1 Review

1.1 Probability

Suppose we have a sample space Ω. Then,

- An element of Ω is denoted by ω, i.e., $\omega \in \Omega$.
- A subset S of Ω is called **event**, i.e., $S \subseteq \Omega$.

Definition 1.1 (Probability Measure)**. Probability measure** on Ω is a function P from subsets of Ω to the real numbers

$$P : \Omega \supseteq S \mapsto \mathbb{R}$$

that satisfies the axioms:

- $P(\Omega) = 1$.
- If $A \subseteq \Omega$, then $P(A) \geq 0$.
- If $A_1, A_2, \dots, A_n, \dots$ are mutually disjoint, then $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$.

More generally, we have the addition law

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

to hold true in all cases.

Theorem 1.1 (Multiplication Principle)**.** If there are p experiments and the ith experiment has n_i possible outcomes. Then there are a total of $n_1 \times n_2 \times \cdots \times n_p$ possible outcomes for the p experiments.

When we calculate permutation, which is sampling r items from n items and list them in order,

- Sampling with Replacement: n^r ways
- Sampling without Replacement: $n(n-1)\cdots(n-r+1)$ ways

When we calculate combination, which is sampling without replcement r items from n items un-orderly, there are

$$\frac{n(n-1)\cdots(n-r+1)}{r!} = \binom{n}{r}$$

ways.

Theorem 1.2 (Multinomial Coefficient)**.** The number of ways that n objects can be grouped into r classes with n_i in the i-th class, $i = 1, \dots, r$, and $\sum_{i=1}^{r} n_i = n$ is

$$\binom{n}{n_1, n_2, \dots, n_r} = \frac{n!}{n_1! n_2! \cdots n_r!}$$

Definition 1.2 (Conditional Probability)**.** Suppose there are two events A and B without a sample space Ω with $P(B) > 0$. The **conditional probability** of A given B is defined to be

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Definition 1.3 (Independence)**.** Two events A and B are said to be independent events if $P(A \cap B) = P(A)P(B)$.

Theorem 1.3 (Law of Total Probability)**.** Let $B_1, B_2, \dots, B$ are a partition of Ω, i.e., $\bigcup_{i=1}^{n} B_i = \Omega$ and $B_i \cap B_j = \varnothing$ for $i \neq j$ with $P(B_i) > 0$ for all i. Then for any event A, we have

$$P(A) = \sum_{i=1}^{n} P(A \cap B_i)$$

Theorem 1.4 (Bayes' Rule)**.** Suppose once more that $B_1, B_2, \dots, B_n$ are a partition of Ω. Then for any event A, we have

$$P(B_j \mid A) = \frac{P(A \mid B_j)P(B_j)}{\sum_{i=1}^{n} P(A \mid B_i)P(B_i)}$$

1.2 Random Variable

Random variable is a function from Ω to the real numbers:

$$X : \Omega \to \mathbb{R}$$

Definition 1.4 (Probability Distribution)**.** The probability distribution of probability measure on Ω which determines the probabilities of the various values of X: $x_1, x_2, \dots,$ with the following properties

- $p(x_i) = P(X = x_i)$
- $\sum_i p(x_i) = 1$

It is called **probability mass function**(pmf) of the random variable X.
Cumulative distribution function(cdf) $F(x)$ is defined as

$$F(x) = P(X \leq x), \quad -\infty < x < \infty$$

The cdf is *non-decreasing* and satisfies

$$\lim_{x \to -\infty} F(x) = 0 \text{ and } \lim_{x \to \infty} F(x) = 1$$

Definition 1.5 (Discrete and Continuous Random Variables)**.** A **discrete** random variable is a random variable that can take on only finite or at most a countably infinite number of values.
A **continuous** random variable is a random variable that can take on a continuum of values.

For a continuous random variable X, the role of frequency function is taken by a **density function**(pdf) $f(x)$, which satisfies the following properties:

- $f(x) \geq 0$
- f is piecewise continuous
- $\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1$.

For a continuous random variable X, for any $a < b$, $P(a < X < b) = \int_a^b f(x)\mathrm{d}x$, hence the probability that rv X takes a *particular value* is 0.

Definition 1.6 (Binomial Distribution)**.** Suppose we have

- n trials, each of which has 2 possible outcomes, namely **success** and **failure**
- Each trial has the same probability of success p
- The n trials are independent.

The binomial random variable $X \sim \mathrm{Bin}(n, p)$ is the total number of successes in the n trials.
The probability distribution is

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n$$

The Bernoulli distribution is the special case of binomial distribution when $n = 1$.

Definition 1.7 (Geometric Distribution)**.** Suppose we have

- Infinite trials, each of which has two possible outcomes, namely success or failure
- Each trial has the same probability of success p
- The trials are independent

Let X be the **total number of trials up to and including the first success**, then $X \sim \mathrm{Geom}(p)$ has geometric distribution.
The proability distribution is

$$p(k) = P(X = k) = (1-p)^{k-1} p, \quad k = 0, 1, \ldots$$

Definition 1.8 (Negative Binomial Distribution)**.** Suppose we have

- The trials are independent
- Each trial has teh same probability of success p.
- Sequence of these trials is performed until there are r successes in all.

Let X be the total number of trials, then $X \sim \mathrm{NB}(r, p)$ has negative binomial distribution.
The probability distribution is

$$P(X = k) = \binom{k-1}{r-1} p^r (1-p)^{k-r}$$

The negative binomial distribution is a generalisation of the geometric distribution.

Definition 1.9 (Poisson Distribution)**.** Random variable $X \sim \mathrm{Poisson}(\lambda)$ follows **Poisson distribution** with parameter $\lambda > 0$ if

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

Theorem 1.5 (Approximation of Binomial using Poisson)**.** $\mathrm{Poisson}(\lambda := np)$ can be derived as the limit of a binomial distribution $\mathrm{Bin}(n, p)$ when n approaches infinity, p approaches 0 with $np = \lambda$.

Definition 1.10 (Uniform Distribution)**.** Let X be a randon variable between a and b where $b > a$. $X \sim U(a, b)$ follows a uniform distribution if the density function of X is

$$f(x) = \begin{cases} \frac{1}{b-a}, & \text{if } a \leq x \leq b \\ 0, & \text{if } x < a \text{ or } x > b \end{cases}$$

Therefore, $F(x) = (x-a)/(b-a)$ on $[a, b]$, 0 on the left of a and 1 on the right of b.

Definition 1.11 (Exponential Distribution)**.** A random variable $X \sim \mathrm{Exp}(\lambda)$ follows an exponential distribution if its density function follows

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0 & x < 0 \end{cases}$$

where $\lambda > 0$.
The cdf of X is

$$F(x) = \begin{cases} 1 - e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Exponential distribution is a special case of gamma distribution.

Definition 1.12 (Gamma Distribution)**.** A random variable $X \sim \Gamma(\alpha, \lambda)$ follows an gamma distribution if its density function follows

$$f(x) = \begin{cases} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, & x \geq 0 \\ 0 & x < 0 \end{cases}$$

where $\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} \mathrm{d}u$ for $x > 0$. Here, we denote α as the **shape parameter** and λ as the **scale parameter**.

Definition 1.13 (Normal Distribution)**.** A random variable $X \sim N(\mu, \gamma^2)$ follows a normal distribution if the density function follows

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

Obviously, we have $f(\mu - x) = f(\mu + x)$.

Theorem 1.6 (Distribution of a Function of Variable)**.** *Suppose $Y \sim g(X)$ where X admits f_X, F_X as pdf and cdf respectively. To calculate f_Y, we first compute*

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = P(XinI)$$

where I is a subset of $\mathbb{R}$. Then take differentiation.

We can easily derive the following result:

Theorem 1.7. If $X \sim N(\mu, \sigma^2)$, and $Y = aX + b$, then $Y \sim N(a\mu + b, a^2\sigma^2)$.

If the function g admits nicer properties, we can have the following theorem

Theorem 1.8. Let X be a continuous rv with density $f(x)$ annd let $Y = g(X)$ where g is a **differentiable, strictly monotonic** function on some interval I. Suppose that $f(x) = 0$ if X is not in I. Then Y has the density function

$$f_Y(y) = f_X(g^{-1}(y))|\frac{\mathrm{d}}{\mathrm{d}y}g^{-1}(y)|$$

for y such that $y = g(x)$ for some x and $f_Y(y) = 0$ for $y \neq g(x)$ for any x in I.

From the theorem, we have the following results:

Theorem 1.9. Let $Z = F(X)$, where X admits a cdf F. Then Z has a uniform distribution on $[0, 1]$.

Theorem 1.10. Let U be uniform on $[0, 1]$, and let $X = F^{-1}(U)$. Then the cdf of X is F.

1.3 Joint Distributions

The joint behaviour of 2 random variable X and Y is determined by the cdf

$$F(x, y) = P(X \leq x, Y \leq y)$$

Theorem 1.11.

$$P(x_1 < X \leq x_2, y_1 < Y \leq y_2) = F(x_2, y_2) - F(x_2, y_1) - F(x_1, y_2) + F(x_1, y_1)$$

Generally, if $X_1, \ldots, X_n$ are jointly distributed random variable, their joint cdf is

$$F(x_1, \ldots, x_n) = P(X_1 \leq x_1, \ldots, X_n \leq x_n)$$

Definition 1.14 (Joint Distribution of Discrete Random Variable)**.** Suppose $X_1, \ldots, X_m$ are discrete random variable defined on same sample space Ω, their joint frequency function $p(x, y)$ is

$$p(x_1, \ldots, x_m) = P(X = x_1, \ldots, X_m = x_m)$$

The marginal frequency function of X_1 is

$$p_{X_1}(x_1) = \sum \cdots \sum_{x_2, \ldots, x_m} p(x_1, \ldots, x_m)$$

Higher dimensional marginal frequency function of X_1 and X_2 can be defined in a similar fashion.

Definition 1.15 (Joint Distribution of Continuous Random Variables)**.** The definition is similar and is omitted. Details can be found in `ST2131` Revision Notes.

Definition 1.16 (Independent Random Variables)**.** Random variables $X_1, \ldots, X_n$ are said to be independent if their joint cdf factors into the product of their marginal cdf's:

$$F(x_1, \ldots, x_n) = F_{X_1}x_1 \cdots F_{X_n}x_n$$

for all $x_1, \ldots, x_n$.

This definition holds for both continuous and discrete random variables.

Definition 1.17 (Discrete Case)**.** X and Y are discrete random variable jointly distributed, If $p_Y(y_j) > 0$, the conditional probability that $X = x_i$, given $Y = y_j$ is

$$p_{X|Y}(x \mid y) := P(X = x_i \mid Y = y_j) = \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)} =$$

Remark: This probability is defined to be zero if $p_Y(y_j) = 0$.

Theorem 1.12. $p_{X|Y}(x \mid y) = p_X(x)$ if X and Y are independent.

Similarly, we define, in the continuous case

$$f_{Y|X}(y \mid x) = \begin{cases} \frac{f_{XY}(x,y)}{f_X(x)} & \text{if } 0 < f_X(x) < \infty \\ 0 & \text{otherwise} \end{cases}$$

Definition 1.18 (Extrema Statistics)**.** Assumer $X_1, \ldots, X_n$ are independent random variable with common cdf F and density f.
Let U be maximum of X_i and V the minimum.
The cdf of U is

$$F_U(u) = [F(u)]^n$$

and density of U is

$$f_U(u) = nf(u)[F(u)]^{n-1}$$

Similarly,

$$F_V(v) = 1 - [1 - F(v)]^n$$

and density of V is

$$f_V(v) = nf(v)[1 - F(v)]^{n-1}$$

Theorem 1.13 (Order Statistics)**.** Let $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$. The density of $X_{(k)}$ is

$$f_k(x) = \frac{n!}{(k-1)!(n-k)!} f(x)F^{k-1}(x)[1 - F(x)]^{n-k}$$

1.4 Expected Values

Definition 1.19 (Expected Value)**.** If X is a discrete random variable with frequency function $p(x)$, the expected value of X, denoted by $E(X)$ is

$$E(X) = \sum_i x_i p(x_i)$$

provided that $\sum_i |x_i| p(x_i) < \infty$. If the sum diverges, the expected is undefined.
Similarly, if X is a continuous random variable with density $f(x)$, then

$$E(X) = \int_{-\infty}^{\infty} x f(x) \mathrm{d}x$$

provided that $\int |x| f(x) \mathrm{d}x < \infty$. If the integral diverges, the expectation is undefined.

Theorem 1.14 (Markov Inequality)**.** If X is a random variable with $P(X \geq 0) = 1$, and for which $E(X)$ exists, then

$$P(X \geq t) \leq \frac{E(X)}{t}$$

Theorem 1.15 (Expectation of Function of Variable)**.** Suppose that $Y = g(X)$,

- If X is discrete with frequency function $p(x)$ then $E(Y) = \sum_x g(x)p(x)$ provided that $\sum |g(x)|p(x) < \infty$.
- If X is continuous with density function $f(x)$ then $E(Y) = \int_{-\infty}^{\infty} g(x)f(x)\mathrm{d}x$ provided that it converges.

Theorem 1.16 (Expectation of Independent Random Variables)**.** If X and Y are independent random variable and g and h are fixed functions, then

$$E[g(X)h(Y)] = E[g(X)]E[h(Y)]$$

provided that the expectations on the right hand side exist.

Theorem 1.17 (Expectation is Linear)**.** If $X_1, \dots, X_n$ are jointly distributed random variable with expectation $E(X_i)$ and Y is a linear function of X_i, i.e., $Y = a + \sum_{i=1}^n b_i X_i$, then

$$E(Y) = a + \sum_{i=1}^n b_i E(X_i)$$

Definition 1.20 (Variance, Standard Deviation)**.** If X is a random variable with expected value $E(X)$, the variance of X is

$$\mathrm{Var}(X) = E\{[X - E(X)]^2\}$$

provided that the expectation exist.
The standard deviation of X is the square root of the variance.
We often use σ^2 to denote variance, and σ for standard deviation.
Therefore,

- If X is discrete, then $\mathrm{Var}(X) = \sum_i (x_i - \mu)^2 p(x_i)$.
- If X is continuous, then $\mathrm{Var}(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\mathrm{d}x$.

Theorem 1.18 (Properties of Variance)**.** We have the following:

1. If $\mathrm{Var}(X)$ exist and $Y = a + bX$, then $\mathrm{Var}(Y) = b^2\mathrm{Var}(X)$.
2. The $\mathrm{Var}(X)$, if exists, may also be calculated as follows:
$$\mathrm{Var}(X) = E(X^2) - [E(X)]^2$$

Theorem 1.19 (Chebyshev's Inequality)**.** Let X be a random variable with mean μ and variance σ^2. Then for any $t > 0$, we have

$$P(|X - \mu| > t) \leq \frac{\sigma^2}{t^2}$$

Definition 1.21 (Model for Measurement Error)**.** Suppose the true value of the quantity being measured is x_0, then teh measurement X is modelled as

$$X = x_0 + \beta + \epsilon$$

where β is a constant error called **bias** and ϵ is the random component of the error.
Here, ϵ is an random variable with $E(\epsilon) = 0$ and $\mathrm{Var}(\epsilon) = \sigma^2$. Hence,

$$E(X) = x_0 + \beta \qquad \mathrm{Var}(X) = \sigma^2$$

A perfect measurement should have $\beta = \epsilon^2 = 0$.

Definition 1.22 (Mean Square Error)**.** The **mean square error** is defined as

$$\mathrm{MSE} = E([X - x_0]^2)$$

It is clear that the mean square error for measurement is $\beta^2 + \epsilon^2$.

Definition 1.23 (Convariance)**.** If X and Y are jointly distributed random variable with means μ_X and μ_Y, respectively, the covariance of X and Y is

$$\mathrm{cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

provided that the expectation exists.

If X and Y is positively(resp. negatively) associated, the convariance will be positive(resp. negative).

Definition 1.24 (Correlation)**.** If X and Y are jointly distributed random variable and the variances of both X and Y exists and non-zero, then the **corelation** of X and Y, denoted by ρ is

$$\rho = \frac{\mathrm{cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

Theorem 1.20 (Properties of Covariance)**.** $-1 \leq \rho \leq 1$. [1] Furthermore, $\rho = \pm 1$ if and only if $P(Y = a + bX) = 1$

[1] This can be shown by considering $\mathrm{Var}(\frac{X}{\sigma_X} \pm \frac{Y}{\sigma_Y})$

for some constant a and b.

Definition 1.25 (Conditional Expectation)**.** Suppose that X and Y are discrete random variable and the conditional frequency function of Y given $X = x$ is $p_{Y|X}(y \mid x)$. The **conditional expectation** of Y given $X = x$ is

$$E(Y \mid X = x) = \sum_y y p_{Y|X}(y \mid x)$$

If X, Y are continuous, then

$$E(h(Y) \mid X = x) = \int h(y) f_{Y|X}(y \mid x) \mathrm{d}y$$

Theorem 1.21 (Expectation and Variance of Conditional Expectation)**.** We have

$$E(Y) = E[E(Y \mid X)]$$

and

$$\mathrm{Var}(Y) = \mathrm{Var}[E(Y \mid X)] + E[\mathrm{Var}(Y \mid X)]$$

Definition 1.26 (Moment Generating Function)**.** The **moment generating function**(mgf) of a random variable X is $M(t) = E(e^{tX})$ if the expectation is defined. Therefore,

- In the discrete case, $M(t) = \sum_x e^{tx} p(x)$ and
- in the continuous case, $M(t) = \int_{-\infty}^{\infty} e^{tx} f(x) \mathrm{d}x$

Theorem 1.22 (Uniqueness of Moment Generating Function)**.** If the mgf exists for t in an open interval containing 0, it uniquely determines the probability distribution.

Theorem 1.23 (Moment Generating Function generates Moment)**.** Let the rth moment of a random variable to be $E(X^r)$ if the expectation exists. If the mgf exists in an open interval containing 0, then

$$M^{(r)}(0) = E(X^r)$$

Theorem 1.24 (Properties of Moment Generating Function)**.** If X has the mgf $M_X(t)$ and $Y = a + bX$, then Y has the mgf $M_Y(t) = e^{at} M_X(bt)$.

If X and Y are independent random variable with mgf's M_X and M_Y and $Z = X + Y$, then

$$M_Z(t) = M_X(t) M_Y(t)$$

on the common interval where both mgf's exist.

1.5 Limit Theorems

Theorem 1.25 (Law of Large Numbers)**.** Let $X_1, \ldots, X_i, . .$ be a sequence of independent random variables with $E(X_i) =$ μ and $\mathrm{Var}(X_i) = \sigma^2$. Let $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$. Then for any $\epsilon > 0$,

$$P(|\bar{X}_n - \mu| > \epsilon) \to 0 \quad \text{as } n \to \infty$$

Definition 1.27 (Converge in Probability, Converge Almost Surely)**.** If a sequence of random variable $\{Z_n\}$ is such that $P(|Z_n - \alpha| > \epsilon) \to 0$ as $n \to \infty$, for any $\epsilon > 0$ and where α is some scalar, then Z_n is said to **converge in probability** to α.

Z_n is said to **converge almost surely** to α if for every $\epsilon > 0$, $|Z_n - \alpha| > \epsilon$ only a finite number of times with probability 1.

Here, converge almost surely is stronger than converge in probability

Definition 1.28 (Converge in Distribution)**.** Let $X_1, \ldots$ be a sequence of random variable with cdf $F_1, \ldots$ and let X be a random variable with distribution function F. We say that X_n converges in distribution to X if

$$\lim_{n \to \infty} F_n(x) = F(x)$$

at every point at which F is continuous.

Definition 1.29 (Continuity Theorem)**.** Let F_n be a sequenc of cdf with the corresponding mgf M_n. Let F be a cdf with the mgf M. If $M_n(t) \to M(t)$ for all t in an open interval containing zero, then $F_n(x) \to F(x)$ at all continuity points of F.

Theorem 1.26 (Central Limit Theorem)**.** Let $X_1, X_2 \ldots$ be a sequence of independent random variable having mean 0 and variance σ^2 and the common distribution function F and mgf M defined in a neighbourhood of 0. Let

$$S_n = \sum_{i=1}^n X_i$$

Then

$$\lim_{n \to \infty} P(\frac{S_n}{\sigma \sqrt{n}} \leq x) = \Phi(x), \quad -\infty < x < \infty$$

2 Normal Distribution and Some Related Distributions

Normal distribution is introduced in section 1.

Theorem 2.1 (Moment Generating Function of Normal Distribution)**.** Suppose $X \sim N(\mu, \sigma^2)$, then

$$M_X(t) = e^{\sigma^2 t^2 + \mu t}$$

Theorem 2.2 (Symmetry of Normal Distribution)**.** Suppose the highest point of the Normal Distribution curve is at $x = \mu$, the normal distribution is symmetric about μ. This implies

- If $x > 0$, the area to the left of $\mu - x$ is the same as the area to the right of $\mu + x$.
- $q_{1-p} = 2\mu - q_p$, where $P(X \leq q_p) = p$.

An empiricial guide to normal distribution is that

- 68% of the data lies within 1 σ interval around μ.
- 95% lies within 2 σ interval
- 99.7% lies within 3σ interval

We have the following results when concerning linear combination of normal random variables;

Theorem 2.3 (Mean of Normal Random Variables)**.** If $X_1, \ldots, X_n$ are independent normal random variable with $X_i \sim N(\mu, \sigma^2)$, then

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n} \sim N(\mu, \frac{\sigma^2}{n})$$

Theorem 2.4 (Linear Combination of Two Normal Random Variable)**.** For any real number a and b, if$X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$ then

$$aX + bY \sim N(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2)$$

This result can be easily shown using moment generating function.
$Z \sim N(0, 1)$ is called standard normal variable, whose cdf is denoted by Φ and density by ϕ. If $X \sim N(\mu, \sigma^2)$, then we can normalise

$$Z = \frac{X - \mu}{\sigma}$$

2.1 χ^2 Distribution

Definition 2.1 (χ_1^2 Distribution)**.** If Z is a standard normal random variable, the distribution of $U = Z^2 \sim \chi_1^2$ is called the chi-square distribution with 1 degree of freedom.

More generally,

Definition 2.2 (χ_n^2 Distribution)**.** If $U_1, \ldots, U_n$ are independent chi-square random variable with 1 degree of freedom, he distribution of $V = U_1 + \cdots + U_n \sim \chi_n^2$ is called the chi-square distribution with n degree of freedom.
χ_n^2 is a gamma distribution with $\alpha = \frac{n}{2}$ and $\lambda = \frac{1}{2}$.
Therefore, density of χ_n^2 is

$$f(v) = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} v^{\frac{n}{2}-1} e^{-\frac{v}{2}}, \qquad v \geq 0$$

And its moment generating function

$$M(t) = (1 - 2t)^{-\frac{n}{2}}$$

From the moment generating function, we can derive that, if $V \sim \chi_n^2$, then

$$E(V) = n \qquad \text{Var}(V) = 2n$$

From definition, if U and V are independent and $U \sim \chi_n^2$ and $V \sim \chi_m^2$, then $U + V \sim \chi_{m+n}^2$.

2.2 t distribution

Definition 2.3 (t Distribution)**.** If $Z \sim N(0, 1)$ and $U \sim \chi_n^2$ and Z and U are independent, then the distribution of $\frac{Z}{\sqrt{\frac{U}{n}}}$ is called the t distribution with n degrees of freedom.
The density function of the t distribution with n degrees of freedom is

$$f(t) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})}(1 + \frac{t^2}{n})^{-\frac{n+1}{2}}$$

From the density, we have $f(t) = f(-t)$, i.e., the t distribution is symmetric about 0.

When the degree of freedom n tends to infinity, the t distribution tends to the standard normal distribution.

2.3 F distribution

Definition 2.4 (F distribution)**.** Let U and V be independent chi-square random variable with m and n df, respectively. The distribution of

$$W = \frac{U/m}{V/n}$$

is called the F distribution with m and n degree of freedom and is denoted by $F_{m,n}$.
The density function of W is given by

$$f(w) = \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})}(\frac{m}{n})^{\frac{m}{2}-1}(1 + \frac{m}{n}w)^{-\frac{m+n}{2}}$$

For $n > 2$, $E(W)$ exists and equals $\frac{n}{n-2}$.

Let $T \sim t_n$, then $T^2 \sim F_{1,n}$.

2.4 Sample Mean and Variance

Definition 2.5 (Sample Statistics)**.** Let $X_1, \ldots, X_n$ be a sample of n independent $N(\mu, \sigma^2)$ random variable.

- $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ is called the sample mean
- $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$ is called teh sample variance

- $E(\bar(X)) = \mu$
- $\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$.

Moreover, $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$.

Theorem 2.5. The rv $\bar{X}$ and the vector of random variable $(X_1 - \bar{X}, \dots, X_n - \bar{X})$ are independent. This is proven here.

From the above theorem, we can show that

Theorem 2.6. $\bar{X}$ and S^2 are independently distributed.

Theorem 2.7. The distribution of $\frac{(n-1)S^2}{\sigma^2}$ is the chi-square distribution with $n-1$ df.
This can be proven by considering $\sum(X_i - \mu)^2 = \sum(X_i - \bar{X} + \bar{X} - \mu)^2$.

Theorem 2.8.

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

3 Survey Sampling

3.1 Population Parameters

Definition 3.1 (Parameter, Statistic)**.** A **parameter** is a numerical summary of the population. It is unknown.
A **statistic** is a summary of a sample taken from the population. We compute it based on the data in our sample. The statistics can either by descriptive or inferential.

We define the following population parameters.

Definition 3.2 (Population Mean, Total, Variance)**.** Assume the population is of size N, then

- Population mean is $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$.
- Population total is $\tau = \sum_{i=1}^{N} x_i$.
- Population variance is $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu^2$.

3.2 Simple Random Sampling

Definition 3.3 (Simple Random Sampling)**.** Suppose we want a sample of size $n < N$ to be collected. Then, a sample is collected via **simple random sampling** if

- Each element in the population has the same chance of being selected.
- Any set of size n from the population have the same chance of being the sample.
- Note that we also assume the sampling is done **without replacement**.

Definition 3.4 (Sample Mean, Variance)**.** Suppose we denote the sample members by $X_1, \dots, X_n$, where each X_i is a random variable representing ith member in the sample. Sample mean $\bar{X} := \frac{1}{n}\sum_{i=1}^{n} X_i$ is an estimate of μ.
Here $\bar{X}$ is a random variable too, whose distribution is called sampling distribution, which determines how accurately $\bar{X}$ estimates μ.

Theorem 3.1 (Expectation and Variance of X_i)**.** Denote the distinct values assumed by the population members by $\zeta_1, \dots, \zeta_m$ and denote the number of population members that have the value ζ_j by n_j where $j = 1, \dots, m$. Then X_i is a discrete random variable with probability mass function

$$P(X_i = \zeta_j) = \frac{n_j}{N}$$

Also,[2]

$$E(X_i) = \mu$$

$$\mathrm{Var}(X_i) = \sigma^2$$

Theorem 3.2 (Unbaised Estimator of μ)**.** With sample random sampling, $E(\bar{X}) = \mu$.

Therefore, with simple random sampling, $E(T) = \tau$, where $T = N\bar{X}$.

Theorem 3.3. For simple random sampling without replacement,[3]

$$\mathrm{cov}(X_i, X_j) = -\frac{\sigma^2}{N-1} \quad \text{if } i \neq j$$

Theorem 3.4 (Variance of $\bar{X}$)**.** With simple random sampling,

$$\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}(\frac{N-n}{N-1}) = \frac{\sigma^2}{n}(1 - \frac{n-1}{N-1})$$

For comparison, $\mathrm{Var}\bar{X} = \frac{\sigma^2}{n}$ when sampling is done *with* replacement.
Therefore, we call factor $(1 - \frac{n-1}{N-1})$ **finite population correction**; we call $\frac{n}{N}$ the **sampling fraction**.

Therefore, with simple random sampling,

$$\mathrm{Var}(T) = N^2\frac{\sigma^2}{n}\frac{N-n}{N-1}$$

3.3 Estimation of σ^2

Within the variance formula of $\bar{X}$, there is one unknown, namely σ^2. Therefore, we would like to estimate $\sigma^2 := \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$ with a function of its sample counterpart $\hat{\sigma}^2 := \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

[2] Proven by calculating expectation and variance using the ζ construction.

[3] This is proven by considering $E(X_iX_j)$ conditional on X_i value.

Theorem 3.5. With simple random sampling, [4]

$$E(\hat{\sigma}^2) = \sigma^2(\frac{n-1}{n})\frac{N}{N-1}$$

Therefore, an **unbiased estimate** of σ^2 is $\frac{N-1}{(n-1)N}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

Combining, we will have

Theorem 3.6 (Unbiased Estimator of $\mathrm{Var}(\bar{X})$)**.** An unbiased estimate of $\mathrm{Var}(\bar{X})$ is

$$s_{\bar{X}}^2 = \frac{s^2}{n}(1 - \frac{n}{N})$$

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

Similarly, an unbiased estimate of $\mathrm{Var}(T)$ is

$$s_T^2 = N^2 s_{\bar{X}}^2$$

3.3.1 Dichotomous Case

In particular, if x_j in the population can only take 1, presence or 0, absence, then we have the special case:

- Population mean $\mu = p$, where p is the porportion of presence.
- Population variance $\sigma^2 = p(1-p)$.
- For a sample of size n, with $X_1, \ldots, X_n$ collected. Then the sample mean $\hat{p}$ is called sample proportion.
- $E(\hat{p}) = p$. Therefore, $\hat{p} := \frac{1}{n}\sum_{i=1}^{n} X_i$ is a unbiased estimate of p.
- $\mathrm{Var}(\hat{p}) = \frac{p(1-p)}{n}(1 - \frac{n-1}{N-1})$.
- An unbiased estimate of $\mathrm{Var}(\hat{p})$ is

$$s_{\hat{p}}^2 = \frac{\hat{p}(1-\hat{p})}{n-1}(1 - \frac{n}{N})$$

Remark: $s_{\bar{X}}, s_T, s_{\hat{p}}$ are called **estimated standard errors**.

3.4 Summary

3.5 Confidence Interval

Definition 3.5 (Confidence Interval)**.** A CI for a parameter θ is a random interval, obtained from the sample, that contain θ with some specified probability.

For example, a $100(1-\alpha)\%$ CI for $Z \sim N(0,1)$ is between $(-z(\alpha/2), z(\alpha/2))$ where z is the quantile function.

Pop	Estimate	Var(Est)	Est(Var)
μ	$\bar{X} = \frac{1}{n}\sum X_i$	$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}(\frac{N-n}{N-1})$	$s_{\bar{X}}^2 = \frac{s^2}{n}(1 - \frac{n}{N})$
p	$\hat{p}$	$\sigma_{\hat{p}}^2 = \frac{p(1-p)}{n}(\frac{N-n}{N-1})$	$s_{\hat{p}}^2 = \frac{\hat{p}(1-\hat{p})}{n-1}(1 - \frac{n}{N}$
τ	$T = N\bar{X}$	$\sigma_T^2 = N^2\sigma_{\bar{X}}^2$	$s_T^2 = N^2 s_{\bar{X}}^2$
σ^2	$(1 - \frac{1}{N})s^2$		

3.6 Ratio Estimate

Definition 3.6 (Population Ratio)**.** We define the population ratio as

$$r := \frac{\mu_y}{\mu_x}$$

The natural estimate of r is $R := \frac{\bar{Y}}{\bar{X}}$. To estimate r, we introduce δ method.

Theorem 3.7 (δ method)**.** Given random variable X with first and second moment known, and $Y = g(X)$ where g non-linear. We will have

$$Y = g(X) \approx g(\mu_X) + (X - \mu_X)g'(\mu_X)$$

Therefore,

$$\mu_Y \approx g(\mu_X)$$

and

$$\sigma_Y^2 \approx \sigma_X^2[g'(\mu_X)]^2$$

up to first order Taylor series expansion around μ_X.
The expansion up to second order

$$Y = g(X) \approx g(X) \approx g(\mu_X) + (X - \mu_X)g'(\mu_X) + \frac{1}{2}(X - \mu_X)^2 g''$$

gives an improvement of $E(Y)$ to

$$E(Y) \approx g(\mu_X) + \frac{1}{2}\sigma_X^2 g''(\mu_X)$$

Theorem 3.8 (δ method of 2 variables)**.** Now suppose $Z = g(X,Y)$ and let $\mu := (\mu_X, \mu_Y)$. With Taylor series expansion to the first order

$$Z = g(X,Y) \approx g(\mu) + (X - \mu_X)\frac{\partial g(\mu)}{\partial x} + (Y - \mu_Y)\frac{\partial g(\mu)}{\partial y}$$

so

$$E(Z) \approx g(\mu)$$

and

$$\mathrm{Var}(Z) \approx \sigma_X^2(\frac{\partial g(\mu)}{\partial x})^2 + \sigma_Y^2(\frac{\partial g(\mu)}{\partial y})^2 + 2\sigma_{XY}(\frac{\partial g(\mu)}{\partial x})(\frac{\partial g(\mu)}{\partial y})$$

whereas Taylor expansion to the second order gives the improved $E(Z)$

$$E(Z) \approx g(\mu) + \frac{1}{2}\sigma_X^2\frac{\partial^2 g(\mu)}{\partial x^2} + \frac{1}{2}\sigma_Y^2\frac{\sigma^2 g(\mu)}{\sigma y^2} + \sigma_{XY}\frac{\partial^2 g(\mu)}{\partial x \partial y}$$

[4] This can be proven by considering $\hat{\sigma}^2 = \frac{1}{n}\sum X_i^2 - \bar{X}^2$ and use variance formula on each term.

Therefore, using the above theorem, we can consider $g(x, y) = \frac{y}{x}$ to approximate $Z = \frac{Y}{X}$. If $\mu_X \neq 0$, we have

$$E(Z) \approx \frac{\mu_Y}{\mu_X} + \sigma_X^2 \frac{\mu_Y}{\mu_X^3} - \frac{\sigma_{XY}}{\mu_X^2} = \frac{\mu_Y}{\mu_X} + \frac{1}{\mu_X^2}(\sigma_X^2 \frac{\mu_Y}{\mu_X} - \rho\sigma_X\sigma_Y)$$

and

$$\text{Var}(Z) \approx \sigma_X^2 \frac{\mu_Y^2}{\mu_X^4} + \frac{\sigma_Y^2}{\mu_X^2} - 2\sigma_{XY}\frac{\mu_Y}{\mu_X^3} = \frac{1}{\mu_X^2}(\sigma_X^2 \frac{\mu_Y^2}{\mu_X^2} + \sigma_Y^2 - 2\rho\sigma_X\sigma_Y\frac{\mu_Y}{\mu_X})$$

Theorem 3.9 (Variance of Ratio)**.** With simple random sampling, the approximate variance of $R = \frac{\bar{Y}}{\bar{X}}$ is

$$\text{Var}(R) \approx \frac{1}{\mu_x^2}(r^2\sigma_{\bar{X}}^2 + \sigma_{\bar{Y}}^2 - 2r\sigma_{\bar{X}\bar{Y}}) = \frac{1}{n}(1 - \frac{n-1}{N-1})\frac{1}{\mu_x^2}(r^2\sigma_x^2 + \sigma_y^2 - 2r\sigma_{xy})$$

Theorem 3.10. With simple random sampling, the expectation of R is given approximately by

$$E(R) \approx r + \frac{1}{n}(1 - \frac{n-1}{N-1})\frac{1}{\mu_x^2}(r\sigma_x^2 - \rho\sigma_x\sigma_y)$$

Using CLT, it can be shown that R is approximately normally distributed. The estimated variance of R is

$$s_R^2 = \frac{1}{n}(1 - \frac{n-1}{N-1})\frac{1}{\bar{X}^2}(R^2 s_x^2 + s_y^2 - 2Rs_{xy})$$

An approximate $100(1-\alpha)\%$ CI for r is $R \pm z(\frac{\alpha}{2})s_R$.

4 Parameter Estimate

There are two main methods to get parameter estimates, namely methods of moments and method of maximum likelihood.

4.1 Method of Moments

Definition 4.1 (kth moment)**.** The kth moment of a probability law is defined as

$$\mu_k = E(X^k)$$

Theorem 4.1 (Natural Estimate of μ_k)**.** If $X_1, \ldots, X_n$ are IID, then the kth sample moment is defined as

$$\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

Theorem 4.2 (Method of Moments)**.** Suppose random variable $X_1, \ldots, X_n$ has joint distribution $f(x \mid \theta)$ dependent on unknown parameter vector θ. We will use the realisation $x_1, \ldots, x_n$ to estimate θ. We define the **bias** to be $E(\hat{\theta}) - \theta$ and **standard error** to be $\sigma_{\hat{\theta}}$.
In method of moments, suppose $\theta = (f_1(\mu_1, \ldots, \mu_m), \ldots, f_k(\mu_1, \ldots, \mu_m))$, then the method of moments estimates of θ is

$$\hat{\theta} = (f_1(\hat{\mu}_1, \ldots, \hat{\mu}_m), \ldots, f_k(\hat{\mu}_1, \ldots, \hat{\mu}_m))$$

Definition 4.2 (Consistency)**.** Let $\hat{\theta}_n$ be an estimate of a parameter θ based on a sample of size n. Then $\hat{\beta}_n$ is said to be **consistent** if $\hat{\theta}_n$ converges in probability to θ as n approaches to infinity. That is, for any $\epsilon > 0$,

$$P(|\hat{\theta}_n - \theta| > \epsilon) \to 0 \text{ as } n \to \infty$$

Since the weak law of large number implies the kth sample moment $\hat{\mu_k}$ converges, in probability to the kth population moment μ_k as sample size $n \to \infty$.

Theorem 4.3 (Consistency of MOM Estimator)**.** MOM estimators are consistent.

Remark. Some MOM estimators are unbiased while other biased.

4.2 Method of Maximum Likelihood

Let $\{f(\cdot \mid \theta) : \theta \in \Theta\}$ be an identifiable parametric family, i.e. there are not $\theta_1 \neq \theta_2$ such that $f(\cdot \mid \theta_1) = f(\cdot \mid \theta_2)$. Suppose $X_1, \ldots, X_n$ are IID random variable with density $f(\cdot \mid \theta_0)$ where $\theta_0 \in \Theta$ is an unknown constant, and $x_1, \ldots, x_n$ realizations of $X_1, \ldots, X_n$. We define **likelihood function** to be

$$\theta \to L(\theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$$

Then we define the **maximum likelihood estimate** of θ_0 to be the value that maximizes the likelihood over Θ, and is denoted as $\hat{\theta}_0$.
We define

- bias := $\text{E}_{\theta_0}(\hat{\theta}_0) = \theta_0$.
- SE $= \text{SD}(\hat{\theta}_0)$

where the subscript θ_0 measn that E and SD are calculated using the density $f(x \mid \theta_0)$.

Theorem 4.4 (MOM vs MLE Estimate)**.** We list down the MOM and MLE Estimate for three main families of probability distribution.

	P1	MOM	MLE	P2	MOM	MLE
Poi	λ	μ_1	$\bar{X}$	$-$	$-$	$-$
Γ	α	$\frac{\mu_1^2}{\mu_2 - \mu_1^2}$	NA	λ	$\frac{\mu_1}{\mu_2 - \mu_1^2}$	$\frac{\alpha}{\bar{X}}$
N	μ	μ_1	$\bar{X}$	σ^2	$\mu_2 - \mu_1^2$	$\frac{1}{n}\sum(x_i - \bar{X})$

Theorem 4.5 (MLE of Multinomial Cell Probability)**.** Suppose an experiment has m possible outcomes $E_1, \ldots, E_m$ with probabilities $p_1, \ldots, p_m$. Let X_i be the number of times E_i occurs in total n independent runs of the experiment. We say $X_1, \ldots, X_m$ follows a multinomial distribution with total cell count n and cell probabilities

$p_1, \dots, p_m$.
The joint pmf function of $X_1, \dots, X_m$ is

$$f(x_1, \dots, x_m \mid p_1, \dots, p_m) = \frac{n!}{\sum_{i=1}^{m} x_i!} \sum_{i=1}^{m} p_i^{x_i}$$

Marginally, each $X_i \sim \text{Bin}(n, p_i)$. However, $\text{cov}(X_i, X_j) = -np_i p_j$, as X_i are not independent.
The MLE of p_j is

$$\hat{p}_j = \frac{x_j}{n}$$

4.3 Large Sample Theory for MLE

Here, we denote the true value of θ by θ_0.

Theorem 4.6 (Consistency of MLE)**.** Under appropriate smoothness conditions on f, the MLE of θ is consistent.

Theorem 4.7 (Fisher Information)**.** Define $I(\theta)$ by

$$I(\theta) = \text{E}\{[\frac{\partial}{\partial\theta} \log f(X \mid \theta)]^2\}$$

Under appropriate smoothness conditions on f, $I(\theta)$ may also be expressed as

$$I(\theta) = -\text{E}[\frac{\partial^2}{\partial\theta^2} \log f(X \mid \theta)]$$

Theorem 4.8 (MLE Asymptotically Unbiased)**.** Let $\hat{\theta}$ be MLE of θ_0. Under smoothness conditions on f, the probability distribution of

$$\sqrt{nI(\theta_0)}(\hat{\theta} - \theta_0)$$

tends to a **standard normal distribution** $Z(0, 1)$.

Furthermore, for an IID sample, the asymptotic variance of MLE is $\frac{1}{nI(\theta)}$.

Theorem 4.9 (Confidence Interval for Normal Distribution)**.** *The MLE of μ and σ^2 from an IID Normal sample are*

$$\hat{\mu} = \bar{X} \textit{ and } \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Since $\frac{\sqrt{n}(\bar{X}-\mu)}{S} \sim t_{n-1}$, where t_{n-1} denotes the t distribution with $n-1$ degrees of freedom and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$, if we use $t_{n-1}(\alpha/2)$ to denote the point beyond which the t_{n-1} distribution has probability $\frac{\alpha}{2}$, then

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

Therefore, the $100(1-\alpha)\%$ CI for μ is

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

For $\hat{\sigma}^2$, we have $\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1}$. Let $\chi^2_{n-1}(\alpha)$ denote the point beyond which the χ^2_{n-1} distribution has probability α, then

$$P(\chi^2_{n-1}(1 - \frac{\alpha}{2}) \leq \frac{n\hat{\sigma}^2}{\sigma^2} \leq \chi^2_{n-1}(\frac{\alpha}{2})) = 1 - \alpha$$

Thus, an exact $100(1-\alpha)\%$ CI for σ^2 is

$$(\frac{n\hat{\sigma}^2}{\chi^2_{n-1}(\alpha/2)}, \frac{n\hat{\sigma}^2}{\chi^2_{n-1}(1 - \alpha/2)})$$

Theorem 4.10 (Approximate CI Using Large Sample Theorem)**.** Since $\sqrt{nI(\hat{\theta})}(\hat{\theta} - \theta_0) \to N(0, 1)$ as $n \to \infty$, we have

$$P(-z(\alpha/2) \leq \sqrt{nI(\hat{\theta})}(\hat{\theta} - \theta_0) \leq z(\alpha/2)) \approx 1 - \alpha$$

Thus, an *approximate* $100(1-\alpha)\%$ CI for θ_0 is given by

$$\hat{\theta} \pm z(\alpha/2)\frac{1}{\sqrt{nI(\hat{\theta})}}$$

4.4 Posterior Distribution

For a given value $\Theta = \theta$, the sample X have probability density $f_{X|\Theta}(x \mid \theta)$. The joint distribution of X and Θ is

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

The marginal distribution of X is $\int f_{X|\Theta}(x \mid \theta) f_\Theta(\theta)$. So the posterior distribution of Θ given data X is

$$f_{\Theta|X}(\theta \mid x) = \frac{f_{X|\Theta}(x \mid \theta) f_\Theta(\theta)}{\int f_{X|\Theta}(x \mid \theta) f_\Theta(\theta \mathrm{d}\theta)}$$

The **posterior mean** is defined to be the mean of the posterior distribution.
The **posterior mode** is the mode of posterior distribution, which is the most probable value of Θ given X.
A Bayesian analogue of 90% CI for θ is the interval from 5th percentile to 95th percentile of posterior distribution.

Theorem 4.11 (Large Sample Normal Approximation to Posterior)**.** Under Weak Conditions, the posterior distribution is approximately normal with mean equal to MLE and posterior variance close to asymptotic variance of MLE if sample size n is large:

$$f_{\Theta|X}(\theta \mid x) \propto \exp(\frac{1}{2}(\theta - \hat{\theta})^2 l''(\hat{\theta}))$$

where the last term is proportional to $N(\hat{\theta}, -\frac{1}{l''(\hat{\theta})})$ density.

5 Efficiency and Sufficiency

The mean square error of an estimator $MSE(\hat{\theta}) = E(\hat{\theta} - \theta_0)^2 = \text{Var}(\hat{\theta}) + [E(\hat{\theta}) - \theta_0]^2$. If $\hat{\theta}$ is unbiased, then the last term diminishes.

Definition 5.1 (Efficiency)**.** Given 2 estimators $\hat{\theta}$ and $\tilde{\theta}$, the **efficiency** of $\hat{\theta}$ relative to $\tilde{\theta}$ is

$$\text{eff}(\hat{\theta}, \tilde{\theta}) = \frac{\text{Var}(\tilde{\theta})}{\text{Var}(\hat{\theta})}$$

If variance cannot be computed exactly, one should use asympototic variance and this efficency is called asymptotic relative efficiency.

Theorem 5.1 (Cramer-Rao Inequality)**.** Let $X_1, \dots, X_n$ be IID with density function $f(x \mid \theta)$. Let $T = t(X_1, \dots, X_n)$ be an unbiased estimate of θ. Then under smoothness assumption on $f(x \mid \theta)$,

$$\text{Var}(T) \geq \frac{1}{nI(\theta)}$$

An unbiased estimate whose variance achieves the lower bound is said to be **efficient**. We say MLE is asymptotically efficient.