Revision notes - ST2132

Ma Hongqiang

November 9, 2018

Contents

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 $A_1, A_2, \ldots, A_n, \ldots$ 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, $i = 1, \ldots, r$, and $\sum_{i=1}^{r} n_i = n$ is

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

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 $B_1, B_2, \dots, B_n$ 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 $X \sim \text{NB}(r,p)$ 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 $X \sim \text{Poisson}(\lambda)$ 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).
$\text{Poisson}(\lambda := np)$ can be derived as the limit of a binomial distribution $\text{Bin}(n,p)$ when n approaches infinity, p approaches 0 with $np = \lambda$.

Definition 1.10 (Uniform Distribution).
Let X be a random 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 $X \sim \text{Exp}(\lambda)$ 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

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

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 $X_1, \ldots, X_n$ 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 $P(Y = a + bX) = 1$ 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

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

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

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 $X_1, \ldots, X_i, \ldots$ be a sequence of independent random variables with $E(X_i) = \mu$ 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 squence 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

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

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 $X_1, \dots, X_n$ 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^2_1 Distribution)**.**
If Z is a standard normal random variable, the distribution of $U = Z^2 \sim \chi^2_1$ is called the chi-square distribution with 1 degree of freedom.

More generally,

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

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

From definition, if U and V are independent and $U \sim \chi^2_n$ and $V \sim \chi^2_m$, then $U + V \sim \chi^2_{m+n}$.

2.2 t distribution

Definition 2.3 (t Distribution)**.**
If $Z \sim N(0,1)$ and $U \sim \chi^2_n$ 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 $(X_1 - \bar{X}, \ldots, X_n - \bar{X})$ 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 $X_1, \ldots, X_n$, 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 $\zeta_1, \ldots, \zeta_m$ and denote the number of population members that have the value ζ_j by n_j where $j = 1, \ldots, m$. Then X_i is a discrete random variable with probability mass function

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

Also,[2]

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

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

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]

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

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

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

For comparison, $\text{Var}\bar{X} = \frac{\sigma^2}{n}$ 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,

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

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

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 $\text{Var}(\bar{X})$)**.**
An unbiased estimate of $\text{Var}(\bar{X})$ 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 $\text{Var}(T)$ is

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

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

[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.

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 $X_1, \dots, X_n$ 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.
- $\text{Var}(\hat{p}) = \frac{p(1-p)}{n}(1 - \frac{n-1}{N-1})$.
- An unbiased estimate of $\text{Var}(\hat{p})$ 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

Population Parameter	Estimate	Variance of Estimate	Estimated Variance
μ	$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$	$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}(\frac{N-n}{N-1})$	$s_{\bar{X}}^2 = \frac{s^2}{n}(1 - \frac{n}{N})$
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	$(1 - \frac{1}{N})s^2$		

3.5 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.

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

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

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

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

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

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

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

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

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 $\widehat{\mu_k}$ 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 $:= \mathrm{E}_{\theta_0}(\hat{\theta}_0) = \theta_0$.
- SE $= \mathrm{SD}(\hat{\theta}_0)$

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.

Family	Parameter 1	MOM Estimate	MLE Estimate	Parameter 2	MOM Estimate	MLE Estim
Poisson	λ	μ_1	$\bar{X}$	$-$	$-$	$-$
Gamma	α	$\frac{\mu_1^2}{\mu_2 - \mu_1^2}$	no closed form	λ	$\frac{\mu_1}{\mu_2 - \mu_1^2}$	$\frac{\alpha}{\bar{X}}$
Normal	μ	μ_1	$\bar{X}$	σ^2	$\mu_2 - \mu_1^2$	$\frac{1}{n}\sum_{i=1}^{n}(x_i -$

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

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

Marginally, each $X_i \sim \text{Bin}(n, p_i)$. However, $\text{cov}(X_i, X_j) = -np_ip_j$, 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

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

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 efficiency is called asymptotic relative efficiency.

Theorem 5.1 (Cramer-Rao Inequality).
Let $X_1, \ldots, X_n$ be IID with density function $f(x \mid \theta)$. Let $T = t(X_1, \ldots, X_n)$ 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.

Definition 5.2 (Sufficiency).
A statistics $T(X_1, \ldots, X_n)$ is said to be sufficient for θ if the conditional distribution of $X_1, \ldots, X_n$, given $T = t$ does not depend on θ for any value of t.

Theorem 5.2 (Factorization Theorem).
A necessary and sufficient condition for the statistic $T(X_1, \ldots, X_n)$ to be sufficient for a parameter θ is that the joint probability function factors in the form

$$f(x_1, \ldots, x_n \mid \theta) = g[T(x_1, \ldots, x_n), \theta]h(x_1, \ldots, x_n)$$

Theorem 5.3. If T is sufficient for θ, the MLE of θ is a function only of T.

Definition 5.3 (1-parameter Member of Exponential Family).
1-parameter members of the exponential family have pdf's or pmf's of the form

$$f(x \mid \theta) = \begin{cases} e^{c(\theta)T(x)+d(\theta)+S(x)}, & \text{if } x \in A, \\ 0, & \text{otherwise} \end{cases}$$

where the set A does not depend on θ.
The 1-parameter exponential family contains many common probability distribution like

- Binomial distribution with known n
- Poisson distribution

- Gamma distribution
- Normal distribution

Theorem 5.4 (Rao-Blackwell Theorem)**.**
Let $\hat{\theta}$ be an estimator of θ with finite second moment $E(\hat{\theta}^2) < \infty$ for all θ. Suppose that T is sufficient for θ, and let $\tilde{\theta} = E(\hat{\theta} \mid T)$. Then for all θ

$$E(\tilde{\theta} - \theta)^2 \leq E(\hat{\theta} - \theta)^2.$$

And the inequality is strict unless $\hat{\theta} = \tilde{\theta}$.

6 Hypothesis Testing

In hypothesis testing, we have two hypotheses:

- First hypothesis is called **null hypothesis** H_0
- Other hypothesis is called **alternative hypothesis** H_A or H_1.

Here, H_1 is usually taken to be the complement of H_0 but there is no guarantee.
The **decision rule** has typically two possible conclusion:

- Reject H_0
- Do not reject H_0.

Definition 6.1 (Type I Error)**.**
Rejecting H_0 when it is true is called **type I error**.
The probability of type I error is the **significance level** of the test and is usually denoted by α.

Definition 6.2 (Type II Error)**.**
Accepting H_0 when it is false is called **type II error**.
The probability of type II error is usually denoted by β.

The decision rule is based on a **test statistic**. The set of values of a test statistics that leads to the rejection of H_0 is called the **rejection region**. The set of values that leads to acceptance of H_0 is called the **acceptance region**.
The probability distribution of the test statistic when H_0 is true is called the **null distribution**.

Definition 6.3 (Likelihood Ratio)**.**
The likelihood ratio is defined to be the ratio of the two likelihood, one under H_0 and the other under H_1.
Let $f_0(x)$ be pmf of X under H_0 and $f_1(x)$ be pmf of X under H_1. Then the likelihood ratio is given by

$$\frac{f_0(x)}{f_1(x)}$$

Given $X_1, \ldots, X_n$ IID with density $f(x \mid \theta)$ and density of X under H_0 is f_0 and under H_1 f_1.
The likelihood ratio of H_0 to H_1 based on $X_1, \ldots, X_n$ is calculated by

$$\Lambda(\mathbf{x}) = \frac{f_0(x_1) \cdots f_0(x_n)}{f_1(x_1) \cdots f_1(x_n)}$$

Note that the smaller ratio means we have more evidence against H_0. Therefore critical region should contain x with smaller ratio.

Theorem 6.1 (Neyman-Pearson Lemma)**.**
Suppose that H_0 and H_1 are simple hypotheses and that the test rejects H_0 when the likelihood ratio is less than c and the significance level α. Then **any other** test for which the significance level is less than or equal to α has power less than or equal to that of the likelihood ratio test.

Do note Neyman-Pearson Lemma only works when both hypothesis are simple. A hypothesis that *does not* completely specify the probability distribution is called a **composite hypothesis**.

6.1 Significance Level α and p-value

Here, α is the probability of falsely rejecting the null hypothesis. However, hypothesis testing always rejects or do not reject null when there is no need for dichrotomous decision. In this case, p-value can be used.

Definition 6.4 (p-value)**.**
Given an sample, p-value is defined to be the smallest significance level at which H_0 would be rejected.
Therefore, the smaller p-value is, the stronger the evidence is against H_0.

It is advisable to choose the simpler hypothesis to be the null hypothesis; it is also advisable to choose the hypothesis that has graver consequence of rejecting to be the null hypothesis, since we can control the probability of it being falsely rejected.

6.2 Uniformly Most Powerful Test

Suppose we are testing $H_0 : \mu = \mu_0$ against $H_1 : \mu > \mu_0$, where the alternative hypothesis is **one-sided**, then the likelihood ratio test is the **uniformly most powerful** test.
However, the likelihood ratio test is not uniformly most powerful when $H_1 : \mu \neq \mu_0$ is two sided.

6.3 Duality of CI and Hypothesis Tests

Theorem 6.2.
Suppose that for every value θ_0 in Θ there is a test at level α of the hypothesis $H_0 : \theta = \theta_0$. Denote the acceptance region of the test by $A(\theta_0)$. The the set

$$C(\boldsymbol{X}) := \{\theta : \boldsymbol{X} \in A(\theta)\}$$

us a $100(1-\alpha)\%$ confidence region for θ.

Theorem 6.3.
Suppose that $C(\boldsymbol{X})$ is a $100(1-\alpha)\%$ confidence region for θ.
Then an acceptance region for a test at level α of the hypothesis $H_0 : \theta = \theta_0$ is

$$A(\theta_0) = \{\boldsymbol{X} \mid \theta_0 \in C(\boldsymbol{X})\}$$

6.4 Generalized Likelihood Ratio Test

Suppose that observations $X = (X_1, \dots, X_n)$ have a joint density $f(\boldsymbol{x} \mid \theta)$.
Let Ω be the set of all possible values of θ.
Let ω_0 and ω_1 be the subsets of Ω such that they form a partition.
The test is of the form:

$$H_0 : \theta \in \omega_0 \text{ vs } H_1 : \theta \in \omega_1$$

We define the generalized likelihood ratio test statistic to be

$$\Lambda^* = \frac{\max_{\theta \in \omega_0}[\text{lik}(\theta)]}{\max_{\theta \in \omega_1}[\text{lik}(\theta)]}$$

and small values of Λ^* tend to discredit H_0.
We define

$$\Lambda = \frac{\max_{\theta \in \omega_0}[\text{lik}(\theta)]}{\max_{\theta \in \Omega}[\text{lik}(\theta)]} = \min\{\Lambda^*, 1\}$$

The generalized likelihood ratio test will reject H_0 if

$$\Lambda \leq \lambda_0$$

where λ_0 is a constant determined by

$$P(\Lambda \leq \lambda_0 \mid H_0) = \alpha$$

Theorem 6.4.
Under smoothness conditions on the probability densities or frequency functions involved, the null distribution of $-2 \log \Lambda$ tends to a **chi-square** distribution with degree of freedom equal to

$$\dim \Omega - \dim \omega_0$$

as sample size n tends to infinity.

We now study LRT for Multinomial distribution. Suppose an experiment can obtain m possible outcomes $E_1, \dots, E_m$ with probabilities $p_1, \dots, p_m$. Let X_i be number of numbers E_i occurs in total n independent runs of the experiemnt. Then $X_1, \dots, X_m$ follow a multinomial distribution with total cell count n and cell probabilities $p_1, \dots, p_m$.
Claerly, joint pmf reads

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

Let us write $\boldsymbol{p} = (p_1, \dots, p_m)$, and we are interested in testing

$$H_0 : p = p(\theta), \theta \in \omega_0$$

where θ is an unknown parameter, against H_1 for any other value of p other than those in H_0.
The numerator of the likelihood ratio is

$$\max_{p \in \omega_0}\left(\frac{n!}{x_1! \cdots x_m!} p_1(\theta)^{x_1} \cdots p_m(\theta)^{x_m}\right)$$

where x_i are the observed cell counts in the m cells. This likelihood is maximized by MLE $\hat{\theta}$.
In the denominator, the probabilities are restricted under Ω, so it is maximized by global MLE $\hat{p_i} = \frac{x_i}{n}$.
Therefore, we have

$$\Lambda = \prod_{i=1}^{m} [\frac{p_i(\hat{\theta})}{\hat{p_i}}]^{x_i}$$

and

$$-2 \log \Lambda = 2 \sum_{i=1}^{m} O_i \log(\frac{O_i}{E_i})$$

where $O_i = n\hat{p_i}, E_i = np_i(\hat{\theta})$ the observed and expected cell counts respectively.
Also, the degree of freedom $df = m - 1 - k$, where k is the degree of freedom in θ.
The generalized LRT rejects the null hypothesis if

$$-2 \log \Lambda > \chi^2_{m-k-1}(\alpha)$$

Theorem 6.5 (Pearson Chi-square Test)**.**
Another test for multinomial distribution is Pearson Chi-square test, where the test statistic is

$$X^2 = \sum_{i=1}^{m} \frac{[x_i - np_i(\hat{\theta})]^2}{np_i(\hat{\theta})}$$

The test statistics follows $X^2 \sim \chi^2_{m-k-1}$ under H_0.

7 Comparing Two Samples

In many experiemets, we have two groups, one as treatment group and one as control group.

7.1 Two Normally Distributed Indepedent Sample

Here, we assume observation from control group are independent random variable with a common distribution F.
Observation from treatment group are indepedent of each other and of the controls and have a common distribution G.
We further assume, $X_1, \dots, X_n$ are IID $N(\mu_X, \sigma^2)$ and $Y_1, \dots, Y_m$ are IID $N(\mu_Y, \sigma^2)$.
We want to **test**

$$H_0 : \mu_X = \mu_Y \text{ vs } H_1 : \mu_X \neq \mu_Y$$

We can think of $\mu_X - \mu_Y$ as the *effect of treatment*.

Theorem 7.1 (Fact)**.**
A *natural estimate* of $\mu_X - \mu_Y$ is $\bar{X} - \bar{Y}$, and indeed it is the MLE of $\mu_X - \mu_Y$.
If $\bar{X}$ independent of $\bar{Y}$, so

$$\bar{X} - \bar{Y} \sim N(\mu_X, \mu_Y, \sigma^2(\frac{1}{n} + \frac{1}{m}))$$

If σ^2 is known, then we have the test statistics

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim N(0, 1)$$

The $100(1 - \alpha)\%$ CI for $\mu_X - \mu_Y$ will be

$$(\bar{X} - \bar{Y}) \pm z(\frac{\alpha}{2})\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}$$

However, if σ^2 unknonw, we need to estimate it via **pooled sample variance** s_p^2:

$$s_p^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{m + n - 2}$$

where $S_X^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$ and $S_Y^2 = \frac{1}{m-1}\sum_{i=1}^m (Y_i - \bar{Y})^2$.
The estimate standard error of $\bar{X} - \bar{Y}$ is then

$$s_{\bar{X}-\bar{Y}} = s_p\sqrt{\frac{1}{n} + \frac{1}{m}}$$

Theorem 7.2 (Theorem A)**.**
Suppose that $X_1, \dots, X_n$ are IID $N(\mu_X, \sigma^2)$ and $Y_1, \dots, Y_m$ are IID $N(\mu_Y, \sigma^2)$ is another independent sample.
The statistic

$$t = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{s_{\bar{X}-\bar{Y}}}$$

follows a t distribution with $df = m + n - 2$.

Therefore, under the asuumptions of Theorem A, a $100(1-\alpha)\%$ cI for $\mu_X - \mu_Y$ is

$$(\bar{X}-\bar{Y}) \pm t_{m+n-2}(\frac{\alpha}{2}) \times s_{\bar{X}-\bar{Y}}$$

Theorem 7.3 (Rejection Region)**.**
Consider the following alternative hypothesis:

- $H_1 : \mu_X \neq \mu_Y$
- $H_2 : \mu_X > \mu_Y$
- $H_3 : \mu_X < \mu_Y$

The null hypothesis is always $H_0 : \mu_X = \mu_Y$ and test statistic $t = \frac{\bar{X}-\bar{Y}}{s_{\bar{X}-\bar{Y}}} \sim t_{m+n-2}$.
The rejection region for the 3 alternatives are:

- $H_1 : |t| > t_{m+n-2}(\frac{\alpha}{2})$
- $H_2 : t > t_{m+n-2}(\alpha)$
- $H_3 : t < t - m + n - 2(\alpha)$.

Alternatively, the test of H_0 against H_1 can be derived as a likelihood ratio test. The unknown parameters are $\theta = (\mu_X, \mu_Y, \sigma)$.
Under H_0, $\theta \in w_0 = \{\mu_X = \mu_Y, 0 < \sigma > \infty\}$ where as under Ω, we have $\mu_X, \mu_Y \in \mathbb{R}$, $\sigma \in \mathbb{R}^+$.
The likelihood for the two sample is

$$\text{lik} = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X_i-\mu_X)^2}{2\sigma^2}} \times \prod_{j=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(Y_j-\mu_Y)^2}{2\sigma^2}}$$

and log-likelihood is

$$l(\mu_X, \mu_Y, \sigma^2) = -\frac{m+n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}[\sum_{i=1}^{n}(X_i-\mu_X)^2 + \sum_{j=1}^{m}(Y_j-\mu_Y)^2]$$

Under ω_0, we have sample of size $m+n$ from a $N(\mu_0, \sigma^2)$ distribution, so MLE of μ_0 and σ_0^2 are

$$\hat{\mu}_0 = \frac{1}{m+n}[\sum_{i=1}^{n} X_i + \sum_{j=1}^{m} Y_j]$$

and

$$\hat{\sigma}_0^2 = \frac{1}{m+n}[\sum_{i=1}^{n}(X_i-\hat{\mu}_0)^2 + \sum_{j=1}^{m}(Y_j-\hat{\mu}_0)^2]$$

so the maximized log-likelihood is

$$l(\hat{\mu}_0, \hat{\sigma}_0^2) = -\frac{m+n}{2}\log\hat{\sigma}_0^2 - \frac{m+n}{2}$$

Under Ω, the MLE is then

$$\hat{\mu}_X = \bar{X}, \hat{\mu}_Y = \bar{Y}$$

and

$$\hat{\sigma}_1^2 = \frac{1}{m+n}[\sum_{i=1}^{n}(X_i - \hat{\mu}_X)^2 + \sum_{j=1}^{m}(Y_j - \hat{\mu}_Y)^2]$$

and maximized log-likelihood is

$$l(\hat{\mu}_X, \hat{\mu}_Y, \hat{\sigma}_1^2) = -\frac{m+n}{2}\log\hat{\sigma}_1^2 - \frac{m+n}{2}$$

and the log of likelihood ratio is thus

$$\log\Lambda = \frac{m+n}{2}\log(\frac{\hat{\sigma}_1^2}{\hat{\sigma}_0^2})$$

The likelihood ratio test rejects for large value of $\log\Lambda$, so then reject for large values of

$$\frac{\hat{\sigma}_1^2}{\hat{\sigma}_0^2} = \frac{\sum_{i=1}^{n}(X_i - \hat{\mu}_0)^2 + \sum_{j=1}^{m}(Y_j - \hat{\mu}_0)^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{j=1}^{m}(Y_j - \bar{Y})^2}$$

We, by using $\sum_{i=1}^{n}(X_i - \hat{\mu}_0)^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 + n(\bar{X} - \hat{\mu}_0)^2$, obtain

$$\bar{X} - \hat{\mu}_0 = \frac{m}{m+n}(\bar{X} - \bar{Y}), \quad \bar{Y} - \hat{\mu}_0 = \frac{n}{m+n}(\bar{Y} - \bar{X})$$

SO the alternative expression for numerator of ratio is then

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{j=1}^{m}(Y_j - \bar{Y})^2 + \frac{mn}{m+n}(\bar{X} - \bar{Y})^2$$

Hence the test reejects for large values of

$$1 + \frac{mn}{m+n} \times \frac{(\bar{X} - \bar{Y})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{j=1}^{m}(Y_j - \bar{Y})^2}$$

or equivalently, the large values of

$$\frac{|\bar{X} - \bar{Y}|}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{j=1}^{m}(Y_j - \bar{Y})^2}}$$

which is $|t|$. Then the likelihood ratio test is equivalent to t test.

7.1.1 Unequal Variance

In the case of unequal variance, then a natural estimate of $\mathrm{Var}(\bar{X}-\bar{Y})$ is

$$\frac{S_X^2}{n}+\frac{S_Y^2}{m}$$

the appropriate test statistic is then

$$\frac{\bar{X}-\bar{Y}}{\sqrt{\frac{S_X^2}{n}+\frac{S_Y^2}{m}}}$$

The null distribution of this statistic can be closely approxiated by the t distribution with

$$df=\frac{(\frac{S_X^2}{n}+\frac{S_Y^2}{m})^2}{\frac{(S_X^2/n)^2}{n-1}+\frac{(S_Y^2/m)^2}{m-1}}$$

Remark: If the underlying distribution **are not normal** and sample size are large, the use of t distribution or normal distribution is justified by CLT. The probability levels of confidence intervals and hypothesis tests are approximately valid.
However, if sample size small and distribution not normal, the conclusion may be invalid.

7.2 Power

Calculation of power are important because it determines how large sample size should be. Note that

$$\text{power} = P(\text{reject } H_9 \mid H_0 \text{ is false})$$

The power of a two-sample t test depends on four factors:

1. The larger the real difference $\delta=\mu_X-\mu_Y$, the greater the power,
2. THe lager the significance level α, the more powerful the test.
3. The population standard deviation, the smaller σ, the larger the power
4. The larger the sample size n and m, the greater the power.

The exact power calculation for t test against alternative hypothesis $H_1:\mu_X-\mu_Y=\delta$ requirees noncentral t. However, one can perform *approximate calculation* based on normal if sample sizes are reasonably large.
Suppose δ,α,σ are iven and sample size $n=m$, then we have

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

The test at level α of $H_0:\mu_X=\mu_Y$ versus $H_1:\mu_X\neq\mu_Y$ is based on test statistic

$$Z=\frac{\bar{X}-\bar{Y}}{\sigma\sqrt{2/n}}$$

and the rejection region for the test is $|Z| > z(\alpha/2)$, or $|\bar{X} - \bar{Y}| > z(\alpha/2)\sigma\sqrt{2/n}$. Otherwise, if $\mu_X - \mu_Y = \delta$, ten

$$\frac{\bar{X} - \bar{Y} - \delta}{\sigma\sqrt{2/n}} \sim N(0,1)$$

therefore,

$$\begin{aligned}\text{power} &= P[|\bar{X} - \bar{Y}| > z(\alpha/2)\sigma\sqrt{2/n} \mid \mu_X - \mu_Y = \delta] \\ &= 1 - \Phi(z(\alpha/2) - \frac{\delta}{\sigma}\sqrt{n/2}) + \Phi(-z(\alpha/2) - \frac{\delta}{\sigma}\sqrt{n/2})\end{aligned}$$

Typically, as δ moves away from ero, one of these terms dominate.

7.3 Paired Samples

In experiments sample can be paired, which casues sample to be dependent. Denote the pairs as $(X_i, Y_i), i = 1, \ldots, n$. Let X, Y have mean μ_X, μ_Y and variance σ_X^2 adn σ_Y^2. Therefore, $\text{cov}(X_i, Y_i) = \rho\sigma_X\sigma_Y$ where ρ is the correlation coefficient.
We assume that *different pairs* are independently distributed.
The differences $D_i = X_i - Y_i$ are independent with mean

$$E(D_i) = \mu_X - \mu_Y$$

and variance

$$\text{Var}(D_i) = \sigma_X^2 + \sigma_Y^2 - 2\rho\sigma_X\sigma_Y$$

If we estimate $\mu_X - \mu_Y$ by $\bar{D} = \bar{X} - \bar{Y}$, then

$$E(\bar{D}) = \mu_X - \mu_Y$$

and

$$\text{Var}(\bar{D}) = \frac{\sigma_X^2 + \sigma_Y^2 - 2\rho\sigma_X\sigma_Y}{n}$$

The variance will be smaller then independent samples, if $\rho > 0$.
With the additional assumption that the differences are a sample from normal distribution with $E(D_i) = \mu_D$ and $\text{Var}(D_i) = \sigma_D^2$, and if σ_D is unknown, inferences will be based on

$$t = \frac{\bar{D} - \mu_D}{s_{\bar{D}}} \sim t_{n-1}$$

Therefore, teh $100(1-\alpha)\%$ CI for μ_D is $\bar{D} \pm t_{n-1}(\alpha/2)s_{\bar{D}}$

7.4 Nonparametric Method

7.4.1 Nonparametric Statistical Methods

Definition 7.1 (Nonparametric Statistical Methods)**.**
Nonparametri statistical methods are inferential methods that do not assume a particular form of distribution for population.

Let $X_1, \dots, X_n$ be IID with cdf F, and $Y_1, \dots, Y_m$ with cdf G.
Consider null hypothesis: $H_0 : F = G$. We are interested whether X are on the whole larger than the Y values or vice-versa.

Theorem 7.4 (Mann-Whitney Test)**.**
A test statistic is calculated in the following way:

- All $(n + m)$ observation are grouped together $Z_1, \dots, Z_{n+m}$ as pooled sample and we assume the values are distinct. We define

  $$\text{Rank}(Z) = i$$

  if Z is the ith **smallest** value within the pooled sample.

- Define rank sum scores

  $$R_X = \sum_{i=1}^{n} \text{Rank}(X_i), R_Y = \sum_{i=1}^{m} \text{Rank}(Y_i)$$

It is obvious that $R_X + R_Y = \frac{1}{2}(m + n)(m + n + 1)$ is fixed. So we should reject either R_X or R_Y is too small or large.
We take the smaller sample, suppose of size $n = \min(n, m)$, and compute the sum ranks R from that sample.
Let $R' = n(m + n + 1) - R$. The Mann-Whitney test statistic is

$$R^* = \min(R, R')$$

and reject H_0 if R^* too small.

7.4.2 Signed Rank Test

We define

$$\text{Rank}(D) = i$$

if D has the ith smallest absolute value within the sample.
Define W_+ to be the sum of ranks among all the positive D_i and W_- the sum of ranks among all negative D_i. We reject H_0 when W_+ is too large or too small.

Note that $W_+ + W_- = \frac{1}{2}n(n + 1)$, so we only need to count one of them.
We just take $W := \min(W_+, W_-)$ and check for small critical values from the table.
If 0 are present, they are discarded. If there are ties, then the D_i's are given an average rank within all the ties, same as Mann-Whitney test.

Both tests are robust to outliers.

8 ANOVA

8.1 One Way ANOVA

One-way layout is an experimental design in which independent measurements are made under each of several treatments. The technique is generalized from comparing 2 independent samples.
We first discuss the ANOVA and F test in the case of I treatments/groups and J measurements each group.
We define Y_{ij} to be the jth observation of the ith group. We consider the observations are corrupted by random errors and that the error in one observation is independent of errors in other observation. Hence, the one-way ANOVA model is

$$Y_{ij} = \mu + \alpha_i + e_{ij}, \quad i = 1, \dots, I; j = 1, \dots, J$$

where μ is the overall mean, α_i the differential effect of the ith treatment and e_{ij} are IID $N(0, \sigma^2)$.
The α_i should be normalized, by

$$\sum_{i=1}^{I} \alpha_i = 0$$

The expected response tot he ith treatment is

$$E(Y_{ij}) = \mu + \alpha_i$$

If $\alpha_i = 0$ for all i, then all treatments will have the same response. In general, $\alpha_i - \alpha_k$ is the difference between the expected values under treatment i and k.
The null hypothesis is

$$H_0 : \alpha_1 = \cdots = \alpha_I$$

Define $\bar{Y_i} = \frac{1}{J}\sum_{j=1}^{J} Y_{ij}$ be the average of the observations under the ith treatment.
Denote $\bar{\bar{Y}} = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} Y_{ij}$ be overall average.
The ANOVA is based on square of deviation of each observation from its overall average.

$$\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - \bar{\bar{Y}})^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - \bar{Y_i})^2 + J\sum_{i=1}^{I}(\bar{Y_i} - \bar{\bar{Y}})^2$$

Or equivalently,

$$SS_{TOT} = SS_W + SS_B$$

It means, sum of squares = sum of squares within groups + sum of squares between groups.

Theorem 8.1.
Let X_i, $i = 1, \dots, n$ be independent random variables with $E(X_i) = \mu_i$ and $\mathrm{Var}(X_i) = \sigma^2$. Then

$$E[(X_i - \bar{X})^2] = (\mu_i - \bar{\mu})^2 + \frac{n-1}{n}\sigma^2$$

where $\bar{\mu} = \frac{1}{n}\sum_{i=1}^{n} \mu_i$.

Theorem 8.2.
Under the assumption of model stated before,

$$\begin{aligned} E(SS_W) &= \sum_{i=1}^{I}\sum_{j=1}^{J} E(Y_{ij} - \bar{Y_i})^2 \\ &= \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{I-J}{J}\sigma^2 = I(J-1)\sigma^2 \end{aligned}$$

$$\begin{aligned} E(SS_B) &= J\sum_{i=1}^{I} E(\bar{Y_i} - \bar{\bar{Y}})^2 \\ &= J\sum_{i=1}^{I}[\alpha_i^2 + \frac{(I-1)\sigma^2}{IJ}] \\ &= J\sum_{i=1}^{I}\alpha_i^2 + (I-1)\sigma^2 \end{aligned}$$

We can have an unbiased estimate for σ^2 to be $s_p^2 := \frac{SS_W}{I(J-1)}$. We can write s_p^2 as

$$s_p^2 = \sum_{i=1}^{I}(J-1)s_i^2$$

where s_i^2 is the sample variance in the ith group. Hence, estiamtes of σ^2 is from the I groups pooling together.
If all $\alpha_i = 0$, we can have $E[\frac{SS_B}{I-1}] = \sigma^2$, so $\frac{SS_W}{I(J-1)}$ and $\frac{SS_B}{I-1}$ should be about equal. Otherwise, SS_B will be inflated.

Theorem 8.3.
If errors are $NID(0, \sigma^2)$, then

$$\frac{SS_W}{\sigma^2} \sim \chi^2_{I(J-1)}$$

If, additionally, $\alpha_i = 0$ for all $i = 1, \dots, I$, then

$$\frac{SS_B}{\sigma^2} \sim \chi^2_{I-1}$$

Furthermore, $\frac{SS_B}{\sigma^2}$ is independent of SS_W.

With this, we se tthe test statistics for the ANOVA null hypothesis $H_0 = \alpha_1 = \cdots = \alpha_I$ to be

$$F = \frac{SS_B/(I-1)}{SS_W/[I(J-1)]} \sim F_{I-1,I(J-1)}$$

We have

- $E[SS_W/[I(J-1)]] = \sigma^2$

- $E[SS_B/(I-1)] = J(I-1)^{-1}\sum_{i=1}^{I}\alpha_i^2 + \sigma^2$

Therefore, if H_0 is true, F should be close to 1. Otherwise, H_0 will be large. Theerfore, we reject large values of F. To be exact, we reject H_0 if $F > F_{I-1,I(J-1)}(\alpha)$. The exact p-value is given by

$$p - \text{value} = P(F_{I-1,I(J-1)} > F)$$

8.2 Groups with Different Size

Assume I groups with different size $J_1, \ldots, J_I$, it can be shown that

$$E(SS_W) = \sigma^2 \sum_{i=1}^{I}(J_i - 1)$$

and

$$E(SS_B) = (I-1)\sigma^2 + \sum_{i=1}^{I} J_i\alpha_i^2$$

Test statistics F still follows F distribution of df $\sum_{i=1}^{I} J_i - 1$ and $I - 1$.
If ANOVA is significant, it means the means are not ALL equal, but ANOVA does not tell us how they differ, and in particular we do not know which pairs are significantly different. To compare pairs or groups and estimate group means and differences, one naive approach is to use t test pairwise. However, we need many tests.