, conclusions

Example: Flip a coin with probability θ of getting heads 50 times.
Questions: What is θ? Is $\theta = 1/2$, i.e., is the coin fair?
Model assumptions: Flips are independent with the same θ.

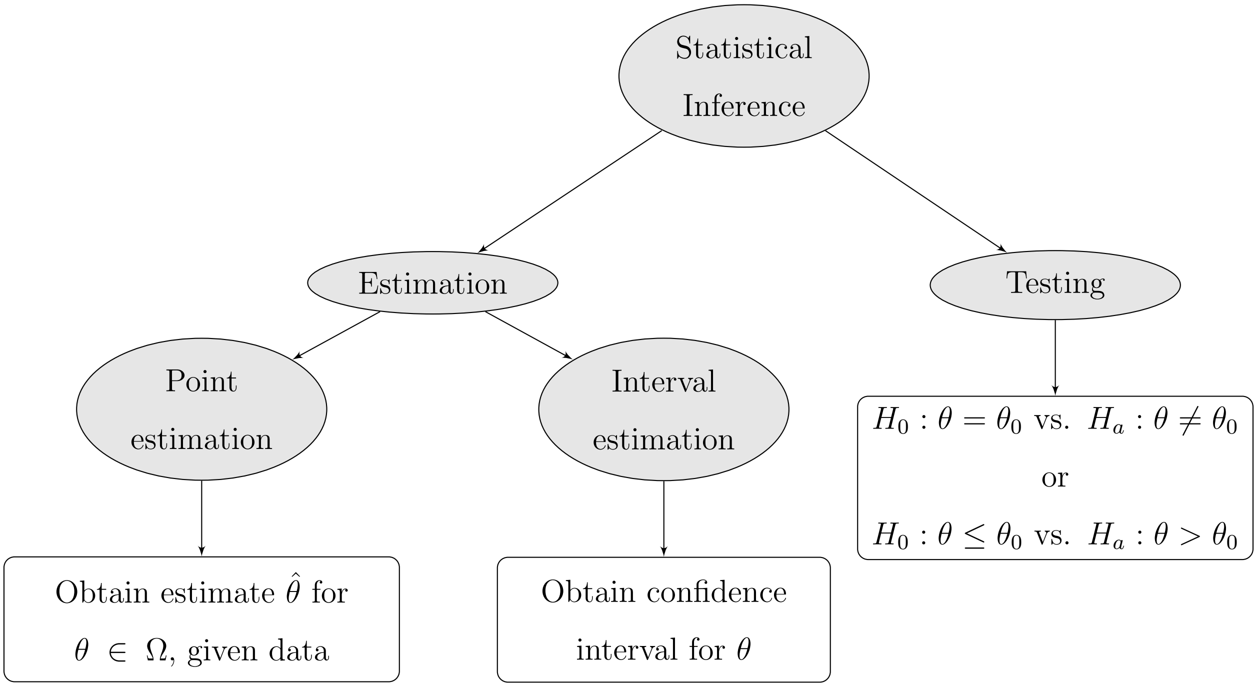
More generally, we study the case where a (probability) *distribution* is known except for a *parameter* θ. An important example is the two-dimensional *parameter vector* $\theta = (\mu, \sigma^2)$ when the distribution is $N(\mu, \sigma^2)$, the *normal distribution* also referred to as *Gaussian distribution* with mean μ and variance σ^2.

Definition: A statistical model is a specification of the distribution of observed data, which depends on a parameter $\theta \in \Omega$, where $\Omega \subset \mathbb{R}^p$. Here Ω is the parameter space, the set of all possible values of the parameter. We write $\{f(x|\theta), \theta \in \Omega\}$ for the model, where

f is the pmf/pdf with known form, and θ is an unknown parameter.

Example: Consider the multivariate normal distribution $N(\mu, \Sigma)$, where μ is a $p-$vector and Σ a $p \times p$ covariance matrix. What is the parameter space Ω?

Example: In the coin flip example, we have $X_1, \ldots, X_n \sim_{i.i.d.}$ Bernoulli(p), where Bernoulli(p) is the same as Binomial$(1, p)$, which we abbreviate as B$(1, p)$ in the following. Here $\theta = p \in [0, 1] = \Omega$.


Statistical Inference
Estimation
Testing
Point estimation
Interval estimation
$H_0 : \theta = \theta_0$ vs. $H_a : \theta \neq \theta_0$
or
$H_0 : \theta \leq \theta_0$ vs. $H_a : \theta > \theta_0$
Obtain estimate $\hat{\theta}$ for $\theta \in \Omega$, given data
Obtain confidence interval for θ

Our default assumption is that θ is fixed and non-random (classical statistics). *Sometimes* θ will be considered as random (Bayesian statistics); whenever we adopt the Bayesian point of view, we will make it very clear that we deal with random parameters.

Inference is based on samples $\{x_1, \ldots, x_n\}$, where we aim to infer the parameter θ.

Definition: Let $X_1, \ldots, X_n$ be n r.v.s (not necessarily independent) and $r : \mathbb{R}^n \to \mathbb{R}^1$ a function. Then a r.v. $T = r(X_1, \ldots, X_n)$ is called a statistic. Here, r.v. stands for *random* *variable*, i.e., a manifestation of a random mechanism that assigns a value to an outcome randomly.

Examples:

1. $T = r(X_1, \dots, X_n) = \frac{1}{n}\sum_{i=1}^n X_i = \overline{X}$ (sample mean)
2. $T = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2$ (sample variance)
3. $T = \frac{1}{2}[\min\{X_1, \dots, X_n\} + \max\{X_1, \dots, X_n\}]$

Definition: Let $X_1, \dots, X_n \sim_{i.i.d.} f(x|\theta), \theta \in \Omega$. Then $\{X_1, \dots, X_n\}$ is called a random sample from $f(x|\theta)$ (or from X, a generic version of the X_i). Here $\sim$ is shorthand for "distributed as " and i.i.d. for "independently and identically distributed".

Example: Radiation counts: Count radioactive particles registered by a Geiger counter near a radioactive probe for consecutive one minute intervals. It is known from physics that these counts are independent and approximately identically distributed (if the radioactive material is decaying very slowly). We then obtain a sample of Poisson counts $(X_1, \dots, X_n) \sim_{i.i.d.}$ Poisson(λ).

Example: Measure the daily temperature in Davis over one year for $n = 365$ daily temperature recordings. These are not independent, as a hot day is more likely to be followed by another hot day than a cold day. Therefore, such data are not i.i.d.

Example: Coin flips,

$$X_i = \begin{cases} 1 & \text{if heads} \\ 0 & \text{if tails} \end{cases}$$

We obtain a sample $\{X_1, \dots, X_n\}$ consisting of the first n flips.

Example: Assume a commuter needs to take a specific bus from a bus stop and the bus runs every 10 minutes. The commuter arrives at a random time at the bus stop, where all times are equally likely. Denoting the waiting times until the bus arrives on consecutive days by $\{X_1, \ldots, X_n\}$, this is a random sample from a $U([0, 10])$ (uniform).

2 Estimation

2.1 MLE

Given a model $\{f(x|\theta) : \theta \in \Omega\}$, for a random sample $X_1, \ldots, X_n \sim_{i.i.d.} f(x|\theta)$ we have for the joint pdf/pmf, which is written as $f_{X_1,\ldots,X_n}$ or as f_n,

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n|\theta) = f_n(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta).$$

Definition: Given a sample $\{X_1, \ldots, X_n\}$, the likelihood function for θ is

$$\tilde{L}(\theta) = \tilde{L}(\theta; X_1, \ldots, X_n) = f_{X_1,\ldots,X_n}(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta).$$

Usually, we work with $L(\theta) = \log \tilde{L}(\theta) = \sum_{i=1}^{n} \log f(x_i|\theta)$, as sums are more convenient. $L(\theta)$ is called the log likelihood function.

The idea of the MLE (maximum likelihood estimator) is to find the parameter value $\theta \in \Omega$ such that the observed data have the highest pdf/pmf, i.e., the parameter for which the observed data are "most likely".

Definition: If $\hat{\theta} = r(X_1, \ldots, X_n)$ is such that $L(\hat{\theta}) = \max_{\theta\in\Omega} L(\theta; X_1, \ldots, X_n)$, then $\hat{\theta}$ is called the MLE of θ. We also write $\hat{\theta} = \text{argmax}_{\theta\in\Omega} L(\theta)$.

How to find the MLE if $\Omega \subset \mathbb{R}^1$:

a) Solve $\frac{dL(\theta)}{d\theta} = 0$ and check whether $\frac{d^2L(\theta)}{d\theta^2} < 0$ (calculus method to find maximum)

b) If $L(\theta)$ is monotone in θ and $\Omega = [a, b]$ for some interval $[a, b]$, then the MLE is a boundary point of $\Omega = [a, b]$, either a or b.

Examples:

1. $X_1, \ldots, X_n \sim_{i.i.d.}$ Poisson(θ) (discrete counts), $\theta > 0$, $\Omega = (0, \infty)$

$$\begin{aligned}
\tilde{L}(\theta) &= \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{x_i}}{x_i!} = \frac{e^{-n\theta}\theta^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!} \\
L(\theta) &= -n\theta + \sum_{i=1}^{n} x_i \log\theta - \sum_{i=1}^{n} \log(x_i!) \\
\frac{dL(\theta)}{d\theta} &= -n + \frac{1}{\theta}\sum_{i=1}^{n} x_i \\
\frac{dL(\theta)}{d\theta} &= 0 \implies \theta = \frac{\sum_{i=1}^{n} x_i}{n} = \overline{x} \text{ (sample mean)}
\end{aligned}$$

Note that $\frac{d^2L(\theta)}{d\theta^2} = -\frac{1}{\theta^2}\sum_{i=1}^{n} x_i < 0$ since $x_i \geq 0$ if at least one $x_i > 0$. If all $x_i = 0$, the MLE does not exist. This shows that existence of the MLE may depend on the exact sample that is recorded.

2. $X_1, \ldots, X_n \sim_{i.i.d.} U([0, \theta])$:

$$\begin{aligned}
f(x|\theta) &= \frac{1}{\theta},\ 0 \leq x \leq \theta,\ f(x|\theta) = 0 \quad \text{otherwise;} \\
\tilde{L}(\theta) &= \left(\frac{1}{\theta}\right)^n,\ 0 \leq x_i \leq \theta \text{ for all } 1 \leq i \leq n, \quad \tilde{L}(\theta) = 0 \quad \text{otherwise;} \\
L(\theta) &= -n\log\theta,\ \text{if } 0 \leq x_i \leq \theta \text{ for all } 1 \leq i \leq n.
\end{aligned}$$

Here we see that $L(\theta)$ is monotone falling as θ increases. Therefore the solution will be the smallest θ that satisfies the constraints $0 \leq x_i \leq \theta$: $\theta = \max_{1 \leq i \leq n} x_i = x_{(n)}$. However, the MLE does not exist if $f(x|\theta) = \frac{1}{\theta},\ 0 < x < \theta$.

3. $X_1, \ldots, X_n \sim_{i.i.d.}$ Poisson(θ) revisited. Now we have prior knowledge that $\theta \geq 3$. The new MLE is

$$\hat{\theta} = \begin{cases} \overline{X} & \text{if } \overline{X} > 3 \\ 3 & \text{if } \overline{X} \leq 3 \end{cases}$$

because $\frac{d^2L(\theta)}{d\theta^2} < 0$ for all θ (so is monotone falling away from $\theta = 3$).

4. *Multivariate case:* Suppose $\theta = (\theta_1, \ldots, \theta_k)^T \in \Omega \subset \mathbb{R}^k$. The strategy for finding

the MLE is to solve the system of equations

$$\frac{\partial L(\theta)}{\partial \theta_j} = 0, \qquad j = 1, \ldots, k.$$

This is a system of k equations with k unknowns. To verify that the solution is indeed a maximum, we must check that the Hessian matrix H is negative definite. The Hessian matrix is defined as

$$H = \left(\frac{\partial^2 L(\theta)}{\partial \theta_j \partial \theta_l} \right).$$

H is negative definite if all of its eigenvalues are negative, or equivalently, if $\mathbf{x}^T H \mathbf{x} < 0$ for all $\mathbf{x} = (x_1, x_2)^T$.

5. Normal distribution with $\theta = (\mu, \sigma^2)$ unknown: Suppose $X_1, \ldots, X_n \sim_{i.i.d.} N(\mu, \sigma^2)$.

$$L(\theta) = \log \left[\prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-(x_i-\mu)^2/2\sigma^2} \right] = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log \sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2.$$

Next set

$$\frac{\partial L(\theta)}{\partial \mu} = 0, \; \frac{\partial L(\theta)}{\partial \sigma^2} = 0 \text{ (likelihood equations)}.$$

Solving the first likelihood equation leads to

$$\frac{\partial L(\theta)}{\partial \mu} = -\frac{1}{2\sigma^2} 2 \sum_{i=1}^n (x_i - \mu)(-1) = 0 \implies \mu = \overline{x}.$$

Now using the second likelihood equation (and plugging in $\mu = \overline{x}$), we have

$$\frac{\partial L(\theta)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} - \frac{1}{2}(-\frac{1}{\sigma^4}) \sum_{i=1}^n (x_i - \overline{x})^2 = 0 \implies \sigma^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})^2.$$

Then the MLE of $\theta = (\mu, \sigma^2)$ is $\hat{\theta} = (\overline{X}, \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2) = (\overline{X}, \hat{\sigma}^2)$, where we still need to show that this is indeed a maximum. For this we will need to show that the Hessian matrix $H(\mu, \sigma^2)$ is negative definite. Note that

$$\begin{aligned}
\frac{\partial^2 L}{\partial \mu^2} &= -\frac{n}{\sigma^2} \\
\frac{\partial^2 L}{\partial (\sigma^2)^2} &= \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n}(x_i - \mu)^2 \\
\frac{\partial^2 L}{\partial \mu \partial \sigma^2} &= -\frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i - \mu).
\end{aligned}$$

Then we have

$$H(\mu, \sigma^2) = \begin{pmatrix} -\frac{n}{\sigma^2} & -\frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i - \mu) \\ -\frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i - \mu) & \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n}(x_i - \mu)^2 \end{pmatrix}.$$

We want to show that $H(\mu, \sigma^2)|_{(\mu,\sigma^2)=(\hat{\mu},\hat{\sigma}^2)}$ is negative definite. To see this, after some cancellation, we have

$$H(\hat{\mu}, \hat{\sigma}^2) = \begin{pmatrix} -\frac{n}{\hat{\sigma}^2} & 0 \\ 0 & -\frac{n}{2\hat{\sigma}^4} \end{pmatrix}.$$

This matrix has two negative eigenvalues (the values on the diagonal), therefore is negative definite. One can also show this directly, noting that for any $\mathbf{x} = (x_1, x_2)^T$, we have

$$\mathbf{x}^T H \mathbf{x} = -\frac{n}{\hat{\sigma}^2}x_1^2 - \frac{n}{2\hat{\sigma}^2}x_2^2 < 0,$$

which implies that H is negative definite.

6. Bernoulli distribution: suppose $X_1, \ldots, X_n \sim_{i.i.d.} B(1, \theta)$ where $\theta \in [0, 1] = \Omega$ is

unknown.

$$
\begin{aligned}
\tilde{L}(\theta) &= \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} \\
L(\theta) &= \sum_{i=1}^{n} x_i \log\theta + (n - \sum_{i=1}^{n} x_i)\log(1-\theta) \\
\frac{dL(\theta)}{d\theta} &= 0 \implies \frac{1}{\theta}\sum_{i=1}^{n} x_i - (n - \sum_{i=1}^{n} x_i)\frac{1}{1-\theta} = 0 \implies \theta = \overline{x}.
\end{aligned}
$$

Show: $\frac{dL^2(\theta)}{d\theta^2} < 0$.

2.2 Properties of the MLE

Invariance of the MLE: Consider the model $\{f(x|\theta), \theta \in \Omega\}$. Assume that $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ is the MLE for θ. Let $\Psi = g(\theta)$ be a function $g : \mathbb{R}^1 \to \mathbb{R}^1$ which is 1:1, i.e., we can write $\theta = g^{-1}(\Psi)$, where g^{-1} is defined on $g(\Omega)$. This is simply a “reparameterization” of the model; the model is then $\{f(x|g^{-1}(\Psi)),\ \Psi \in g(\Omega)\}$. Then the MLE $\hat{\Psi}$ for Ψ is $g(\hat{\theta})$, where $\hat{\theta}$ is the MLE for θ. This is because

$$\max_{\Psi \in g(\Omega)} f_n(x_1, \ldots, x_n | g^{-1}(\Psi)) = f_n(x_1, \ldots, x_n | \hat{\theta}).$$

Therefore $\hat{\Psi} = g(\hat{\theta})$. Note that the 1:1 property for g can be relaxed:

Theorem. For a function g, $\hat{\psi} = g(\hat{\theta})$ is the MLE of $\psi = g(\theta)$.

Defining $G_\psi = \{\theta : g(\theta) = \psi\}$, $L^*(\psi) = \max_{\theta \in G_\psi} \log f_n(\mathbf{X}|\theta)$, one finds that $\hat{\psi} = \mathrm{argmax}_{\psi \in g(\Omega)} L^*(\psi)$ is the MLE of ψ.

Applications: Consider the normal distribution, where $\theta = (\mu, \sigma^2)$.

a) An alternative parameterization is $\theta = (\mu, \sigma)$. If we want to find the MLE of σ, we have $g(\sigma^2) = \sqrt{\sigma^2} = \sigma$ and if $\hat{\sigma}^2$ is the MLE for σ^2, then $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$ is the MLE for σ.

b) If we want the MLE of $EX^2 = \sigma^2 + \mu^2$, we have $\Psi = g(\theta) = g(\mu, \sigma^2) = \sigma^2 + \mu^2$ is not invertible. However, it still holds that $\hat{\Psi} = \hat{\sigma}^2 + \hat{\mu}^2$ is the MLE of Ψ.

Consistency of the MLE: Under mild assumptions the MLE *converges in probability* to the true parameter θ. For this, we write

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

Convergence in probability is an *asymptotic property* which we will discuss in more detail later.

Problems with the MLE:

- The MLE *may not exist* in some cases. For example, we have seen that the MLE does not exist when the sampling distribution is $U((0,\theta))$, where $0 < X < \theta$ and for the Poisson distribution when all outcomes are zero (since we require the parameter to satisfy $\theta > 0$.)

- It also *may not be unique.* To see this, consider the case when the sample comes from $U([\theta, \theta+1])$. In this case, we have $L(\theta) = 1$, for all θ with $\theta \leq x_{(1)}$, $x_{(n)} \leq \theta+1$, where $x_{(1)} = \min(x_1, \ldots, x_n)$, $x_{(n)} = \max(x_1, \ldots, x_n)$. This implies that the MLE is not unique, as will be many θ which satisfy this.

- There may also be *numerical difficulties* in finding the MLE. For example the Gamma(α, β) has pdf $f(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, x > 0$. Consider the case where $\beta = 1$, then

$$\begin{aligned} f_n(\mathbf{x}|\alpha) &= \frac{1}{\Gamma(\alpha)^n} \prod_{i=1}^{n} x_i^{\alpha-1} e^{-\sum_{i=1}^n x_i} \\ L(\alpha) &= -n \log \Gamma(\alpha) + (\alpha - 1) \sum_{i=1}^{n} \log x_i - \sum_{i=1}^{n} x_i \\ \frac{dL(\alpha)}{d\alpha} &= -n \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} + \sum_{i=1}^{n} \log x_i = 0 \implies \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} = \frac{1}{n} \sum_{i=1}^{n} \log x_i, \end{aligned}$$

which cannot be solved explicitly; we require numerical methods.

Newton-Raphson algorithm: The Newton-Raphson algorithm is a standard method to find the maximum of a differentiable function g, including finding the MLE. In the following, f is the derivative of the function g, or $f = g'$. Let $f : \mathbb{R} \to \mathbb{R}$ for which we want to solve $f(\theta) = 0$. Let θ_0 be an initial guess for the solution, then update the solution by

$$\begin{aligned} \theta_1 &= \theta_0 - \frac{f(\theta_0)}{f'(\theta_0)}, \\ \theta_{k+1} &= \theta_k - \frac{f(\theta_k)}{f'(\theta_k)} \end{aligned}$$

and iterate over k until convergence. This procedure is justified by a Taylor expansion:

$$0 = f(\theta) \approx f(\theta_0) + f'(\theta_0)(\theta - \theta_0) \text{ (plus remainder terms that are neglected),}$$

which implies that $\theta \approx \theta_0 - \frac{f(\theta_0)}{f'(\theta_0)}$. This procedure may fail if $f'(\theta_0) \approx 0$. There is also an analogous multivariate version, where f' is replaced by a gradient vector. For the initial value θ_0, one can use the method of moments.

Profile Likelihood. Let $\theta = (\alpha, \beta)$ and consider $f(x \mid \theta)$, $\theta \in \Omega$, a family of distributions. If for any given α the MLE of β exists, denoted by $h(\alpha) = \beta_\alpha$, we plug this function into the log likelihood function $l(\alpha, \beta)$ to arrive at $l^*(\alpha) = l(\alpha, h(\alpha))$, which is called the "log profile likelihood" and which contains only the parameter α. This makes it a lot easier to to locate the argmax numerically (or explicitly). We find

$$\max_\alpha l^*(\alpha) = \max_\alpha l(\alpha, h(\alpha)) = \max_\alpha l(\alpha, \beta_\alpha) = \max_\alpha \max_\beta l(\alpha, \beta). \tag{2.1}$$

Hence if $\hat{\alpha}^*$ maximizes $l^*(\alpha)$, then $(\hat{\alpha}^*, h(\hat{\alpha}^*))$ maximizes $l(\alpha, \beta)$.

The profiles likelihood approach, if feasible (that is, when a close-form solution for β_α exists if the value of α is given), simplifies the MLE approach. This is especially attractive when β is high or infinite dimensional. An example is the Cox proportional hazards model when the baseline hazard function is unknown and modeled nonparametrically.

2.3 Method of Moments

Suppose that we have $X_1, \ldots, X_n \sim_{i.i.d.} f(x|\theta)$. Then define *population* and *sample* moments, $\mu_j(\theta)$ and m_j, respectively, as

$$\mu_j(\theta) = \frac{1}{n}\sum_{i=1}^{n} E_\theta(X_i^j) = E_\theta X^j$$
$$m_j = \frac{1}{n}\sum_{i=1}^{n} X_i^j,$$

for $j \geq 1$. Assume we aim to estimate $q(\theta)$, a function of θ. Suppose $q(\theta) = g(\mu_1(\theta), \ldots, \mu_r(\theta))$ for a function g.

Example: $q(\theta) = \mathrm{Var}(X) = \sigma^2 = EX^2 - (EX)^2 = \mu_2(\theta) - \mu_1(\theta)^2$.

Then the estimate $\hat{q}(\theta) = g(m_1, \ldots, m_r)$ of $q(\theta)$ is the *Methods of Moments* estimate.

Examples:

1. $\hat{q}(\theta) = \hat{\sigma}^2 = m_2 - m_1^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - (\overline{X})^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2 = \hat{\sigma}^2$. Note that this is the same as the MLE for the Gaussian model, but the methods of moments makes no distributional assumption.

2. Suppose $X_1, \ldots, X_n \sim_{i.i.d.} N(\mu, \sigma^2)$. Then the method of moment estimators for μ, σ^2 are $\overline{X}, \hat{\sigma}^2$.

3. Suppose $X_1, \ldots, X_n \sim_{i.i.d.}$ Bernoulli(θ). We know that the mean for this distribution is θ, i.e., $\mu_1(\theta) = EX_1 = \theta$, so $\overline{X}$ is a method of moments estimate for θ (same as MLE). Now if we wish to estimate the variance, we have

$$\mathrm{Var}(X_1) = \theta(1-\theta) = q(\theta),$$

so we obtain $\widehat{\mathrm{Var}}(X) = q(m_1) = \overline{X}(1 - \overline{X})$ as a method of moments estimator for $\mathrm{Var}(X)$ based on the first moment. A second methods of moments estimator can be derived from example 1., which gives $\widehat{\mathrm{Var}}(X) = m_2 - m_1^2$.

4. Suppose $X_1, \dots, X_n \sim_{i.i.d.}$ Poisson(θ). Then $EX_1 = \theta$ and $\mathrm{Var}(X_1) = \theta$. This means that (a) $\mu_1(\theta) = \theta$, leading to $\hat{\theta} = m_1 = \overline{X}$ is a method of moments estimator of θ, but also (b) $\mu_2(\theta) - \mu_1(\theta)^2 = \theta$, leading to $\hat{\theta} = m_2 - m_1^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - (\frac{1}{n}\sum_{i=1}^n X_i)^2$ as a second method of moments estimator.

   The conclusion is that method of moments estimators are not unique so need to be exactly specified on a case-by-case basis.

5. Suppose $X_1, \dots, X_n \sim_{i.i.d.} U([0, \theta])$. Here, $\mu_1(\theta) = EX_1 = \frac{\theta}{2}$, which leads to $\hat{\theta} = 2\overline{X}$ as a method of moments estimator. This estimator is unreasonable, since if $X_{(n)} = \max X_i > 2\overline{X}$, this estimator makes no sense.

6. Suppose $X_1, \dots, X_n \sim_{i.i.d.}$ Gamma(α, β). Here, we have $EX_1 = \frac{\alpha}{\beta} = \mu_1(\theta)$ and $\mathrm{Var}(X_1) = \frac{\alpha}{\beta^2} = \mu_2 - \mu_1^2 \implies \mu_2 = \frac{\alpha(\alpha+1)}{\beta^2}$. To obtain method of moments estimators, we have two equations and two unknowns:

$$\begin{aligned} m_1 &= \frac{\alpha}{\beta} \\ m_2 &= \frac{\alpha(\alpha+1)}{\beta^2}. \end{aligned}$$

   To solve this system, substitute $\beta = \frac{\alpha}{m_1}$ to obtain $\alpha = \frac{m_1^2}{m_2 - m_1^2}$. One can then plug this in to obtain an expression for β.

Notes: – There are often several methods of moment estimates for the same $q(\theta)$.

– MoM estimators are usually easy to compute and are valuable as preliminary estimates.

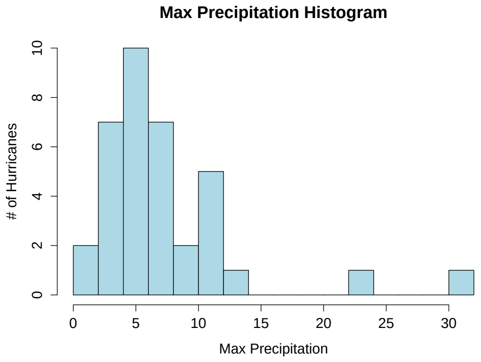
– MoM estimators do not always work as the uniform distribution example shows.

– Under mild assumptions, MoM estimators are consistent (where consistency is as defined above and defined as convergence in probability to the target), but they are often not efficient (a property of estimators to be discussed later).

Example

Although hurricanes generally strike only the eastern and southern coastal regions of the United States, they do occasionally sweep inland before completely dissipating. In the period from 1900 to 1969 a total of 36 hurricanes moved as far as the Appalachians. The table below lists a subset of the 36 maximum 24-hour precipitation levels recorded.

Year	Name	Location	Max Precipitation (in)
1969	*Camille*	Tye River, Va.	31.00
1968	*Candy*	Hickley, N.Y.	2.82
1965	*Betsy*	Haywood Gap, N.C.	3.98
1960	*Brenda*	Cairo, N.Y.	4.02
⋮	⋮	⋮	⋮
1900		St. Johnsbury, Vt.	0.67



The histogram of the data shows that because of its skewed shape, Y (the maximum number of 24-hour precipitation) might be well approximated by a member of the gamma family,

$$f_Y(x|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \text{ for } x > 0.$$

Here α and β are the parameters to be estimated. The complexity of $f_Y(x|\alpha,\beta)$ makes the method of maximum likelihood unwieldy. As an alternative, we will find the method-of-moments estimates.

For the data, the first two sample moments are

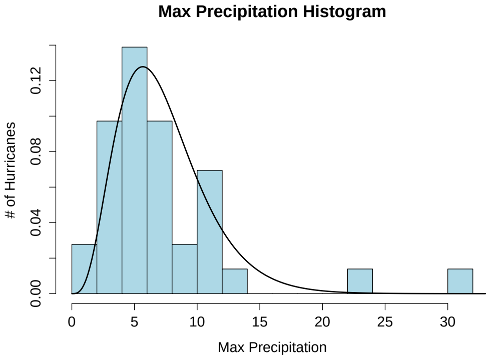
$$m_1 = \frac{1}{36}\sum_{i=1}^{36} y_i = \frac{1}{36}(262.35) = 7.29, \quad m_2 = \frac{1}{36}\sum_{i=1}^{36} y_i^2 = \frac{1}{36}(3081.2177) = 85.89.$$

We have seen that for the Gamma distribution

$$\hat{\alpha} = \frac{m_1^2}{m_2 - m_1^2}$$

and by plugging in m_1 and m_2 from the data this leads to $\hat{\alpha} = 1.63$ and then $\hat{\beta} = \frac{\hat{\alpha}}{m_1}$; plugging in m_1 and $\hat{\alpha}$ this yields $\hat{\beta} = 0.225$.

The figure below shows the fitted model superimposed over the original data, where the pdf $f_Y(x|\hat{\alpha}, \hat{\beta})$ is displayed, overlaid with the histogram.



The agreement between data and fit is seen to be quite good.

3 Bayesian Methods

3.1 Priors and Posteriors

Prior and *posterior* distributions are key quantities in Bayesian statistics. Bayesians view parameters θ as random variables, rather than fixed quantities. So far, we have viewed θ

as *fixed* but *unknown.*

When viewing θ as a r.v. there is a "prior" belief about likely values of θ and where they occur; this belief is subjective and not based on data.

The assumed distribution of θ is the prior distribution. If θ follows a distribution, then $f(x|\theta)$ is the conditional pmf or pdf of $X = x|\Theta = \theta$. We denote the prior distribution as $\theta \sim \xi,\ \theta \in \Omega$, where ξ is a pmf or pdf.

In the Bayesian framework, random samples $X_1, \ldots, X_n$ are *conditionally* i.i.d. given θ, i.e.,

$$f_n(\mathbf{x}|\theta) = \prod_{i=1}^{n} f(x_i|\theta).$$

This implies that the joint pdf of $(X_1, \ldots, X_n)$ is

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \int_\Omega f_n(\mathbf{x}|\theta)\xi(\theta)d\theta,$$

since the joint distribution of $(X_1, \ldots, X_n, \theta)$ has pdf/pmf

$$f_{X_1,\ldots,X_n,\theta}(\mathbf{x}, \theta) = f_n(\mathbf{x}|\theta)\xi(\theta).$$

Recall from 200A:

a) $f_{X,Y}(x, y) = f_{X|Y}(x|y)f_Y(y) = f_{Y|X}(y|x)f_X(x)$

b) $f_Y(y) = \int f_{X,Y}(x, y)dx; \quad f_X(x) = \int f_{X,Y}(x, y)dy$

The posterior distribution is defined by the pmf/pdf of θ given $\mathbf{X} = (X_1, \ldots, X_n)$ (the observed sample). We denote the posterior pdf/pmf as $\xi_{\theta|\mathbf{X}}$:

$$\begin{aligned}\xi_{\theta|\mathbf{X}}(\theta|\mathbf{x}) &= \frac{f_{X_1,\ldots,X_n,\theta}(\mathbf{x}, \theta)}{f_{X_1,\ldots,X_n}(\mathbf{x})} \\ &= \frac{f_n(\mathbf{x}|\theta)\xi(\theta)}{\int_\Omega f_n(\mathbf{x}|\theta)\xi(\theta)d\theta}.\end{aligned}$$

Note that

1. The distribution of θ has changed from the "prior" distribution (before observing data) to the "posterior" distribution after observing the data.

2. $\xi_{\theta|\mathbf{X}}(\theta|\mathbf{x}) \propto f_n(\mathbf{x}|\theta)\xi(\theta)$ as a function of θ where $\propto$ means "proportional to". Since for each given $\mathbf{x}$, $\xi_{\theta|\mathbf{X}}(\theta|\mathbf{x})$ is a pdf/pmf with argument θ, we have

$$\int_\Omega \xi_{\theta|\mathbf{X}}(\theta|\mathbf{x})d\theta = 1.$$

Example: Assume in a clinical trial for a treatment of acne with a skin cream, the chance of success (acne is resolved) is θ. In this case, the observed data is a sequence of Bernoulli trials where $X_i = 1$ if acne is resolved and $X_i = 0$ otherwise, and the probability of $X_i = 1$ is θ. In other words, $X_i|\theta \sim_{i.i.d.}$ Bernoulli(θ). Then the joint pmf for the sample, given $\theta \in \Omega = [0, 1]$ is

$$f_{\mathbf{X}|\theta}(\mathbf{x}|\theta) = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}.$$

What is a possible prior distribution for θ?

a) Uniform($[0, 1]$). This is a case where we assume all θ are equally likely to be encountered.

b) Beta(α, β). The pdf for Beta(α, β) is given by

$$f(x|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1} \text{ for } 0 < x < 1.$$

When $\alpha = \beta = 1$, this is the $U(0, 1)$ pdf, which is a special case. Note that if $X \sim$ Beta(α, β), then $EX = \frac{\alpha}{\alpha+\beta}$ and $\text{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

c) Beta distributions are a popular prior for θ, where

$$\xi(\theta) = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \text{ for } 0 < \theta < 1,$$

with $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ referred to as the *Beta function*. This is a commonly used prior when θ takes values on $(0,1)$. The resulting models are sometimes referred to as beta-binomial models.

Using a beta prior, we find

$$\begin{aligned} f_{\mathbf{X},\theta}(\mathbf{x},\theta) &= f_{\mathbf{X}|\theta}(\mathbf{x}|\theta)\xi(\theta) \\ &= \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}\frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= Constant \times \theta^{\sum_{i=1}^n x_i+\alpha-1}(1-\theta)^{n-\sum_{i=1}^n x_i+\beta-1}. \end{aligned}$$

The posterior distribution

$$\xi_{\theta|\mathbf{X}}(\theta|\mathbf{x}) = \frac{f_{X_1,\ldots,X_n,\theta}(\mathbf{x},\theta)}{f_{X_1,\ldots,X_n}(\mathbf{x})}$$

therefore is a Beta($\sum_{i=1}^n x_i + \alpha, n - \sum_{i=1}^n +\beta$) distribution.

If the prior is such that the prior distribution and the posterior distribution belong to the same distribution family for a statistical model and random sample, it is called a conjugate prior. We have shown that the beta distribution is a conjugate prior for the Bernoulli model. Conjugate priors are convenient for mathematical calculations but not necessarily better than other priors.

Priors can be a "subjective" choice. In the empirical Bayes framework the prior selection itself is based on data samples.

The larger the sample size, the less does the posterior distribution depend on the choice of the prior distribution.

Summary of Common Conjugate Priors:

Model $f(x\|\theta)$	Conjugate Prior $\xi(\theta)$	Posterior Distribution $\xi(\theta\|\mathbf{X})$
Bernoulli$(n, \theta), n$ known	Beta(α, β)	Beta$(\sum_{i=1}^n X_i + \alpha, n - \sum_{i=1}^n X_i + \beta)$
Negative Binomial$(r, \theta), r$ known	Beta(α, β)	Beta$(\alpha + nr, \beta + \sum_{i=1}^n X_i)$
Poisson(θ)	Gamma$(\alpha, \beta), \alpha =$ shape	Gamma$(\alpha + \sum_{i=1}^n X_i, \beta + n)$
Normal$(\theta, \sigma^2), \sigma^2$ known	Normal(τ, ν^2)	Normal$(\frac{\sigma^2\tau + n\nu^2\overline{X}}{\sigma^2 + n\nu^2}, \frac{\sigma^2\nu^2}{\sigma^2 + n\nu^2})$
Exponential(θ)	Gamma(α, β)	Gamma$(\alpha + n, \beta + \sum_{i=1}^n X_i)$
Uniform$(0, \theta)$	Pareto(λ, r)	Pareto$(\max\{\lambda, X_{(n)}\}, n + r)$

Notes: (1) The p.d.f. of the Pareto(λ, r) distribution is given by

$$f(x|\lambda, r) = \frac{r\lambda^r}{x^{r+1}}, \text{ for } x > \lambda.$$

(2) The negative binomial distribution with parameters r and θ describes the number of failures before the r^{th} success, where the probability of success is θ.

3.2 Bayes Estimators

Suppose we have the statistical model $\{f(x|\theta),\ \theta \in \Omega\}$ for a sample $\mathbf{X}$ with a prior distribution $\xi(\theta)$.

Definition: A <u>loss function</u> $L(\theta, a)$ is a function that measures the cost if one estimates θ by $a = a(\mathbf{X})$ (a statistic).

Examples:

1. $L(\theta, a) = (a - \theta)^2$ is known as "squared error loss" (most popular loss function)
2. $L(\theta, a) = |a - \theta|$ is known as "absolute error loss"

When θ is random, it makes sense to minimize the *expected loss, conditional on* $\mathbf{X}$:

$$E[L(\theta, a(\mathbf{X}))|\mathbf{X}] = \int_\Omega L(\theta, a(\mathbf{X}))\xi_{\theta|\mathbf{X}}(\theta|\mathbf{X})d\theta.$$

An estimator $\delta^*(\mathbf{X}) = a(\mathbf{X})$ such that $E[L(\theta, a(\mathbf{X}))|\mathbf{X}]$ is minimized for each sample $\mathbf{X}$ is called a Bayes estimator of θ, i.e.,

$$E[L(\theta, \delta^*(\mathbf{X}))|\mathbf{X}] = \min_a E[L(\theta, a)|\mathbf{X}]$$

for all $\mathbf{X}$. Thus, the Bayes estimator depends on the loss function L and the prior $\xi(\theta)$.

Theorem.

a) The Bayes estimator for the squared error loss is

$$E(\theta|\mathbf{X}) = \int_\Omega \theta\xi_{\theta|\mathbf{X}}(\theta|\mathbf{X})d\theta,$$

the posterior mean.

b) The Bayes estimator for the absolute error loss function is the median of the posterior distribution, i.e., in the case of a continuous posterior distribution, the point m where

$$\int_{-\infty}^{m} \xi_{\theta|\mathbf{X}}(\theta|\mathbf{x})d\theta = \frac{1}{2}.$$

Example: Suppose $X_1, \ldots, X_n \sim_{i.i.d.}$ Binomial$(1, \theta)$ (Bernoulli), and that $\theta \sim$ Beta(α, β).

This is a conjugate prior and we have seen

$$\theta|\mathbf{X} \sim \text{Beta}(\sum_{i=1}^{n} X_i + \alpha, n - \sum_{i=1}^{n} X_i + \beta).$$

The mean of Beta(α, β) is $\frac{\alpha}{\alpha+\beta}$, which implies that the Bayes estimator under squared error loss is the posterior mean,

$$\hat{\theta} = \frac{\alpha + \sum_{i=1}^{n} X_i}{n + \alpha + \beta} = \frac{\overline{X} + \alpha/n}{1 + (\alpha + \beta)/n},$$

which will essentially be $\overline{X}$ for large n. For the absolute error loss, we need to obtain the posterior median, where $\hat{\theta} = m$ if

$$\frac{1}{B(\sum_{i=1}^{n} X_i + \alpha, n - \sum_{i=1}^{n} X_i + \beta)} \int_0^m t^{\sum_{i=1}^{n} X_i + \alpha - 1} (1-t)^{n - \sum_{i=1}^{n} X_i + \beta - 1} dt = \frac{1}{2},$$

which requires a numerical solution.

Example: Suppose $X_1, \ldots, X_n \sim_{i.i.d.} N(\theta, \sigma^2)$, where θ is unknown and σ^2 is known. If we use a normal prior distribution on θ, i.e.,

$$\xi(\theta) = \frac{1}{\sqrt{2\pi}\nu_0} e^{\frac{-(\theta-\mu_0)^2}{2\nu_0^2}},$$

or $\theta \sim N(\mu_0, \nu_0^2)$, then the posterior distribution will also be normal as $N(\mu_1, \nu_1^2)$, where

$$\begin{aligned} \mu_1 &= \frac{\sigma^2 \mu_0 + n\nu_0^2 \overline{X}_n}{\sigma^2 + n\nu_0^2} \\ \nu_1^2 &= \frac{\sigma^2 \nu_0^2}{\sigma^2 + n\nu_0^2}. \end{aligned}$$

Notice that μ_1 is a "weighted mean" between μ_0 and $\overline{X}_n$: more towards μ_0 if σ^2 is large; more toward $\overline{X}_n$ if $n\nu_0^2$ is large. The Bayes estimator under squared error loss is

$$\delta^*(\mathbf{X}) = \mu_1 = \frac{\sigma^2 \mu_0 / n + \nu_0^2 \overline{X}_n}{\sigma^2/n + \nu_0^2} = \frac{\overline{X}_n + \frac{\sigma^2 \mu_0}{n\nu_0^2}}{1 + \frac{\sigma^2}{n\nu_0^2}},$$

which becomes indistinguishable from $\overline{X}$ as n gets large.

Notes about Bayes estimators

1. They may not exist
2. They may be sensitive to the choice of the prior for small/modest n
3. It is difficult to use a prior for parameters θ that are vector-valued
4. The choice of the prior distribution can be controversial: subjective (Bayesian) versus empirical (empirical Bayesian which is considered a non-Bayesian approach as the prior is learned from previous data)

4 Sufficient Statistics

4.1 Data Reduction and the Sufficiency Principle

Assume for the model $\{f(x, \theta), \theta \in \Omega$ we have a sample $\mathbf{X} = \{X_1, \ldots, X_n\}$ and aim at inference for θ. How do we summarize the information contained in this sample so that

- We do not need to deal with the original data sample of size n
- We have an efficient approach to extract information from the original data

Goal: Summarize the information contained in $\mathbf{X} \quad \longrightarrow$ Data Reduction.
Retain information relevant for inference about the parameter θ, discard information that is irrelevant to infer θ.

Example: Suppose we have a sample of Bernoulli trials $X_1, \ldots, X_n \sim_{i.i.d.} \text{Binomial}(1, \theta)$. For estimation of θ, we use $\overline{X}$, so we only need to know $T = T(\mathbf{X}) = \sum_{i=1}^n X_i =$ total number of successes. In other words, T summarizes the necessary information in the sample and is known as a sufficient statistic (recall that any function of the data is called a statistic).

For *big data*, data reduction is essential and sufficient statistics can deliver this. Why don't we need the entire sample $X_1, \ldots, X_n$? Because

$$f_\theta(\mathbf{X} = \mathbf{x} | T = t) = \frac{1}{\binom{n}{t}}$$

does not involve θ, and therefore carries no information about θ.

Definition: Let $X_1, \ldots, X_n$ be a random sample from the model $\{f(x|\theta), \theta \in \Omega\}$. Then a statistic $T = T(\mathbf{X})$ is a sufficient statistic for θ if the conditional distribution of $\mathbf{X} = (X_1, \ldots, X_n)$ given $T = t$ does not depend on θ, i.e., $f_{X|T}(\mathbf{x}|t) = f(\mathbf{x}|t)$ does not depend on θ.

Note: The distribution of T depends on θ, as T carries the essential information about θ. The concept of sufficiency is attributed to Fisher (1922).

Example: Suppose $X_1, \ldots, X_n \sim_{i.i.d.} U(0, \theta)$. What would be a sufficient statistic for θ? The maximum of the sample, $X_{(n)}$. To prove this, we need the factorization theorem.

There are two ways to check sufficiency:
(a) through the definition
(b) through the factorization theorem (below).

4.2 The Factorization Theorem

As before, denote by $\mathbf{X} = \{X_1, \ldots, X_n\}$ a random sample from the model $\{f(x \mid \theta), \quad \theta \in \Omega\}$ and by $f_n(\mathbf{x} \mid \theta)$ the joint p.d.f. or p.m.f. of $\mathbf{X}$. Let $T = T(X_1, \ldots, X_n) = r(X_1, \ldots, X_n)$ be a statistic with p.d.f. or p.m.f. $f_T(t \mid \theta)$. Note that we use $T(\cdot), r(\cdot)$ interchangeably.

Sufficiency Lemma. A statistic $T(X_1, \ldots, X_n)$ is sufficient for θ if, for every value $\mathbf{x}$ of $\mathbf{X}$, the ratio $\frac{f_n(\mathbf{x}|\theta)}{f_T(T(\mathbf{x})|\theta)}$ does not depend on θ.

Proof: The proof requires careful handling of the event $E = \{T(\mathbf{X}) = T(\mathbf{x})\}$ as typically $P(E) = 0$ for a continuous distribution (pdf case); details omitted. For the discrete (pmf) case, $\{\mathbf{X} = \mathbf{x}\} \subset \{T(\mathbf{X}) = T(\mathbf{x})\}$, hence

$$\begin{aligned} & f_\theta(\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = T(\mathbf{x})) \\ & \quad = \frac{f_\theta(\mathbf{X} = \mathbf{x}, T(\mathbf{X}) = T(\mathbf{x}))}{f_\theta(T(\mathbf{X}) = T(\mathbf{x}))} = \frac{f_n(\mathbf{X} = \mathbf{x} \mid \theta)}{f_\theta(T(\mathbf{X}) = T(\mathbf{x}))} = \frac{f_n(\mathbf{x} \mid \theta)}{f_T(T(\mathbf{x}) \mid \theta)}, \end{aligned}$$

where by assumption the latter does not depend on θ. Hence T is sufficient by definition of sufficiency.

Factorization Theorem (Neyman-Fisher). Let $X_1, \ldots, X_n$ be a random sample from $f(\mathbf{x}|\theta)$, $\theta \in \Omega$. A statistic $T = T(X_1, \ldots, X_n) = r(X_1, \ldots, X_n)$ is sufficient for θ if and only if

$$f_n(\mathbf{x}|\theta) = u(\mathbf{x}) v(r(\mathbf{x}), \theta)$$

for functions $u(\cdot), v(\cdot, \cdot)$ where u does not depend on θ and v depends on $\mathbf{x}$ only through $T = T(\mathbf{x}) = r(\mathbf{x})$.

Proof: The continuous case is omitted – it requires more advanced probabilistic arguments.
$\Rightarrow$: If T is sufficient, set $u(\mathbf{x}) = f_{\mathbf{X}|T}(\mathbf{x} \mid t)$; $v(r(\mathbf{x}), \theta) = f_T(t \mid \theta) = f(r(\mathbf{x}), \theta)$.
$\Leftarrow$: If factorization holds, with $A_{T(\mathbf{x})} = \{\mathbf{y} \in \mathbb{R}^n : \; T(\mathbf{y}) = T(\mathbf{x})\}$,

$$\begin{aligned} \frac{f_n(\mathbf{x} \mid \theta)}{f_T(T(\mathbf{x}) \mid \theta)} &= \frac{u(\mathbf{x}) v(T(\mathbf{x}) \mid \theta)}{\sum_{\mathbf{y} \in A_{T(\mathbf{x})}} u(\mathbf{y}) v(T(\mathbf{y}) \mid \theta)} \\ &= \frac{u(\mathbf{x}) v(T(\mathbf{x}) \mid \theta)}{v(T(\mathbf{x}) \mid \theta) \sum_{\mathbf{y} \in A_{T(\mathbf{x})}} u(\mathbf{y})} \\ &= \frac{u(\mathbf{x})}{\sum_{\mathbf{y} \in A_{T(\mathbf{x})}} u(\mathbf{y})}, \end{aligned}$$

which does not depend on θ. By the sufficiency lemma T is sufficient.

Examples:

1. Suppose we have a sequence of Bernoulli trials. In this case, the joint density is

$$f_n(\mathbf{x}|\theta) = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}.$$

   Letting $u = 1$ and $v(t, \theta) = \theta^t(1-\theta)^{n-t}$ we see that $T = \sum_{i=1}^n X_i$ is sufficient.

2. Suppose we have a uniform sample where

$$f_n(\mathbf{x}|\theta) = \begin{cases} \frac{1}{\theta^n} & \text{if } 0 \leq x_i \leq \theta \text{ for all } i \\ 0 & \text{otherwise} \end{cases}$$

   Note that the conditions for the joint pdf being non-zero are equivalent to $0 \leq X_{(1)} \leq X_{(n)} \leq \theta$. A more convenient way to write this joint pdf is using *indicator functions.* An indicator function $1_A(x)$ is defined as

$$1_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases}$$

   We can then use indicator functions to rewrite the joint pdf as

$$f_n(\mathbf{x}|\theta) = \frac{1}{\theta^n} 1_{[0,\theta]}(x_{(n)}) 1_{[0,\infty]}(x_{(1)}).$$

   If we let $u(\mathbf{x}) = 1_{[0,\infty]}(x_{(1)})$ and $v(t, \theta) = \frac{1}{\theta^n} 1_{[0,\theta]}(x_{(n)})$, we see that u does not depend on θ and v only depends on the data through $X_{(n)}$, which implies that $X_{(n)}$ is sufficient.

3. Suppose we have a random sample from a $N(\theta, \sigma^2)$ distribution where σ^2 is known.

The joint distribution in this case is

$$\begin{aligned} f_n(\mathbf{x}|\theta) &= \frac{1}{(\sqrt{2\pi}\sigma)^n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\theta)^2} \\ &= \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2\right) \exp\left(\frac{2\theta}{2\sigma^2}\sum_{i=1}^n x_i - \frac{n\theta^2}{2\sigma^2}\right). \end{aligned}$$

Letting $u(\mathbf{x}) = \frac{1}{(\sqrt{2\pi}\sigma)^n}\exp(-\frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2)$ and $v(t,\theta) = \exp(\frac{2\theta}{2\sigma^2}\sum_{i=1}^n x_i - \frac{n\theta^2}{2\sigma^2})$, we see that a sufficient statistic is $T = \sum_{i=1}^n X_i$. Note that the sufficient statistic is not unique; we can also choose $\overline{X}$ or $\frac{1}{2}\overline{X}$, for example.

Theorem on Functions of Sufficient Statistics. If T is sufficient for θ, any $1:1$ function $g(T) = T^*$ of T is also sufficient for θ.

Proof: Define the function $v^*(t^*,\theta) = v(g^{-1}(t^*),\theta) = v(t,\theta)$. Since T is sufficient, by the factorization theorem,

$$\begin{aligned} f_n(\mathbf{x}|\theta) &= u(\mathbf{x})v(t,\theta) \\ &= u(\mathbf{x})v(g^{-1}(t^*),\theta) \\ &= u(\mathbf{x})v^*(t^*,\theta), \end{aligned}$$

which implies that T^* is sufficient, again by the factorization theorem.

Examples:

1. Let $X_1,\ldots,X_n \sim_{i.i.d.} N(\mu,\sigma^2)$ where μ is known and σ^2 is unknown. Then $T = \sum_{i=1}^n (X_i-\mu)^2$ is sufficient for σ^2, since

$$f_n(\mathbf{x}|\theta) = \frac{1}{(\sqrt{2\pi}\sigma)^n}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2\right),$$

setting $u(\mathbf{x}) = 1$. But by the above theorem, $\frac{1}{n-1}\sum_{i=1}^n (X_i-\mu)^2$ and $\frac{1}{n}\sum_{i=1}^n (X_i-\mu)^2$ are also sufficient for σ^2.

2. Let $X_1, \ldots, X_n \sim_{i.i.d.} N(\theta, \theta^2)$. Then the joint pdf is

$$\begin{aligned} f_n(\mathbf{x}|\theta) &= \left(\frac{1}{\sqrt{2\pi}\theta}\right)^n \exp\left(-\frac{1}{2\theta^2}\sum_{i=1}^n (x_i - \theta)^2\right) \\ &= \left(\frac{1}{\sqrt{2\pi}\theta}\right)^n \exp\left(-\frac{1}{2\theta^2}\left\{\sum_{i=1}^n x_i^2 - 2\theta\sum_{i=1}^n x_i\right\}\right) \exp\left(-\frac{n}{2}\right). \end{aligned}$$

In this case, there is not a single sufficient statistic; we need both $\sum_{i=1}^n x_i$ and $\sum_{i=1}^n x_i^2$. This also holds when $X_1, \ldots, X_n \sim_{i.i.d.} N(\mu, \sigma^2)$, both μ and σ^2 unknown, where we also need both $\sum_{i=1}^n x_i$ and $\sum_{i=1}^n x_i^2$.

4.3 Jointly Sufficient Statistics

Definition: Let $X_1, \ldots, X_n \sim_{i.i.d.} f(x|\theta)$ for $\theta \in \Omega \subset \mathbb{R}$ or $\Omega \subset \mathbb{R}^p$ for $p > 1$. $T_j = r_j(\mathbf{X}),$ for $j = 1, \ldots, k$ is jointly sufficient for θ if and only if the conditional distribution of $\mathbf{X} = (X_1, \ldots, X_n)$ given $T = (T_1, \ldots, T_k)$ does not depend on θ, or

$$f_{\mathbf{X}|T_1,\ldots,T_k}(\mathbf{x}|t_1, \ldots, t_k)$$

does not depend on θ.

Factorization theorem in this case: $T_1, \ldots, T_k$ are jointly sufficient for θ if and only if

$$f_n(\mathbf{x}|\theta) = u(\mathbf{x})v(r_1(\mathbf{x}), \ldots, r_k(\mathbf{x}), \theta),$$

where $u(\mathbf{x})$ does not depend on θ, $v(\cdot)$ depends on $\mathbf{x}$ only through $r_j(\mathbf{x})$ and $T_j = r_j(\mathbf{X}),\ j = 1 \ldots, k$.

Example: Suppose $X_1, \ldots, X_n \sim_{i.i.d.} N(\theta, \theta^2)$. In this case, $(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ are jointly sufficient for θ (also the case for $N(\mu, \sigma^2)$). Notice that the map

$$(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2) \rightarrow (\overline{X}, \sum_{i=1}^n (X_i - \overline{X})^2)$$

is 1 : 1. To see this, note that the function

$$g(a, b) = \left(\frac{1}{n}a,\, b - \frac{1}{n}a^2\right) = (c, d)$$

has inverse

$$a = nc$$

$$b = d + nc^2.$$

We can also consider statistics like $(\overline{X}, \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2)$ and $(\overline{X}, \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2)$ which are both jointly sufficient.

Exponential Family. This is an important class of distributions $\{f(x \mid \theta),\, \theta \in \Omega\}$, where $\Omega \subset \mathcal{R}^k$ with p.d.f. or p.m.f.

$$f(x \mid \theta) = a(\theta)b(x)\exp[\sum_{j=1}^k c_j(\theta)d_j(x)].$$

It follows from the Factorization Theorem that

$$T = r(\mathbf{X}) = (\sum_{i=1}^n d_1(X_i), \ldots, \sum_{i=1}^n d_k(X_i))$$

is a sufficient statistic for θ in the exponential family. This is a very attractive property of the exponential family as it facilitates dimension reduction. When the data do not come from an exponential family or even a parametric family, data reduction may be challenging.

Example: Suppose $X_1, \ldots, X_n \sim_{i.i.d.} f(x|\theta)$, $\theta \in \Omega$, where f is a pdf. The order statistics $X_{(1)}, \ldots, X_{(n)}$ are always sufficient: One can record data in increasing order without losing information about θ. This seems obvious because the order of the X_i does not contain information about the parameter θ, i.e., $f(\mathbf{x} \mid x_{(1)}, \ldots, x_{(n)}) = \frac{1}{n!}$, since all $n!$ arrangements are equal likely.

A more formal proof: We have that

$$f_n(\mathbf{x}|\theta) = \prod_{i=1}^{n} f(x_i|\theta) = \prod_{i=1}^{n} f(x_{(i)}|\theta) = u(\mathbf{x})v(t_1, \ldots, t_n, \theta),$$

where $u(\mathbf{x}) = 1$ and $t_j = r_j(\mathbf{x}) = x_{(j)}$. Then by the factorization theorem, $(T_1, \ldots, T_n) = (X_{(1)}, \ldots, X_{(n)})$ are jointly sufficient.

Example: For the $N(\mu, \sigma^2)$ case, μ, σ^2 unknown, sufficient statistics include $(X_{(1)}, \ldots, X_{(n)})$ and $(\overline{X}, S^2)$. It seems clear that the latter pair of statistics is a much better summary of the data. In this case we need to identify the *simplest* sufficient statistic.

Definition: A statistic $T = (T_1, \ldots, T_k)$ is a <u>minimal sufficient</u> statistic of θ if T is sufficient and is a function of every other sufficient statistic.

In other words, T cannot be reduced further without destroying the property of sufficiency. This means that a function $\Psi(T)$ of a minimal sufficient statistic is minimal sufficient if and only if Ψ is a $1:1$ function: If T is minimal sufficient, then there must exist a function g s.t. $T = g(\psi(T))$, so that $g = \psi^{-1}$ and ψ is invertible. On the other hand, if ψ is invertible, then, for any sufficient statistic S, there must exist a function h s.t. $T = h(S)$ and then $\psi(T) = \psi(h(S))$, so that $\psi(T)$ is minimal sufficient.

Note that sufficient statistics obtained from the factorization theorem in the most concise way are usually minimal sufficient. How can we prove minimal sufficiency? In general, it is not easy except in the case of the MLE and Bayes estimators. Sometimes one can apply the following result (proof omitted).

Lehmann-Scheffé Theorem on Minimal Sufficient Statistics. Let $f_n(\mathbf{x}|\theta)$ be the p.d.f./p.m.f. of a sample $\mathbf{X}$. A statistic $T(\mathbf{X})$ is minimal sufficient if the following holds: For every two sample points $\mathbf{x}$ and $\mathbf{y}$ the ratio $\frac{f_n(\mathbf{x}|\theta)}{f_n(\mathbf{y}|\theta)}$ is constant as a function of θ if and only if $T(\mathbf{x}) = T(\mathbf{y})$.

Example. Consider a random sample $\mathbf{X}$ from a uniform distribution on the interval $[\theta, \theta+1]$, $-\infty < \theta < \infty$. Find a minimal sufficient statistic for θ.

Solution: The joint p.d.f. of $\mathbf{X}$ is

$$f_n(\mathbf{x}|\theta) \quad = \quad \begin{cases} 1, & \text{if } \theta \leq x_i \leq \theta+1,\ 1 \leq i \leq n \\ 0, & \text{otherwise,} \end{cases}$$

which is equivalent to

$$f_n(\mathbf{x}|\theta) \quad = \quad \begin{cases} 1, & \text{if } x_{(n)} - 1 \leq \theta \leq x_{(1)} \\ 0, & \text{otherwise.} \end{cases}$$

Then for any two $\mathbf{x}$ and $\mathbf{y}$ the ratio $\frac{f_n(\mathbf{x}|\theta)}{f_n(\mathbf{y}|\theta)}$ will not involve θ if and only if $x_{(1)} = y_{(1)}$ and $x_{(n)} = y_{(n)}$. Thus, $X_{(1)}$ and $X_{(n)}$ are minimal (joint) sufficient statistics. Note that here the minimal sufficient statistic has dimension two while the parameter is only one dimensional.

A minimal sufficient statistic is not unique as any one-to-one function of it carries exactly the same information so is also minimal sufficient. Although a sufficient statistic contains all the relevant information in the data about θ, this does not mean that we should discard the original data once we have extracted the information contained in the sufficient statistic. One reason is that in order to do statistical inference, or for uncertainty quantification, we may need information beyond sufficient statistics.

An easier way to obtain minimal sufficient statistics is to go through MLE or Bayes estimators.

Theorem. For the model $\{f(x|\theta),\, \theta \in \Omega\}$, if a sufficient statistic (or jointly sufficient statistic) T exists, and the MLE $\hat{\theta}$ exists, then $\hat{\theta}$ depends on the observations only through

T, i.e., $\hat{\theta} = h(T)$ for some function h.

Proof: By the factorization theorem, for any sufficient statistic $T = r(\mathbf{x})$,

$$\tilde{L}(\theta) = f(\mathbf{x}|\theta) = u(\mathbf{x})v(r(\mathbf{x}), \theta).$$

Then maximizing $\tilde{L}(\theta)$ in terms of θ is equivalent to maximizing $v(r(\mathbf{x}), \theta)$ in terms of θ. This implies that $\hat{\theta}$ is a function of $r(\mathbf{X})$. The same holds for Bayes estimators because

$$f(\theta|\mathbf{x}) \propto f(\mathbf{x}|\theta)\xi(\theta) \propto v(r(\mathbf{x}), \theta)\xi(\theta),$$

which implies that any Bayes estimator $\hat{\theta}$ must be a function of $T = r(\mathbf{X})$.

Corollary: If the MLE (or Bayes estimator) $\hat{\theta}$ is sufficient, then $\hat{\theta}$ is minimal sufficient.

Proof: As above, the factorization theorem applied to any sufficient statistic $T = r(\mathbf{x})$ implies that $\hat{\theta}$ is a function of T, so if $\hat{\theta}$ is sufficient, it is minimal sufficient.

Examples:

1. Suppose that we have a random sample from a $N(\mu, \sigma^2)$ distribution, where both μ and σ^2 are unknown. Then the MLE $(\overline{X}, \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2)$ is sufficient by the factorization theorem, which implies that it is minimal sufficient because it is also the MLE. We also know that $(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ is minimal sufficient, because it is a $1:1$ function of $(\overline{X}, \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2)$.

2. Suppose $X_1, \ldots, X_n \sim U(\theta_1, \theta_2)$, θ_1, θ_2 unknown. Here $\theta_1 \leq X_{(1)} \leq \cdots \leq X_{(n)} \leq \theta_2$ and

   $$f_n(\mathbf{x}|\theta_1, \theta_2) = \frac{1}{(\theta_2 - \theta_1)^n} 1_{(-\infty, \theta_2]}(x_{(n)}) 1_{[\theta_1, \infty)}(x_{(1)}).$$

   From the joint pdf we can see that

   a) $(X_{(1)}, X_{(n)})$ are jointly sufficient for (θ_1, θ_2) (from the factorization theorem)

   b) $\tilde{L}(\theta_1, \theta_2)$ is maximized when $\theta_2 - \theta_1$ is minimized. Therefore the MLEs for (θ_1, θ_2) are $(X_{(1)}, X_{(n)})$.

c) Then, since the MLE is sufficient, it is minimal sufficient for (θ_1, θ_2).

4.4 Improving an Estimator

Rao–Blackwell theory: Consider the model $f(\mathbf{x}|\theta),\ \theta \in \Omega$, a sample $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(\mathbf{x}|\theta)$, an estimator $\delta(X_1, \ldots, X_n)$ for θ, and a new estimator $\delta_0(T) = E(\delta(X_1, \ldots, X_n)|T)$, where $T = r(X_1, \ldots, X_n)$ is a sufficient statistic for θ.

Note: Since the definition of $\delta_0(T)$ is a conditional expectation conditioned on a sufficient statistic T, its distribution does not depend on θ. This is called "Rao–Blackwellization".
Note: For $Y \sim f(\cdot|\theta)$, $E_\theta Y = \int y f(y|\theta) dy$.
For any r.v.s Y and Z, $E(Y|Z) = \int y f_{Y|Z}(y, Z) dy = h(Z)$ for a function h.

Mean Squared Error as measure of quality of an estimator:
$\text{MSE}_\theta(\delta) = R(\theta, \delta) = E_\theta(\delta - \theta)^2$
for an estimator δ of θ; MSE is a measure of risk.

Rao–Blackwell theorem: $R(\theta, \delta_0) \leq R(\theta, \delta)$ for all $\theta \in \Omega$,
and if δ is not a function of T and $R(\theta, \delta) < \infty$, then $R(\theta, \delta_0) < R(\theta, \delta)$ for all $\theta \in \Omega$.

Note: The theorem suggests to look for estimators that are functions of a minimal sufficient statistic T. When starting with an arbitrary estimator δ, $E(\delta|T)$ is often difficult to compute.

Definition (Admissibility) An estimator δ is inadmissible if and only if there exists another estimator δ_0 such that
(a) $R(\theta, \delta_0) \leq R(\theta, \delta)$ for all $\theta \in \Omega$; and
(b) $R(\theta, \delta_0) < R(\theta, \delta)$ for at least one $\theta \in \Omega$.
If δ is inadmissible, δ_0 dominates δ with respect to the risk $R(\theta, \delta)$. An estimator δ_0 is admissible if there is no other estimator that dominates it.

Note: The Rao–Blackwell theorem implies that any estimator δ that is not a function of any given sufficient statistic T must be inadmissible, because it is dominated by $\delta_0 = E(\delta(\mathbf{X})|T)$.

Note: Admissible estimators must be a function of all sufficient statistics so must be a minimally sufficient statistic. But this is not sufficient for admissibility. General goal is to avoid inadmissible estimators, while not all admissible estimators have good properties.

5 Properties of Estimators

5.1 Unbiasedness

Definition An estimator $\hat{\theta} = \delta(X_1, \ldots, X_n)$ is an *unbiased* estimator of θ if and only if $E(\hat{\theta}) = \theta$.

Bias: $\text{bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$

Example. $E(\bar{X}) = E(n^{-1}\sum_{i=1}^n X_i) = n^{-1}\sum_{i=1}^n E(X_i) = \mu$. This implies $\bar{X}$ is an unbiased estimator for μ.

Example. $m_k = n^{-1}\sum_{i=1}^n X_i^k$ is an unbiased estimator of $\mu_k = E(X_1^k)$.

Example. Method of moment estimators may be biased: Consider MOM of σ^2,

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \bar{X}^2 = m_2 - m_1^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2.$$

This is also the MLE. Can show that $E(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2$, which implies $\text{bias}(\hat{\sigma}^2) = E(\hat{\sigma}^2) - \sigma^2 = -\frac{\sigma^2}{n}$ (the bias is small for large samples). Hence, $E(\frac{n}{n-1}\hat{\sigma}^2) = E(\frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2) = \sigma^2$ and therefore the estimator $s^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$ is unbiased.

Example. The MLE of σ^2 when μ is known in $N(\mu, \sigma^2)$ is $\hat{\sigma}^2 = n^{-1} \sum_{i=1}^n (X_i - \mu)^2$. This MLE is unbiased for σ^2, in this specific situation, because $E(X_i - \mu)^2 = \text{var}(X_1) = \sigma^2$.

Note: Unbiased estimators are not unique.

Example. $X_1, \dots, X_n \sim \text{Pois}(\theta)$ implies $\mu = \sigma^2 = \theta = \lambda$. Hence $\bar{X}$ and s^2 are both unbiased for θ, which implies all estimators $\alpha\bar{X} + (1-\alpha)s^2$ are unbiased, for all $\alpha \in \mathbb{R}$. Which estimator is the best among all unbiased estimators?
→ the one with minimum variance

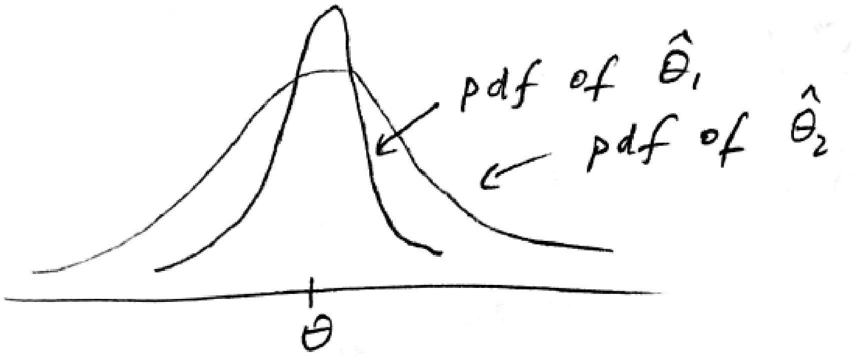
Example. There is no unbiased estimator for $\sqrt{\theta}$ in a Bernoulli(θ) model, $0 < \theta < 1$:
For $n = 1$ suppose $\hat{\theta} = \delta(X)$, $X \in \{0, 1\}$ is an unbiased estimator. This implies
$E(\hat{\theta}) = \delta(1)\theta + \delta(0)(1-\theta) = \sqrt{\theta}$
needs to hold for all θ. The left side is linear in θ, which cannot equal to $\sqrt{\theta}$ for all θ. This shows that unbiased estimators may not exist.

Example. Let $X = \#$ of failures until the first success in a sequence of Bernoulli trials Bernoulli(θ). The target is θ. Then $P(X = k) = (1-\theta)^k\theta$, $k = 0, 1, 2, \dots$ (geometric distribution). Assume $\hat{\theta} = \delta(X)$ is unbiased. Then $E(\hat{\theta}) = \sum_{k=0}^\infty \delta(k)(1-\theta)^k\theta = \theta$, implying $\sum_{k=0}^\infty \delta(k)(1-\theta)^k = 1$, whence $\delta(0) = 1, \quad \delta(k) = 0$ for $k \geq 1$. This means the unbiased estimator is $\hat{\theta} = \delta(X) = 1$ if the first trial is a success (no waiting to the first success so this corresponds to $k = 0$) , and $\hat{\theta} = \delta(X) = 0$ if the first trial is a failure, which means that $k \geq 1$. This unbiased estimator is clearly unreasonable, so unbiased estimators are not always desirable.

5.2 Uniformly Minimum Variance Unbiased Estimators and Completeness

How can we compare various unbiased estimators? Obviously an estimator that has smaller variance for all $\theta \in \Omega$ is better:

Assume two estimators $\hat\theta_1$, $\hat\theta_2$ are unbiased and $\mathrm{var}(\hat\theta_1) \leq \mathrm{var}(\hat\theta_2)$ for all θ. This implies $\hat\theta_1$ is better than $\hat\theta_2$, as it tends to be closer to the true value θ.



Definition If an unbiased estimator θ^* of θ is such that $\mathrm{var}(\theta^*) \leq \mathrm{var}(\hat\theta)$ for all $\theta \in \Omega$, and all estimators $\hat\theta$, then θ^* is a *Uniformly Minimum Variance Unbiased Estimator* (UMVUE) of θ.

Let $X_1, \dots X_n \sim f(x|\theta)$ be a random sample from the model $f(x|\theta)$, $\theta \in \Omega$. A statistic $T = T(\mathbf{X})$ is *complete* if for any function g and for all $\theta \in \Omega$ it holds that if $E_\theta(g(T)) = 0$ for all $\theta \in \Omega$ then $P_\theta(g(T) = 0) = 1$ for all $\theta \in \Omega$.

It means that the statistic T is in some sense maximally compressed.

Example. Consider a random sample $X_1, \dots X_n$ from the Bernoulli model $\mathrm{B}(p)$, where $0 < p < 1$ and $T = \sum_{i=1}^n X_i$. Then T is complete: With $z = p/(1-p)$, for an arbitrary function g, if $E_p(g(T)) = 0$, then

$$E_p(g(T)) = \sum_{t=0}^{n} g(t) \binom{n}{t} p^t (1-p)^{n-t} = (1-p)^n \sum_{t=0}^{n} g(t) \binom{n}{t} z^t = 0.$$

A polynomial of order n is zero on its entire domain only when its coefficients are zero, so that $g(t) = 0$ for all $0 \leq t \leq n$, i.e., $P_\theta(g(T) = 0) = 1$.

Generally, it is difficult to show completeness of a statistic. A general result is

Theorem on Complete Statistics in the Exponential Family. Assume a random sample $X_1, \ldots, X_n$ is obtained from a distribution in the exponential family

$$f(x \mid \theta) = a(\theta)b(x)\exp[\sum_{j=1}^{k} c_j(\theta)d_j(x)],\ \theta \in \Omega.$$

If Ω contains an open set, the statistic $T = (\sum_{i=1}^{n} d_1(X_i), \ldots, \sum_{i=1}^{n} d_k(X_i))$ is complete.

Note: The above example for $T(\mathbf{X}) = \sum_{i=1}^{n} X_i$ being a complete statistic in the Bernoulli model is a consequence of this theorem. The theorem does not apply to models such as $N(\theta, \theta^2)$ since there the set $(\mu, \sigma^2) = (\theta, \theta^2)$ is a curve in $\mathcal{R}^2$ that does not contain an open set.

Complete statistics are connected to minimal sufficient statistics as follows:

Theorem. If a minimal sufficient statistic exists, then any complete statistic is also a minimal sufficient statistic.

The next result shows how to find the UMVUE through a complete sufficient statistic.

Lehmann-Scheffé Theorem on UMVUE. Let $T = T(\mathbf{X})$ be a complete sufficient statistic for θ in the model $f(x|\theta)$ and $W = W(\mathbf{X})$ an unbiased estimator of a function of the parameter of interest $g(\theta)$. Then $\phi(T) = E_\theta(W \mid T)$ is the unique UMVUE of $g(\theta)$.

Proof. Because $E_\theta\phi(T) = E_\theta W = g(\theta)$, $\phi(T)$ is unbiased. By the Rao-Blackwell Theorem, since T is sufficient, $\text{var}_\theta(\phi(T)) \leq \text{var}_\theta(W)$ for all θ. Let S be any other unbiased estimator and $\phi^*(T) = E(S \mid T)$, then $E_\theta[\phi(T) - \phi^*(T)] = 0$ for all θ. Therefore, by the completeness of T, $\quad P(\phi(T) = \phi^*(T)) = 1$ for all θ. Hence $\phi(T)$ is the unique UMVUE.

The Lehmann-Scheffé theorem on UMVUE implies that when T is complete and sufficient, there is at most one function of T that is unbiased for $g(\theta)$.

How to apply the Lehmann-Scheffé Theorem on UMVUE:

1. Find a complete sufficent statistic T (straightforward if the model belongs to the exponential family, otherwise can be difficult).

2. If you can find an unbiased estimator $\phi(T)$, this is UMVUE, since $E_\theta(\phi(T) \mid T) = \phi(T)$.

3. Otherwise, find any unbiased estimator $W(\mathbf{X})$ and then compute $\phi(T) = E(W \mid T)$.

Example. Let $X_1, \ldots, X_n$ be a random sample from a Poisson distribution with parameter λ, i.e. $f(x \mid \lambda) = e^{-\lambda}\frac{\lambda^x}{x!} = \frac{1}{x!}e^{-\lambda}e^{x\log\lambda}$.

Since the Poisson distribution belongs to the exponential family,

$f(x \mid \theta) = a(\theta)b(x)\exp[c_1(\theta)T_1(x)]$,

with $a(\theta) = e^{-\lambda}, b(x) = \frac{1}{x!}, c_1(\theta) = \log\lambda, T_1(x) = x$ and the parameter space includes an open interval, e.g. (0,1), the theorem on complete statistics in the exponential family implies that $\sum_{i=1}^n X_i$ is a complete sufficient statistic T.

The sample mean $\bar{X}$ is a function of $\sum_{i=1}^n X_i$ and is an unbiased estimator of λ , so the L-S Theorem implies that $\bar{X} = E_\theta(\bar{X}|T)$ is the UMVUE of λ.

The L-S theorem also implies that $\bar{X}$ is the only unbiased estimator that is a function of $\bar{X}$ or $\sum_{i=1}^n X_i$. Note that this is not in contradiction to the fact that $\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$ is also an unbiased estimator of λ (the variance of X is λ).

5.3 Mean Squared Error of Estimators

How to compare two estimators (not necessarily unbiased)? One may have larger variance, the other larger bias, etc. A good criterion for such comparisons is the Mean Squared Error (MSE):

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

Estimator $\hat{\theta}_1$ is better than $\hat{\theta}_2$ w.r.t. MSE if

$$\text{MSE}(\hat{\theta}_1) \leq \text{MSE}(\hat{\theta}_2) \quad \text{for all} \quad \theta \in \Omega,$$

with strict inequality for at least one θ.

Theorem. $\mathrm{MSE}(\hat\theta) = [\,\mathrm{bias}(\hat\theta)]^2 + \mathrm{var}(\hat\theta)$.

Proof.

$$\begin{aligned}\mathrm{MSE}(\hat\theta) &= E(\hat\theta-\theta)^2 = E\{[\hat\theta - E(\hat\theta)] + [E(\hat\theta)-\theta]\}^2 \\ &= E(\hat\theta - E(\hat\theta))^2 + [\,\mathrm{bias}(\hat\theta)]^2 + 2E\{[\hat\theta - E(\hat\theta)][E(\hat\theta)-\theta]\} \\ &= \mathrm{var}(\hat\theta) + [\,\mathrm{bias}(\hat\theta)]^2 + 2(E(\hat\theta)-\theta)\underbrace{E(\hat\theta - E(\hat\theta))}_{=0} \\ &= \mathrm{var}(\hat\theta) + [\,\mathrm{bias}(\hat\theta)]^2.\end{aligned}$$

Note: If both $\hat\theta_1$ and $\hat\theta_2$ are unbiased, then $\mathrm{MSE}(\hat\theta_j) = \mathrm{var}(\hat\theta_j)$, $j = 1, 2$, and we only need to compare the variances.

Note: Squared errors are just one criterion to compare estimators, can also use absolute errors, $E|\hat\theta - \theta|$ or could use $E[\,\exp(\hat\theta-\theta)^2]$ etc.

Example. The methods of moments estimator for the variance σ^2 is

$$\hat\sigma^2 = n^{-1}\sum_{i=1}^n (X_i - \bar X)^2 = n^{-1}\sum_{i=1}^n X_i^2 - (n^{-1}\sum_{i=1}^n X_i)^2 = m_2 - m_1^2,$$

while

$$s^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar X)^2,$$

is the “sample variance”.

Then s^2 is unbiased, but $\mathrm{MSE}(\hat\sigma^2) \le \mathrm{MSE}(s^2)$. (Show this).

6 Fisher Information and Efficient Estimation

Fisher information quantifies the amount of information in a sample about a parameter θ in a model $\{f(x,\theta),\ \theta\in\Omega\}$.

Basic assumptions:

(A1) $X \sim f(X|\theta),\ \theta \in \Omega \subset \mathbb{R}$

(A2) $f(x|\theta) > 0$ for $x \in S,\ \theta \in \Omega$, where the domain S does not depend on θ

(A3) $f(x|\theta)$ is twice differentiable as a function of θ.

Note: (A2) excludes models like $U(0,\theta)$, where $S = (0,\theta)$ depends on θ.

Definition Under assumptions (A1)–(A3), the Fisher information in a single r.v. X (or a sample with a single observation) is defined as

$$I(\theta) = E_\theta\left\{\left[\frac{\partial \log f(X|\theta)}{\partial\theta}\right]^2\right\} = \int_S \left[\frac{\partial \log f(x|\theta)}{\partial\theta}\right]^2 f(x|\theta)dx.$$

Remember: $Eg(X) = \int g(x)f_X(x)dx$, if $X \sim f_X$, $E_\theta g(X) = \int g(x)f(x|\theta)dx$.
In the case of a discrete r.v., the integral becomes a sum over the possible outcomes.

Theorem. Under (A1)–(A3),

(a) $0 \le I(\theta) \le \infty$

(b) $E(\frac{\partial \log f(X|\theta)}{\partial\theta}) = 0$ and $\mathrm{var}(\frac{\partial \log f(X|\theta)}{\partial\theta}) = I(\theta)$

(c) $I(\theta) = -E(\frac{\partial^2 \log f(X|\theta)}{\partial\theta^2}) = -\int_S \frac{\partial^2 \log f(x|\theta)}{\partial\theta^2} f(x|\theta)d\theta.$

Proof.

(a) Obvious

(b) $\int f(x|\theta)dx = 1$ for all $\theta \implies 0 = \frac{\partial}{\partial\theta}\int_S f(x|\theta)dx \underbrace{=}_{\text{by assumptions}} \int_S \frac{\partial}{\partial\theta} f(x|\theta)dx = \int_S \frac{\partial \log f(x|\theta)}{\partial\theta} f(x|\theta)dx = E\left(\frac{\partial \log f(X|\theta)}{\partial\theta}\right).$

(c) First observe

$$0 = \frac{\partial^2}{\partial\theta^2}\int_S f(x|\theta)dx \underbrace{=}_{\text{by assumptions}} \int_S \frac{\partial^2}{\partial\theta^2} f(x|\theta)dx \qquad (*)$$

$$\frac{\partial^2 \log f(x|\theta)}{\partial\theta^2} = \frac{\partial}{\partial\theta}\left[\frac{\frac{\partial}{\partial\theta}f(x|\theta)}{f(x|\theta)}\right] = \frac{\left[\frac{\partial^2}{\partial\theta^2}f(x|\theta)\right] f(x|\theta) - \left[\frac{\partial}{\partial\theta}f(x|\theta)\right]^2}{f(x|\theta)^2} \qquad (**)$$

Then

$$\begin{aligned} E\left(\frac{\partial^2 \log f(X|\theta)}{\partial\theta^2}\right) &= \int_S \left(\frac{\partial^2 \log f(x|\theta)}{\partial\theta^2}\right) f(x|\theta)dx \\ &\overset{(**)}{=} \int_S \frac{\partial^2}{\partial\theta^2} f(x|\theta)dx - \int_S \left[\frac{\partial}{\partial\theta}\log f(x|\theta)\right]^2 f(x|\theta)dx \overset{(*)}{=} -I(\theta) \implies (c) \end{aligned}$$

Definition Fisher information for θ in a random sample $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x|\theta)$:

$$\begin{aligned} I_n(\theta) &= E_\theta\left[\frac{\partial}{\partial\theta}\log f_n(\mathbf{x}|\theta)\right]^2 \\ &= \underbrace{\int_S \cdots \int_S}_{n \text{ times}} \left(\frac{\partial}{\partial\theta}\log f_n(\mathbf{x}|\theta)\right)^2 f_n(\mathbf{x}|\theta)dx_1 \ldots dx_n \\ &= \sum_{i=1}^n \int_S \left(\frac{\partial}{\partial\theta}\log f(x_i|\theta)\right)^2 f(x_i|\theta)dx_i = nI(\theta). \end{aligned}$$

Theorem. Under (A1)–(A3),

(a) $0 \le I_n(\theta) \le \infty$

(b) $I_n(\theta) = \text{ var}\left(\frac{\partial}{\partial\theta}\log f_n(\mathbf{X}|\theta)\right)$

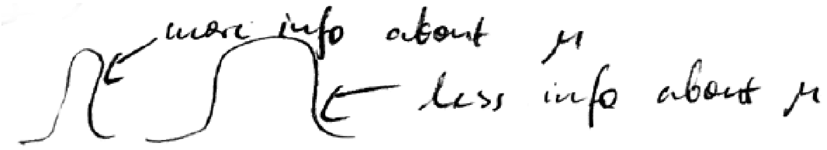
(c) $I_n(\theta) = -E_\theta\left(\frac{\partial^2}{\partial\theta^2}\log f_n(\mathbf{X}|\theta)\right) = nI(\theta)$

Proof: Similar to one-dimensional case.

Example $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$: A. σ^2 is known, $\theta = \mu$ is unknown parameter of interest.

$$f(x|\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\theta)^2}{2\sigma^2}\right)$$
$$\log f(x|\theta) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(x-\theta)^2}{2\sigma^2}$$
$$\frac{\partial \log f(x|\theta)}{\partial\theta} = \frac{(x-\theta)}{\sigma^2}, \qquad \frac{\partial^2 \log f(x|\theta)}{\partial\theta^2} = -\frac{1}{\sigma^2}$$
$$\implies I(\theta) = I(\mu) = -(-\frac{1}{\sigma^2}) = \frac{1}{\sigma^2}.$$

This means the larger σ^2, the less information one has about μ. (why?)



Note: $I_n(\theta) = \frac{n}{\sigma^2}$, where one also has $\operatorname{var}(\bar{X}) = \sigma^2/n$.

B. μ is known, σ^2 is unknown: Set $\theta^* = \sigma^2$. Then

$$\log f(x|\sigma^2) = -\frac{1}{2}\log(2\pi\theta^*) - \frac{(x-\mu)^2}{2\theta^*}, \quad \frac{\partial}{\partial\theta^*}\log f(x|\theta^*) = -\frac{1}{2\theta^*} + \frac{(x-\mu)^2}{2(\theta^*)^2},$$
$$\frac{\partial^2}{\partial\theta^{*2}}\log f(x|\theta^*) = \frac{1}{2\theta^{*2}} - \frac{(x-\mu)^2}{\theta^{*3}} \quad \text{and observing} \quad E(X-\mu)^2 = \theta^*,$$
$$E\left(\frac{\partial^2}{\partial\theta^{*2}}\log f(X|\theta^*)\right) = -\frac{1}{2\theta^{*2}}$$
$$I(\theta) = -E\left(\frac{\partial^2}{\partial\theta^{*2}}\log f(X|\theta^*)\right) = \frac{1}{2\sigma^4}, \quad I_n(\theta) = \frac{n}{2\sigma^4}.$$

Also observe that as the above theorem says, indeed $E\left(\frac{\partial}{\partial\theta^*}\log f(X|\theta^*)\right) = 0$.

Cramér–Rao (Information) Inequality/Bound: Let $T = r(X_1, \ldots, X_n) = r(\mathbf{X})$ be a statistic with finite variance, $\operatorname{var}_\theta(T) < \infty$. Assume $m(\theta) = E_\theta(T)$ is differentiable in θ. Then:

$$\operatorname{var}_\theta(T) \geq \frac{m'(\theta)^2}{nI(\theta)} = \frac{m'(\theta)^2}{I_n(\theta)},$$

with equality if and only if

$$\left[\operatorname{corr}\left(T, \frac{\partial \log f_n(\mathbf{X}|\theta)}{\partial\theta}\right)\right]^2 = 1, \quad \text{or equivalently} \quad T = u(\theta)\frac{\partial \log f_n(\mathbf{X}|\theta)}{\partial\theta} + v(\theta)$$

for some functions u, v.

Proof:

$$\begin{aligned}
\operatorname{cov}\left(T, \frac{\partial \log f_n(\mathbf{X}|\theta)}{\partial\theta}\right) &= E\left(T\frac{\partial \log f_n(\mathbf{X}|\theta)}{\partial\theta}\right) \quad \left(\text{since } E\left(\frac{\partial \log f_n(X|\theta)}{\partial\theta}\right) = 0\right) \\
&= \int_S \cdots \int_S r(\mathbf{x}) \underbrace{\frac{\partial}{\partial\theta}\left[\log f_n(\mathbf{x}|\theta)\right]}_{\frac{\partial}{\partial\theta} f_n(\mathbf{x}|\theta)/f_n(\mathbf{x}|\theta)} f_n(\mathbf{x}|\theta) dx_1 \ldots dx_n \\
&= \underbrace{\frac{\partial}{\partial\theta} \int_S \cdots \int_S r(\mathbf{x}) f_n(\mathbf{x}|\theta) dx_1 \ldots dx_n}_{E_\theta(T) = m(\theta)} = m'(\theta).
\end{aligned}$$

Now use $\operatorname{cov}(X,Y)^2 \leq \operatorname{var}(X)\operatorname{var}(Y)$ (book Theorem 4.6.3, Cauchy-Schwarz inequality) for any r.v. X, Y, since $|\rho| \leq 1$ (correlation). This implies

$$\begin{aligned}
m'(\theta)^2 &\leq \operatorname{var}_\theta(T)\operatorname{var}\left(\frac{\partial \log f_n(\mathbf{X}|\theta)}{\partial\theta}\right) \\
&= \operatorname{var}_\theta(T) n I(\theta),
\end{aligned}$$

where the case of equality (last statement) is handled via Ex. 17 Section 4.6 of the book.

Corollary. If T is an unbiased estimator of θ, then $m(\theta) = \theta$, $m'(\theta) = 1$, which implies

$$\operatorname{var}_\theta(T) \geq \frac{1}{nI(\theta)}.$$

This lower bound on the variance of unbiased estimators implies that the best MSE that an unbiased estimator can achieve is the Cramér–Rao bound.

Definition An estimator T is an efficient estimator of its expected value $m(\theta)$ if and only if its variance achieves the Cramér–Rao bound$\frac{[m'(\theta)]^2}{nI(\theta)}$.

An unbiased estimator T of θ is efficient $\iff$ $\text{var}_\theta(T) = \frac{1}{nI(\theta)}$.
If T is unbiased and efficient then T is an UMVUE of θ.

Example $X_1, \dots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, $\mu = \theta$ unknown, σ^2 known. We calculated $I(\mu) = \sigma^{-2}$. Consider the MLE $= \bar{X}$ which is also the MOM since $E\bar{X} = \mu$ (unbiased) $\implies m'(\theta) = 1$. $\text{var}(\bar{X}) = \sigma^2/n = 1/(nI(\mu)) \implies \bar{X}$ is efficient and is UMVUE.
Additionally we know from the above theorem that there are functions $u(\theta), v(\theta)$ such that

$$\bar{X} = u(\theta)\frac{\partial \log f_n(\mathbf{X}|\theta)}{\partial\theta} + v(\theta) = \frac{\sigma^2}{n}\left[\frac{1}{\sigma^2}\sum_{i=1}^n X_i - \frac{n\theta}{\sigma^2}\right] + \theta$$

for $u(\theta) = \sigma^2/n$, $v(\theta) = \theta$.

Example $X_1, \dots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, where μ is known and σ^2 is unknown, and let $\theta = \sigma^2$. To find efficient estimators, we use the linear representation of T in terms of functions u, v as per the Cramér–Rao bound:

$$\begin{aligned}
\frac{\partial \log f_n(\mathbf{X}|\theta)}{\partial\theta} &= -\frac{n}{2\theta} + \frac{\sum_{i=1}^n (X_i - \mu)^2}{2\theta^2} \\
&\implies \sum_{i=1}^n (X_i - \mu)^2 = \underbrace{2\theta^2}_{u(\theta)} \frac{\partial \log f_n(\mathbf{X}|\theta)}{\partial\theta} + \underbrace{n\theta}_{v(\theta)} \\
&\implies T = \sum_{i=1}^n (X_i - \mu)^2 \text{ is efficient for } ET = n\theta \\
&\implies \frac{T}{n} = \frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2 \text{ is efficient for } \theta = \sigma^2.
\end{aligned}$$

Alternatively we can verify that T/n is efficient by directly calculating its variance:

$$\text{var}(T/n) = \frac{n}{n^2}\,\text{var}[(X_1 - \mu)^2] = \frac{1}{n}[3\sigma^4 - \sigma^4] = \frac{2\sigma^4}{n}$$

since $E(X_1-\mu)^4 = 3\sigma^4$ for Gaussian r.v.s (book §5.6). On the other hand,

$$I(\theta) = -E\left(\frac{\partial^2}{\partial\theta^2}\log f(X|\theta)\right) = -\left[-\frac{1}{2\theta^2} - \frac{E(X_1-\mu)^2}{\theta^3}\right] = \frac{1}{2\sigma^2}$$

as calculated above, which means that $I_n(\theta) = \frac{n}{2\sigma^2}$ and since $\operatorname{var}(T/n) = \frac{1}{nI(\theta)} = \frac{2\theta^2}{n}$, T/n achieves the Cramér–Rao bound. Also $E(T/n) = \sigma^2$ so it is unbiased, which means T/n is a UMVUE.

7 Sampling Distributions

7.1 Sampling Distribution of a Statistic and Consistency

Goal: Uncertainty/variance quantification of the estimators.
For this, need distribution of $\hat{\theta}$, $\quad \hat{\theta} = r(X_1,\dots,X_n)$, the "sampling distribution" of $\hat{\theta}$.

Notation. Here and in the following, ϕ is the pdf of a standard normal r.v. and Φ is its distribution function.

Definition. For a statistic $T = r(X_1,\dots,X_n)$, a function of the random sample $(X_1,\dots,X_n)$, the distribution of T as a r.v. is the "sampling distribution" of T.

Example 1a. $X_1,\dots,X_n \overset{\text{iid}}{\sim} f(x|\theta)$ with $EX_1 = \mu$, $\operatorname{var}(X_1) = \sigma^2 < \infty$. Central Limit Theorem $\Longrightarrow$ $\sqrt{n}(\frac{\bar{X}-\mu}{\sigma}) \xrightarrow{D} N(0,1)$ $\iff$ $P(\sqrt{n}\left(\frac{\bar{X}-\mu}{\sigma}\right) \leq x) \overset{n\to\infty}{\to} \int_{-\infty}^{x}\phi(v)dv = \Phi(x)$ at all $x \in \mathbb{R}$ $\Longrightarrow$ Sampling distribution of $\sqrt{n}(\bar{X}-\mu)$ converges to $N(0,\sigma^2)$ in distribution as $n \to \infty$.

Definition. $\hat{\theta}$ is consistent for θ if for all $\delta > 0$ it holds that $P(|\hat{\theta}-\theta| > \delta) \to 0$ as $n \to \infty$.

Example 1b. When $f(x|\theta)$ is normal, $\sqrt{n}(\bar{X}-\mu) \sim N(0,\sigma^2)$ for all n, not only in the limit as $n \to \infty$. This implies: $\bar{X} \xrightarrow{P} \mu$ ($\bar{X}$ is consistent for μ), i.e. $P(|\bar{X}-\mu| > \delta) \overset{n\to\infty}{\longrightarrow} 0$

for all $\delta > 0$.

Example 2. Markov inequality: For any r.v. X,

$$P(|X| > \delta) \leq \frac{EX^2}{\delta^2},$$

which implies

$$P(|\hat{\theta} - \theta| > \delta) \leq \frac{E(\hat{\theta} - \theta)^2}{\delta^2} \to 0,$$

i.e., consistency if $\text{MSE}(\hat{\theta}) \to 0$.

Example 3. $X_{(n)}$ is consistent for θ in the model $U(0, \theta)$:

$$P(|X_{(n)} - \theta| > \delta) = P(X_{(n)} \leq \theta - \delta) = P(\text{all } X_i \leq \theta - \delta) = \left(\frac{\theta - \delta}{\theta}\right)^n = \left(1 - \frac{\delta}{\theta}\right)^n \overset{n\to\infty}{\longrightarrow} 0$$

Example 4. Consider variance estimation:

$$\begin{aligned}
\hat{\sigma}_0^2 &= \hat{\sigma}_{0n}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2 \quad \mu \text{ known}, \\
\hat{\sigma}_1^2 &= \hat{\sigma}_{1n}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2 \quad \mu \text{ unknown} \\
\hat{\sigma}_2^2 &= \hat{\sigma}_{2n}^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2 = s_n^2 \quad \mu \text{ unknown.}
\end{aligned}$$

Consider a sample $(X_1, \ldots, X_n) \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, then the *standardized* data are $(X_i - \mu)/\sigma \overset{\text{iid}}{\sim} N(0, 1)$. Then

$$\frac{n}{\sigma^2}\hat{\sigma}_0^2 = \sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi_n^2.$$

Definition. χ^2 distribution: The χ^2 distribution with m d.f. (degrees of freedom) is Gamma$(m/2, 1/2)$, written as χ_m^2.

Reminder: $X \sim$ Gamma(α, β), if it has pdf $f(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \quad \alpha, \beta > 0, x \geq 0$, with domain $(0, \infty)$ and $EX = \alpha/\beta$, $\text{var}(X) = \alpha/\beta^2$.

Properties:

1. $E\chi^2_m = \frac{m/2}{1/2} = m,\ \mathrm{var}(\chi^2_m) = \frac{m/2}{(1/2)^2} = 2m.$

2. mgf (moment generating function) of χ^2_m:

$$\psi(t) \underbrace{=}_{\text{gamma}} \left(\frac{\beta}{\beta - t}\right)^{\alpha} = \left(\frac{1}{1-2t}\right)^{m/2} \quad \text{for} \quad \alpha = \frac{m}{2},\, \beta = \frac{1}{2}.$$

3. If $X_i \sim$ Gamma(α_i, β) are independent, then $\sum_{i=1}^n X_i \sim$ Gamma$(\sum_{i=1}^n \alpha_i, \beta)$ ($\implies$ sum of independent χ^2 is also χ^2) (because mgf of sum of r.v.s is the product of their mgfs)

4. By the Central Limit Theorem (CLT),

$$\frac{1}{\sqrt{2n}}(\chi^2_n - n)\xrightarrow{D}N(0,1)$$

5. $\chi^2_2 \sim$ Exp$(1/2)$

6. If $X \sim N(0,1) \implies X^2 \sim \chi^2_1$ (use brute force method instead or transformation formula)

7. If $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ where μ is known, then $\frac{n}{\sigma^2}\hat{\sigma}^2_0 \sim \chi^2_n$.

7.2 Joint Distribution of $\bar{X}$ and $\hat{\sigma}^2_1$

As before, $\bar{X} = \bar{X}_n = n^{-1}\sum_{i=1}^n X_i, \quad \hat{\sigma}^2_1 = \hat{\sigma}^2_{1n} = n^{-1}\sum_{i=1}^n (X_i - \bar{X})^2.$

Theorem. If $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, then

(i) $\bar{X}_n$ and $\hat{\sigma}^2_{1n}$ are independent

(ii) $\bar{X}_n \sim N(\mu, \sigma^2/n), \quad \frac{n}{\sigma^2}\hat{\sigma}^2_{1n} \sim \chi^2_{n-1}$

Proof (A). Consider a $n \times n$ matrix $A = (a_{ij})_{1 \le i,j \le n}$. The following are equivalent:

$$A \text{ is orthogonal} \quad \iff \quad A' = A^{-1} \iff AA' = A'A = \underbrace{I_n}_{n \times n \text{ identity matrix}} = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}$$

$$\iff \quad \sum_{l=1}^{n} a_{kl}^2 = 1 \quad \text{and} \quad \sum_{l=1}^{n} a_{kl} a_{k'l} = 0] \quad \forall\, 1 \le k, k' \le n$$

$$\iff \quad \sum_{k=1}^{n} a_{kl}^2 = 1 \quad \text{and} \quad \sum_{k=1}^{n} a_{kl} a_{kl'} = 0 \quad \forall\, 1 \le l, l' \le n.$$

Lemma 1 A is orthogonal implies $|\det(A)| = 1$. For $\mathbf{y} = A\mathbf{x}$, $\|\mathbf{y}\| = \|\mathbf{x}\|$ (Euclidean norm) and A is a rotation matrix.

Lemma 2 $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(0,1)$, A orthogonal $n \times n$: For $\mathbf{y} = A\mathbf{x}$, $y_1, \ldots y_n$ are iid $N(0,1,$ $\sum_{i=1}^n y_i^2 = \sum_{i=1}^n x_i^2$.

Proof. $\mathbf{y} = A\mathbf{x}$, $\mathbf{x} = A^{-1}\mathbf{y} = A'\mathbf{y}$ implies that the Jacobian is $J = A'$, $|\det(J)| = 1$ and

$$f_{\mathbf{y}}(\mathbf{y}) = \frac{1}{(2\pi)^{n/2}} e^{-\sum_{i=1}^n y_i^2/2} \cdot 1 \implies y_1, \ldots, y_n \overset{\text{iid}}{\sim} N(0,1).$$

Book Theorem 3.9.6: transformation theorem gives

$$f_{\mathbf{y}}(\mathbf{y}) = f_{\mathbf{x}}(A^{-1}\mathbf{y})|\det(A^{-1})| = 1/(2\pi)^{n/2} \prod e^{-x_i^2/2} = 1/(2\pi)^{n/2} \prod e^{-y_i^2/2}.$$

Proof of Theorem continued (B). $\bar{X}_n = (1/n \ldots 1/n)\mathbf{x}$, where $\mathbf{X} = (X_1 \ldots X_n)$. By Gram–Schmidt orthogonalization, construct orthogonal matrix A with first row $(1/\sqrt{n} \ldots 1/\sqrt{n})$.
For $\mathbf{y} = A\mathbf{x}$, $\quad y_1, \ldots, y_n \overset{\text{iid}}{\sim} N(0,1)$ by lemma 2, and
$y_1 = (1/\sqrt{n} \ldots 1/\sqrt{n})\mathbf{x} = \sqrt{n}\bar{X}_n, \quad \sum_{i=1}^n X_i^2 = \sum_{i=1}^n y_i^2$.
Therefore
$\sum_{i=2}^n y_i^2 = \sum_{i=1}^n X_i^2 - y_1^2 = \sum_{i=1}^n X_i^2 - n\bar{X}^2 = \sum_{i=1}^n (X_i - \bar{X})^2 \sim \chi^2_{n-1}$
and
$\sum_{i=2}^n y_i^2$ is independent of $y_1 = \sqrt{n}\bar{X}$.

Therefore

$\bar{X}_n$ is independent of $\sum_{i=1}^n (X_i - \bar{X})^2 \implies \bar{X}_n, \hat{\sigma}^2_{1n}$ are independent.

Note: The assumption was that we start with $N(0,1)$ data X_i.

Now $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2) \implies Z_i = \frac{X_i - \mu}{\sigma} \sim N(0,1)$ and therefore

$$\begin{aligned} &\bar{Z}_n = \frac{\bar{X} - \mu}{\sigma}, \quad \sum_{i=1}^n (Z_i - \bar{Z}_n)^2 \text{ are independent.} \\ \implies &\frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X}_n)^2 = \sum_{i=1}^n (Z_i - \bar{Z}_n)^2 \sim \chi^2_{n-1} \\ \implies &\bar{X}_n, \hat{\sigma}^2_{1n} \text{ are independent.} \end{aligned}$$

Note: $E(\frac{n}{\sigma^2}\hat{\sigma}^2_{1n}) = E(\chi^2_{n-1}) \implies E(\hat{\sigma}^2_{1n}) = \sigma^2(n-1)/n = \sigma^2(1-1/n) \implies \hat{\sigma}^2_{2n}$ unbiased.

<u>t-distribution</u>

$Y \sim N(0,1)$, $Z \sim \chi^2_n$ independent. Define distribution of $X = Y/\sqrt{Z/n} \overset{\text{def}}{\sim} t_n$, t-distribution with n df.

Finding the pdf of t_n: Set $W = Z$, then

$$\begin{pmatrix} Y \\ Z \end{pmatrix} \longrightarrow \begin{pmatrix} X \\ W \end{pmatrix}, \text{ inverse:} \begin{pmatrix} X \\ W \end{pmatrix} \longrightarrow \begin{pmatrix} Y \\ Z \end{pmatrix}$$

From the definition of the transformation, $Y = \frac{1}{\sqrt{n}} X \sqrt{W}, \quad Z = W$

$$\underbrace{\implies}_{\text{Jacobian of inverse map}} \quad J = \det \begin{bmatrix} \frac{\partial y}{\partial x} & \frac{\partial y}{\partial w} \\ \frac{\partial z}{\partial x} & \frac{\partial z}{\partial w} \end{bmatrix} = \begin{vmatrix} \frac{w}{n} & \frac{x}{2\sqrt{n}\sqrt{w}} \\ 0 & 1 \end{vmatrix} = \sqrt{w/n}.$$

Book: transformations of r.v.s Theorem 3.9.5. It follows that

with the transformation approach one can derive the pdf of the t-distribution.

Properties

(a) $f_X(x)$ is symmetric around 0, with maximum at 0.

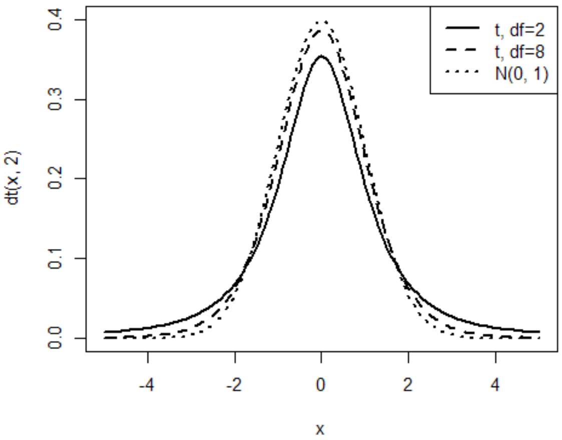
(b) $n = 1 \implies t_1$ is Cauchy distribution, with pdf $f(x) = \frac{1}{\pi(1+x^2)}$ and expected value $= \infty$

(c) $t_n \overset{n\to\infty}{\longrightarrow} N(0,1)$

Proof For any given x,

$$\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}} = \underbrace{\left(1+\frac{x^2}{n}\right)^{-\frac{n}{2}}}_{\to e^{-x^2/2}} \underbrace{\left(1+\frac{x^2}{n}\right)^{1/2}}_{\to 1} \to e^{-x^2/2},$$

since $(1+\alpha/n)^{\zeta n} \to e^{\alpha\zeta}$.



(d) $X \sim t_n$, $n > 1$ has $EX = 0$ and $\operatorname{var}(X) = \frac{n}{n-2}$ for $n > 2$

$E(|X|^k) < \infty$ for $k < n$, $E(|X|^k) = \infty$ for $k \geq n$: only the first $(k-1)$ moments exist

(e) Note that from $\bar{X}_n \sim N(\mu, \frac{\sigma^2}{n})$ and the above theorem it follows that

$$\left.\begin{aligned} Y &= \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \\ Z &= \frac{n}{\sigma^2}\hat{\sigma}^2 \sim \chi^2_{n-1} \end{aligned}\right\} Y, Z \text{ independent}$$

$$\implies U = Y/\left(\frac{Z}{n-1}\right)^{1/2} \sim t_{n-1}$$

$$U = \frac{\sqrt{n}(\bar{X}-\mu)/\sigma}{\left(\frac{n\hat{\sigma}^2_{1n}}{\sigma^2(n-1)}\right)^{1/2}} = \frac{\bar{X}-\mu}{\hat{\sigma}^2_{2n}/\sqrt{n}}, \quad \hat{\sigma}^2_{2n} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = s_n^2$$

(f) Standardizing and studentizing:

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1) \quad \text{“standardizing” } \bar{X}$$

$$\frac{\bar{X}-\mu}{\hat{\sigma}_2/\sqrt{n}} \sim t_{n-1} \quad \text{“studentizing” } \bar{X}$$

The t-distribution was discovered by Gosset, who published it under the pseudonym “Student”, therefore it is known as “Student’s t-distribution”.

8 Interval Estimation/Confidence Intervals

Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} f(x|\theta),\ \theta \in \Theta$.
Note: If f is a pdf (continuous r.v.) then $P(\hat{\theta} = \theta) = 0$.
Goal: Find statistics A, B with $A < B$ that define an interval (A, B) such that $P(A < \theta < B) \geq \gamma$ for some $0 < \gamma < 1$, for all $\theta \in \Theta$.

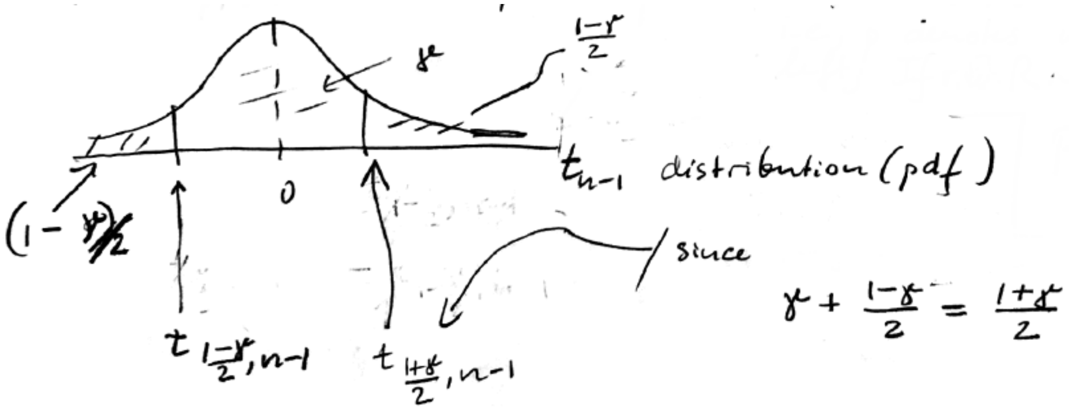
Definition. We call (A, B) a $100\gamma\%$ c.i. for θ: $P(A < \theta < B)$ is the coverage probability for the interval (A, B), γ is the confidence level of the interval, $l = B - A$ is the length of the interval (for the same coverage probability, a shorter interval is preferred).

Interpretation: Redoing the experiment many times, $A < \theta < B$ will happen $100\gamma\%$ of the time in the long run.

<u>How to construct ci's</u>: *Example*

$$f(x|\theta) = N(\mu, \sigma^2), \quad \frac{\bar{X} - \mu}{\hat{\sigma}_{2n}/\sqrt{n}} \sim t_{n-1}.$$

Let $t_{p,n-1}$ be the pth quantile of t_{n-1}, i.e., p denotes mass to the left; If r.v. $R \sim t_{n-1}$, then $P(R \leq t_{p,n-1}) = p$.


$\frac{1-\gamma}{2}$
γ
t_{n-1} distribution (pdf)
0
$(1-\gamma)/2$
since
$\gamma + \frac{1-\gamma}{2} = \frac{1+\gamma}{2}$
$t_{\frac{1-\gamma}{2}, n-1}$
$t_{\frac{1+\gamma}{2}, n-1}$

Note: $-t_{(1+\gamma)/2,n-1} = t_{(1-\gamma)/2,n-1}$ (symmetry)

Require: $P(|\frac{\bar{X}-\mu}{\tilde{\sigma}/\sqrt{n}} < c) \geq \gamma$ for the variance estimator $\tilde{\sigma}^2 = \hat{\sigma}^2_{2n} = (n-1)^{-1}\sum_{i=1}^n (X_i - \bar{X})^2$.

Calculate:

$$\begin{aligned}
P\left(\left|\frac{\bar{X}-\mu}{\tilde{\sigma}/\sqrt{n}}\right| < c\right) &= P(-c < \frac{\bar{X}-\mu}{\tilde{\sigma}/\sqrt{n}} < c) \\
&= P(-c\frac{\tilde{\sigma}}{\sqrt{n}} < \bar{X} - \mu < c\frac{\tilde{\sigma}}{\sqrt{n}}) \\
&= P(-\bar{X} - c\frac{\tilde{\sigma}}{\sqrt{n}} < -\mu < -\bar{X} + c\frac{\tilde{\sigma}}{\sqrt{n}}) \\
&= P(\bar{X} - c\frac{\tilde{\sigma}}{\sqrt{n}} < \mu < \bar{X} + c\frac{\tilde{\sigma}}{\sqrt{n}}) \\
&= \gamma \quad \text{if } c = t_{(1+\gamma)/2}
\end{aligned}$$

Notes:

(a) Level γ c.i. for θ in shorthand: $\bar{X} \pm t_{(1+\gamma)/2,n-1}\frac{\tilde{\sigma}}{\sqrt{n}}$.

(b) Level γ c.i.'s are not unique: Could use $[\bar{X} - t_{\gamma_1,n-1}\frac{\tilde{\sigma}}{\sqrt{n}}, \bar{X} + t_{\gamma_2,n-1}\frac{\tilde{\sigma}}{\sqrt{n}}]$ as long as $\gamma_2 - \gamma_=\gamma$ this is still a $100\gamma\%$ c.i. Choosing $\gamma_1 = 1 - \gamma_2$ as above (where we choose $\gamma_1 = (1-\gamma)/2$, $\gamma_2 = (1+\gamma)/2$) gives the shortest length interval among all these and therefore is the best choice. [why?]

(c) The distribution of $\frac{\bar{X}-\mu}{\tilde{\sigma}/\sqrt{n}}$ does not depend on the parameter $\theta = (\mu, \sigma^2)$: Thus $\frac{\bar{X}-\mu}{\tilde{\sigma}/\sqrt{n}}$ is a pivotal quantity which is key to find the c.i.

Method of pivots: In general, to find a c.i. for θ: Find a function g of $\mathbf{X} = (X_1, \dots, X_n)$ s.t. the distribution of $g(\mathbf{X}, \theta)$ does not involve θ. Then the function g is a pivot.

Example: $g(\mathbf{X}, \theta) = \frac{\bar{X}-\theta}{\tilde{\sigma}/\sqrt{n}}$ in the normal modal.

General: Let C, D be such that $P(C < g(\mathbf{X}, \theta) < D) = \gamma$. Then solve

$$\underbrace{g(\mathbf{X}, \theta) = C}_{\Longrightarrow A = A(\mathbf{X})}, \quad \underbrace{g(\mathbf{X}, \theta) = D}_{\Longrightarrow B = B(\mathbf{X})} \text{ for } \theta.$$

It follows that $P(A < \theta < B) = \gamma$, thus $[A, B]$ is the $100\gamma\%$ c.i.

Example The above derivation of c.i. for μ when σ^2 is unknown, for Gaussian samples.

Example $X_1, \dots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, where σ^2 is known. Parameter $\theta = \mu$. Then use the standardized version of $\bar{X}$ as pivot:

$$g(\mathbf{X}, \mu) = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = Z \sim N(0, 1).$$

As required, the distribution of this pivot does not depend on the unknown parameter $\theta = \mu$.

To apply the method of pivots:

- The starting point is to find C, D such that

$$\gamma = P(C < g(\mathbf{X}, \mu) < D) = P(C < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < D)$$

- Then choose

$$C = z_{(1-\gamma)/2}, \quad D = z_{(1+\gamma)/2},$$

where z_p is the pth quantile of $N(0,1)$ ("z value"), i.e., for $0 < p < 1$, $P(Z < z_p) = p$. This means

$$P(C < Z < D) = \gamma$$

for any r.v. $Z \sim N(0,1)$, including $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$.

- Solve the equations

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = z_{(1-\gamma)/2}, \quad \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = z_{(1+\gamma)/2}$$

for μ, which leads to

$$A = A(\mathbf{X}) = \bar{X} - z_{(1+\gamma)/2}\frac{\sigma}{\sqrt{n}}, \quad B = B(\mathbf{X}) = \bar{X} + z_{(1+\gamma)/2}\frac{\sigma}{\sqrt{n}}$$

and the $100\gamma\%$ CI:

$$\bar{X} \pm z_{(1+\gamma)/2}\frac{\sigma}{\sqrt{n}}.$$

Assuming $\tilde{\sigma} = \sigma$, this interval is generally shorter than the corresponding interval

$$\bar{X} \pm t_{(1+\gamma)/2,n-1}\frac{\tilde{\sigma}}{\sqrt{n}} \text{ as } z_{(1+\gamma)/2} < t_{(1+\gamma)/2,n-1} \quad (n \geq 3)$$

where the difference gets smaller for larger n.

Example $\gamma = 0.95$, $z_{(1+\gamma)/2} = 1.96$, whereas

$$t_{(1+\gamma)/2,n-1} \approx \begin{array}{ll} 4.30 & n=3 \\ 2.23 & n=10 \\ 2.04 & n=30 \\ 1.98 & n=120 \\ 1.96 & n=\infty\,(z_{0.975}=1.96) \end{array}$$

Example Gaussian model $n = 16$, $\bar{X} = 7.5$, $\tilde{\sigma} = 1.8$
$\implies$ 95% CI for μ is $7.5 \pm \frac{1.8}{\sqrt{16}} \underbrace{2.11}_{t_{.975,15}}$.

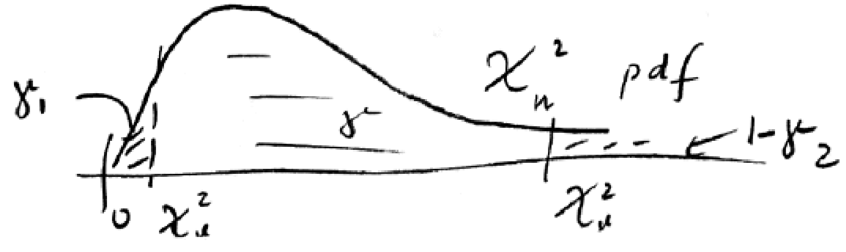
Interpretation of CI: Before a sample is taken, there is a probability γ that the interval $[A, B]$ will include θ, since A, B are r.v.s. When repeating the construction of a CI across samples, the result is that "on average" a fraction γ of the CI will include the true parameter θ.

Example Gaussian model, confidence interval for σ^2
(a) $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, μ known: Pivot

$$g(\mathbf{X}, \theta) = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 \sim \chi^2_n.$$

$$\implies P(C < \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2 < D) = \gamma \quad \text{if}$$

$C = \chi^2_{\gamma_1,n}$, $D = \chi^2_{\gamma_2,n}$ such that $\gamma_2 - \gamma_1 = \gamma$. Again, choose $\gamma_1 = (1 - \gamma)/2$ and $\gamma_2 = (1 + \gamma)/2$.



$$\begin{aligned}\implies \gamma &= P\Big(\frac{C}{\sum_{i=1}^n (X_i-\mu)^2} < \frac{1}{\sigma^2} < \frac{D}{\sum_{i=1}^n (X_i-\mu)^2}\Big)\\ &= P\Big(\underbrace{\frac{\sum_{i=1}^n (X_i-\mu)^2}{D}}_{A} < \sigma^2 < \underbrace{\frac{\sum_{i=1}^n (X_i-\mu)^2}{C}}_{B}\Big)\\ \implies &\text{level } \gamma \text{ CI for } \sigma^2 \text{ is } \left[\frac{\sum_{i=1}^n (X_i-\mu)^2}{\chi^2_{\gamma_2,n}}, \frac{\sum_{i=1}^n (X_i-\mu)^2}{\chi^2_{\gamma_1,n}}\right]\end{aligned}$$

(b) μ unknown, replaced by $\bar{X}$:

$$g(\mathbf{X},\theta) = \sum_{i=1}^n \left(\frac{X_i-\bar{X}}{\sigma}\right)^2 \sim \chi^2_{n-1} \quad \text{is pivotal.}$$

Obtain CI

$$\left(\frac{\sum_{i=1}^n (X_i-\bar{X})^2}{\chi^2_{\gamma_2,n-1}}, \frac{\sum_{i=1}^n (X_i-\bar{X})^2}{\chi^2_{\gamma_1,n-1}}\right)$$

as long as $0<\gamma_1<\gamma_2<1$, $\gamma_2-\gamma_1=\gamma$. [What is the shortest interval?]

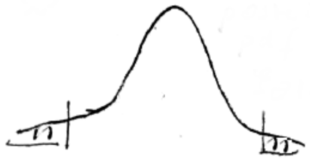
Default choice: $\gamma_1 = \frac{1-\gamma}{2}$, $\gamma_2 = \frac{1+\gamma}{2}$.

Bayesian CI

Assume θ has prior distribution $\xi(\theta)$, with $\xi_{\theta|\mathbf{X}}(\theta|\mathbf{X})$ as posterior distribution. Then find $A(\mathbf{X})$, $B(\mathbf{X})$ as quantiles from the posterior distribution such that

$$\gamma = \Pr(A(\mathbf{X}) < \theta < B(\mathbf{X})|\mathbf{X}) = \int_{A(\mathbf{X})}^{B(\mathbf{X})} \xi_{\theta|\mathbf{X}}(\theta|\mathbf{X})d\theta,$$

from which $(A(\mathbf{X}), B(\mathbf{X}))$ emerges as Bayesian CI.



Posterior pdf $\xi_{\theta|\mathbf{X}}$

Example $X_1,\ldots,X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(1,\theta)$, $\theta \sim \text{Beta}(\alpha,\beta) \implies$ can choose a prior ci by finding $\tilde{C}, \tilde{D}$ such that

$$\underbrace{\Pr(\tilde{C} < \theta \leq \tilde{D})}_{\text{Prior CI for } \theta} = \gamma.$$

The posterior distribution for θ is (as seen before) given by

$$\theta|\mathbf{X} \sim \text{Beta}(\alpha + \sum_{i=1}^{n} X_i,\ \beta + n - \sum_{i=1}^{n} X_i)$$

and with

$$\begin{aligned} C &= \gamma_1\text{th quantile of Beta}(\alpha + \sum_{i=1}^{n} X_i,\ \beta + n - \sum_{i=1}^{n} X_i) \\ D &= \gamma_2\text{th quantile of Beta}(\alpha + \sum_{i=1}^{n} X_i,\ \beta + n - \sum_{i=1}^{n} X_i) \end{aligned}$$

such that $0 < \gamma_1 < \gamma_2 < 1$, $\gamma = \gamma_2 - \gamma_1$, we obtain

$$P(C < \theta < D|\mathbf{X}) = \gamma.$$

The default choice is $\gamma_1 = 1 - \gamma_2$ (symmetry).

Then $[C, D]$ is the Bayesian ci at level γ for θ.

Example. $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, where μ is unknown and σ^2 is known. Prior and posterior distributions for $\theta = \mu$ are:

$$\begin{aligned} \theta &\sim N(\tau, \nu^2) \\ \underbrace{\xi(\theta|\mathbf{X})}_{\text{posterior}} &\sim N(\underbrace{\frac{\sigma^2\tau + n\nu^2\bar{X}_n}{\sigma^2 + n\nu^2}}_{=\zeta}, \underbrace{\frac{\sigma^2\nu^2}{\sigma^2 + n\nu^2}}_{=\rho^2}) \end{aligned}$$

Then the Bayes ci at level $\gamma = 0.95$ $\mu = \theta$ is

$$\zeta \pm \rho z_{(1+\gamma)/2}.$$

Example: $\tau = 1$, $\nu^2 = 1$, $\sigma^2 = 4$, $n = 100$, $\bar{X_n} = 4$. Then $\zeta = \frac{101}{26}, \quad \rho = \frac{1}{26}$.
Plugging in to obtain the Bayes c.i. at level $\gamma = 0.95$ for μ:
$[\frac{101}{26} - 1.96\frac{1}{26}, \frac{101}{26} + 1.96\frac{1}{26}]$ or $[3.809, 3.96]$.

In contrast, the classical ci at level $\gamma = 0.95$ for μ is as follows:
Use $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ as pivot (Note that the distribution of this pivot does not depend on the parameters). Then obtain the ci
$[\bar{X} + \frac{\sigma}{\sqrt{n}} z_{0.025}, \bar{X} + \frac{\sigma}{\sqrt{n}} z_{0.975}]$, which gives $[3.608, 4.392]$.
We find that the Bayesian ci is narrower (is this expected?)

Interpretation of confidence intervals:

- Bayesian: As quantiles of posterior distribution that enclose the random parameter θ.
- Classical: Concerning the ci for parameter θ at level γ, e.g., $\gamma = 0.95$: We say that we are 95% confident that the ci encloses θ.

9 Large Sample Properties of Estimators

9.1 Convergence of Random Variables

Let $\{Z_n\}$ be a sequence of random variables. We characterize the convergence of such sequences as follows.

Modes of Convergence, Definitions

1. $Z_n \longrightarrow Z$ in probability $\Longleftrightarrow Z_n - Z \longrightarrow 0$ in probability.

   $\Longleftrightarrow$ For any $\epsilon > 0, P(|Z_n - Z| > \epsilon) \longrightarrow 0$, as $n \longrightarrow \infty$.

2. $Z_n \longrightarrow Z$ in distribution $\Longleftrightarrow Z_n - Z \longrightarrow 0$ in distribution

   $\Longleftrightarrow F_{Z_n}(x) \longrightarrow F_Z(x)$, for all arguments x at which the *cumulative distribution function* $F_Z(x) = P(Z \le x)$ is continuous.

Note: In case Z has a pdf f_Z, $F_Z(x) = \int_{-\infty}^{x} f_Z(u)du$ and then all x are continuity points of F_Z.

3. $Z_n \longrightarrow Z$ in quadratic mean (or in mean square) $\Longleftrightarrow E\{(Z_n - Z)^2\} \longrightarrow 0$.

Some Properties of Convergence of Random Variables

1. $Z_n \longrightarrow Z$ in quadratic mean $\Longrightarrow Z_n \longrightarrow Z$ in probability.

2. $Z_n \longrightarrow Z$ in probability, and g is a continuous function $\Longrightarrow g(Z_n) \longrightarrow g(Z)$ in probability.

3. $Z_n \longrightarrow Z$ in distribution, and g is a continuous function $\Longrightarrow g(Z_n) \Longrightarrow g(Z)$ in distribution.

4. $X_n \longrightarrow X$ in probability, $Y_n \longrightarrow Y$ in probability $\Longrightarrow$

   $X_n \pm Y_n \longrightarrow X \pm Y$ in probability;

   $X_n Y_n \longrightarrow XY$ in probability;

   and $X_n/Y_n \longrightarrow X/Y$, if $P(Y = 0) = 0$.

5. *Slutsky's Theorem.* $X_n \longrightarrow X$ in distribution, $Y_n \longrightarrow c$ (a constant) $\Longrightarrow$

   $X_n \pm Y_n \longrightarrow X \pm c$ in distribution;

   $X_n Y_n \longrightarrow cX$ in distribution;

   and $X_n/Y_n \longrightarrow X/c$, if $c \ne 0$.

   This is an important tool when determining the asymptotic distribution of estimators.

9.2 Consistency of Estimators

Let $\hat{\theta}_n = T(X_1, \dots, X_n)$ be a statistic, $\hat{\theta}_n$ is called a consistent estimator of θ if $\hat{\theta}_n \xrightarrow{P} \theta$, where $\xrightarrow{P}$ means convergence in probability, i.e. for all $\varepsilon > 0$, $P(|\hat{\theta}_n - \theta| > \varepsilon) \xrightarrow{P} 0$.

The consistency defined above is *weak consistency* as opposed to *strong consistency*, which means that $P(\hat{\theta}_n \to \theta) = 1$.
Strong consistency implies weak consistency.

Example. By the Law of Large Number (LLN) $\bar{X} \xrightarrow{P} \mu$, if $\mu < \infty$.
So $\bar{X}$ is a consistent estimator of μ.

Using the strong law of large numbers (SLLN) one can show that any moment estimator also has the strong consistency property for the population moment it targets.

Theorem (Convergence in probability is preserved under a continuous transformation).

(a) If a sequence of r.v.s Z_n converges in probability to another r.v. Z (i.e. $Z_n \xrightarrow{P} Z$) and g is continuous, then $g(Z_n) \xrightarrow{P} g(c)$.

(b) $Z_n \xrightarrow{P} c$ and g is continuous at c, then $g(Z_n) \xrightarrow{P} g(c)$.

(c) If $\text{MSE}(\hat{\theta}) = E(\hat{\theta}_n - \theta)^2 \to 0$, then $\hat{\theta}_n \xrightarrow{P} \theta$.

Here $E(\hat{\theta}_n - \theta)^2 \to 0$ is called convergence in quadratic mean, which then implies convergence in probability.

Proof: The proof of (a) and (b) is an exercise in calculus. For (c), by Markov's inequality, for any $\varepsilon > 0$, $P(|\hat{\theta}_n - \theta| > \varepsilon) \leq \frac{E(\hat{\theta}-\theta)^2}{\varepsilon^2} \to 0$.

Example. The LNN directly implies that $m_k(\theta)$, the kth sample moment, is a consistent estimator of $\mu_k = E(X^k)$.
Then the above theorem implies that any MoM estimator $q(m_1, \ldots, m_k)$ is consistent for its target $q(\mu_1, \ldots, \mu_k)$ if q is continuous at θ.

Part (c) of the theorem is often useful to show consistency of an estimator, which is obtained by showing $\mathrm{MSE}(\hat{\theta}) = E(\hat{\theta}_n - \theta)^2 \to 0$.
For this, use $E(\hat{\theta}_n - \theta)^2 = \mathrm{bias}^2(\hat{\theta}_n) + \mathrm{var}(\hat{\theta}_n)$, then show
bias $\to 0$ and var $\to 0 \Rightarrow \mathrm{MSE}(\hat{\theta}_n) \to 0$.

Example. Consider the Gamma-Poisson Bayesian model: Let $X_1, \ldots, X_n \sim \mathrm{Pois}(\theta)$, where $\theta \sim \mathrm{Gamma}(\alpha, \beta)$ with $E(\theta) = \alpha/\beta$.
This is a conjugate prior with posterior $\xi(\theta \mid \mathbf{x}) \sim \mathrm{Gamma}(\alpha + \sum_i x_i, n + \beta)$.

The Bayes estimator for θ w.r.t sq error loss is $\hat{\theta} =$ mean of $\mathrm{Gamma}(\alpha + \sum_i x_i, n + \beta)$, which is

$$\begin{array}{cccccc} \hat{\theta} = & \frac{\alpha}{\beta+n} & + & \frac{\sum x_i/n}{1+\beta/n} & \\ & \downarrow & & \downarrow & \\ & 0 & \text{as } n \to \infty & 1 & \text{(for the denominator)} \end{array}$$

Assume the true parameter is θ_0 (non-random), then by LLN $\bar{X} \xrightarrow{P} \theta_0$.
Now we use properties of convergence 5. (section 9.1) to conclude
$\frac{1}{1+\beta/n}\bar{X} \xrightarrow{P} \theta_0$, since $c_n = \frac{1}{1+\beta/n} \to 1$.

Specifically, we apply

$$\begin{aligned} X_n \xrightarrow{P} X, Y_n \xrightarrow{P} Y \quad \Rightarrow \quad & X_n \pm Y_n \xrightarrow{P} X \pm Y \text{ } (Y_n,\, Y \text{ can be constants}) \\ & X_n Y_n \xrightarrow{P} XY \\ & \frac{X_n}{Y_n} \xrightarrow{P} \frac{X}{Y}, \text{ if } P(Y=0)=0 \end{aligned}$$

and conclude that the Bayes estimator is consistent for the true parameter θ_0. This is because for large samples it behaves like $\bar{X}$, which is consistent by the LLN.

Example. Consider $X_1, \ldots, X_n \sim f(x \mid \beta) = \beta e^{-\beta x}, x > 0$
Then $E(X_i) = \frac{1}{\beta}$ and the MLE of β is $\hat{\beta} = \frac{1}{\bar{X}}$.
Since $\bar{X} \xrightarrow{P} \mu = \frac{1}{\beta}$, $g(\bar{X}) = \frac{1}{\bar{X}} \xrightarrow{P} \beta$, showing the consistency of the MLE.

Note: Sometimes it is easier to use the definition to prove consistency as the next example shows.

Example. For the model $X_1, \ldots, X_n \sim U(0, \theta)$, show that the MLE $\hat{\theta} = X_{(n)}$ is consistent:

$$\begin{aligned}
P(|\hat{\theta} - \theta| > \varepsilon) &= P(\hat{\theta} - \theta > \varepsilon \text{ or } \hat{\theta} - \theta < -\varepsilon) \\
&= P(\underbrace{\hat{\theta} > \theta + \varepsilon}_{\text{impossible}} \text{ or } \hat{\theta} - \theta < -\varepsilon) \\
&= P(\hat{\theta} - \theta < -\varepsilon) \\
&= P(\hat{\theta} < \theta - \varepsilon) \\
&= P(X_i < \theta - \varepsilon, \quad i = 1, \ldots, n) \\
&= \prod_i P(X_i < \theta - \varepsilon), \text{ since } X_i \text{ are independent} \\
&= \left(\frac{\theta - \varepsilon}{\theta}\right)^n \\
&\to 0 \quad \text{as} \quad n \to \infty.
\end{aligned}$$

As in the above examples, MLEs and Bayes estimators are usually consistent under mild assumptions.