

An Introduction to Statistical Inference

International Undergraduate Summer Enrichment Program 2026

Bei Jiang

Department of Mathematical and Statistical Sciences
University of Alberta

- Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, which is a measure space with total measure one, where
 - Ω is a set (sometimes called a sample space in elementary probability). $\omega \in \Omega$ is called an outcome.
 - $\mathcal{F}$ is a σ -algebra (or σ -field) of subsets of Ω . $A \in \mathcal{F}$ is called an event.
 - a σ -algebra on a set Ω is a nonempty collection of subsets of Ω , that is closed under complement, countable unions, and countable intersections.
 - $\mathbb{P}$ is a function from $\mathcal{F}$ to $[0, 1]$ with $\mathbb{P}(\Omega) = 1$ and $\mathbb{P}(A)$ is called the probability of A .
- A random variable (vector) is a function $X : \Omega \rightarrow \mathbb{R}$ ($\mathbb{R}^n$). We say that X has distribution Q , denoted by $X \sim Q$, if $\mathbb{P}(X \in B) = \mathbb{P}(\{\omega : X(\omega) \in B\}) = Q(B)$ for $B \in \mathcal{B}$.
- An expectation is an integral with respect to $\mathbb{P}$. $\mathbb{E}[X] = \int_{\Omega} X(\omega) d\mathbb{P}(\omega) = \int \cdots \int_{\mathbb{R}^n} x dQ(x)$.

- A statistical model is a family of candidate probability distributions. $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ for some observed data $X \sim P_\theta$. θ is called the parameter.
- Goal of Estimation: Observe $X \sim P_\theta$ and guess the value of $g(\theta)$ (estimand).
- A statistic is any function $T(X)$ of data X . An estimator $\delta(X)$ of $g(\theta)$ is a statistic meant to guess $g(\theta)$.

Example. Flip a biased coin n times. θ is the probability of landing heads and X is the number of heads after n flips. $\Theta = [0, 1]$. $X \sim \text{Binomial}(n, \theta)$, with $p_\theta(x) = P(X = x) = \theta^x(1 - \theta)^{n-x} \binom{n}{x}$ for $x \in \{0, \dots, n\}$.

In the example, a natural estimator is $\delta_0(X) = X/n$. Is this a good estimator?

- A loss function $L(\theta, d)$ measures the “badness” of the guess. For example, $L(\theta, d) = (d - g(\theta))^2$ is the squared error.
- Typical properties: (1) $L(\theta, d) \geq 0$ for all θ, d , and (2) $L(\theta, g(\theta)) = 0$ for all θ .
- The risk function is $R(\theta, \delta(\cdot)) = \mathbb{E}_\theta[L(\theta, \delta(X))]$. For example, if $L(\theta, d) = (d - g(\theta))^2$, then $R(\theta, \delta) = \mathbb{E}_\theta [(\delta(X) - g(\theta))^2]$ (the Mean Squared Error, MSE).

Example A commonly chosen loss function for parameter estimation is the squared error loss, defined by $l(\theta, \delta(X)) = (\theta - \delta(X))^2$. In this case, we have

$$\begin{aligned} R(\theta, \delta) &= E_{\theta} l(\theta, \delta(X)) \\ &= E_{\theta} (\theta - \delta(X))^2 \\ &= E_{\theta} (\theta - E_{\theta} \delta(X) + E_{\theta} \delta(X) - \delta(X))^2 \\ &= \underbrace{(\theta - E_{\theta} \delta(X))^2}_{\text{Bias}^2} + \underbrace{E_{\theta} (\delta(X) - E_{\theta} \delta(X))^2}_{\text{Variance}} \end{aligned}$$

This result allows a frequentist to analyze the variance and bias of an estimator separately, and can be used to motivate frequentist ideas, e.g. minimum variance unbiased estimators (MVUE).

- $X_1, \dots, X_n$: Random variables identically and independently distributed (iid) with probability density (or mass) function $f_X(x \mid \theta)$.
- $x_1, \dots, x_n$: Realization of random variables $X_1, \dots, X_n$.
- $\mathbf{X} = (X_1, \dots, X_n)$ is a random sample of a population (typically iid), and the characteristics of this population are described by $f_{\mathbf{X}}(\mathbf{x} \mid \theta)$.
- The joint pdf (or pmf) of $\mathbf{X} = (X_1, \dots, X_n)$ (assuming iid) is

$$f_{\mathbf{X}}(\mathbf{x} \mid \theta) = \prod_{i=1}^n f_X(x_i \mid \theta)$$

An example

For a sample size n , let $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \text{Bernoulli}(p_0)$. What is the probability of $\sum_{i=1}^n X_i \leq m$?

$$\sum_{i=1}^n X_i \sim \text{Binomial}(n, p_0)$$
$$\Pr\left(\sum_{i=1}^n X_i \leq m\right) = \sum_{k=0}^m \binom{n}{k} p_0^k (1 - p_0)^{n-k}$$

Comparing the risks of different estimators

Example. $X \sim \text{Binomial}(n, \theta)$, so $p_\theta(x) = \theta^x(1 - \theta)^{n-x} \binom{n}{x}$. An estimator for θ is $\delta_0(X) = X/n$. The expectation of the estimator is $\mathbb{E}_\theta[X/n] = \theta$ (it is unbiased). So, $R(\theta, \delta) = \text{var}_\theta(X/n) = \theta(1 - \theta)/n$. Other choices:

$$\delta_1(X) = \frac{X + 3}{n},$$

$$\delta_2(X) = \frac{X + 3}{n + 6}.$$

$R(\theta, \delta_1)$ is always greater than $R(\theta, \delta_0)$ because δ_1 has the same variance as δ_0 , but more bias. $R(\theta, \delta_2)$ is smaller than $R(\theta, \delta_0)$ when θ is close to $1/2$.

Resolve the issue of ambiguity

Recall the binomial example: δ_1 is definitely bad, but the comparison between δ_0 and δ_2 depends on the true value of θ (unknown).

- Average-case risk (Bayes risk): for some measure Λ , minimize $R_\pi(\theta) = \int_{\Theta} R(\theta, \delta) d\pi(\theta)$. This is called the **Bayes estimator**, and π is the prior.
- Worst-case risk (minimax risk): minimize $\sup_{\theta \in \Theta} R(\theta, \delta)$; in this case, the estimator $\hat{\theta}$ is minimax (minimax optimal).
- Constrain the choice of estimator: for example, only consider unbiased δ . $\mathbb{E}_\theta[\delta(X)] = g(\theta) \forall \theta \in \Theta$.

Suppose we observe $x_1, \dots, x_n$ from a population with probability density function $f_X(x|\theta)$. The goal is to define a function of data

$$T(x_1, \dots, x_n) : \mathbb{R}^n \rightarrow \mathbb{R}^d$$

We wish this summary statistic to be able to:

- Be simpler than the original data, e.g. $d \leq n$.
- Keep all the information about θ that is contained in the original data $x_1, \dots, x_n$

An Example: A Statistic as Data Reduction

Suppose

$$X_1, X_2, X_3 \stackrel{\text{i.i.d.}}{\sim} \text{Bernoulli}(p), \quad 0 < p < 1.$$

A *statistic* is any function of the observed data that does not depend on the unknown parameter. Consider the statistic

$$T(X_1, X_2, X_3) = X_1 + X_2 + X_3,$$

which maps the sample space $\{0, 1\}^3$ to the set $\{0, 1, 2, 3\}$.

Partition	X_1	X_2	X_3	$T(\mathbf{X})$
A_0	0	0	0	0
A_1	0	0	1	1
	0	1	0	1
	1	0	0	1
A_2	0	1	1	2
	1	0	1	2
	1	1	0	2
A_3	1	1	1	3

where

An example

The partition of the sample space based on $T(\mathbf{X})$ is **coarser** than the original sample space.

- There are 8 elements in the sample space $\mathcal{X}$.
- They are partitioned into 4 subsets
- Thus, $T(\mathbf{X})$ is simpler (or coarser) than $\mathbf{X}$.

- A statistic $T(\mathbf{X})$ is a **sufficient statistic** for θ if the conditional distribution of sample $\mathbf{X}$ given the value of $T(\mathbf{X})$ does not depend on θ . That is, the conditional pdf or pmf of $\mathbf{X}$ given $T = t$, $f_{\mathbf{X}}(\mathbf{x} \mid T(\mathbf{X}) = t) = h(\mathbf{x})$ does not depend on θ

An example

- Suppose $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \text{Bernoulli}(p), 0 < p < 1$.
- In fact, $T(X_1, \dots, X_n) = \sum_{i=1}^n X_i$ is a sufficient statistic for p .
- $T(\mathbf{X}) = \sum_{i=1}^n X_i \sim \text{Binomial}(n, p)$
- Need to find the conditional pmf of $\mathbf{X}$ given $T = t$, and show that the distribution does not depend on p .

Definition

$X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} f_X(x | \theta)$. The joint distribution of $\mathbf{X} = (X_1, \dots, X_n)$ is

$$f_{\mathbf{X}}(\mathbf{x} | \theta) = \prod_{i=1}^n f_X(x_i | \theta)$$

Given that $\mathbf{X} = \mathbf{x}$ is observed, the function of θ defined by $L(\theta | \mathbf{x}) = f(\mathbf{x} | \theta)$ is called the likelihood function.

Point Estimation: Ingredients

- Data: $\mathbf{x} = (x_1, \dots, x_n)$ - realizations of random variables $(X_1, \dots, X_n)$.
- $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} f_X(x \mid \theta)$.
- Assume a model $\mathcal{P} = \{f_X(x \mid \theta) : \theta \in \Omega \subset \mathbb{R}^p\}$ where the functional form of $f_X(x \mid \theta)$ is known, but θ is unknown.
- Task is to use data $\mathbf{x}$ to make inference on θ

Definition

If we use a function of sample $w(X_1, \dots, X_n)$ as a "guess" of $\tau(\theta)$, where $\tau(\theta)$ is a function of true parameter θ . Then $w(\mathbf{X}) = w(X_1, \dots, X_n)$ is called a point estimator of $\tau(\theta)$. The realization of the estimation, $w(\mathbf{x}) = w(x_1, \dots, x_n)$ is called the estimate of $\tau(\theta)$.

Example:

- $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, 1)$, where $\theta \in \Omega \in \mathbb{R}$. Suppose $n = 6$, and one observes $(x_1, \dots, x_6) = (2.0, 2.1, 2.9, 2.6, 1.2, 1.8)$. Define $w_1(X_1, \dots, X_n) = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X} = 2.1$. Define $w_2(X_1, \dots, X_n) = X_{(1)} = 1.2$.

A method to equate sample moments to population moments and solve equations.

Sample moments	Population moments
$m_1 = \frac{1}{n} \sum_{i=1}^n X_i$	$\mu'_1 = E[X \theta] = \mu'_1(\theta)$
$m_2 = \frac{1}{n} \sum_{i=1}^n X_i^2$	$\mu'_2 = E[X^2 \theta] = \mu'_2(\theta)$
$m_3 = \frac{1}{n} \sum_{i=1}^n X_i^3$	$\mu'_3 = E[X^3 \theta] = \mu'_3(\theta)$
$\vdots$	$\vdots$

Point estimator of $T(\theta)$ is obtained by solving equations like this.

$$m_1 = \mu'_1(\theta)$$

$$m_2 = \mu'_2(\theta)$$

$$\vdots$$

$$m_k = \mu'_k(\theta)$$

Definition

- For a given sample point $\mathbf{x} = (x_1, \dots, x_n)$,
- let $\hat{\theta}(\mathbf{x})$ be the value such that $L(\theta | \mathbf{x})$ attains its maximum.
- More formally, $L(\hat{\theta}(\mathbf{x}) | \mathbf{x}) \geq L(\theta | \mathbf{x}) \forall \theta \in \Omega$ where $\hat{\theta}(\mathbf{x}) \in \Omega$.
- Then $\hat{\theta}(\mathbf{x})$ is called the maximum likelihood estimate of θ based on data $\mathbf{x}$,
- and $\hat{\theta}(\mathbf{X})$ is the maximum likelihood estimator (MLE) of θ .

Example of MLE - Exponential Distribution

Problem

Let $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \text{Exponential}(\beta)$. Find MLE of β .

Solution

$$\begin{aligned} L(\beta \mid \mathbf{x}) &= f_{\mathbf{x}}(\mathbf{x} \mid \theta) = \prod_{i=1}^n f_X(x_i \mid \theta) \\ &= \prod_{i=1}^n \left[\frac{1}{\beta} e^{-x_i/\beta} \right] = \frac{1}{\beta^n} \exp \left(- \sum_{i=1}^n \frac{x_i}{\beta} \right) \end{aligned}$$

where $\beta > 0$.

Use the derivative to find potential MLE

To maximize the likelihood function $L(\beta \mid \mathbf{x})$ is equivalent to maximize the log-likelihood function

$$l(\beta \mid \mathbf{x}) = \log L(\beta \mid \mathbf{x}) = \log \left[\frac{1}{\beta^n} \exp \left(- \sum_{i=1}^n \frac{x_i}{\beta} \right) \right]$$

$$= - \frac{\sum_{i=1}^n x_i}{\beta} - n \log \beta$$

$$\frac{\partial l}{\partial \beta} = \frac{\sum_{i=1}^n x_i}{\beta^2} - \frac{n}{\beta} = 0$$

$$\sum_{i=1}^n x_i = n\beta$$

$$\hat{\beta} = \frac{\sum_{i=1}^n x_i}{n} = \bar{x}$$

Use the double derivative to confirm local maximum

$$\begin{aligned}\frac{\partial^2 l}{\partial \beta^2} \Big|_{\beta=\bar{x}} &= -2 \frac{\sum_{i=1}^n x_i}{\beta^3} + \frac{n}{\beta^2} \Big|_{\beta=\bar{x}} \\ &= \frac{1}{\beta^2} \left(-\frac{2 \sum_{i=1}^n x_i}{\beta} + n \right) \Big|_{\beta=\bar{x}} \\ &= \frac{1}{\bar{x}^2} \left(-\frac{2n\bar{x}}{\bar{x}} + n \right) \\ &= \frac{1}{\bar{x}^2} (-n) < 0\end{aligned}$$

Therefore, we can conclude that $\hat{\beta}(\mathbf{X}) = \bar{X}$ is unique local maximum on the interval

Check boundary and confirm global maximum

$\beta \in (0, \infty)$. If $\beta \rightarrow \infty$

$$l(\beta \mid \mathbf{x}) = -\frac{\sum_{i=1}^n x_i}{\beta} - n \log \beta \rightarrow -\infty$$

$$L(\beta \mid \mathbf{x}) \rightarrow 0$$

If $\beta \rightarrow 0$,

$$\begin{aligned} l(\beta \mid \mathbf{x}) &= -\frac{\sum_{i=1}^n x_i}{\beta} - n \log \beta \\ &= -\frac{\sum_{i=1}^n x_i}{\beta} - n \left(\frac{1}{\beta} \log \beta^\beta - 1 \right) \\ &= -\frac{\sum_{i=1}^n x_i - n(\beta \log \beta - 1)}{\beta} \rightarrow -\infty \end{aligned}$$

$$L(\beta \mid \mathbf{x}) \rightarrow 0$$

- $\frac{\partial l}{\partial \beta} = 0$ at $\hat{\beta} = \bar{x}$
- $\frac{\partial^2 l}{\partial \beta^2} < 0$ at $\hat{\beta} = \bar{x}$
- $L(\beta \mid \mathbf{x}) \rightarrow 0$ (lowest bound) when β approaches the boundary

Therefore $l(\beta \mid \mathbf{x})$ and $L(\beta \mid \mathbf{x})$ attains the global maximum when $\hat{\beta} = \bar{x}$, $\hat{\beta}(\mathbf{X}) = \bar{X}$ is the MLE of β .

How do we find MLE?

If the function is differentiable with respect to θ .

- Find candidates that makes first order derivative to be zero
- Check second-order derivative to check local maximum.
 - ① For one-dimensional parameter, negative second order derivative implies local maximum.
 - ② For two-dimensional parameter, suppose $L(\theta_1, \theta_2)$ is the likelihood function. Then we need to show
 - $\partial^2 L(\theta_1, \theta_2) / \partial \theta_1^2 < 0$ or $\partial^2 L(\theta_1, \theta_2) / \partial \theta_2^2 < 0$.
 - Determinant of second-order derivative is positive
- Check boundary points to see whether boundary gives global maximum.

If the function is NOT differentiable with respect to θ .

- Use numerical methods
- Or perform directly maximization, using inequalities, or properties of the function.

Method of Evaluating Estimators

Definition: Unbiasedness

Suppose $\hat{\theta}$ is an estimator for θ , then the bias of θ is defined as

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

If the bias is equal to 0, then $\hat{\theta}$ is an unbiased estimator for θ .

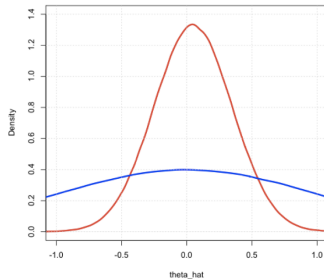
Example

$X_1, \dots, X_n$ are iid samples from a distribution with mean μ . Let $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ be an estimator of μ . The bias is

$$\begin{aligned} \text{Bias}(\hat{\mu}) &= E(\bar{X}) - \mu \\ &= E\left(\frac{1}{n} \sum_{i=1}^n X_i\right) - \mu = \frac{1}{n} \sum_{i=1}^n E(X_i) - \mu = \mu - \mu = 0 \end{aligned}$$

Therefore $\bar{X}$ is an unbiased estimator for μ .

How important is unbiased



- $\hat{\theta}_1$ (blue) is unbiased but has a chance to be very far away from $\theta = 0$.
- $\hat{\theta}_2$ (red) is slightly biased but more likely to be closer to the true θ than $\hat{\theta}_1$.

Definition

Mean Squared Error (MSE) of an estimator $\hat{\theta}$ is defined as

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

Property of MSE

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

Example

Suppose $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, 1)$: 1) $\hat{\mu}_1 = 1$ and 2) $\hat{\mu}_2 = \bar{X}$.

$$\text{MSE}(\hat{\mu}_1) = E(\hat{\mu}_1 - \mu)^2 = (1 - \mu)^2$$

$$\text{MSE}(\hat{\mu}_2) = E(\bar{X} - \mu)^2 = \text{Var}(\bar{X}) = \frac{1}{n}$$

- Suppose that the true $\mu = 1$, then $\text{MSE}(\hat{\mu}_1) = 0 < \text{MSE}(\hat{\mu}_2)$, and no estimator can beat $\hat{\mu}_1$ in terms of MSE when true $\mu = 1$.
- Therefore, we cannot find an estimator that is uniformly the best in terms of MSE across all $\theta \in \Omega$ among all estimators
- Restrict the class of estimators, and find the "best" estimator within the small class.

Frequentist's Framework

$$\mathcal{P} = \{\mathbf{X} \sim f_{\mathbf{X}}(\mathbf{x} \mid \theta), \theta \in \Omega\}$$

Bayesian Statistic

- Parameter θ is considered as a random quantity
- Distribution of θ can be described by probability distribution, referred to as prior distribution
- A sample is taken from a population indexed by θ , and the prior distribution is updated using information from the sample to get posterior distribution of θ given the sample.

- Prior distribution of θ : $\theta \sim \pi(\theta)$.
- Sample distribution of $\mathbf{X}$ given θ .

$$\mathbf{X} \mid \theta \sim f(\mathbf{x} \mid \theta)$$

- Joint distribution $\mathbf{X}$ and θ

$$f(\mathbf{x}, \theta) = \pi(\theta)f(\mathbf{x} \mid \theta)$$

- Marginal distribution of $\mathbf{X}$.

$$m(\mathbf{x}) = \int_{\theta \in \Omega} f(\mathbf{x}, \theta) d\theta = \int_{\theta \in \Omega} f(\mathbf{x} \mid \theta) \pi(\theta) d\theta$$

- Posterior distribution of θ (conditional distribution of θ given $\mathbf{X}$)

$$\pi(\theta \mid \mathbf{x}) = \frac{f(\mathbf{x}, \theta)}{m(\mathbf{x})} = \frac{f(\mathbf{x} \mid \theta) \pi(\theta)}{m(\mathbf{x})} \quad (\text{Bayes' Rule})$$

Advantages and Drawbacks of Bayesian Inference

Advantages over Frequentist's Framework

- Allows making inference on the distribution of θ given data.
- Available information about θ can be utilized.
- Uncertainty and information can be quantified probabilistically.

Drawbacks of Bayesian Inference

- Misleading prior can result in misleading inference.
- Bayesian inference is often (but not always) prone to be “subjective”
- See Larry Wasserman “Frequentist Bayes is Objective” (2006) Bayesian Analysis 3:451-456.
- Bayesian inference could be sometimes unnecessarily complicated to interpret, compared to Frequentist's inference.

Definition

Bayes Estimator of θ is defined as the posterior mean of θ .

$$E(\theta \mid \mathbf{x}) = \int_{\theta \in \Omega} \theta \pi(\theta \mid \mathbf{x}) d\theta$$

Example Problem $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)$ where $0 \leq p \leq 1$. Assume that the prior distribution of p is $\text{Beta}(\alpha, \beta)$. Find the posterior distribution of p and the Bayes estimator of p , assuming α and β are known.

Solution (1/4)

Prior distribution of p is

$$\pi(p) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1} (1-p)^{\beta-1}$$

Sampling distribution of $\mathbf{X}$ given p is

$$f_{\mathbf{X}}(\mathbf{x} | p) = \prod_{i=1}^n \{p^{x_i} (1-p)^{1-x_i}\}$$

Joint distribution of $\mathbf{X}$ and p is

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}, p) &= f_{\mathbf{X}}(\mathbf{x} | p) \pi(p) \\ &= \prod_{i=1}^n \{p^{x_i} (1-p)^{1-x_i}\} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1} (1-p)^{\beta-1} \end{aligned}$$

Solution (2/4)

The marginal distribution of $\mathbf{X}$ is

$$\begin{aligned} m(\mathbf{x}) &= \int f(\mathbf{x}, p) dp = \int_0^1 \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\sum_{i=1}^n x_i + \alpha - 1} (1 - p)^{n - \sum_{i=1}^n x_i + \beta - 1} dp \\ &= \int_0^1 \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\sum x_i + \alpha) \Gamma(n - \sum x_i + \beta)}{\Gamma(\alpha + \beta + n)} \\ &\quad \times \frac{\Gamma(\sum x_i + \alpha + n - \sum x_i + \beta)}{\Gamma(\sum x_i + \alpha) \Gamma(n - \sum x_i + \beta)} p^{\sum x_i + \alpha - 1} (1 - p)^{n - \sum x_i + \beta - 1} dp \\ &= \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\sum_{i=1}^n x_i + \alpha) \Gamma(n - \sum_{i=1}^n x_i + \beta)}{\Gamma(\alpha + \beta + n)} \\ &\quad \times \int_0^1 f_{\text{Beta}}\left(\sum x_i + \alpha, n - \sum x_i + \beta\right)(p) dp \\ &= \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\sum_{i=1}^n x_i + \alpha) \Gamma(n - \sum_{i=1}^n x_i + \beta)}{\Gamma(\alpha + \beta + n)} \end{aligned}$$

The posterior distribution of $\theta \mid \mathbf{x}$:

$$\begin{aligned}
 \pi(\theta \mid \mathbf{x}) &= \frac{f(\mathbf{x}, p)}{m(\mathbf{x})} \\
 &= \frac{\left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\sum x_i + \alpha - 1} (1-p)^{n - \sum x_i + \beta - 1} \right]}{\left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\sum x_i + \alpha) \Gamma(n - \sum x_i + \beta)}{\Gamma(\alpha + \beta + n)} \right]} \\
 &= \frac{\Gamma(\alpha + \beta + n)}{\Gamma(\sum x_i + \alpha) \Gamma(n - \sum x_i + \beta)} p^{\sum x_i + \alpha - 1} (1-p)^{n - \sum x_i + \beta - 1}
 \end{aligned}$$

The Bayes estimator of p is

$$\begin{aligned}\hat{p} &= \frac{\sum_{i=1}^n x_i + \alpha}{\sum_{i=1}^n x_i + \alpha + n - \sum_{i=1}^n x_i + \beta} = \frac{\sum_{i=1}^n x_i + \alpha}{\alpha + \beta + n} \\ &= \frac{\sum_{i=1}^n x_i}{n} \frac{n}{\alpha + \beta + n} + \frac{\alpha}{\alpha + \beta} \frac{\alpha + \beta}{\alpha + \beta + n} \\ &= [\text{Guess about } p \text{ from data}] \cdot \text{weight}_1 \\ &\quad + [\text{Guess about } p \text{ from prior}] \cdot \text{weight}_2\end{aligned}$$

As n increase, weight in $_1 = \frac{n}{\alpha + \beta + n} = \frac{1}{\frac{\alpha + \beta}{n} + 1}$ becomes bigger and bigger and approaches to 1 . In other words, influence of data is increasing, and the influence of prior knowledge is decreasing.

Is the Bayes estimator unbiased?

$$E \left[\frac{\sum_{i=1}^n x_i + \alpha}{\alpha + \beta + n} \right] = \frac{np + \alpha}{\alpha + \beta + n} \neq p$$

Unless $\frac{\alpha}{\alpha + \beta} = p$.

$$\text{Bias} = \frac{np + \alpha}{\alpha + \beta + n} - p = \frac{\alpha - (\alpha + \beta)p}{\alpha + \beta + n}$$

As n increases, the bias approaches to zero.

Posterior conditioning on sufficient statistics

If $T(\mathbf{X})$ is a sufficient statistic, then the posterior distribution of θ given $\mathbf{X}$ is the same to the posterior distribution given $T(\mathbf{X})$. In other words,

$$\pi(\theta \mid \mathbf{x}) = \pi(\theta \mid T(\mathbf{x}))$$

Definition 7.2.15

Let $\mathcal{F}$ denote the class of pdfs or pmfs for $f(x \mid \theta)$. A class Π of prior distributions is a conjugate family of $\mathcal{F}$, if the posterior distribution is the class Π for all $f \in \mathcal{F}$, and all priors in Π , and all $x \in \mathcal{X}$.

Example: Beta-Binomial conjugate

Let

- $x_1, \dots, x_n \mid p \sim \text{Binomial}(m, p)$
- $\pi(p) \sim \text{Beta}(\alpha, \beta)$

where m, α, β is known. The posterior distribution is

$$\pi(p \mid \mathbf{x}) \sim \text{Beta} \left(\sum_{i=1}^n x_i + \alpha, mn - \sum_{i=1}^n x_i + \beta \right)$$

Example: Gamma-Poisson conjugate

- $X_1, \dots, X_n \mid \lambda \sim \text{Poisson}(\lambda)$
- $\pi(\lambda) \sim \text{Gamma}(\alpha, \beta)$
- Prior:

$$\pi(\lambda) = \frac{1}{\Gamma(\alpha)\beta^\alpha} \lambda^{\alpha-1} e^{-\lambda/\beta}$$

- Sampling distribution

$$\mathbf{X} \mid \lambda \stackrel{\text{i.i.d.}}{\sim} \frac{e^{-\lambda} \lambda^x}{x!}$$
$$f_{\mathbf{X}}(\mathbf{x} \mid \lambda) = \prod_{i=1}^n \frac{e^{-\lambda} \lambda^{x_i}}{x_i!}$$

- Joint distribution of $\mathbf{X}$ and λ .

$$\begin{aligned} f(\mathbf{x} \mid \lambda) \pi(\lambda) &= \left[\prod_{i=1}^n \frac{e^{-\lambda} \lambda^{x_i}}{x_i!} \right] \frac{1}{\Gamma(\alpha) \beta^\alpha} \lambda^{\alpha-1} e^{-\lambda/\beta} \\ &= e^{-n\lambda - \lambda/\beta} \lambda^{\sum x_i + \alpha - 1} \frac{1}{\prod_{i=1}^n x_i!} \frac{1}{\Gamma(\alpha) \beta^\alpha} \end{aligned}$$

- Marginal distribution

$$m(\mathbf{x}) = \int f(\mathbf{x} \mid \lambda) \pi(\lambda) d\lambda$$

- Posterior distribution (proportional to the joint distribution)

$$\begin{aligned}\pi(\lambda \mid \mathbf{x}) &= \frac{f(\mathbf{x} \mid \lambda)\pi(\lambda)}{m(\mathbf{x})} \\ &= e^{-n\lambda - \lambda/\beta} \lambda^{\sum x_i + \alpha - 1} \frac{1}{\Gamma(\sum x_i + \alpha) \left(\frac{1}{n + \frac{1}{\beta}}\right)^{\sum x_i + \alpha}}\end{aligned}$$

- So, the posterior distribution is Gamma $\left(\sum x_i + \alpha, \left(n + \frac{1}{\beta}\right)^{-1}\right)$.

Example: Normal Bayes Estimators

Let $x_1, \dots, x_n$ be i.i.d. $\sim \mathcal{N}(\theta, \sigma^2)$ and suppose that the prior distribution of θ is $\mathcal{N}(\mu, \tau^2)$. Assuming that σ^2, μ^2, τ^2 are all known, the posterior distribution of θ also becomes normal, with mean and variance given by

$$\begin{aligned} \mathbb{E}[\theta \mid \mathbf{x}] &= \frac{\tau^2}{n\tau^2 + \sigma^2} \bar{x} + \frac{\sigma^2}{\sigma^2 + n\tau^2} \mu \\ \text{Var}(\theta \mid \mathbf{x}) &= \frac{\sigma^2 \tau^2}{\sigma^2 + n\tau^2} \end{aligned}$$

where $\mathbf{x} = (x_1, \dots, x_n)^T$ and $\bar{x} = 1/n \sum_{i=1}^n x_i$.

- The normal family is its own conjugate family.
- The Bayes estimator for θ is a linear combination of the prior and sample means
- As the prior variance τ^2 approaches to infinity, the Bayes estimator tends toward to sample mean
- As the prior information becomes more vague, the Bayes estimator tends to give more weight to the sample information

Hypothesis Testing

A hypothesis is a statement about a population parameter

Without any hypothesis,

$$\theta \in \Omega = \Omega_0 \cup \Omega_0^c.$$

Two complementary statements about θ

- Null hypothesis : $H_0 : \theta \in \Omega_0$
- Alternative hypothesis : $H_1 : \theta \in \Omega_0^c$

Simple and composite hypothesis

Simple hypothesis

Both H_0 and H_1 consist of only one parameter value.

- $H_0 : \theta = \theta_0 \in \Omega_0$
- $H_1 : \theta = \theta_1 \in \Omega_0^c$

Composite hypothesis

One or both of H_0 and H_1 consist more than one parameter values.

- One-sided hypothesis: $H_0 : \theta \leq \theta_0$ vs $H_1 : \theta > \theta_0$.
- One-sided hypothesis: $H_0 : \theta \geq \theta_0$ vs $H_1 : \theta < \theta_0$.
- Two-sided hypothesis: $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$.

An Example of Hypothesis

$$X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, 1)$$

Let X_i is the change in blood pressure after a treatment.

$$\begin{aligned} H_0 : \quad & \theta = 0 \quad (\text{no effect}) \\ H_1 : \quad & \theta \neq 0 \quad (\text{some effect}) \end{aligned}$$

Two-sided composite hypothesis.

Another Example of Hypothesis

- Let θ denotes the proportion of defective items from a machine.
- One may want the proportion to be less than a specified maximum acceptable proportion θ_0 .
- We want to test whether the products produced by the machine is acceptable.

$$H_0 : \theta \leq \theta_0 \quad (\text{acceptable})$$

$$H_1 : \theta > \theta_0 \quad (\text{unacceptable})$$

A hypothesis testing procedure is a rule that specifies:

- 1 For which sample points H_0 is accepted as true (the subset of the sample space for which H_0 is accepted is called the acceptable region)
- 2 For which sample points H_0 is rejected and H_1 is accepted as true (the subset of sample space for which H_0 is rejected is called the rejection region or critical region).

Rejection region (R) on a hypothesis is usually defined through a test statistic $W(\mathbf{X})$. For example,

$$R_1 = \{\mathbf{x} : W(\mathbf{x}) > c, \mathbf{x} \in \mathcal{X}\}$$

$$R_2 = \{\mathbf{x} : W(\mathbf{x}) \leq c, \mathbf{x} \in \mathcal{X}\}$$

Example of hypothesis testing

$X_1, X_2, X_3 \stackrel{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)$. Consider hypothesis tests

$$H_0 : p \leq 0.5$$

$$H_1 : p > 0.5$$

① Test 1 : Reject H_0 if $\mathbf{x} \in \{(1, 1, 1)\}$

$$\iff \text{rejection region} = \{(1, 1, 1)\}$$

$$\iff \text{rejection region} = \left\{ \mathbf{x} : \sum x_i > 2 \right\}$$

② Test 2 : Reject H_0 if $\mathbf{x} \in \{(1, 1, 0), (1, 0, 1), (0, 1, 1), (1, 1, 1)\}$

$$\iff \text{rejection region} = \{(1, 1, 0), (1, 0, 1), (0, 1, 1), (1, 1, 1)\}$$

$$\iff \text{rejection region} = \left\{ \mathbf{x} : \sum x_i > 1 \right\}$$

Example

Let $X_1, \dots, X_n$ be changes in blood pressure after a treatment.

$$H_0 : \theta = 0$$

$$H_1 : \theta \neq 0$$

An example rejection region $R = \left\{ \mathbf{x} : \frac{\bar{x}}{s_x/\sqrt{n}} > 3 \right\}$.

		Decision	
Truth		Accept H_0	Reject H_0
	H_0	Correct Decision	Type I error
	H_1	Type II error	Correct Decision

Type I and Type II error

Type I error

If $\theta \in \Omega_0$ (if the null hypothesis is true), the probability of making a type I error is

$$\Pr(\mathbf{X} \in R \mid \theta)$$

Type II error

If $\theta \in \Omega_0^c$ (if the alternative hypothesis is true), the probability of making a type II error is

$$\Pr(\mathbf{X} \notin R \mid \theta) = 1 - \Pr(\mathbf{X} \in R \mid \theta)$$

Definition - The power function

The power function of a hypothesis test with rejection region R is the function of θ defined by

$$\beta(\theta) = \Pr(\mathbf{X} \in R \mid \theta) = \Pr(\text{reject } H_0 \mid \theta)$$

If $\theta \in \Omega_0^c$ (alternative is true), the probability of rejecting H_0 is called the power of test for this particular value of θ .

- Probability of type I error = $\beta(\theta)$ if $\theta \in \Omega_0$.
- Probability of type II error = $1 - \beta(\theta)$ if $\theta \in \Omega_0^c$.

An ideal test should have power function satisfying $\beta(\theta) = 0$ for all $\theta \in \Omega_0$, $\beta(\theta) = 1$ for all $\theta \in \Omega_0^c$, which is typically not possible in practice.