Emory University

MATH 362 Mathematical Statistics II

Learning Notes

Jiuru Lyu

June 18, 2025

Contents

1 Estimation

1.1 Introduction

Definition 1.1.1 (Model). A *model* is a distribution with certain parameters.

Example 1.1.2 The normal distribution: $N(\mu, \sigma^2)$.

Definition 1.1.3 (Population). The *population* is all the objects in the experiment.
Definition 1.1.4 (Data, Sample, and Random Sample). *Data* refers to observed value from sample. The *sample* is a subset of the population. A *random sample* is a sequence of independent, identical ($i.i.d.$) random variables.
Definition 1.1.5 (Statistics). *Statistics* refers to a function of the random sample.

Example 1.1.6 The sample mean is a function of the sample:

$$\overline{Y} = \frac{1}{n}(Y_1 + \cdots + Y_n).$$

Example 1.1.7 Central Limit Theorem

We randomly toss $n = 200$ fair coins on the table. Calculate, using the central limit theorem, the probability that at least 110 coins have turned on the same side.

$$\overline{X} = \frac{X_1 + \cdots + X_{200}}{200} \quad \overset{\text{CLT}}{\sim} \quad N\left(\mu, \sigma^2\right),$$

where

$$\mu = \mathbf{E}\left(\overline{X}\right) = \frac{\sum_{i=1}^{200} \mathbf{E}(X_i)}{200},$$

$$\sigma^2 = \mathbf{Var}\left(\overline{X}\right) = \mathbf{Var}\left(\frac{X_1 + \cdots + X_{200}}{200}\right) = \frac{\sum_{i=1}^{200} \mathbf{Var}(X_i)}{200^2}.$$

Definition 1.1.8 (Statistical Inference). The process of *statistical inference* is defined to be the process of using data from a sample to gain information about the population.

Example 1.1.9 Goals in statistical inference

1. **Definition 1.1.10 (Estimation).** To obtain values of the parameters from the data.
2. **Definition 1.1.11 (Hypothesis Testing).** To test a conjecture about the parameters.
3. **Definition 1.1.12 (Goodness of Fit).** How well does the data fit a given distribution.
4. Linear Regression

1.2 The Method of Maximum Likelihood and the Method of Moments

Example 1.2.1 Given an unfair coin, or p-coin, such that

$$X = \begin{cases} 1 & \text{head with probability } p, \\ 0 & \text{tail with probability } 1-p. \end{cases}$$

How can we determine the value p?

Solution 1.

1. Try to flip the coin several times, say, three times. Suppose we get HHT.
2. Draw a conclusion from the experiment.

Key idea: The choice of the parameter p should be the value that maximizes the probability of the sample.

$$\mathbf{P}(X_1 = 1, X_2 = 1, X_3 = 0) = \mathbf{P}(X_1 = 1)\mathbf{P}(X_2 = 1)\mathbf{P}(X_3 = 0) = p^2(1-p) \coloneqq f(p).$$

Solving the optimization problem $\max_{p>0} f(p)$, we find it is most likely that $p = \frac{2}{3}$. This method is called the *likelihood maximization method.* □

Definition 1.2.2 (Likelihood Function). For a random sample of size n from the discrete (or continuous) pdf $p_X(k;\theta)$ (or $f_Y(y;\theta)$), the *likelihood function*, $\mathbf{L}(\theta)$, is the product of the pdf evaluated at $X_i = k_i$ (or $Y_i = y_i$). That is,

$$\mathbf{L}(\theta) \coloneqq \prod_{i=1}^{n} p_X(k_i;\theta) \quad \text{or} \quad \mathbf{L}(\theta) \coloneqq \prod_{i=1}^{n} f_Y(y_i;\theta).$$

Definition 1.2.3 (Maximum Likelihood Estimate). Let $\mathbf{L}(\theta)$ be as defined in Definition 1.2.2. If θ_e is a value of the parameter such that $\mathbf{L}(\theta_e) \geq \mathbf{L}(\theta)$ for all possible values of θ, then we call θ_e the *maximum likelihood estimate* for θ.

Theorem 1.2.4 The Method of Maximum Likelihood
Given random samples $X_1, \ldots, X_N$ and a density function $p_X(x)$ (or $f_X(x)$), then we have the likelihood function defined as

$$\begin{aligned} \mathbf{L}(\theta) = p_X(X;\theta) &= \mathbf{P}(X_1, X_2, \ldots, X_N) \\ &= \mathbf{P}(X_1)\mathbf{P}(X_2)\cdots\mathbf{P}(X_N) && [independent] \\ &= \prod_{i=1}^{N} p_X(X_i;\theta) && [identical] \end{aligned}$$

Then, the maximum likelihood estimate for θ is given by

$$\theta^* = \arg\max_{\theta} L(\theta),$$

where

$$\mathbf{L}\Big(\arg\max_{\theta} L(\theta)\Big) = \mathbf{L}^*(\theta) = \max_{\theta} \mathbf{L}(\theta).$$

Example 1.2.5 Consider the Poisson distribution $X = 0, 1, \ldots,$ with $\lambda > 0$. Then, the pdf is given by

$$p_X(k,\lambda) = e^{-\lambda}\frac{\lambda^k}{k!}, \quad k = 0, 1, \ldots$$

Given data $k_1, \ldots, k_n$, we have the likelihood function

$$\mathbf{L}(\lambda) = \prod_{i=1}^{n} p_X(X = k; \lambda) = \prod_{i=1}^{n} e^{-\lambda}\frac{\lambda^{k_i}}{k_i!} = e^{-n\lambda}\frac{\lambda^{\sum k_i}}{k_1!\cdots k_n!}$$

Then, to find the maximum likelihood estimate of λ, we need to $\max_{\lambda} \mathbf{L}(\lambda)$. That is to solve $\frac{\partial \mathbf{L}(\lambda)}{\partial \lambda} = 0$ and $\frac{\partial^2 \mathbf{L}(\lambda)}{\partial \lambda^2} < 0$.

Example 1.2.6 Waiting Time.

Consider the exponential distribution $f_Y(y) = \lambda e^{-\lambda y}$ for $y \geq 0$. Find the MLE λ_e of λ.

Solution 2.

The likelihood function of the exponential distribution is given by

$$\mathbf{L}(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda y_i} = \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} y_i\right).$$

Now, define

$$\ell(\lambda) = \ln \mathbf{L}(\lambda) = n \ln \lambda - \lambda \sum_{i=1}^{n} y_i.$$

To optimize $\ell(\lambda)$, we compute

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\ell(\lambda) = \frac{n}{\lambda} - \sum_{i=1}^{n} y_i \overset{set}{=} 0$$

So,

$$\frac{n}{\lambda} = \sum_{i=1}^{n} y_i \implies \lambda_e = \frac{n}{\sum_{i=1}^{n} y_i} =: \frac{1}{\overline{y}},$$

where $\overline{y}$ is the sample mean. □

Example 1.2.7 Given the exponential distribution $f_Y(y) = \lambda e^{-\lambda y}$ for $y \geq 0$. Find the MLE of λ^2.

Solution 3.

Define $\tau = \lambda^2$. Then, $\lambda = \sqrt{\tau}$, and so

$$f_Y(y) = \sqrt{\tau} e^{-\sqrt{\tau} y}, \quad y \geq 0.$$

Then, the likelihood function becomes

$$\mathbf{L}(\tau) = \prod_{i=1}^{n} f_Y(y) = \tau^{\frac{n}{2}} \exp\left(-\sqrt{\tau} \sum_{i=1}^{n} y_i\right).$$

Similarly, after maximization, we find

$$\tau_e = \frac{1}{(\overline{y})^2}.$$

□

Theorem 1.2.8 Invariant Property for MLE
Suppose λ_e is the MLE of λ. Define $\tau := h(\lambda)$. Then, $\tau_e = h(\lambda_e)$.

Proof 4. In this proof, we will prove the case when h is a one-to-one function. The case of h being a many-to-one function is beyond the scope of this course.

Suppose $h(\cdot)$ is a one-to-one function. Then, $\lambda = h^{-1}(\tau)$ is well-defined. Then,

$$\max_{\lambda} \mathbf{L}(\lambda; y_1, \ldots, y_n) = \max_{\tau} \mathbf{L}\big(h^{-1}(\tau); y_1, \ldots, y_n\big) = \max_{\tau} \mathbf{L}(\tau; y_1, \ldots, y_n).$$

■

Example 1.2.9 Waiting Time with an unknown Threshold.

Let $\lambda = 1$ in exponential but there is an unknown threshold θ, that, is $f_Y(y) = e^{-(y-\theta)}$ for $y \geq \theta,\ \theta > 0$.

Solution 5.

Note that the likelihood function is given by

$$\begin{aligned} \mathbf{L}(\theta; y_1, \ldots, y_n) = \prod_{i=1}^{n} f_Y(y_1) &= \exp\left(-\sum_{i=1}^{n}(y_i - \theta)\right), \quad y_i \geq \theta,\ \theta > 0 \\ &= \exp\left(-\sum_{i=1}^{n}(y_i - \theta)\right) \cdot \mathbb{1}_{[y_i \geq 0,\ \theta > 0]}, \end{aligned}$$

where

$$\mathbb{1}_{x \in A} = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A. \end{cases}$$

Using order statistics,

$$\begin{aligned} \mathbf{L}(\theta) &= \exp\left(-\sum_{i=1}^{n}(y_i - \theta)\right) \cdot \mathbb{1}_{\left[y_{(n)} \geq y_{(n-1)} \geq \cdots \geq y_{(1)} \geq \theta,\ \theta > 0\right]} \\ &= \exp\left(-\sum_{i=1}^{n} y_i + n\theta\right) \mathbb{1}_{\left[y_{(n)} \geq \cdots \geq y_{(1)} \geq \theta,\ \theta > 0\right]}. \end{aligned}$$

So, we know $\theta \leq y_{(1)} = y_{\min}$.

To maximize the likelihood function, we want to maximize $-\sum y_i + n\theta$. That is, to maximize θ, as $\theta \leq y_{\min}$, it must be that $\theta_{\max} = y_{\min}$. Therefore, the MLE is $\theta^* = y_{\min}$. □

Example 1.2.10 Suppose $Y_1, \ldots, Y_n \sim$ Uniform$[0, a]$. That is, $f_Y(y; a) = \dfrac{1}{a}$ for $y \in [0, a]$. Find MLE a_e of a.

Solution 6.

Note that

$$\begin{aligned} f_Y(y; a) &= \frac{1}{a} \cdot \mathbb{1}_{\{y \in [0,a]\}} \\ &= \frac{1}{a} \cdot \mathbb{1}_{\left\{0 \le y_{(1)} \le \cdots \le y_{(n)} \le a\right\}} \qquad \text{where } y_{(1)} = \min y_i \text{ and } y_{(n)} = \max y_i \end{aligned}$$

Then,

$$\mathbf{L}(a) = \frac{1}{a^n} \mathbb{1}_{\left\{0 \le y_{(1)} \le \cdots \le y_{(n)} \le a\right\}}$$

To maximize $\mathbf{L}(a)$, we want to minimize a^n. Since $a \ge y_{(n)}$, it must be that $a_e = y_{(n)}$. Here, we call $a_e = y_{(n)}$ an *estimate*, and $\widehat{a_{\text{MLE}}} = Y_{(n)}$ an *estimator*. □

Example 1.2.11 MLE that Does Not Esist

Suppose $f_Y(y; a) = \dfrac{1}{a}, \quad y \in [0, a)$. Find the MLE.

Solution 7.

The likelihood function is the same:

$$\mathbf{L}(a) = \frac{1}{a^n} \mathbb{1}_{\left\{0 \le y_{(1)} \le \cdots \le y_{(n)} < a\right\}}.$$

However, since $[0, a)$ is not a closed set, the optimization problem $\max\limits_{a \in [0,a)} \mathbf{L}(a)$ does not have a solution. Hence, the estimate does not exist. □

Remark 1.1 *MLE may not be unique all the time.*

Example 1.2.12 Multiple MLE Values

Suppose $X_1, \ldots, X_n \sim$ Uniform$\left[a - \dfrac{1}{2}, a + \dfrac{1}{2}\right]$, where $f_X(x; a) = 1, \ x \in \left[a - \dfrac{1}{2}, a + \dfrac{1}{2}\right]$. Find the MLE.

Solution 8.

In the indicator function notation, we can rewrite the pdf to be

$$f_X(x; a) = \mathbb{1}_{\left\{a - \frac{1}{2} \le x \le a + \frac{1}{2}\right\}} = \mathbb{1}_{\left\{a - \frac{1}{2} \le x_{(1)} \le \cdots \le x_{(n)} \le a + \frac{1}{2}\right\}}.$$

So, the likelihood function will be

$$\mathbf{L}(a) = \prod_{i=1}^{n} f_x(x_i; a) = \begin{cases} 1, & a \in \left[x_{(n)} - \frac{1}{2}, x_{(1)} + \frac{1}{2}\right] \\ 0, & \text{otherwise}. \end{cases}$$

So, the $\mathbf{L}(a)$ will be maximized whenever $a \in \left[x_{(n)} - \frac{1}{2}, x_{(1)} + \frac{1}{2}\right]$. Therefore, MLE can be any value in the range $\left[x_{(n)} - \frac{1}{2}, x_{(1)} + \frac{1}{2}\right]$. Say,

$$a_e = x_{(n)} - \frac{1}{2} \quad \text{or} \quad a_e = x_{(1)} - \frac{1}{2} \quad \text{or} \quad a_e = \frac{x_{(n)} - \frac{1}{2} + x_{(1)} + \frac{1}{2}}{2} = \frac{x_{(n)} + x_{(1)}}{2}.$$

□

Theorem 1.2.13 MLE for Multiple Parameters
In general, we have the likelihood function $\mathbf{L}(\theta)$, where $\theta = (\theta_1, \dots, \theta_p)$. To find the MLE, we need

$$\frac{\partial \mathbf{L}(\theta)}{\partial \theta_i} = 0 \quad i = 1, \dots, p,$$

and the Hessian matrix

$$\left(\frac{\partial^2 \mathbf{L}(\theta)}{\partial \theta_i \partial \theta_j}\right)_{i,j=1,\dots,p} := \begin{pmatrix} \frac{\partial^2 \mathbf{L}(\theta)}{\partial \theta_1^2} & \cdots & \frac{\partial^2 \mathbf{L}(\theta)}{\partial \theta_1 \partial \theta_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 \mathbf{L}(\theta)}{\partial \theta_p \partial \theta_1} & \cdots & \frac{\partial^2 \mathbf{L}(\theta)}{\partial \theta_p^2} \end{pmatrix}$$

should be negative dfinite.

Example 1.2.14 MLE for Multiple Parameters: Normal Distribution
Suppose $Y_1, \dots, Y_n \sim N(\mu, \sigma)$. Then,

$$f_{Y_i}(u; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(y_i - \mu)^2/\left(2\sigma^2\right)}.$$

Find the MLE for μ and σ.

Solution 9.

The likelihood function will be

$$\mathbf{L}(\mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-(y_i-\mu)^2/\left(2\sigma^2\right)}.$$

Then, we define

$$\ell(\mu, \sigma) = \ln \mathbf{L}(\mu, \sigma) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln \sigma^2 - \frac{1}{2}\left(\sigma^2\right)^{-}1 \sum_{i=1}^{n} (y_i - \mu)^2.$$

Set

$$\begin{cases} \dfrac{\partial \ell(\mu, \sigma)}{\partial \mu} = 0 & \text{①} \\ \dfrac{\partial \ell(\mu, \sigma)}{\partial \sigma} = 0 & \text{②} \end{cases}$$

From ①, we have

$$\frac{1}{\sigma^2} \sum_{i=1}^{n} (y_1 - \mu) = 0$$

$$\sum_{i=1}^{n} y_i = n\mu \implies \boxed{\mu_e = \frac{\sum y_i}{n} = \overline{y}}$$

From ②, by the invariant property of MLE, we instead set

$$\frac{\partial \ell(\mu, \sigma)}{\partial \sigma^2} = 0$$

$$-\frac{n}{2} \cdot \frac{1}{\sigma^2} + \frac{1}{2}\left(\frac{1}{\sigma^2}\right)^2 \sum_{i=1}^{n} (y_i - \mu)^2 = 0$$

$$\frac{1}{2\sigma^2}\left(-n + \frac{1}{\sigma^2}\sum_{i=1}^{n} (y_i - \mu)^2\right) = 0$$

$$-n\sigma^2 + \sum_{i=1}^{n} (y_i - \mu)^2 = 0 \qquad (\mu_e = \overline{y})$$

$$\sum_{i=1}^{n} (y_i - \overline{y})^2 = n\sigma^2$$

$$\sigma_e^2 = \frac{1}{n}\sum_{i=1}^{n} (y_i - \overline{y})^2 \implies \boxed{\sigma_e = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \overline{y})^2}}$$

□

1.3 The Method of Moment

Definition 1.3.1 (Moment Generating Function). The *Moment Generating Function (MGF)* is defined as

$$\mathbf{M}_X(t) = \mathbf{E}\left[e^{tX}\right],$$

and it uniquely determines a probability distribution.

Definition 1.3.2 (Moment). The k*-th order moment* of X is $\mathbf{E}\left[X^k\right]$.

Example 1.3.3 Meaning of Different Moments

- $\mathbf{E}[X]$: location of a distribution
- $\mathbf{E}[X^2] = \mathbf{Var}(X) - \mathbf{E}[X]^2$: width of a distribution
- $\mathbf{E}[X^3]$: skewness – positively skewed / negatively skewed
- $\mathbf{E}[X^4]$: kurtosis / tailedness – speed decaying to 0.

Example 1.3.4 Moment Estimate: Moments of Population and Sample

Population	**Sample,** $X_1, \ldots, X_n$
$\mathbf{E}[X] = \mu$	$\widehat{\mu} = \overline{X} = \dfrac{X_1 + \cdots + X_n}{n}$
$\mathbf{E}[X^2] = \mu^2 + \sigma^2$	$\widehat{\mu}^2 + \widehat{\sigma}^2 = \dfrac{X_1^2 + \cdots + X_n^2}{n}$
$\vdots$	$\vdots$
$\mathbf{E}\left[X^k\right]$	$\dfrac{X_1^k + \cdots + X_n^k}{n}$

Rationale: The population moments should be close to the sample moments.

Example 1.3.5

- Consider $N(\mu, \sigma^2)$, where σ is given. Estimate μ.

  By the method of moment estimate, we have $\mu_e = \overline{X}$.

- Consider $N(\mu, \sigma^2)$. Estimate μ and σ.

  We have $\mu_e = \overline{X}$ and $\mu_e^2 + \sigma_e^2 = \dfrac{X_1^2 + \cdots + X_n^2}{n}$.

- Consider $N(\theta, \sigma^2)$. Given $\mathbf{E}(X^4) = 3\sigma^4$, estimate μ and σ.

  We have $\mu_e = \overline{X}$, $\mu_e^2 + \sigma_e^2 = \dfrac{X_1^2 + \cdots + X_n^2}{n}$, and $3\sigma^4 = \dfrac{X_1^4 + \cdots + X_n^4}{n}$. We have three equations but only two unknowns, then a solution is not guaranteed. So, we need some restrictions on this method (see Remark 1.2).

Theorem 1.3.6 Method of Moments Estimates
For a random sample of size n from the discrete (or continuous) population/pdf $p_X(k; \theta_1, \ldots, \theta_s)$ (or $f_Y(y; \theta_1, \ldots, \theta_s)$), solutions to the system

$$\begin{cases} \mathbf{E}(Y) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} y_i \\ \qquad \vdots \\ \mathbf{E}(Y^s) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} y_i^s \end{cases}$$

which are denoted by $\theta_{1e}, \ldots, \theta_{se}$, are called the **method of moments estimates** of $\theta_1, \ldots, \theta_s$.

Remark 1.2 *To estimate k parameters with the method of moments estimates, we will only match the first k orders of moments.*

Example 1.3.7 Consider the Gamma distribution:

$$f_Y(y; r, \lambda) = \frac{\lambda^r}{\Gamma(r)} y^{r-1} e^{-\lambda y} \quad \text{for } y \geq 0.$$

Given $\mathbf{E}(Y) = \dfrac{r}{\lambda}$ and $\mathbf{E}(Y^2) = \dfrac{r}{\lambda^2} + \dfrac{r^2}{\lambda^2}$. Estimate r and λ.

Solution 1.

$$\mathbf{E}(Y) = \frac{r}{\lambda} \implies \frac{r_e}{\lambda_e} = \frac{y_1 + \cdots + y_n}{n} = \overline{y} \qquad ①$$

$$\mathbf{E}\big(Y^2\big) = \frac{r}{\lambda^2} + \frac{r^2}{\lambda^2} \implies \frac{r_e}{\lambda_e^2} + \frac{r_e^2}{\lambda_e^2} = \frac{y_1^2 + \cdots + y_n^2}{n} \qquad ②$$

Substitute ① into ②, we have

$$\frac{\overline{y}}{\lambda_e} + (\overline{y})^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 \implies \boxed{\lambda_e = \frac{\overline{y}}{\frac{1}{n}\sum y_i^2 - \overline{y}^2}} \qquad ③$$

Substitute ③ into ①, we have

$$r_e = \overline{y}\lambda_e = \boxed{\frac{\overline{y}^2}{\frac{1}{n}\sum y_i^2 - \overline{y}^2}}.$$

□

Remark 1.3 *The* sample variance *is defined as*

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(y_i^2 - 2y_i\overline{y} + \overline{y}^2\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} y_i^2 - 2\overline{y}\cdot\frac{\sum y_i}{n} + \frac{1}{n}\cdot n\overline{y}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} y_i^2 - 2\overline{y}^2 + \overline{y}^2 & \overline{y} = \frac{\sum y_i}{n} \\
&= \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \overline{y}^2.
\end{aligned}$$

So, in Example 1.3.7, if we define $\widehat{\sigma}^2$ to be the sample variance, we can further simply our estimate as follows:

$$\lambda_e = \frac{\overline{y}}{\widehat{\sigma}^2}, \qquad r_e = \frac{\overline{y}^2}{\widehat{\sigma}^2}.$$

1.4 Interval Estimation

Example 1.4.1 Estimate μ, where $X \sim N(\mu, 1)$.

We take some samples and compute their sample means:

$$\overline{X}^1 = \frac{x_1 + \cdots + x_n}{n}, \overline{X}^2 = \frac{\widetilde{x_1} + \cdots + \widetilde{x_n}}{n}, \cdots$$

Finding the distribution of $\overline{X}$, we can find an interval $\left[\widehat{\theta}_L, \widehat{\theta}_U\right]$ such that

$$\mathbf{P}\left(\widehat{\theta}_L \le \overline{X} \le \widehat{\theta}_U\right) = 1 - \alpha.$$

Remark 1.4 *By using the variance of the estimator, one can construct an interval such that with a high probability that the interval contains the unknown parameter.*

Definition 1.4.2 (Confidence Interval). The interval, $\left[\widehat{\theta}_L, \widehat{\theta}_U\right]$ is called the *confidence interval*, and the high probability is $1-\alpha$, where α is given.

Remark 1.5 *Take $\alpha = 5\%$, then $\left[\widehat{\theta}_L, \widehat{\theta}_U\right]$ is the 95% confidence interval of μ. It does not mean that μ has 95% chance to be in $\left[\widehat{\theta}_L, \widehat{\theta}_U\right]$. However, if we construct 1000 such intervals, 950 of them will contain μ.*

Example 1.4.3 A random sample of size 4, $(Y_1 = 6.5,\ Y_2 = 9.2,\ Y_3 = 9.9,\ Y_4 = 12.4)$, from a normal population:

$$f_Y(y;\mu) = \frac{1}{\sqrt{2\pi}0.8} e^{-\frac{1}{2}\left(\frac{y-\mu}{0.8}\right)^2} \sim N(\mu, \sigma^2 = 0.64).$$

Both MLE and MME give $\mu_e = \overline{y} = 9.5$. The estimator $\widehat{\mu} = \overline{Y}$ follows normal distribution. Construct 95%-confidence interval for μ.

Solution 1.

$\mathbf{E}\left(\overline{Y}\right) = \mu$ and $\mathbf{Var}\left(\overline{Y}\right) = \frac{\sigma^2}{n} = \frac{0.64}{4}$. By the Central Limit Theorem, $\overline{Y}$ approximately follow $N\left(\mu, \frac{\sigma^2}{n}\right)$. So, $\frac{\overline{Y}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$. Then,

$$\mathbf{P}\left(z_1 \leq \frac{\overline{Y}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \leq z_2\right) = 0.95 \implies \mathbf{P}\left(\overline{Y} - z_2\sqrt{\frac{\sigma^2}{n}} \leq \mu \leq \overline{Y} - z_1\sqrt{\frac{\sigma^2}{n}}\right) = 0.95$$

There are infinite many ways to construct a confidence interval by selecting different z_1 and z_2. However, since we don't have any prior knowledge on μ, it is good for us to choose z_1 and z_2 symmetrically. Moreover, symmetric z_1 and z_2 will yield a smaller interval. We know the symmetric z_1, z_2 pair will be $z_1 = -1.96$ and $z_2 = 1.96$. Therefore,

$$\mathbf{P}\left(\overline{Y} - 1.96\sqrt{\frac{0.64}{4}} \leq \mu \leq \overline{Y} + 1.96\sqrt{\frac{0.64}{4}}\right) = 0.95.$$

Then, 95% confidence interval is $[9.5 - 1.96 \times 0.4,\ 9.5 + 1.96 \times 0.4]$. □

Theorem 1.4.4 Confidence Interval

In general, for a normal population with σ known, the $100(1-\alpha)\%$ *two-sided confidence interval* for μ is

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

Theorem 1.4.5 Variation of Confidence Interval

- One-sided interval:

$$\left(\overline{y} - z_\alpha \frac{\sigma}{\sqrt{n}},\ \overline{y}\right) \text{ or } \left(\overline{y},\ \overline{y} + z_\alpha \frac{\sigma}{\sqrt{n}}\right)$$

- σ is unknown and sample size is small: z-score $\to$ t-score.
- σ is unknown and sample size is large: z-score by CLT.
- Non Gaussian population but sample size is large: z-score by CLT.

Theorem 1.4.6
Let k be the number of successes in n independent trials, where n is large and $p = \mathbf{P}(\text{success})$ is unknown. An approximate $100(1-\alpha)\%$ confidence interval for p is the set of numbers

$$\left(\frac{k}{n} - z_{\alpha/2}\sqrt{\frac{(k/n)(1-k/n)}{n}},\ \frac{k}{n} + z_{\alpha/2}\sqrt{\frac{(k/n)(1-k/n)}{n}}\right).$$

Definition 1.4.7 (Margin of Error). The *margin of error*, denoted by d, is the quantity

$$d = z_{\alpha/2}\sqrt{\frac{(k/n)(1-k/n)}{n}}.$$

Remark 1.6 *Stating the sample mean and the margin of error is equivalent to stating the confidence interval. Note that* $\mathrm{C.I.} = \widehat{p} \pm d$.

Theorem 1.4.8 Estimate Margin of Error
When p is close to $\frac{1}{2}$, then $d \approx d_m = \frac{z_{\alpha/2}}{2\sqrt{n}}$, which is equivalent to $\sigma_n \approx \frac{1}{2\sqrt{n}}$. However, if p is away from $\frac{1}{2}$, d and d_m are very different.

Remark 1.7 *Theorem 1.4.8 gives aconservative estimation of the margin of error, which is* d_m.

Proposition 1.9 : Given d, we can estimate the sample size.

Proof 2.

$$d = z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \implies n \approx \widehat{p}(1-\widehat{p})/\left(\frac{d}{z_{\alpha/2}}\right)^2.$$

However, since n is unknown, $\widehat{p}$ is also unknown. We, therefore, need information on the actual p to conclude an estimation of the sample size.

- If p is known,

$$n = \frac{p(1-p)}{\left(\frac{d}{z_{\alpha/2}}\right)^2}.$$

- If p is unknown. Let $f(p) = p(1-p)$. f will be maximized when $p = 0.5$. So, $f(p) = p(1-p) \leq 0.25$. Then,

$$n \leq \frac{0.25}{\left(\frac{d}{z_{\alpha/2}}\right)^2}.$$

  Since we are conservative, take $n = \frac{\frac{1}{4}z_{\alpha/2}^2}{d^2} = \frac{z_{\alpha/2}^2}{4d^2}$. This estimation is a conservative estimation of the sample size.

■

1.5 Properties of Estimation

The main question is that estimators are not unique in general. How do we choose a good estimator?

Definition 1.5.1 (Unbiasedness). Given a random sample of size n when whose population distribution depends on an unknown parameter θ. Let $\widehat{\theta}$ be an estimator of θ. Then,

- $\widehat{\theta}$ is called *unbiased* if $\mathbf{E}(\widehat{\theta}) = \theta$.
- $\widehat{\theta}$ is called *asymptotically unbiased* if $\lim_{n\to\infty} \mathbf{E}(\widehat{\theta}) = \theta$.
- If θ is biased, then the *bias* is given by the quantity $\mathbf{B}(\widehat{\theta}) = \mathbf{E}(\widehat{\theta}) - \theta$.

Example 1.5.2 Consider the exponential distribution: $f_Y(y;\lambda) = \lambda e^{-\lambda y}$ for $y \geq 0$. Determine if the estimator $\widehat{\lambda} = \frac{1}{\overline{Y}}$ is biased or not.

Hint: $n\overline{Y} = \sum_{i=1}^{n} Y_i \sim$ *Gamma*(n, λ).

Solution 1.

Recall that $\mathbf{E}[g(x)] = \int_x g(x) f_X(x)\,\mathrm{d}x$. Define $X = \sum_{i=1}^{n} Y_i \sim \text{Gamma}(n, \lambda)$. Also, recall the following facts:

$$\Gamma(n) = (n-1)! = (n-1)\Gamma(n-1)$$

and the integration over any probability density function will yield a result of 1 by definition.

Then,

$$\begin{aligned}
\mathbf{E}(\widehat{\lambda}) = \mathbf{E}\left(\frac{1}{\overline{\overline{Y}}}\right) = \mathbf{E}\left(\frac{n}{\sum Y_i}\right) &= n\mathbf{E}\left(\frac{1}{\sum Y_i}\right) \\
&= n\mathbf{E}\left(\frac{1}{X}\right) \\
&= n\int_x \frac{1}{x}\cdot\frac{\lambda^n}{\Gamma(n)}x^{n-1}e^{-\lambda x}\,\mathrm{d}x \\
&= n\int_x \frac{\lambda^n}{(n-1)!}x^{n-2}e^{-\lambda x}\,\mathrm{d}x \\
&= \frac{n\lambda}{(n-1)}\underbrace{\int_x \frac{\lambda^{n-1}}{\Gamma(n-1)}x^{n-2}e^{-\lambda x}\,\mathrm{d}x}_{=1} \\
&= \frac{n}{n-1}\lambda.
\end{aligned}$$

Therefore, $\mathbf{E}(\widehat{\lambda}) \neq \lambda$, and so $\widehat{\lambda}$ is biased. However, note that

$$\lim_{n\to\infty}\mathbf{E}(\widehat{\lambda}) = \lim_{n\to\infty}\frac{n}{n-1}\lambda = \lambda.$$

By definition, then $\widehat{\lambda}$ is asymptotically unbiased. □

Example 1.5.3 Consider the exponential distribution $f(y;\theta) = \frac{1}{\theta}e^{-y/\theta}$ for $y \geq 0$. Then, $\widehat{\theta} = \overline{Y}$ is unbiased.

Remark 1.8 *Suppose* $\{X_1,\ldots,X_n\}$ *are* $i.i.d.$ *random variables, and* $\mathbf{E}(X_i) = \mu$ *for* $i = 1,\ldots,n$ *. Then,* $\overline{X}$*, the sample mean, is always an unbiased estimator:*

$$\mathbf{E}(\overline{X}) = \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}(X_i) = \frac{1}{n}\cdot n\cdot\mu = \mu.$$

Theorem 1.5.4 Sample Variance is Biased
Suppose $\{X_1,\ldots,X_n\}$ are $i.i.d.$ random variables, and $\mathbf{E}(X_i) = \mu$, $\mathbf{Var}(X_i) = \sigma^2$ for $i = 1,\ldots,n$. Then, the sample variance $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ is biased.

Proof 2. Note that

$$
\begin{aligned}
\mathbf{E}(\widehat{\sigma}^2) &= \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right) \\
&= \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \mu + \mu - \overline{X}\right)^2\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\left[(X_i-\mu)^2 + \left(\mu - \overline{X}\right)^2 + 2(X_i - \mu)(\mu - \overline{X})\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\left\{\underbrace{\mathbf{E}(X_i-\mu)^2}_{\mathbf{Var}(X_i)} + \mathbf{E}\left(\mu - \overline{X}\right)^2 + 2\mathbf{E}\left[(\mu - \overline{X})(X_i - \mu)\right]\right\}
\end{aligned}
$$

Hint: $\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu) = \frac{1}{n}\sum_{i=1}^{n}X_i - \frac{1}{n}\sum_{i=1}^{n}\mu = \overline{X} - \mu$

$$
\begin{aligned}
&= \frac{1}{n}\sum_{i=1}^{n}\mathbf{Var}(X_i) + \frac{1}{n}\cdot n\mathbf{E}\left(\mu - \overline{X}\right)^2 + 2\mathbf{E}\left[(\mu - \overline{X})\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\sigma^2 + \mathbf{E}\left(\mu - \overline{X}\right)^2 + 2\mathbf{E}\left[(\mu - \overline{X})(\overline{X} - \mu)\right] \\
&= \frac{1}{n}\cdot n \cdot \sigma^2 + \mathbf{E}\left(\mu - \overline{X}\right)^2 - 2\mathbf{E}\left[(\mu - \overline{X})^2\right] \\
&= \sigma^2 - \mathbf{E}\left(\mu - \overline{X}\right)^2 \\
&= \sigma^2 - \underbrace{\mathbf{E}\left(\overline{X} - \mu\right)^2}_{=\mathbf{Var}(\overline{X})} \\
&= \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2 \neq \sigma^2
\end{aligned}
$$

Therefore, $\widehat{\sigma}^2$ is not an unbiased estimator. ■

Theorem 1.5.5 Adjusted Sample Variance is Unbiased
With the same set up in Theorem 1.5.4, define the adjusted sample variance to be

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

Then, S^2 is an unbiased estimator of σ^2.

Definition 1.5.6 (Decision Theory). Minimize the error of an estimator (sample statistics) relative to the true parameter (population parameter) using a loss function.

Definition 1.5.7 (Mean Squared Error). The *mean squared error* (MSE) is defined by

$$\mathbf{MSE}(\widehat{\theta}) = \mathbf{E}\left[(\widehat{\theta} - \theta)^2\right]$$

Theorem 1.5.8 Decomposition of MSE
Generally,

$$\mathbf{MSE}(\theta) = \mathbf{Var}(\widehat{\theta}) + \mathbf{B}\left(\widehat{\theta}\right)^2$$

If $\widehat{\theta}$ is unbiased, $\mathbf{MSE}(\widehat{\theta}) = \mathbf{Var}(\widehat{\theta})$. $\mathbf{Var}(\theta)$ measures the precision of the estimator.

Proof 3. Note that we will the following:

$$\begin{aligned}
\mathbf{MSE}(\widehat{\theta}) &= \mathbf{E}\left[(\widehat{\theta} - \theta)^2\right] \\
&= \mathbf{E}(\widehat{\theta}^2 + \theta^2 - 2\widehat{\theta}\theta) \\
&= \mathbf{E}(\widehat{\theta}) - 2\theta\mathbf{E}(\widehat{\theta}) + \theta^2 \\
&= \underbrace{\mathbf{E}(\widehat{\theta}^2) - \mathbf{E}(\widehat{\theta})^2}_{} + \underbrace{\mathbf{E}(\widehat{\theta})^2 - 2\theta\mathbf{E}(\widehat{\theta}) + \theta^2}_{} \\
&= \mathbf{Var}(\widehat{\theta}) + \left[\mathbf{E}(\widehat{\theta}) - \theta\right]^2 \\
&= \mathbf{Var}(\theta) + \mathbf{B}(\widehat{\theta})^2
\end{aligned}$$

If $\widehat{\theta}$ is unbiased, $\mathbf{B}(\widehat{\theta}) = 0$, and so $\mathbf{MSE}(\widehat{\theta}) = \mathbf{Var}(\widehat{\theta})$. ■

Definition 1.5.9 (Efficiency). Let $\widehat{\theta}_1$ and $\widehat{\theta}_2$ be two unbiased estimators for a parameter θ. If we have $\mathbf{Var}(\widehat{\theta}_1) < \mathbf{Var}(\widehat{\theta}_2)$, then we say that $\widehat{\theta}_1$ is *more efficient* than $\widehat{\theta}_2$. The *relative efficiency* of $\widehat{\theta}_1$ with respect to $\widehat{\theta}_2$ is the ratio $\dfrac{\mathbf{Var}(\widehat{\theta}_2)}{\mathbf{Var}(\widehat{\theta}_1)}$.

1.6 Best Unbiased Estimator

Definition 1.6.1 (Best/Minimum-Variance Estimator). Let Θ be the set of all estimators $\widehat{\theta}$ that are unbiased for the parameter θ. We way that $\widehat{\theta}^*$ is a *best* or *minimum-variance estimator* (MVE) if $\widehat{\theta}^* \in \Theta$ and $\mathbf{Var}(\widehat{\theta}^*) \leq \mathbf{Var}(\widehat{\theta}) \quad \forall\, \widehat{\theta} \in \Theta$.

Definition 1.6.2 (Fisher's Information). The *Fisher's information* of a continuous random variable Y with pdf $f_Y(y;\theta)$ is defined as

$$\mathbf{I}(\theta) = \mathbf{E}\left[\left(\frac{\partial \ln f_Y(y;\theta)}{\partial\theta}\right)^2\right] = -\mathbf{E}\left[\frac{\partial^2}{\partial\theta^2} \ln f_Y(y;\theta)\right].$$

Remark 1.9 *The Fisher's information measures the amount of information that a sample Y contains about the unknown parameter θ. If $\mathbf{I}(\theta)$ is big, then the curvature of $f_Y(y;\theta)$ is big, and* *thus it is more likely that we can find a region where* $\widehat{\theta}$ *is concentrated.*

Extension 1.1 (Joint Fisher's Information) *Suppose* $Y_1, \ldots, Y_n$ *are continuous* $i.i.d.$ *random variables, each has a Fisher's information of* $\mathbf{I}(\theta)$. *Then,*

$$\mathbf{E}\left[\left(\frac{\partial}{\partial\theta}\ln f_{Y_1,\ldots,Y_n}(y_1,\ldots,y_n;\theta)\right)^2\right] = n\mathbf{I}(\theta).$$

Theorem 1.6.3 Properties of Fisher's Information

Define the *Fisher's Score Function* $\frac{\partial}{\partial\theta}\ln f_Y(y;\theta)$. Then,

$$\mathbf{E}_Y\left[\frac{\partial}{\partial\theta}\ln f_Y(y;\theta)\right] = 0.$$

Proof 1. Note that by chain rule, we have

$$\begin{aligned}
\mathbf{E}_Y\left[\frac{\partial}{\partial\theta}\ln f_Y(y;\theta)\right] &= \int_Y \left(\frac{\partial}{\partial\theta}\ln f_Y(y;\theta)\right) f_Y(y;\theta)\,\mathrm{d}y \\
&= \int_Y \frac{1}{f_Y(y;\theta)}\left(\frac{\partial}{\partial\theta}f_Y(y;\theta)\right) f_Y(y;\theta)\,\mathrm{d}y \\
&= \int_Y \frac{\partial}{\partial\theta}f_Y(y;\theta)\,\mathrm{d}y \\
&= \frac{\partial}{\partial\theta}\int_Y f_Y(y;\theta)\,\mathrm{d}y = \frac{\partial}{\partial\theta}(1) = 0.
\end{aligned}$$

■

Corollary 1.4 :

$$\mathbf{I}(\theta) = \mathbf{Var}\left(\frac{\partial}{\partial\theta}\ln f_Y(y;\theta)\right).$$

Proof 2. By definition, we have

$$\begin{aligned}
\mathbf{Var}\left(\frac{\partial}{\partial\theta}\ln f_Y(y;\theta)\right) &= \mathbf{E}\left[\left(\frac{\partial}{\partial\theta}\ln f_Y(y;\theta)\right)^2\right] - \left(\underbrace{\mathbf{E}\left(\frac{\partial}{\partial\theta}\ln f_Y(y;\theta)\right)}_{=0,\text{ by Theorem 1.6.3.}}\right)^2 \\
&= \mathbf{E}\left[\left(\frac{\partial}{\partial\theta}\ln f_Y(y;\theta)\right)^2\right] \\
&= \mathbf{I}(\theta).
\end{aligned}$$

■

Theorem 1.6.5 Cramér-Rao Inequality
Under regular condition, let $Y_1, \dots, Y_n$ be a random sample of size n form the continuous population pdf $f_Y(y;\theta)$. Let $\widehat{\theta} = \widehat{\theta}(Y_1, \dots, Y_n)$ be any unbiased estimator for θ. Then,

$$\mathbf{Var}(\widehat{\theta}) \geq \frac{1}{n\mathbf{I}(\theta)}.$$

Remark 1.10 *A similar statement holds for the discrete case* $p_X(k;\theta)$.

Definition 1.6.6 (Efficiency of Unbiased Estimator). An unbiased estimator $\widehat{\theta}$ is *efficient* if $\mathbf{Var}(\widehat{\theta})$ is equal to the Cramér-Rao lower bound. That is, $\mathbf{Var}(\widehat{\theta}) = (n\mathbf{I}(\theta))^{-1}$. Such an estimator is the MVE defined in Definition 1.6.1. The *efficiency* of an unbiased estimator $\widehat{\theta}$ is defined to be the quantity

$$\left(n\mathbf{I}(\theta)\mathbf{Var}(\widehat{\theta})\right)^{-1}.$$

Example 1.6.7 Suppose $X \sim$ Bernoulli(p). Is $\widehat{p} = \overline{X}$ efficient?

Solution 3.

Note that we have the following

$$\begin{aligned}
f_X(x;p) &= p^x(1-p)^{1-x}, \quad x = 0, 1 \\
\ln f_X(x;p) &= x \ln p + (1-x)\ln(1-p) \\
\frac{\partial}{\partial p} \ln f_X(x;p) &= \frac{x}{p} - \frac{1-x}{1-p} \\
\frac{\partial^2}{\partial p^2} \ln f_X(x;p) &= -\frac{x}{p^2} - \frac{1-x}{(1-p)^2}
\end{aligned}$$

Therefore, the Fisher's information can be computed by

$$\begin{aligned}
\mathbf{I}(p) = -\mathbf{E}\left[\frac{\partial^2}{\partial p^2} \ln f_X(x;p)\right] &= -\mathbf{E}\left[-\frac{x}{p^2} - \frac{1-x}{(1-p)^2}\right] \\
&= \mathbf{E}\left[\frac{x}{p^2}\right] + \mathbf{E}\left[\frac{1-x}{(1-p)^2}\right] \\
&= \frac{\mathbf{E}(x)}{p^2} + \frac{1-\mathbf{E}(x)}{(1-p)^2} \\
&= \frac{p}{p^2} + \frac{1-p}{(1-p)^2} = \frac{1}{p} + \frac{1}{1-p} = \frac{1}{p(1-p)}.
\end{aligned}$$

Note that

$$\mathbf{Var}(\overline{X}) = \mathbf{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathbf{Var}(X_i) = \frac{1}{n}\mathbf{Var}(X_i) = \frac{1}{n} \cdot p(1-p).$$

So, we have

$$\mathbf{Var}(\overline{X}) = \frac{p(1-p)}{n} = \frac{1}{n\left(\frac{1}{p(1-p)}\right)} = \frac{1}{n\mathbf{I}(p)}.$$

Therefore, $\widehat{p}$ is efficient. □

Example 1.6.8 Suppose $X \sim N(\mu, \sigma^2)$, with σ^2 is known. What is $\mathbf{I}(\mu)$?

Solution 4.

Note that

$$\frac{\mathrm{d}^2}{\mathrm{d}\mu^2} \ln f_X(x;\mu) = -\frac{1}{\sigma^2}.$$

Then,

$$\mathbf{I}(\mu) = -\mathbf{E}\left[\frac{\mathrm{d}^2}{\mathrm{d}\mu^2} \ln f_X(x;\mu)\right] = -\mathbf{E}\left[-\frac{1}{\sigma^2}\right] = \frac{1}{\sigma^2}.$$

□

1.7 Sufficiency

Remark 1.11 *Use Likelihood Function to Define Fisher's Information*

- *We can define the* score function *as* $\dfrac{\partial \ln \mathbf{L}(Y_1,\ldots,y_n;\theta)}{\partial\theta} = 0 \implies$ *MLE.*
- $\mathbf{E}\left[\dfrac{\partial \ln \mathbf{L}(Y;\theta)}{\partial\theta}\right] = 0$
- $\mathbf{I}(\theta) = \mathbf{E}\left[\left(\dfrac{\partial \ln \mathbf{L}(Y;\theta)}{\partial\theta}\right)^2\right] = -\mathbf{E}_Y\left[\dfrac{\partial^2 \ln \mathbf{L}(Y;\theta)}{\partial\theta^2}\right]$
- $-\mathbf{E}_Y\left[\dfrac{\partial^2 \ln \mathbf{L}(Y_1,\ldots,Y_n;\theta)}{\partial\theta^2}\right] = n\mathbf{I}(\theta).$

Proof 1.

$$\begin{aligned}
-\mathbf{E}_Y\left[\frac{\partial^2 \ln \mathbf{L}(Y_1,\ldots,Y_n;\theta)}{\partial\theta^2}\right] &= -\mathbf{E}_Y\left[\frac{\partial^2}{\partial\theta^2} \ln \mathbf{L}(Y_1,\ldots,Y_m;\theta)\right] \\
&= -\mathbf{E}_Y\left[\frac{\partial^2}{\partial\theta^2} \ln\left(\prod_{i=1}^n f_Y(Y_i;\theta)\right)\right] \\
&= -\mathbf{E}_Y\left[\frac{\partial^2}{\partial\theta^2} \sum_{i=1}^n f_Y(y_i;\theta)\right] = \sum_{i=1}^n\left(-\mathbf{E}_Y\left[\frac{\partial^2}{\partial\theta^2} f_Y(y_i;\theta)\right]\right) = n\mathbf{I}(\theta)
\end{aligned}$$

■

- $\widehat{\theta_{MLE}} \xrightarrow{n\to\infty} N\left(\theta, \dfrac{1}{\mathbf{I}(\theta)}\right)$. *Note that* $\dfrac{1}{\mathbf{I}(\theta)}$ *is the C-R lower bound. We see that* $\widehat{\theta_{MLE}}$ *is asymptotically efficient.*

Remark 1.12 (Sufficiency Intuition) *Sufficiency tells us how much information can we get out of the data.*

Rationale *Let* $\widehat{\theta}$ *be an estimator to the unknown parameter* θ*. Does* $\widehat{\theta}$ *contain all information about* θ*?* e.g., The data itself is a sufficient estimator.

Definition 1.7.1 (Sufficiency). Let $(X_1, \ldots, X_n)$ be a random sample of size n from a continuous population with an unknown parameter θ. We call θ is *sufficient* if

$$f_{Y_1,\ldots,Y_n|\widehat{\theta}}\left(Y_1, \ldots, Y_n \mid \widehat{\theta} = \theta_e\right) = b(y_1, \ldots, y_n),$$

where $b(y_1, \ldots, y_n)$ is independent of θ ($\perp\!\!\!\perp \theta$). Also, $\widehat{\theta} = h(Y_1, \ldots, Y_n)$ and $\theta_e = h(y_1, \ldots, y_n)$. In this case, $\widehat{\theta}$ contains all the information about θ from $\{y_1, \ldots, y_n\}$.

Example 1.7.2

- Toss a coin 5 times and get 3 heads. Estimate $p =$ probability of H.

  Solution 2.

$$\mathbf{P}\left(HHHTT \mid p_e = \frac{3}{5}\right) = \frac{1}{\binom{3}{5}} \perp\!\!\!\perp p \implies \text{sufficient}$$

□

- A random sample of size n from Bernoulli(p). Check the sufficiency of $p = \sum_{i=1}^{n} X_i$.

  Solution 3.

  Suppose the random sample is $\{X_1, \ldots, X_n\}$. Then, consider

$$\mathbf{P}(X_1 = x_1, \ldots, X_n = x_n, \sum_{i=1}^{n} X_i = C \mid \widehat{p} = C) = \frac{\mathbf{P}(X_1 = x_1, \ldots, X_n = x_n, \sum_{i=1}^{n} X_i = C)}{\mathbf{P}(\widehat{p} = C)}.$$

  What new information can $\sum_{i=1}^{n} X_i = C$ tell us? $X_n = C - \sum_{i=1}^{n-1} X_i$.

Note that $\mathbf{P}(\widehat{p} = C) = \mathbf{P}\left(\sum_{i=1}^{n} X_i = C\right)$. Since the summation of Bernoulli(p) random variables is a Binomial(n, p) random variable, we have $\mathbf{P}(\widehat{p} = C) = \binom{n}{C} p^C (1-p)^{n-C}$.

Case I Suppose $\sum_{i=1}^{n} X_i = C$. Then,

$$\begin{aligned}
&\frac{\mathbf{P}(X_1 = x_1, \ldots, X_n = x_n, \sum_{i=1}^{n} X_i = C)}{\mathbf{P}(\widehat{p} = C)} \\
&= \frac{\left(\prod_{i=1}^{n-1}\right) p^{X_i}(1-p)^{1-X_i} p^{C - \sum_{i=1}^{n-1} X_i} (1-p)^{\left(1-C+\sum_{i=1}^{n-1} X_i\right)}}{\binom{n}{C} p^C (1-p)^{n-C}} \\
&= \frac{p^{\sum_{i=1}^{n-1} X_i + C - \sum_{i=1}^{n-1} X_i} (1-p)^{(n-1) - \sum_{i=1}^{n-1} X_i + 1 - C + \sum_{i=1}^{n-1} X_i}}{\binom{n}{C} p^C (1-p)^{n-C}} \\
&= \frac{p^C (1-p)^{n-C}}{\binom{n}{C} p^C (1-p)^{n-C}} = \frac{1}{\binom{n}{C}} \perp\!\!\!\perp p \implies \text{sufficient}
\end{aligned}$$

Case II Suppose $\sum_{i=1}^{n} X_i \neq C$. Then,

$$\frac{\mathbf{P}(X_1 = x_1, \ldots, X_n = x_n, \sum_{i=1}^{n} X_i = C)}{\mathbf{P}(\widehat{p} = C)} = \frac{0}{\mathbf{P}(\widehat{p} = C)} = 0 \perp\!\!\!\perp p \implies \text{sufficient}$$

□

Theorem 1.7.3 Factorization Property
$\widehat{\theta}$ is sufficient if and only if the likelihood can be factorized as

$$\mathbf{L}(\theta) = \underbrace{g(\theta_e;\theta)}_{\theta_e=h(y_1,\dots,y_n)\ \&\ \theta} \cdot \underbrace{u(y_1,\dots,y_n)}_{\perp\!\!\!\perp\theta}.$$

1.8 Consistency

Definition 1.8.1 (Consistency). An estimator $\widehat{\theta}_n = h(W_1,\dots,W_n)$ is said to be *consistent* if it converges to θ in probability; i.e., for any $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbf{P}\left(\left|\widehat{\theta}_n - \theta\right| < \varepsilon\right) = 1.$$

Remark 1.13 *1. Consistency is an asymptotical property (defined in a large sample limit).*

2. n= sample size. $\left|\widehat{\theta}_n - \theta\right|$ is the distance between estimator and true θ.

Lemma 1.2 Markov Inequality: Suppose $X \geq 0$ is a random variable and $a > 0$ is a constant. Then,

$$\mathbf{P}(X \geq a) \leq \frac{\mathbf{E}(X)}{a}.$$

Remark 1.14 *Markov inequality is good for determining extreme values. If $\mathbf{E}(X)$ is small, then it is very unlikely that X will take some extremely large numbers.*

Theorem 1.8.3 Chebyshev Inequality
Let W be some random variable with finite mean μ and variance σ^2. Then, for any $\varepsilon > 0$, we have

$$\mathbf{P}(|W-\mu| < \varepsilon) \leq 1 - \frac{\sigma^2}{\varepsilon^2}$$

or, equivalently,

$$\mathbf{P}(|W-\mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}.$$

Proof 1. Consider the random variable $|W-\mu|$. Then, by Markov Inequality,

$$\begin{aligned}\mathbf{P}(|X-\mu| \geq \varepsilon) &= \mathbf{P}\left(|X-\mu|^2 \geq \varepsilon^2\right) \\ &= \mathbf{P}\left((X-\mu)^2 \geq \varepsilon^2\right) \leq \frac{\mathbf{E}[(X-\mu)^2]}{\varepsilon^2} = \frac{\sigma^2}{\varepsilon^2}\end{aligned}$$

■

Corollary 1.4 : The sample mean $\widehat{\mu}_n = \dfrac{1}{n}\sum_{i=1}^{n} W_i$ is a consistent estimator for $\mathbf{E}(W) = \mu$, provided that the population W has finite mean μ and variance σ^2.

Proposition 1.5 : If $\widehat{\theta}_n$ is an unbiased estimator of θ, then $\widehat{\theta}_n$ is consistent if

$$\lim_{n\to\infty} \mathbf{Var}\left(\widehat{\theta}_n\right) = 0.$$

Proof 2. Suppose $\widehat{\theta}_n$ is an unbiased estimator of θ. Then, $\mathbf{E}\left(\widehat{\theta}_n\right) = \theta$. So, by Chebyshev Inequality, we have

$$\mathbf{P}\left(\left|\widehat{\theta}_n \theta\right| \geq \varepsilon\right) = \mathbf{P}\left(\left|\widehat{\theta}_n - \mathbf{E}\left(\widehat{\theta}_n\right)\right| \geq \varepsilon\right) \leq \frac{\mathbf{E}\left[\left(\widehat{\theta}_n - \mathbf{E}\left(\widehat{\theta}_n\right)\right)^2\right]}{\varepsilon^2} = \frac{\mathbf{Var}\left(\widehat{\theta}_n\right)}{\varepsilon^2}.$$

If we have $\mathbf{Var}\left(\widehat{\theta}_n\right) \to 0$ when $n \to \infty$, then

$$\lim_{n\to\infty} \mathbf{P}\left(\left|\widehat{\theta}_n - \theta\right| \geq \varepsilon\right) \leq \lim_{n\to\infty} \frac{\mathbf{Var}\left(\widehat{\theta}_n\right)}{\varepsilon^2} = \frac{0}{\varepsilon} = 0.$$

Therefore, it must be that $\lim_{n\to\infty} \mathbf{P}\left(\left|\widehat{\theta}_n - \theta\right| \geq \varepsilon\right) = 0$ as probability cannot take negative values. Hence,

$$\begin{aligned}
\lim_{n\to\infty} \mathbf{P}\left(\left|\widehat{\theta}_n - \theta\right| < \varepsilon\right) &= \lim_{n\to\infty} \left(1 - \mathbf{P}\left(\left|\widehat{\theta}_n - \theta\right| \geq \varepsilon\right)\right) \\
&= 1 - \lim_{n\to\infty} \mathbf{P}\left(\left|\widehat{\theta}_n - \theta\right| \geq \varepsilon\right) \\
&= 1 - 0 = 1.
\end{aligned}$$

Then, by definition, $\widehat{\theta}_n$ is consistent. ■

1.9 Bayesian Estimator

Theorem 1.9.1 Bayes' Rule

$$\mathbf{P}(A \mid B) = \frac{\mathbf{P}(B \mid A)\mathbf{P}(A)}{\mathbf{P}(B \mid A)\mathbf{P}(A) + \mathbf{P}(B \mid A^C)\mathbf{P}(A^C)}.$$

$$\mathbf{P}(A \mid B^C) = 1 - \mathbf{P}(A \mid B) = \frac{\mathbf{P}(B^C \mid A)\mathbf{P}(A)}{\mathbf{P}(B^C \mid A)\mathbf{P}(A) + \mathbf{P}(B^C \mid A^C)\mathbf{P}(A^C)}.$$

Rationale Let W be an estimator dependent on a parameter θ.

1. Frequentists view θ as a parameter whose exact value to be estimated (θ is fixed).

2. Bayesians view θ is the value of a random variable Θ. (θ *is uncertain and has its known parameter distribution*).

Data Generation The following procedure generates data with an additional layer of randomness.

1. θ is sampled from a distribution.
2. Under this θ, we sample the data.

Definition 1.9.2 (Prior distribution, Posterior distribution). Our prior knowledge on Θ is called the *prior distribution*: $p_\Theta(\theta)$. The conditional distribution of the data given the parameter is the *likelihood*: $p(X \mid \Theta)$. Then, the Bayes' Rule will be

$$\underbrace{\mathbf{P}(\Theta \mid X)}_{\text{posterior distribution given the observation}} = \frac{\overbrace{\mathbf{P}(X \mid \Theta)}^{\text{likelihood}} \cdot \overbrace{\mathbf{P}(\Theta)}^{\text{prior distribution}}}{\underbrace{\mathbf{P}(X)}_{\text{margin distirbution of data}}}$$

Theorem 1.9.3 Bayesian Estimator

$$g_\Theta(\theta \mid W = w) = \begin{cases} \dfrac{p_W(w \mid \Theta = \theta) p_\Theta(\theta)}{p_W(w)} & \text{if } W \text{ and } \Theta \text{ are discrete} \\ \\ \dfrac{f_W(w \mid \Theta = \theta) f_\Theta(\theta)}{f_W(w)} & \text{if } W \text{ and } \Theta \text{ are constinuous,} \end{cases}$$

where

$$\begin{aligned} f_W(x) &= \int_H f_{W,\Theta}(w, \theta) \, \mathrm{d}\theta \quad \text{for } \theta \in H \\ &= \int_H f_W(w \mid \Theta = \theta) f_\Theta(\theta) \, \mathrm{d}\theta. \end{aligned}$$

Further, let $A = f_W(w) = \int_H f_W(w \mid \Theta = \theta) f_\Theta(\theta) \, \mathrm{d}\theta$. Then, A normalizes $\text{likelihood} \times \text{prior}$:

$$1 = \int \frac{f_W(w \mid \Theta = \theta) f_\Theta(\theta)}{A} \, \mathrm{d}\theta.$$

So,

$$g_\Theta(\theta \mid W = w) = \text{constant} \cdot f_W(w \mid \Theta = \theta) f_\Theta(\theta) \quad \text{or} \quad \text{posterior} \propto \text{likelihood} \times \text{prior}.$$

Example 1.9.4 A call center. Let X =number of calls coming into the center. Then we know that $X \sim$ Poisson(λ). This particular call center believes that Λ is distributed with pdf

$$p_\Lambda(8) = 0.25 \quad \text{and} \quad p_\Lambda(10) = 0.75.$$

The call center believes that the number of calls coming into the center has recently changed, so they pick an hour and observe that $X = 7$ calls come in.

Solution 1.

We want to find: $\mathbf{P}(\Lambda = 8 \mid X = 7)$ and $\mathbf{P}(\Lambda = 10 \mid X = 7)$. By Bayes' Rule:

$$\begin{aligned}
\mathbf{P}(\Lambda = 8 \mid X = 7) &= \frac{\mathbf{P}(X = 7 \mid \Lambda = 8)\mathbf{P}(\Lambda = 8)}{\mathbf{P}(X = 7)} \\
&= \frac{\mathbf{P}(X = 7 \mid \Lambda = 8)\mathbf{P}(\Lambda = 8)}{\mathbf{P}(X = 7 \mid \Lambda = 8)\mathbf{P}(\Lambda = 8) + \mathbf{P}(X = 7 \mid \Lambda = 10)\mathbf{P}(\Lambda = 10)} \\
&= \frac{e^{-8}\left(\frac{8^7}{7!}\right)(0.25)}{e^{-8}\left(\frac{8^7}{7!}\right)(0.25) + e^{-10}\left(\frac{10^7}{7!}\right)(0.75)} \approx 0.66
\end{aligned}$$

Then, $\mathbf{P}(\Lambda = 10 \mid X = 7) = 1 - \mathbf{P}(\Lambda = 8 \mid X = 7) = 1 - 0.66 = 0.34$. Or, alternatively, we can use the Bayes' Rule again. $\square$

Table 1: Convention of Picking a Prior Distribution

Parameter	Prior Distribution
Bernoulli(p)	Beta
Binomial(p)	Beta
Poisson(λ)	Gamma
Exponential(λ)	Gamma
Normal(μ)	Normal
Normal(σ^2)	Inverse Gamma

Remark 1.15 *When we have no prior knowledge on the belief, we choose a uniform distribution.*

Example 1.9.5 Consider an unfair coin Θ (a random variable indicating the probability of getting head). Flip the coin n times, $X =$ number of heads. Find the posterior distribution.

Solution 2.

By the Bayes' rule,

$$f_{\Theta|X}(\theta \mid X = x) = \frac{f_\Theta(\theta)\mathbf{P}(X = k \mid \theta)}{\mathbf{P}(X = k)}.$$

We know $\theta \in [0, 1]$, so $\Theta \sim$ Uniform$[0, 1]$ and $f_\Theta(\theta) = 1$. So,

$$f_{\Theta|X}(\theta \mid X = x) = \frac{1 \cdot \binom{n}{k} \cdot \theta^k (1-\theta)^{n-k}}{\mathbf{P}(X = k)} = \underbrace{\frac{1 \cdot \binom{n}{k}}{\mathbf{P}(X = k)}}_{\text{constant}} \theta^k (1-\theta)^{n-k}$$

Definition 1.9.6 (Beta Distribution). For a distribution Beta(α, β), the pdf is given by

$$f_Y(y; \alpha, \beta) = \frac{y^{\alpha-1}(1-y)^{\beta-1}}{\mathbf{B}(\alpha, \beta)} \quad \text{for } y \in [0,1] \text{ and } \alpha, \beta > 0,$$

where

$$\mathbf{B}(\alpha, \beta) \coloneqq \int_0^1 y^{\alpha-1}(1-y)^{\beta-1}\,\mathrm{d}y = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}, \quad \alpha, \beta > 0.$$

The expectation of $X \sim$ Beta$(\alpha, beta)$ is given by

$$\mathbf{E}(X) = \frac{\alpha}{\alpha + \beta}.$$

Disregarding the constant, $\theta^k(1-\theta)^{n-k}$ is part of the Beta distribution with $\alpha = k+1$ and $\beta = n-k+1$. So, $\Theta \sim$ Beta$(k+1, n-k+1)$. To form a distribution, the constant must, therefore, be

$$\begin{aligned}
\frac{\binom{n}{k}}{\mathbf{P}(X = k)} = \frac{1}{\mathbf{B}(k+1, n-k+1)} &= \frac{\Gamma(k+1+n-k+1)}{\Gamma(k+1)\Gamma(n-k+1)} \\
&= \frac{\Gamma(n+2)}{\Gamma(k+1)\Gamma(n-k+1)} \\
&= \frac{(n+1)!}{k!(n-k)!} \qquad \qquad If\ n \in \mathbb{N},\ then\ \Gamma(n) = (n-1)!
\end{aligned}$$

Note that Beta$(\alpha = 1, \beta = 1) =$ Unform$(0, 1)$. So, in this example,

$$\text{Beta}(1,1) \xrightarrow{\text{Data}} \text{Beta}(k+1, n-k+1).$$

Moreover, $\mathbf{E}(\Theta) = \dfrac{k+1}{k+1+n-k+1} = \dfrac{k+1}{n+2}$. □

Example 1.9.7 Let $X_1, \ldots, X_n$ be a random sample form Bernoulli(θ): $p_X(k;\theta) = \theta^k(1-\theta)^{1-k}$ for $k = 0, 1$. Let $X = \sum_{i=1}^{n} X_i$. Then, X follows Binomial(n, θ). Consider the prior distribution $\Theta \sim$ Beta(r, s), i.e., $f_\Theta(\theta) = \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)}\theta^{r-1}(1-\theta)^{s-1}$ for $\theta \in [0, 1]$. Then, the posterior distribution is

$$\Theta \mid X \sim \text{Beta}(r+k, s+n-k).$$

Proof 3. Note that

$$\begin{aligned}
f_{\Theta|X}(\theta \mid X = x) &= \frac{p_X(X = k \mid \theta) f_\Theta(\theta)}{\int_0^1 p_X(X = k \mid \theta) f_\Theta(\theta)\,\mathrm{d}\theta} \\
&= \frac{\binom{n}{k}\theta^k(1-\theta)^{n-k}\frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)}\theta^{r-1}(1-\theta)^{s-1}}{\int_0^1 \binom{n}{k}\theta^k(1-\theta)^{n-k}\frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)}\theta^{r-1}(1-\theta)^{s-1}\,\mathrm{d}\theta} \\
&= \frac{\binom{n}{k}\frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)}\theta^{k+r-1}(1-\theta)^{n-k+s-1}}{\binom{n}{k}\frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)}\int_0^1 \theta^{k+r-1}(1-\theta)^{n-k+s-1}\,\mathrm{d}\theta}
\end{aligned}$$

Note that $\theta^{k+r-1}(1-\theta)^{n-k+s-1}$ is part of Beta$(k+r, n-k+s)$. So,

$$\begin{aligned}
1 &= \int_0^1 \frac{\Gamma(k+r+n-k+s)}{\Gamma(k+r)\Gamma(n-k+s)}\theta^{k+r-1}(1-\theta)^{n-k+s-1}\,\mathrm{d}\theta \\
1 &= \frac{\Gamma(r+n+s)}{\Gamma(k+r)\Gamma(n-k+s)}\int_0^1 \theta^{k+r-1}(1-\theta)^{n-k+s-1}\,\mathrm{d}\theta \\
\int_0^1 \theta^{k+r-1}(1-\theta)^{n-k+s-1}\,\mathrm{d}\theta &= \frac{\Gamma(k+r)\Gamma(n-k+s)}{\Gamma(r+n+s)}.
\end{aligned}$$

Therefore,

$$f_{\Theta|X}(\theta \mid X = x) = \frac{\theta^{k+r-1}(1-\theta)^{n-k+s-1}}{\frac{\Gamma(k+r)\Gamma(n-k+s)}{\Gamma(r+n+s)}} = \frac{\Gamma(r+n+s)}{\Gamma(k+r)\Gamma(n-k+s)}\theta^{k+r-1}(1-\theta)^{n-k+s-1}.$$

This is exactly a Beta distribution with parameter $\alpha = k + r$ and $\beta = n - k + s$. ■

Definition 1.9.8 (Conjugate Prior). If the posterior distributions $p(\Theta \mid X)$ are in the sample probability distribution family as the prior probability distribution $p(\Theta)$, the prior and posterior are called *conjugate distributions* and the prior is called a *conjugate prior* for the likelihood function.

Remark 1.16 *Common Conjugate Priors*

- *Beta distributions are conjugate priors for Bernoulli, Binomial, Negative binomial, and Geometric likelihood.*
- *Gamma distributions are conjugate priors for Poisson and Exponential likelihood*

Definition 1.9.9 (Bayesian Point Estimation). Given the posterior $f_{\Theta|W}(\theta \mid W = w)$, how can one calculate the appropriate point estimate θ_e?

Definition 1.9.10 (Loss Function). Let θ_e be an estimate for θ based on a statistic W. The *loss function* associated with θ_e is denoted $\mathbf{L}(\theta_e, \theta)$, where $\mathbf{L}(\theta_e, \theta) \geq 0$ and $\mathbf{L}(\theta, \theta) = 0$.

- The lost function is $\mathbf{E}\left[\mathbf{L}(\widehat{\theta}, \theta)\right]$.
- The MSE, mean square error, is $\mathbf{E}\left[\left(\widehat{\theta} - \theta\right)^2\right]$.

  1. If we have not data, then notice that

  $$\mathbf{E}\left[(\theta - c)^2\right] = \mathbf{E}\left(\theta^2\right) + \mathbf{E}\left(c^2\right) - 2c\mathbf{E}(\theta)$$

  is minimized at $c = \mathbf{E}(\theta)$. Therefore,

  $$\min \mathbf{E}\left[(\theta - \widehat{\theta})^2\right] = \mathbf{E}[(\theta - \mathbf{E}(\theta))]^2 = \mathbf{Var}(\theta).$$

  So, $\widehat{\theta}^* = \mathbf{E}(\theta)$, the prior expectation.

  2. If we have data $X = x$, then

  $$\min \mathbf{E}\left[(\theta - \widehat{\theta})^2 \mid X = x\right] \implies \widehat{\theta}^* = \mathbf{E}[\theta \mid X = x].$$

  This $\widehat{\theta}^*$ is called the posterior expectation.

Theorem 1.9.11 Squared-Loss Bayesian Estimation

Step 1. Solve the posterior distribution.

Step 2. Calculate the posterior expectation.

Generally, if we know the posterior pdf $f_\Theta(\theta \mid X = x)$, the point estimate is

$$\mathbf{E}[\theta \mid X = x] = \int_\Theta \theta f_\Theta(\theta \mid X = x)\,\mathrm{d}\theta.$$

Theorem 1.9.12
Let $f_\Theta(\theta \mid W = w)$ be the posterior distribution of the random variable Θ.

- If $\mathbf{L}(\theta_e, \theta) = |\theta_e - \theta|$, then the Bayesian point estimate for θ is the median of the posterior distribution $f_\Theta(\theta \mid W = w)$;
- If $\mathbf{L}(\theta_e, \theta) = (\theta_e - \theta)^2$, then the Bayesian point estimate for θ is the mean of the posterior distribution $f_\Theta(\theta \mid W = w)$.

2 Inference Based on Normal

2.1 Sample Variance and Chi-Square Distribution

Recall that if $Y \sim \text{Normal}(\mu, \sigma^2)$, we have MLEs defined as

$$\widehat{\mu} = \overline{Y} \quad \text{and} \quad \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2.$$

If σ is known, we can do the interval estimation:

$$Z := \frac{\overline{Y} - \mathbf{E}(\overline{Y})}{\sqrt{\mathbf{Var}(\overline{Y})}} \sim N(0,1).$$

However, what if we don't know σ? We will have to estimate it with a sample variance.
Definition 2.1.1 (Sample Variance). To estimate σ^2, we define the following unbiased *sample variance*:

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

Remark 2.1 *We often compute S^2 using the fact that*

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n} y_i^2 - n\overline{y}^2 \quad \textit{i.e.}, S^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - n\overline{y}^2\right]$$

Definition 2.1.2 (Chi-Squared Distribution). Suppose $W_k \sim \chi^2(k)$, the *chi-squared distribution with degree of freedom* k. Then,

$$W_k = Z_1^2 + Z_2^2 + \cdots + Z_k^2, \text{ where } Z_i \overset{i.i.d.}{\sim} N(0,1).$$

k is called the *degree of freedom* of the chi-squared distribution and is denoted as $df = k$.

Theorem 2.1.3 Chi-Squared Distribution and Gamma Distribution
$\chi^2(1)$ is equivalent to Gamma$\left(\frac{1}{2}, \frac{1}{2}\right)$. Hence, $\chi^2(n)$ is equivalent to Gamma$\left(\frac{n}{2}, \frac{1}{2}\right)$.

Proof 1. Recall: For $Y_1 \sim \text{Gamma}(n, \lambda)$ and $Y_2 \sim \text{Gamma}(m, \lambda)$, we have the following sum rule

$$Y_1 + Y_2 \sim \text{Gamma}(n + m, \lambda).$$

Then, as $Z_1^2 \sim \chi^2(1) = \text{Gamma}\left(\frac{1}{2}, \frac{1}{2}\right)$, we have

$$Z_1^2 + Z_2^2 + \cdots + Z_n^2 \sim \chi^2(n) = \text{Gamma}\left(\frac{1}{2} + \cdots + \frac{1}{2}, \frac{1}{2}\right) = \text{Gamma}\left(\frac{n}{2}, \frac{1}{2}\right).$$

■

Theorem 2.1.4 Expectation and Variance of $\chi^2(n)$
If $W_n \sim \chi^2(n)$, then

$$\mathbf{E}(W_n) = n = df \quad \text{and} \quad \mathbf{Var}(W_n) = 2n$$

Proof 2. For $Y \sim \text{Gamma}(n, \lambda)$, $\mathbf{E}(Y) = \frac{n}{\lambda}$ and $\mathbf{Var}(Y) = \frac{n}{\lambda^2}$. As $W_n \sim \chi^2(n) = \text{Gamma}\left(\frac{n}{2}, \frac{1}{2}\right)$, we have

$$\mathbf{E}(W_n) = \frac{n/2}{1/2} = n \quad \text{and} \quad \mathbf{Var}(W_n) = \frac{n/2}{1/4} = 2n.$$

■

Theorem 2.1.5
Consider a random sample $Y_1, \ldots, Y_n$ drawn from $N(0, 1)$. Let S^2 be the sample variance and $\overline{Y}$ be the sample mean. Then,

- S^2 and $\overline{Y}$ are independent;
- $\frac{(n-1)}{\sigma^2} S^2 \sim \chi^2(n-1)$

Remark 2.2 *We can think of the second bullet point as the following rationale: knowing $\overline{Y}$, we only need $(n-1)$ data, and we can calculate Y_n from $\overline{Y}$ and $Y_1, \ldots, Y_{n-1}$. This explains why the chi-squared distribution is of $df = n - 1$.*

Proof 3.(informally)

1. We will prove the case when $n = 2$.

$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2$. If $n = 2$, $\overline{Y} = \frac{Y_1 + Y_2}{2}$, then

$$\begin{aligned} S^2 &= \left(Y_1 - \overline{Y}\right)^2 + \left(Y_2 - \overline{Y}\right)^2 \\ &= \left(Y_1 - \frac{Y_1 + Y_2}{2}\right)^2 + \left(Y_2 - \frac{Y_1 + Y_2}{2}\right) \\ &= \left(\frac{Y_1 - Y_2}{2}\right)^2 + \left(\frac{Y_2 - Y_1}{2}\right)^2 \\ &= \frac{1}{2}(Y_1 - Y_2)^2. \end{aligned}$$

Claim. Recall that if X_1 and X_2 are independent, then

$$\mathbf{E}(X_1 X_2) = \mathbf{E}(X_1)\mathbf{E}(X_2). \tag{1}$$

The backward implication is not true in general, but specially for normal distributions. That is, if (1) holds and $X_1,\ X_2$ normal are normal, then $X_1 \perp\!\!\!\perp X_2$.

As $Y_1 - Y_2$ and $Y_1 + Y_2$ are both normal distributed, to show they are independent of each other, we only need to show that

$$\mathbf{E}[(Y_1 - Y_2)(Y_1 + Y_2)] = \mathbf{E}(Y_1 - Y_2)\mathbf{E}(Y_1 + Y_2).$$

The detailed proof is omitted, but the equality holds.

2. Show that $\frac{(n-1)}{\sigma^2}S^2 \sim \chi^2_{n-1}$. Note that $Y_i \sim N(\mu, \sigma)$. Then,

$$\frac{Y_i - \mu}{\sigma} \sim N(0,1) \quad \text{and} \quad \frac{\overline{Y} - \mu}{\sqrt{\sigma^2/n}} \sim N(0,1).$$

So,

$$\frac{(Y_i - \mu)^2}{\sigma^2} \sim \chi^2_1 \implies \frac{\sum_{i=1}^{n}(Y_i - \mu)^2}{\sigma^2} \sim \chi^2_n \quad \text{and} \quad \frac{(\overline{Y} - \mu)^2}{\sigma^2/n} \sim \chi^2_1.$$

Claim. If $U_1 \sim \chi^2(m)$ and $U_2 \sim \chi^2(n)$ with $U_1 \perp\!\!\!\perp U_2$, then $U_1 + U_2 \sim \chi^2(m+n)$ by the summation rule of Gamma.

Therefore, by the Claim, we have

$$\begin{aligned}\frac{\sum_{i=1}^{n}(Y_i-\mu)^2}{\sigma^2} &= \frac{\sum_{i=1}^{n}\left(Y_i-\overline{Y}+\overline{Y}-\mu\right)^2}{\sigma^2}\\ &\sim \frac{\sum_{i=1}^{n}(Y_i-\overline{Y})^2+\sum_{u=1}^{n}(\overline{Y}-\mu)^2}{\sigma^2}\\ &= \frac{(n-1)S^2}{\sigma^2}+\frac{\sum_{i=1}^{n}(\overline{Y}-\mu)^2}{\sigma^2}.\end{aligned}$$

Note that $\frac{\sum_{i=1}^{n}(Y_i-\mu)^2}{\sigma^2}\sim\chi_n^2$ and $\frac{\sum_{i=1}^{n}(\overline{Y}-\mu)^2}{\sigma^2}\sim\chi_1^2$. So, it must be that $\frac{(n-1)S^2}{\sigma^2}\sim\chi_{m-1}^2$.

■

2.2 Inference on μ and σ

Definition 2.2.1 (Sampling Distribution). The *sampling distributions* are defined as the distributions of functions of random sample of given size.

Aim: Determine distributions for the following statistics:

Statistics	Distribution
(Sample Variance) $S^2 := \frac{1}{n-1}\sum_{n=1}^{n}(Y_1-\overline{Y})^2$	Chi-square distribution
$T := \frac{\overline{Y}-\mu}{S/\sqrt{n}}$	Student t distribution
$\frac{S_1^2}{\sigma_1^2}\Big/\frac{S_2^2}{\sigma_2^2}$	F distribution

Definition 2.2.2 (The Test Statistic). The *test statistic* is defined as

$$T := \frac{\overline{Y}-\mu}{S/\sqrt{n}},$$

with $\overline{Y}=\frac{1}{n}\sum_{i=1}^{n}Y_i$ and $S^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2$.

Definition 2.2.3 (Student t-Ratio). Consider

- $Z := \dfrac{\sqrt{\mu}}{\sigma}(\overline{Y} - \mu) \sim N(0,1)$
- $V \sim \chi^2_n$
- $Z \perp\!\!\!\perp V$

Then, we define the *student t-ratio* with n degrees of freedom as

$$T_n := \frac{Z}{\sqrt{V/n}}.$$

Note that $Z \sim N(0,1)$ and $\sqrt{V/n} \sim \sqrt{\dfrac{\chi^2_n}{n}}$.

Theorem 2.2.4 Distribution of $\dfrac{\overline{Y} - \mu}{S/\sqrt{n}}$
Consider $Y_1, \ldots, Y_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$. Let S^2 to be the sample variance. Then,

$$\frac{\overline{Y} - \mu}{S/\sqrt{n}} \sim T_{n-1}.$$

Proof 1. Note that

$$\frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \tag{2}$$

and

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1} \tag{3}$$

Then, consider

$$\begin{aligned}
\frac{\overline{Y} - \mu}{S/\sqrt{n}} = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \cdot \frac{\sigma}{S} &= \frac{\dfrac{\overline{Y} - \mu}{\sigma/\sqrt{n}}}{\sqrt{\dfrac{S^2}{\sigma^2}}} \\
&= \frac{\dfrac{\overline{Y} - \mu}{\sigma/\sqrt{n}}}{\sqrt{\dfrac{(n-1)S^2}{\sigma^2} \cdot \dfrac{1}{\sqrt{n-1}}}} \\
&= \frac{\dfrac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)}{\sqrt{\dfrac{(n-1)S^2/\sigma^2}{n-1}} \sim \chi^2_{n-1}} \qquad S^2 \perp\!\!\!\perp \overline{Y} \\
&\sim T_{n-1}.
\end{aligned}$$

■

Theorem 2.2.5 Connection Between $N(0,1)$ and t
T distribution is flatter/more spread out than $N(0,1)$. It has heavier tails.

Proof 2. Note that

- $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$ is an unbiased estimator of σ^2.
- S_n^2 is a consistent estimator of σ^2.

So, $\mathbf{Var}(S_n^2) \to 0$ as $n \to \infty$. This implies that the difference between T and $N(0,1)$ is significant when n is small. ■

Theorem 2.2.6 Inference on μ
If σ^2 is known, we inference μ using $Z = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}}$. We use z-score and z_α table to construct the $100(1-\alpha)\%$ CI as $\left(\overline{y} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{y} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$. Alternatively, if σ^2 is unknown, we use $T_{n-1} = \frac{\overline{Y} - \mu}{S/\sqrt{n}}$. We apply t_{n-1} score and $t_{\alpha,n-1}$ table to construct a similar CI.

Theorem 2.2.7 Inference on σ
A two-sided $100(1-\alpha)\%$ CI on σ will be given by

$$\left(\sqrt{\frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}}, \sqrt{\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}}\right).$$

Proof 3. Note that

$$X_n := \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Then,

$$\mathbf{P}(x_a \leq X_n \leq x_b) = 100(1-\alpha)\%.$$

To construct a two-sided CI, since chi-square distribution is not symmetric, we can choose the two points that have the same density value (this will ensure a short CI). However, this method is very numerically expensive. To save computational cost, we will still choose the two points that covers the $\alpha/2\%$ and $(1-\alpha/2)\%$ distribution. It is also known as to find $\chi^2_{\alpha/2,n-1}$ from the

χ^2 table. Hence,

$$\mathbf{P}(\chi^2_{\alpha/2,n-1} \leq X_n \leq \chi^2_{1-\alpha/2,n-1}) = 100(1-\alpha)\%$$
$$\mathbf{P}(\chi^2_{\alpha/2,n-1} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{1-\alpha/2,n-1}) = 100(1-\alpha)\%$$
$$\implies \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}$$

So, $100(1-\alpha)\%$ CI of σ^2 is

$$\left(\frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}, \frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}\right)$$

and a $100(1-\alpha)\%$ CI of σ is

$$\left(\sqrt{\frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}}, \sqrt{\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}}\right).$$

■

3 Hypothesis Testing

3.1 Decision Rules

Definition 3.1.1 (Hypotheses). We define

- H_0: null hypothesis (status quo), and
- H_1: alternative hypothesis

where H_0 and H_1 are such that

- H_0 and H_1 are disjoint sets
- H_0 always include an equal sign.

Example 3.1.2 We have a breath analyzer for DUI test. Do we need to calibrate the breath analyzer? Collect 30 samples whose alcohol level are known as 12.6. Measurements from the breath analyzer is $Y_1, \dots, Y_{30} \overset{i.i.d.}{\sim} N(\mu, \sigma = 0.4)$. Set up a hypothesis testing.

Solution 1.

If $\mu_0 = 12.6$m then the breath analyzer is accurate. So,

- H_0: $\mu_0 = 12.6$, and
- H_1: $\mu_0 \neq 12.6$.

Assume H_0 is true, then $\mathbf{P}\left(\left|\frac{\overline{Y} - \mu_0}{\sigma/\sqrt{n}}\right| > m \,\middle|\, H_0\right)$ should be small. i.e.,

$$\mathbf{P}\left(\left|\frac{\overline{Y} - \mu_0}{\sigma/\sqrt{n}}\right| > m \,\middle|\, H_0 : \mu_0 = 12.6\right) \leq \alpha \overset{\text{set}}{=} 0.05,$$

where α is called the *significance level.* Then,

$$\mathbf{P}\left(\left|\frac{\overline{Y} - 12.6}{0.4/\sqrt{30}}\right| > 1.96 \,\middle|\, H_0 : \overline{Y} \sim N(12.6, 0.4)\right) = 0.05, \; (m = z_{\alpha/2} = 1.96)$$

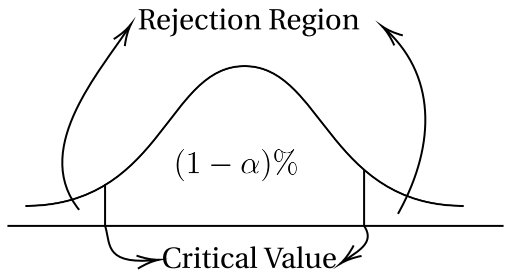
Simply, we get

$$\mathbf{P}\left(\left|\overline{Y} - 12.6\right| \geq 0.14\right) = 0.05.$$

So, if $\overline{Y} \geq 12.74$ or $\overline{Y} \leq 12.46$, we will *reject* H_0. If $\overline{Y} \in [12.46, 12.74]$, we will *tail to reject* H_0 (or, data is not sufficient to reject H_0). □

Definition 3.1.3 (Test Statistic). Any function of the observed data whose numerical value dictates whether H_0 is accepted or rejected.

Definition 3.1.4 (Critical Region/Rejection Region/C). The set of values for the test statistic that result in the null hypothesis being rejected.



Definition 3.1.5 (Critical Value). The particular point in C that separates the rejection region from the acceptance region.

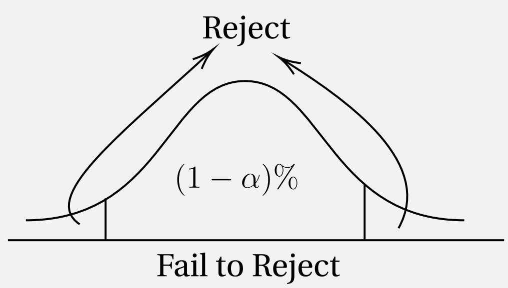
Definition 3.1.6 (Level of Significance/α). The probability that the test statistic lies in the critical region C under H_0.

Theorem 3.1.7 Procedure of Hypothesis Testing

1. Set up the hypotheses: H_0, H_1
2. Find the test statistics:

$$X_i \overset{i.i.d.}{\sim} N(\mu, \sigma^2), \quad \text{where } \sigma^2 \text{ is known} \implies Z := \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}.$$

3. Determine the decision rule



Remark 3.1 (Different Alternative Hypotheses) *For the same $H_0 : \theta = \theta_0$, we have have multiple different alternative hypotheses:*

$$H_1 : \begin{cases} \theta \neq \theta_0 \implies \mathbf{P}(z \neq [c_1, c_2] \mid H_0) = \alpha & z \in [c_1, c_2] \\ \theta < \theta_0 \implies \mathbf{P}(z < c \mid H_0) = \alpha & z \geq c \\ \theta > \theta_0 \implies \mathbf{P}(z > c \mid H_0) = \alpha & z \leq c \end{cases}$$

Definition 3.1.8 (Simple/Composite Hypothesis). *Simple hypothesis* is any hypothesis which specifies the population distribution completely. *Composite hypothesis* is any hypothesis which does not specify the population distribution completely.

Example 3.1.9 Suppose $H_0 : \mu = 120$ and $H_1 : \mu > 120$. Let $Y_1, \ldots, Y_n$ be samples. Suppose $\sigma = 12, n = 50, \overline{y} = 125.2$. Set up a hypothesis testing at significance level of $\alpha = 0.05$.

Solution 2.

Define test statistics: $\dfrac{\overline{Y} - \mu}{\sigma/\sqrt{n}}$. Then, under H_0,

$$Z := \frac{\overline{Y} - 120}{12/\sqrt{50}} \sim N(0, 1).$$

So,

$$z_{\text{obs}} = \frac{\overline{y} - 120}{12/\sqrt{50}} = \frac{125.2 - 120}{12/\sqrt{50}} = 3.06.$$

The *p-value* is given by

$$\mathbf{P}(Z \geq 3.06) \approx 0.001 < \alpha \implies \text{reject } H_0$$

□

Theorem 3.1.10 Summary: Testing

	Proportion	Mean		Variance
		σ^2 known	σ^2 unknown	
Distribution	Binomial(n, p)	Normal(μ, σ^2)	$N(\mu, \sigma^2)$	χ^2_n
Test Statistics	$\dfrac{k - np}{\sqrt{np(1-p)}}$	$Z := \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$	$T_{n-1} := \dfrac{\overline{X} - \mu}{S/\sqrt{n}}$	$\dfrac{(n-1)S^2}{\sigma^2}$

3.2 Types of Errors

Definition 3.2.1 (Type I and Type II Errors). *Type I Error* is $\mathbf{P}(\text{reject } H_0 \mid H_0 \text{ is true}) = \alpha$. *Type II Error* is given by $\mathbf{P}(\text{fail to reject } H_0 \mid H_1 \text{ is true}) = \beta$.

Remark 3.2 (Possible Situations of Type I Error) *If $\overline{Y}_{obs}$ falls into the rejection region:*

- $\overline{Y}$ *does not follow the distribution in* H_0
- $\overline{Y}$ *happens to take the extreme/unlikely values even when* $\overline{Y}$ *follows the distribution in* H_0.

Decision/Truth	H_0	H_1
H_0	*No Errors*	β, *Type II Error*
H_1	α, *Type I Error*	*No Errors*

Example 3.2.2 Example 3.1.2 Revisit. Calculate β, probability of type II error occurs. Recall that $Y \sim N(\mu, \sigma^2 = 0.16), H_0 : \mu = 12.6, H_1 : \mu = \mu_1$.

Solution 1.

$$
\begin{aligned}
\mathbf{P}(\text{Type II Error}) &= \mathbf{P}(\text{fail to reject } H_0 \mid H_1 \text{ is true}) \\
&= \mathbf{P}\left(\left|\frac{\overline{Y} - \mu_0}{\sigma/\sqrt{n}}\right| < z_{\alpha/2} \mid \mu_1\right) \\
&= \mathbf{P}\left(\mu_0 - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \overline{Y} \leq \mu_0 + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \mid \mu_1\right) \\
&= \mathbf{P}\left(\mu_0 - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \overline{Y} \leq \mu_0 + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \mid \overline{Y} \sim N\left(\mu_1, \frac{\sigma^2}{n}\right)\right) \\
&= \mathbf{P}\left(-z_{\alpha/2} + \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} \leq \frac{\overline{Y} - \mu_1}{\sigma/\sqrt{n}} \leq z_{\alpha/2} + \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} \mid H_1\right) \\
\beta(\mu_1) &\coloneqq \Phi\left(z_{\alpha/2} + \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}}\right) - \Phi\left(-z_{\alpha/2} + \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}}\right),
\end{aligned}
$$

where $\Phi(z) \coloneqq \mathbf{P}(Z \leq z)$. □

Definition 3.2.3 (p-Hacking). Post-hoc adjustment on data or on the tests. For example, after seeing the data,

- Adjust the side of the a one-sided test,
- Collect more data until H_0 is rejected. *When sample size n increases, CI becomes narrower, and thus the rejection region is wider.*
- Adjust the significance level.

3.3 Two-Sample Inferences

Theorem 3.3.1 Test Statistics
Suppose $X_1, \ldots, X_n \sim N(\mu_X, \sigma_X^2)$ and $Y_1, \ldots, Y_m \sim N(\mu_Y, \sigma_Y^2)$. Assume that $\sigma_X = \sigma_Y = \sigma$ and $\overline{X}$ and $\overline{Y}$ be the sample mean, respectively. Then

$$\frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sqrt{\left(\frac{1}{n} + \frac{1}{m}\right)\left(\sum_{i=1}^{n} (X_i - \overline{X})^2 + \sum_{i=1}^{m} (Y_i - \overline{Y})^2\right) \Big/ (m+n-2)}} \sim T_{m+n-2}$$

Proof 1. Note that, by CLT,

$$Z := \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sqrt{\sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)}} \sim N(0,1).$$

Further, since

$$\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2} \sim \chi^2_{n-1} \quad \text{and} \quad \frac{\sum_{i=1}^{m}(Y_i - \overline{Y})^2}{\sigma^2} \sim \chi^2_{m-1},$$

we know

$$V := \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2 + \sum_{i=1}^{m}(Y_i - \overline{Y})^2}{\sigma^2} \sim \chi^2_{m+n-2}$$

Therefore,

$$T = \frac{Z}{\sqrt{V/(m+n-2)}}$$

■

Theorem 3.3.2 Hypothesis Testing
Suppose $H_0 : \mu_X = \mu_Y$ and $H_1 : \mu_X \neq \mu_Y$. Under H_0 $(\mu_X = \mu_Y)$, the test statistics

$$t = \frac{\overline{X} - \overline{Y}}{\sqrt{\left(\frac{1}{n} + \frac{1}{m}\right) S_p^2}} \sim T_{m+n-2},$$

where

$$S_p^2 := \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2 + \sum_{i=1}^{m}(Y_i - \overline{Y})^2}{m + n - 2}.$$

Example 3.3.3 Let $X = \#$ of successes in n-trials and $Y = \#$ of successes in m trials, then $X \sim$ Binomial(n, p_X) and $Y \sim$ Binomial(m, p_Y). We want to test $H_0 : p_X = p_y$ versus a valid H_1 such as $p_X \neq p_Y$. If n and m are large enough, then by the CLT, we have

$$\frac{X}{n} \sim N(p_X, p_x(1 - p_X)/n) \quad \text{and} \quad \frac{Y}{m} \sim N(p_Y, p_Y(1 - p_Y)/m).$$

This is not quite the two sample t-test because there are only two parameter. But, under $H_0 : p_x = p_y = p$:

$$\frac{X}{n} - \frac{Y}{m} \sim N\left(0, \left(\frac{1}{n} + \frac{1}{m}\right)p(1 - p).\right)$$

So, define

$$Z := \frac{\frac{X}{n} - \frac{Y}{m}}{\sqrt{\left(\frac{1}{n} + \frac{1}{m}\right)p(1 - p)}} \sim N(0, 1).$$

We can then estimate p with $\widehat{p} = \frac{X + Y}{n + m}$, which gives us

$$Z := \frac{\frac{X}{n} - \frac{Y}{m}}{\sqrt{\left(\frac{1}{n} + \frac{1}{m}\right)\widehat{p}(1 - \widehat{p})}} \sim N(0, 1) \quad \text{as } n, m \to \infty$$

4 Regression Analysis

4.1 Introduction to Regression

Theorem 4.1.1 How to fit a Regression Model

1. Plot data (x_i, y_i)
2. Find a line $y = ax + b$
3. Draw inference on a, b, and $y \mid x$.

To find a best fit line, let's minimize the discripency:

$$\min_{a,b} y_i - (ax_i + b).$$

Mean Squared Error Note that we want to solve

$$\min_{a,b} \mathbf{E}\big[(Y_i - aX_i - b)^2\big].$$

The solutions are given by

$$a^* = \arg\min_a \mathbf{E}\big[(Y_i - aX_i - b)^2\big] \quad \text{and} \quad b^* = \arg\min_b \mathbf{E}\big[(Y_i - aX_i - b)^2\big].$$

A Probability View Since (x, y) is a pair of random variables, we simply the situation by placing all of the uncertainty on the y_i's and assume that the x_i's are controlled by the experimenter. Recall that for any two random variables X and Y, the conditional expectation of Y on X, namely

$$f(x) = \mathbf{E}[Y \mid X = x]$$

minimizes the mean squared error

$$\mathbf{E}\big[(Y - f(X))^2\big]$$

- Difficulties: The regression curve $Y = \mathbf{E}[Y \mid x]$ is complicated and hard to obtain.
- Compromise: Assume that $f(x) = a + bx$ (i.e., the first order approximation).

A Statistics View Let $\mathbf{L}(a, b) = \sum_{i=1}^{n}(y_i - (ax_i + b))^2 = \sum_{i=1}^{n}(y_i - ax_i - b)^2$. For the *least square method*, we choose a and b so that we minimize $\mathbf{L}$. That is, $\dfrac{\partial \mathbf{L}}{\partial b} = \dfrac{\partial \mathbf{L}}{\partial a} = 0$. By solving, we find that

$$a^* = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} \quad \text{and} \quad b = \overline{y} - a\overline{x}.$$

Theorem 4.1.2 Rationale
The following statements are equivalent:

- $\mathbf{Cov}(X, Y) \neq 0$.
- $\exists\, b_0, b_1 \in \mathbb{R}\ s.t.$
$$\mathbf{E}\big[(Y - b_0 - b_1 X)^2\big] < \mathbf{E}\big[(Y - \mu_Y)^2\big]$$
- $\exists\, b_1 \in \mathbb{R}\ s.t.\ \mathbf{Var}(Y - b_1 X) < \mathbf{Var}(Y)$.

Definition 4.1.3 (Regression). Suppose (X_i, Y_i) is a pair of random variable. The regression is defined as

$$Y_i = f(X_i) + \varepsilon_i, \quad \text{where } \mathbf{E}(\varepsilon_i) = 0.$$

Remark 4.1 *If* $\mathbf{E}(\varepsilon_i) \neq 0$ *and suppose* $\mathbf{E}(\varepsilon_i) = a$. *Then, we can fit* $Y_i = f(X_i) + a + \mathbf{E}(\varepsilon_i')$, *where* $\mathbf{E}(\varepsilon_i') = 0$.

Suppose $a = f(X_i)$, then we want to solve

$$\min_a \mathbf{E}\big[(Y_i - a)^2 \mid X_i = x\big].$$

Define $h(a) = \mathbf{E}[(Y_i - a)^2 \mid X_i = x] = \mathbf{E}[Y_i^2 \mid X_i = x] - 2a\mathbf{E}[Y_i \mid X_i = x] + a^2$, a quadratic function. Then, by the first-order optimality condition, we set

$$\frac{\partial h}{\partial a} = -2\mathbf{E}[Y_i \mid X_1 = x] + 2a = 0,$$

and so

$$a^* = \arg\min_a h(a) = \mathbf{E}[Y_i \mid X_i = x].$$

4.2 Linear Regression Model

Definition 4.2.1 (Linear Regression/Simple Linear Model). Assume $\mathbf{E}[Y_i \mid X_i = x] = \beta_0 + \beta_1 X_i$, so we have

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

which is the *linear regression model* (a.k.a *simple linear model*).

How to Find the Best β_0 and β_1? Let's consider the mean squared error.

$$\mathbf{MSE}(\beta_0, \beta_1) = \mathbf{E}\big[(Y - \beta_0 - \beta_1 X)^2\big].$$

Then, by the first-order optimality condition, we have

$$\frac{\partial \mathbf{MSE}}{\partial \beta_0} = -2\mathbf{E}[Y - \beta_0 - \beta_1 X] \overset{\text{set}}{=} 0$$

$$\beta_0^* = \mathbf{E}[Y] - \beta_1 \mathbf{E}[X] =: \mu_Y - \beta_1 \mu_X$$

Meanwhile, we have

$$\begin{aligned}
\frac{\partial \mathbf{MSE}}{\partial \beta_1} &= -2\mathbf{E}[X(Y - \beta_0 - \beta_1 X)] \\
&= -2\mathbf{E}[XY] + 2\beta_0 \mathbf{E}[X] + 2\beta_1 \mathbf{E}[X^2] \\
&= -2\mathbf{E}[XY] + 2(\mathbf{E}[Y] - \beta_1 \mathbf{E}[X])\mathbf{E}[X] + 2\beta_1 \mathbf{E}[X^2] \\
&= -2\mathbf{E}[XY] + 2\mathbf{E}[X]\mathbf{E}[Y] - 2\beta_1 \mathbf{E}[X]\mathbf{E}[X] + 2\beta_1 \mathbf{E}[X^2] \\
&= -2\left[\underbrace{(\mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y])}_{\mathbf{Cov}(X,Y)} + \beta_1 \underbrace{\left(\mathbf{E}[X]^2 - \mathbf{E}[X^2]\right)}_{\mathbf{Var}(X,Y)}\right] \\
&= -2(\mathbf{Cov}(X, Y) + \beta \mathbf{Var}(X)) \\
&\overset{\text{set}}{=} 0 \\
\beta_1^* &= \frac{\mathbf{Cov}(X, Y)}{\mathbf{Var}(X)}.
\end{aligned}$$

How to represent the best β_0 and β_1 in terms of X_i and Y_i? Suppose we have data

$$(X_1, Y_1), \ldots, (X_n, Y_n).$$

Then, theoretically,

$$\beta_0^* = \mu_Y - \beta_1^* \mu_X \quad \text{and} \quad \beta_1^* = \frac{\mathbf{Cov}(X, Y)}{\mathbf{Var}(X)}.$$

To have sample estimate of them, let's define

$$\overline{x} = \frac{1}{n}\sum_{i=1}^n x_i; \quad \overline{Y} = \frac{1}{n}\sum_{i=1}^n Y_i; \quad S_{XX} = \sum_{i=1}^n (x_i - \overline{x})^2 = \sum_{i=1}^n x_i^2 - n\overline{x}^2$$

$$S_{YY} = \sum_{i=1}^n (Y_i - \overline{Y})^2 = \sum_{i=1}^n Y_i^2 - n\overline{Y}^2; \quad S_{XY} = \sum_{i=1}^n (x_i - \overline{x})(Y_i - \overline{Y}) = \sum_{i=1}^n x_i Y_i - n\overline{x}\overline{Y}$$

Then, the estimate

$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{x} \quad \text{and} \quad \widehat{\beta}_1 = \frac{S_{SY}/n}{S_{XX}/n} = \frac{S_{XY}}{S_{XX}}$$

are called the *Ordinary Least Square (OLS) Estimate.*

What is the Error Term, ε_i? We know that

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i.$$

Assume that $\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$ (That is, ε_i is independent of data). If X_i is given, $\beta_0 + \beta_1 X_i$ is just a constant, and so

$$Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2),$$

where we call Y_i the *response* or *dependent variable* and X_i the *exploratory* or *independent variable.*

Although we know $\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$, but σ^2 is unknown parameter. What is the MLE of σ^2? As $Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$, then

$$f_Y(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-\beta_0-\beta_1 x_i)^2/2\sigma^2}$$

Then, the likelihood function $\mathbf{L}(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} f_Y(y_i)$. Consider

$$\ell(\beta_0, \beta_1, \sigma^2) = \ln \mathbf{L}(\beta_0, \beta_1, \sigma^2) = -\frac{n}{2} \ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 X_i)^2$$

Solving $\dfrac{\partial \ell}{\partial \sigma} = 0$, we get

$$\widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (\underbrace{\overbrace{y_i}^{\text{true data}} - \overbrace{\widehat{\beta}_0 - \widehat{\beta}_1 X_i}^{\widehat{y}_i\text{, our model}}}_{\text{residual}}).$$

This is a MLE of multiple parameters.

Summary In the following theorems, we assume that we fit a simple linear model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad \text{with } \varepsilon_i \overset{i.i.d.}{=} N(0, \sigma^2).$$

Theorem 4.2.2 OLS Estimate of β_1 and β_0
Let $(x_1, Y_1), \ldots, (x_n, Y_n)$ be a set of points satisfying the linear model $\mathbf{E}[Y \mid x] = \beta_0 + \beta_1 x$ (That is, let $Y_1, \ldots, Y_n$ be independent random variables where $Y_i \sim N(\beta_0, \beta_1 x_i, \sigma^2)$ with x_i are known and β_0, β_1, and σ^2 are unknown). The maximum likelihood estimators for β_0, β_1, and σ^2 are given by

$$\widehat{\beta}_1 = \frac{n \sum_{i=1}^{n} x_i Y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n}\right) Y_i}{n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2} \quad \text{and} \quad \widehat{\beta}_0 = \frac{\sum_{i=1}^{n} Y_i - \widehat{\beta}_1 \sum_{i=1}^{n} x_i}{n} = \overline{Y} - \widehat{\beta}_1 \overline{x}$$

$$\widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \widehat{Y}_i\right)^2, \quad \widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i.$$

Theorem 4.2.3 Distributions of $\widehat{\beta}_0$ and $\widehat{\beta}_1$

- $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are both normally distributed.
- $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are unbiased. That is, $\mathbf{E}[\widehat{\beta}_0] = \beta_0$ and $\mathbf{E}[\widehat{\beta}_1] = \beta_1$.
- $\widehat{\beta}_1$, $\overline{Y}$ and $\widehat{\sigma}^2$ are mutually independent.
- $\dfrac{n\widehat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$. That is, $\mathbf{E}[\widehat{\sigma}^2] = \dfrac{n-2}{n}\sigma^2$.

Remark 4.2 *The best fit line is a linear way to model the data but some data are nonlinear.*

1. ***Polynomial Data*** $y = b + a_1 x + \cdots + a_m x^m$*: minimize* $m+1$ *equations to find* $a_1, \ldots, a_m, b$.
2. ***Exponential Data*** $y = Be^{ax}$*: apply linear technique to* $(x_i, \ln y_i)$ *since* $\ln y = \ln B + ax$.
3. ***Log Data*** $y = Bx^a$*: apply linear technique to* $(\ln x_i, \ln y_i)$ *since* $\ln y = \ln B + a \ln x$.
4. *If* $y = \dfrac{L}{1+e^{a+bx}}$*, then* $\ln\left(\dfrac{L-y}{y}\right)$ *is linear with* x.
5. *If* $y = \dfrac{1}{a+bx}$*, then* $\dfrac{1}{y}$ *is linear with* x.
6. *If* $y = \dfrac{x}{a+bx}$*, then* $\dfrac{1}{y}$ *is linear with* x.
7. *If* $y = 1 - e^{-x^b/a}$*, then* $\ln\left(\ln\left(\dfrac{1}{1-y}\right)\right)$ *is linear with* $\ln x$