Statistical Models and Data Analysis

Phu Sakulwongtana

1 Too Many Distributions (And Its related Quantities)

1.1 Normal Distribution and Friends

Definition 1.1. **(Normal Distribution)** We define the normal distribution to be:

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Definition 1.2. **(Cumulative Normal Distribution)** We define CDF of normal distribution as:

$$\mathcal{N}(x \le y|\mu,\sigma^2) = \Phi\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right] \qquad \text{where} \qquad \operatorname{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x \exp(-t^2)\ \mathrm{d}t$$

Definition 1.3. **(Multinomial Cell Probabilities)** We consider $X_1, \dots, X_m$ the counts in cells $1, \dots, m$ follows multinomial distribution with total count of n and cell probabilities $p_1, \dots, p_m$ as we have:

$$p(X_1, \dots, X_m|p_1, \dots, p_m) = \frac{n!}{\prod_{i=1}^m X_i!}\prod_{i=1}^m p_i^{X_i}$$

The marginal distribution of each X_i that is binomial (n, p_i), and the joint frequency function isn't product of marignal frequency function.

1.2 Statistical Properties

Definition 1.4. **(Mean/Variance)** Mean and Variance of a random variable x are defined as:

$$\mathbb{E}[f(x)] = \int f(x)p(x)\ \mathrm{d}x \qquad \operatorname{var}(x) = \mathbb{E}[(x - \mathbb{E}[x])^2]$$

Definition 1.5. **(Covariance/Correlation Coefficient)** Covariance and Correlation coefficient between 2 variables are defined as:

$$\operatorname{cov}(x,y) = \mathbb{E}[(x-\mathbb{E}[x])(y-\mathbb{E}[y])] \qquad \rho = \frac{\operatorname{cov}(x,y)}{\sqrt{\operatorname{var}(x)\operatorname{var}(y)}}$$

Theorem 1.1. ***(Markov's Inequality)*** *If X is a random variable with $P(X \ge 0) = 1$ and for which $\mathbb{E}[X]$ exists then:*

$$\mathbb{P}(X \ge t) \le \frac{\mathbb{E}[X]}{t}$$

Proof. Consider the expectation:

$$\begin{aligned}\mathbb{E}[X] &= \int xp(x)\ \mathrm{d}x \\ &= \int_{x<t} xp(X)\ \mathrm{d}x + \int_{x \ge t} xp(X)\ \mathrm{d}x\end{aligned}$$

All the terms in the integral are non-negative because X takes only non-negative value, and so:

$$\begin{aligned}\mathbb{E}[X] &\geq \int_{x\geq t} xp(X)\ \mathrm{d}x \\ &\geq \int_{x\geq t} tp(x) = t\mathbb{P}(X\geq t)\end{aligned}$$

□

Theorem 1.2. ***(Chebyshev's Inequality)*** *Let X be a random variable with mean μ and σ^2. Then for any $t > 0$:*

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

Proof. We let $Y = (X-\mu)^2$. Then $\mathbb{E}[Y] = \sigma^2$ and this result follows from Markov inequality to Y. □

Theorem 1.3. ***(Law of Large Number)*** *Let $X_1, X_2, \dots, X_i, \dots$ be sequence of independent random variables with $\mathbb{E}[X_i] = \mu$ and $\operatorname{var}(X_i) = \sigma^2$. Let $\bar{X}_n = 1/n\sum_{i=1}^n X_i$. Then for any $\varepsilon > 0$:*

$$\mathbb{P}(\left|\bar{X}_n - \mu\right| > \varepsilon) \to 0 \qquad \textit{as} \qquad n\to\infty$$

Proof. Let's find the $\mathbb{E}[\bar{X}_n]$ and $\operatorname{var}(\bar{X}_n)$, and since X_i are independent

$$\mathbb{E}[\bar{X}_n] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_i] = \mu \qquad \operatorname{var}(\bar{X}_n) = \frac{1}{n^2}\sum_{i=1}^n \operatorname{var}(X_i) = \frac{\sigma^2}{n}$$

This follows from Chebyshev's inequality, which is:

$$\mathbb{P}(\left|\bar{X}_n - \mu\right| > \varepsilon) \leq \frac{\operatorname{var}(\bar{X}_n)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \to 0$$

as $n \to \infty$. Thus the thoerem is proven. □

Definition 1.6. (Convergence of Distribution Function) Let $X_1, X_2, \dots$ be a sequence of random variable with CDF $F_1, F_2, \dots$ and let X be random variable with distribution F. We say that X_n converge to X if:

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

at every point at which F is continuous.

Theorem 1.4. ***(Continuiy Theorem)*** *Let F_n be a sequence of CDF with the corresponding momement generating function M_n. Let F be a CDF with momement-generating funcion 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 1.5. ***(Central Limit Theorem)*** *Let $X_1, X_2, \dots$ be a sequence of independent random variable having mean 0 and variance σ^2 and the common distribution function F and momement-generating function M defined in a neighborhood of zero. Let:*

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

Then, we have:

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{S_n}{\sigma\sqrt{n}} \leq x\right) = \Phi(x) \qquad -\infty < x < \infty$$

Proof. Let $Z_n = S_n/(\sigma\sqrt{n})$. We will show that the mgf of Z_n tends to the mgf of the standard normal distribution. Since S_n is the sum of independent random variable:

$$M_{S_n}(t) = [M(t)]^n \qquad M_{Z_n}(t) = \left[M\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n$$

Consider tthe Taylor series expansion about zero, as we have:

$$M(s) = M(0) + sM'(0) + \frac{1}{2}s^2M''(0) + \varepsilon_s$$

Please note that $\varepsilon_s/s^2 \to 0$ as $s \to 0$. Since $\mathbb{E}[X] = 0, M'(0) = 0$ and $M''(0) = \sigma^2$. As $n \to \infty$, and $t/(\sigma\sqrt{n}) \to 0$ and:

$$M\left(\frac{t}{\sigma\sqrt{n}}\right) = 1 + \frac{1}{2}\sigma^2\sigma^2\left(\frac{t}{\sigma\sqrt{n}}\right)^2 + \varepsilon_n$$

Please note that $\varepsilon_n/(t^2/(n\sigma^2)) \to 0$ as $n \to \infty$, and we have:

$$M_{Z_n}(t) = \left(1 + \frac{t^2}{2n} + \varepsilon_n\right)^n$$

It can be shown that if $a_n \to a$, then we have:

$$\lim_{n\to\infty}\left(1 + \frac{a_n}{n}\right)^n = \exp(a)$$

From this result it follows that:

$$M_{Z_n}(t) \to \exp(t^2/2) \quad \text{as} \quad n \to \infty$$

And, so $\exp(t^2/2)$ is the mgf of the standard normal distribution, as we have shown. □

1.3 Quantities

Definition 1.7. (Sample Mean and Sample Variance) Let $X_1, \dots, X_n$ be independent $\mathcal{N}(\mu, \sigma^2)$ random variable. We refer to them as sample, and we denote sample mean $\bar{X}$ and sample variance S^2 to be:

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

We have $\mathbb{E}[\bar{X}] = \mu$ and $\text{var}(\bar{X}) = \sigma^2/n$.

Theorem 1.6. *The random variable $\bar{X}$ and the vector of random variables $(X_1 - \bar{X}, X_2 - \bar{X}, \dots, X_n - \bar{X})$ are independent. And so, $\bar{X}$ and S^2 are independently distributed.*

Proof. The proof will be based on momement-generating function:

$$M(s, t_1, \dots, t_n) = \mathbb{E}\left\{\exp\left[s\bar{X} + t_1(X_1 - \bar{X}) + \cdots + t_n(X_n - \bar{X})\right]\right\}$$

We observe that since:

$$\sum_{i=1}^n t_i(X_i - \bar{X}) = \sum_{i=1}^n t_iX_i - n\bar{X}\bar{t}$$

Then, we have:

$$s\bar{X} + \sum_{i=1}^n t_i(X_i - \bar{X}) = \sum_{i=1}^n \left[\frac{s}{n} + (t_i - \bar{t})\right]X_i = \sum_{i=1}^n a_iX_i$$

where we have $a_i = s/n + (t_i - \bar{t})$. Furthermore, we observe that:

$$\sum_{i=1}^n a_i = s \qquad \sum_{i=1}^n a_i^2 = \frac{s^2}{n} + \sum_{i=1}^n (t_i - \bar{t})^2$$

Now, we have $M(s, t_1, \dots, t_n) = M_{X_1,\dots,X_n}(a_1, \dots, a_n)$. Since X_i are independent normal random variable, we have:

$$\begin{aligned} M(s, t_1, \dots, t_n) &= \prod_{i=1}^n M_{X_i}(a_i) = \prod_{i=1}^n \exp\left(\mu a_i + \frac{\sigma^2}{2} a_i^2\right) \\ &= \exp\left(\mu \sum_{i=1}^n a_i + \frac{\sigma^2}{2} \sum_{i=1}^n a_i^2\right) \\ &= \exp\left[\mu s + \frac{\sigma^2}{2}\left(\frac{s^2}{n}\right) + \frac{\sigma^2}{2} \sum_{i=1}^n (t_i - \bar{t})^2\right] \\ &= \exp\left(\mu s + \frac{\sigma^2}{2n} s^2\right) \exp\left[\frac{\sigma^2}{2} \sum_{i=1}^n (t_i - \bar{t})^2\right] \end{aligned}$$

We can see that the first factor is mgf of $\bar{X}$. Since the mgf of the vector $(X_1 - \bar{X}, \dots, X_n - \bar{X})$ can be obtained by setting $s = 0$ in M, the factor is this mgf. Thus the prove is shown. □

1.4 Distribution from Normal Distribution

Definition 1.8. (χ^2-Distribution)

- If Z is a standard normal random variable, the distribution of $U = Z^2$ is called the chi-square distribution with 1 degree of freedom.
- If $U_1, U_2, \dots, U_n$ are independent 1 degree of freedom, the distribution of $V = U_1 + U_2 + \cdots + U_n$ is called χ^2-distribution with n degrees of freedom and it is denoted by χ^2_n.

We can see that the χ^2-square n-degree of is gamma distribution with $\alpha = n/2$ and $\lambda = 1/2$, so pdf is:

$$p(v) = \frac{1}{2^{n/2}\Gamma(n/2)} v^{(n/2)-1} \exp(-v/2)$$

for $v \geq 0$, and so $\mathbb{E}[V] = n$ and $\text{var}(V) = 2n$. Finally, it is clear that if $U \sim \chi^2_n$ and $V \sim \chi^2_m$, then we have $U + V \sim \chi^2_{m+n}$

Definition 1.9. (*T*-Distribution) If $Z \sim \mathcal{N}(0, 1)$ and $U \sim \chi^2_n$ and Z and U are independent, then the distribution of $Z/\sqrt{U/n}$ is called the t-distribution with n degrees of freedom. The density function of the t distribution with n degrees of freedom is:

$$p(t) = \frac{\Gamma[(n+1)/2]}{\sqrt{n\pi}\Gamma(n/2)} \left(1 + \frac{t^2}{n}\right)^{-(n+1)/2}$$

It is clear that $f(t) = f(-t)$, and so it is symmetric about zero. As the number of degree of freedom appraoches ∞ the t-distribution tends to standard normal distribution.

Definition 1.10. (*F*-Distribution) Let U and V be independent χ^2-distribution with m and n degrees of freedom. The distribution of:

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

is called F-distribution with m and n degrees of freedom, and is denoted by $F_{m,n}$, where its pdf is:

$$p(w) = \frac{\Gamma[(m+n)/2]}{\Gamma(m/2)\Gamma(n/2)} \left(\frac{m}{n}\right)^{m/2} w^{m/2-1} \left(1 + \frac{m}{n} w\right)^{-(m+n)/2}$$

One can show that, for $n > 2$ as $\mathbb{E}[W]$ exists and equal $n/(n-2)$. Finally, from the definition of t_n random variable follows an $F_{1,n}$ distribution.

Theorem 1.7. *The distribution of* $(n-1)S^2/\sigma^2$ *is* χ^2_{n-1}*-distribution*

Proof. Please note that:

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2 = \sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 \sim \chi^2_n$$

And, note that:

$$\begin{aligned}\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i=\mu)^2 &= \frac{1}{\sigma^2}\sum_{i=1}^{n}[(X_i-\bar{X})+(\bar{X}-\mu)]^2 \\ &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\bar{X})^2 + \left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)^2\end{aligned}$$

Note that $\sum_{i=1}^{n}(X_i-\bar{X}) = 0$. Now this relation is like $W = U + V$, as U and V are independent, we have $M_W(t) = M_U(t)M_V(t)$ as both W and V are χ^2-distribution, we have:

$$M_U(t) = \frac{M_W(t)}{M_V(t)} = \frac{(1-2t)^{-n/2}}{(1-2t)^{-1/2}} = (1-2t)^{-(n-1)/2}$$

The last expression is the mgf of a random variable with a χ^2_{n-1} distribution. □

Corollary 1.1. *We can show that:*

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

Proof. We can show that it is equivalent to the following ratio:

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} = \frac{\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{S^2/\sigma^2}}$$

The latter ratio is $\mathcal{N}(0,1)$ and the square root of χ^2_{n-1} distribution. And so from the definition is t_{n-1}. □

2 Estimation of Parameters

2.1 Method of Moments

Remark 1. Given the set of data $X_1, \dots, X_n$ sampled from a known distribution family but unknown parameter $P(x|\theta)$, we would like to this parameter.

Definition 2.1. (Moments) The k-th moment of probability is defined as $\mu_k = \mathbb{E}[X^k]$, where X is random variable following distribution. The sample moment is:

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

Definition 2.2. (Method of Moments) The method of moments estimates parameters by finding expression for them in terms of lowest possible order moments and substituting sample moments into the expression. Suppose there are 2 parameters, which can be expressed in terms of 2 moments as:

$$\theta_1 = f_1(\mu_1, \mu_2) \qquad \theta_2 = f_2(\mu_1, \mu_2)$$

Then method moments simply substitute the sample moment of the functions getting the parameter $\hat{\theta}_1, \hat{\theta}_2$.

Definition 2.3. (Sampling Distribution/Standard Error) It is natural question to aks to the distribution of the estimate, which is called *sampling distribution*, or the approximation to that distribution. The standard error is the standard deviation of sampling distribution.

Example 2.1. *We will consider the use of method of moments in 3 difference kinds of distribution:*

- *Poisson Distribution: This is simple as $\lambda = \mathbb{E}[X]$, so the parameter is set to:*

$$\hat{\lambda} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

To consider the sampling distribution, we have:

$$p(\hat{\lambda} = n) = p(S = nv) = \frac{(n\lambda_0)^{nv}\exp(-n\lambda_0)}{(nv)!}$$

Since $S = \sum_i X_i$ is Poisson, the mean and variance are both $n\lambda_0$, so we have $\mathbb{E}[\hat{\lambda}] = 1/n\mathbb{E}[S] = \lambda_0$ and $\text{var}(\hat{\lambda}) = \lambda_0/n$, and so the standard error is the square root of the variance.

- ***Normal Distribution:*** *We can see that $\mathbb{E}[X] = \mu$ and $\mathbb{E}[X^2] = \mu^2 + \sigma^2$, and so, we have:*

$$\begin{aligned} \hat{\mu} &= \hat{\mu}_1 = \bar{X} \\ \hat{\sigma}^2 &= \hat{\mu}_2 - \hat{\mu}_1^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 \end{aligned}$$

We can see that the sampling distribution of $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$ and $n\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-1}$.

- ***Gamma Distribution:*** *We can see that the first 2 moments are given as $\mathbb{E}[X] = \alpha/\lambda$ and $\mathbb{E}[X^2] = (\alpha(\alpha+1))/\lambda^2$. From the second equation: $\mu_2 = \mu_1^2 + \mu_1/\lambda$, and so we have:*

$$\hat{\lambda} = \frac{\hat{\mu_1}}{\hat{\mu}_2 - \hat{\mu}_1^2} \qquad \hat{\alpha} = \frac{\hat{\mu}_1^2}{\hat{\mu}_2 - \hat{\mu}_1^2}$$

The sampling distribution can be hard to find. We will have to use boostrapping to do this.

Definition 2.4. (Bootstrap) We can samping *with replacement* of the data, and we calculate the parameter via any mean. The distribution of the parameter is the approximation of the sampling distribution. This method is called bootstrap.

Definition 2.5. (Consistent) Let $\hat{\theta}_n$ be an estimate of parameter θ based on sample of size n. Then $\hat{\theta}_n$ is said to be consistent in probability if $\hat{\theta}_n$ converges in probability to θ as n appraoches infinity, that is for any $\varepsilon > 0$:

$$\mathbb{P}\left(\left|\hat{\theta}_n - \theta\right| > \varepsilon\right) \to 0 \quad \text{as} \quad n \to \infty$$

The weak law of large number implies the sample moment converge in probability to population moment.

2.2 Maximum Likelihood

Definition 2.6. (Maximum Likelihood) We will assume the data X_i to be iid, and so the log-likelihood:

$$l(\theta) = \log\prod_{i=1}^{n} p(X_i|\theta) = \sum_{i=1}^{n}\log f(X_i|\theta)$$

Example 2.2. *We will consider difference distributions and its maximum likelihood estimate:*

- ***Poisson Distribution:*** *The log-likelihood of the Poisson distribution is:*

$$l(\lambda) = \sum_{i=1}^{n}(X_i \log \lambda - \lambda - \log X_i!) = \log \lambda \sum_{i=1}^{n} X_i - n\lambda - \sum_{i=1}^{n} \log X_i!$$

We can see that its derivative is given as:

$$l'(\lambda) = \frac{1}{\lambda}\sum_{i=1}^{n} X_i - n = 0$$

The MLE is equal to $\hat{\lambda} = \bar{X}$

- ***Normal Distribution:*** *The log-likelihood is given by*

$$\begin{aligned} l(\mu, \sigma^2) &= \log \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right) \\ &= -n \log \sigma - \frac{n}{2}\log 2\pi - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2 \end{aligned}$$

This leads to the following derivative:

$$\frac{\partial l}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu) \qquad \frac{\partial l}{\partial \sigma} = -\frac{n}{\sigma} + \sigma^{-3}\sum_{i=1}^{n}(X_i - \mu)^2$$

Setting the derivative to zero, and we have $\hat{\mu} = \bar{X}$ *and we substitute the MLE for* μ *for* σ *as we have* $\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$. *The sampling distribution is the same as method of moment.*

- ***Gamma Distribution:*** *The log-likelihood is given by:*

$$\begin{aligned} l(\alpha, \lambda) &= \log \prod_{i=1}^{n} \frac{1}{\Gamma(\alpha)} \lambda^{\alpha} X_i^{\alpha - 1} \exp(-\lambda X_i) \\ &= n\alpha \log \lambda + (\alpha - 1)\sum_{i=1}^{n} \log X_i - \lambda \sum_{i=1}^{n} X_i - \lambda \sum_{i=1}^{n} X_i - n \log \Gamma(\alpha) \end{aligned}$$

for $0 \le x < \infty$. *Now, we have the following derivative:*

$$\frac{\partial l}{\partial \alpha} = n \log \lambda + \sum_{i=1}^{n} \log X_i - n\frac{\Gamma'(\alpha)}{\Gamma} \qquad \frac{\partial l}{\partial \lambda} = \frac{n\alpha}{\lambda} - \sum_{i=1}^{n} X_i$$

Setting the second partial to zero as $\hat{\lambda} = (n\hat{\alpha})/(\sum_{i=1}^{n} X_i) = \hat{\alpha}/\bar{X}$. *Now* α *can be solved by non-linear equation via iterative method:*

$$n \log \hat{\alpha} - n \log \bar{X} + \sum_{i=1}^{n} \log X_i - n\frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})} = 0$$

The sampling distribution can be found by bootstrapping.

- ***Multinomial-Cell Distribution:*** *We have the following log-likelihood to be:*

$$l(p_1, \dots, p_m) = \log \frac{n!}{\prod_{i=1}^{m} X_i!} \prod_{i=1}^{m} p_i^{X_i} = \log n! - \sum_{i=1}^{m} \log X_i! + \sum_{i=1}^{m} x_i \log p_i$$

Maximizing the likelihood would be subject to contraint as we have have the following Lagragian:

$$\mathcal{L}(p_1, \dots, p_m) = \log n! - \sum_{i=1}^{m} \log x_i! + \sum_{i=1}^{m} x_i \log p_i + \lambda\left(\sum_{i=1}^{m} p_i - 1\right)$$

Setting the partial derivative to be equal to zero:

- *As we have the following system of equation:* $\hat{p}_j = -x_j/\lambda$ *for* $j = 1, \ldots, m$ *summing both equation as we have:* $1 = -n/\lambda$ *or* $\lambda = -n$ *and so* $\hat{p}_j = x_j/n$*.*
- *The sampling distribution of* $\hat{p}_j$ *is determined by the distribution of* x_j*, which is biomial.*

Theorem 2.1. *Under appropriate smoothness conditions on* f*, the MLE from an iid sample is consistent.*

Proof. Consider maximizing the following values, given the $X_1, X_2, \ldots, X_n \sim p(X|\theta_0)$:

$$\frac{1}{n}l(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log p(X_i|\theta)$$

as n tends to infinity, the law of large number implies that:

$$\begin{aligned}\frac{1}{n}l(\theta) &\to \mathbb{E}_{X\sim p(X|\theta_0)}[\log p(X|\theta)] \\ &= \int p(x|\theta_0) \log p(x|\theta) \ \mathrm{d}x\end{aligned}$$

The θ that maximizes $l(\theta)$ should be closed to the θ that maximizes $\mathbb{E}[\log f(X|\theta)]$ (again not shown). We consider the derivative:

$$\frac{\partial}{\partial\theta}\int p(x|\theta_0)\log p(x|\theta) \ \mathrm{d}x = \int p(x|\theta)\frac{p(x|\theta_0)}{p(x|\theta)}\frac{\partial}{\partial\theta} \ \mathrm{d}x$$

If $\theta = \theta_0$, this equation becomes:

$$\int \frac{\partial}{\partial\theta}p(x|\theta_0) \ \mathrm{d}x = \frac{\partial}{\partial\theta}\int p(x|\theta_0) \ \mathrm{d}x = \frac{\partial}{\partial\theta}(1) = 0$$

This shows that θ_0 is stationary and (hopefully) it is a maximum. The assumption of smoothness on f must be strong enough to justify this. □

Lemma 2.1. *Define* $I(\theta)$ *by:*

$$I(\theta) = \mathbb{E}\left[\frac{\partial}{\partial\theta}\log p(X|\theta)\right]^2 = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log p(X|\theta)\right]$$

Under appropriate smoothness conditions on p, this can be expressed on the right-hand side.

Proof. Observe that $\int p(x|\theta) \ \mathrm{d}x = 1$, and so we have, the following observation:

$$\frac{\partial}{\partial\theta}\int p(X|\theta) \ \mathrm{d}x = 0 \qquad \frac{\partial}{\partial\theta}p(x|\theta) = p(x|\theta)\left[\frac{\partial}{\partial\theta}\log p(x|\theta)\right]$$

Combinding this with identity, as we have (take the second derivative to be):

$$\begin{aligned}0 = \frac{\partial}{\partial\theta}\int p(x|\theta) \ \mathrm{d}x &= \int \left[\frac{\partial}{\partial\theta}\log p(x|\theta)\right]p(x|\theta) \ \mathrm{d}x \\ &= \int \left[\frac{\partial^2}{\partial\theta^2}\log p(x|\theta)\right]p(x|\theta) \ \mathrm{d}x + \int \left[\frac{\partial}{\partial\theta}\log p(x|\theta)\right]^2 p(x|\theta) \ \mathrm{d}x\end{aligned}$$

And so we have the lemma is proven. □

Theorem 2.2. *Under smoothness condition on* f*, the probability distribution of* $\sqrt{nI(\theta_0)}(\hat{\theta} - \theta_0)$ *thends to a standard normal distribution*

Proof. The following is sketch of proof. Consider the Taylor series expansion (of $l'(\hat{\theta})$), as we have:

$$\begin{aligned} 0 = l'(\hat{\theta}) &\approx l'(\theta_0) + (\hat{\theta} - \theta_0)l''(\theta) \\ (\hat{\theta} - \theta_0) &\approx \frac{-l'(\theta_0)}{l''(\theta_0)} \\ \sqrt{n}(\hat{\theta} - \theta_0) &\approx \frac{-n^{-1/2}l'(\theta_0)}{n^{-1}l''(\theta_0)} \end{aligned}$$

We consider the numeraor of this last expression. Its expectation is given as:

$$\mathbb{E}\left[n^{-1/2}l'(\theta_0)\right] = n^{-1/2}\sum_{i=1}^{n}\mathbb{E}\left[\frac{\partial}{\partial\theta}\log p(X_i|\theta_0)\right] = 0$$

As we have θ_0, which is the fixed point (see thoerem above). Now, consider the variance of the quantity:

$$\text{var}\left[n^{-1/2}l'(\theta_0)\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log p(x_i|\theta_0)\right]^2 = I(\theta_0)$$

Consider the denominator to be. Together with the law of large number, the expression converges to:

$$\frac{1}{n}l''(\theta_0) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta^2}\log p(x_i|\theta_0) \longrightarrow \mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log p(x|\theta_0)\right] = -I(\theta_0)$$

Thus, we have:

$$n^{1/2}(\hat{\theta} - \theta_0) \approx \frac{n^{-1/2}l'(\theta_0)}{I(\theta_0)}$$

We have the following mean and variance of the ratio to be:

$$\begin{aligned} \mathbb{E}[n^{1/2}(\hat{\theta} - \theta_0)] &\approx 0 \\ \text{var}[n^{1/2}(\hat{\theta} - \theta_0)] &\approx \frac{I(\theta_0)}{I^2(\theta_0)} = \frac{1}{I(\theta_0)} \end{aligned}$$

And so we have $\text{var}(\hat{\theta} - \theta_0) \approx 1/(nI(\theta_0))$. Thus the equation is proven. □

Remark 2. For an iid sample, the MLE is the maximizer of the log-likelihood function $l(\theta) = \sum_{i=1}^{n}\log p(X_i|\theta)$ has the asymptotic variance that is given as:

$$\frac{1}{nI(\theta_0)} = -\frac{1}{\mathbb{E}[l''(\theta_0)]}$$

When $\mathbb{E}[l''(\theta_0)]$ is large, meaning that $l(\theta)$ is changing very rapidly in a vincinity of θ_0 and the variance of the maximizer is small.

Remark 3. **(Confidence Interval for Mean and Variance Estimate)** Consider the maximum likelihood estimate of μ and σ^2 from an iid normal sample to be:

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

There are various confidence interval on each of the likelihood estimation as we have:

- Confidence interval of μ is based on:

$$\frac{\sqrt{n}(\bar{X} - \mu)}{S} \sim t_{n-1} \qquad \text{where} \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Let $t_{n-1}(\alpha/2)$ denote the point beyound which t distribution with $n-1$ degree of freedom has probability $\alpha/2$, to be:

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

The inequality can be manipulated to yields:

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

The probability that μ lies in the interval is $1-\alpha$.

- Let's consider the conditional interval σ^2, as we have the following distribution:

$$\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1}$$

Let $\chi^2_m(\alpha)$ denote the point beyound which the chi-square distribution with m degree of freedom that has probability α:

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

Manipulation of the inequality yields:

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

- For a general maximum likelihood methods, one can consider the distribution of $\sqrt{nI(\hat{\theta})}(\hat{\theta}-\theta_0)$, where it is normally distributed, and so we have the following intervales:

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

which we can yields the confidence interval, as we have:

$$\mathbb{P}\left(-\frac{z(\alpha/2)}{\sqrt{nI(\hat{\theta})}} \leq \theta_0 \leq \frac{z(\alpha/2)}{\sqrt{nI(\hat{\theta})}}\right)$$

- For the estimation for *random multinomial*. The counts are not iid, so the variance of the parameter estimate is of the form $1/[nI(\theta)]$ can't be used. It can be shown that:

$$\text{var}(\hat{\theta}) \approx \frac{1}{\mathbb{E}[l'(\theta_0)^2]} = -\frac{1}{\mathbb{E}[l''(\theta_0)]}$$

Please note that this is used to construct the confidence interval instead of above.

2.3 Cramer-Rao Lower Bound

Definition 2.7. (Efficiency of Estimates) Given 2 estimates $\hat{\theta}$ and $\tilde{\theta}$ of a parameter θ, the efficiency of $\hat{\theta}$ and $\tilde{\theta}$ is defined to be:

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

Theorem 2.3. *Let $X_1, \dots, X_n$ be iid with density function $p(x|\theta)$. Let $T = t(X_1, \dots, X_n)$ be unbiased estimate of θ. Then under smoothness assumption on $p(x|\theta)$, we have:*

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

Proof. Let the following value:

$$Z = \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log p(X_i|\theta) = \sum_{i=1}^{n} \frac{1}{p(X_i|\theta)} \frac{\partial}{\partial \theta} p(X_i|\theta)$$

We already show that $\mathbb{E}[Z] = 0$. Because the correlation coefficient of Z and T is less than or equal to 1 in absolute value as:

$$\text{cov}^2(Z, T) \le \text{var}(Z)\,\text{var}(T)$$

Furthermore, we have shown that (from the lemma of $I(\theta)$):

$$\text{var}\left[\frac{\partial}{\partial \theta} \log p(X|\theta)\right] = I(\theta)$$

and so $\text{var}(Z) = nI(\theta)$. The proof will be complete by showing that $\text{cov}(Z, T) = 1$. Please note that (follows product rule):

$$\left(\sum_{i=1}^{n} \frac{1}{p(X_i|\theta)} \frac{\partial}{\partial \theta} p(X_i|\theta)\right)\left(\prod_{j=1}^{n} f(x_j|\theta)\right) = \frac{\partial}{\partial \theta} \prod_{i=1}^{n} f(x_i|\theta)$$

Since Z has mean of 0, we have:

$$\begin{aligned}
\text{cov}(Z, T) &= \mathbb{E}[ZT] \\
&= \int \cdots \int t(x_1, \dots, x_n) \left[\sum_{i=1}^{n} \frac{1}{p(X_i|\theta)} \frac{\partial}{\partial \theta} p(X_i|\theta)\right] \prod_{j=1}^{n} f(x_j|\theta) \ \mathrm{d}x_j \\
&= \int \cdots \int t(x_1, \dots, x_n) \frac{\partial}{\partial \theta} \prod_{i=1}^{n} f(x_i|\theta) \ \mathrm{d}x_i \\
&= \frac{\partial}{\partial \theta} \int \cdots \int t(x_1, \dots, x_n) \prod_{i=1}^{n} f(x_i|\theta) \ \mathrm{d}x_i \\
&= \frac{\partial}{\partial \theta} \mathbb{E}[T] = \frac{\partial}{\partial \theta} \theta = 1
\end{aligned}$$

This proves the inequality as we have. □

Definition 2.8. (Efficient) The unbiased estimate whose variance achieves this lower bound is said to be efficient. Since the asymptotic variance of maximum likelihood estimate is equal to lower bound, it is said to be asymptotically efficient.

2.4 Sufficient Statistics

Definition 2.9. A statistics $T(X_1, \dots, X_n)$ is said to be sufficient for θ if conditional distribution of $X_1, \dots, X_n$ given $T = t$ doesn't depends on θ or any value of t.

Theorem 2.4. *A necessary and sufficient condition for $T(X_1, \dots, X_n)$ to be sufficient for a parameter θ is the joint probability function factors in the form of:*

$$p(x_1, \dots, x_n|\theta) = g[T(x_1, \dots, x_n), \theta] h(x_1, \dots, x_n)$$

Proof. We will consider it to be in discrete case. Suppsoe that the frequency function factors. To simplify notation, we let $\boldsymbol{X}$ denotes $(X_1, \dots, X_n)$ and $\boldsymbol{x}$ denotes $(x_1, \dots, x_n)$. We have:

$$\begin{aligned}
P(T = t) &= \sum_{T(x)=t} P(\boldsymbol{X} = \boldsymbol{x}) \\
&= g(t, \theta) \sum_{T(x)=t} h(\boldsymbol{x})
\end{aligned}$$

We then have:

$$P(\boldsymbol{X} = \boldsymbol{x}|T = t) = \frac{P(\boldsymbol{X} = \boldsymbol{x}, T = t)}{P(T = t)} = \frac{h(\boldsymbol{x})}{\sum_{T(\boldsymbol{X})=t} h(\boldsymbol{x})}$$

This conditional distributed doesn't depend on θ. To show that the conclusion holds in other direction, suppose that the conditional distribution of $\boldsymbol{X}$ given T is independent of θ. Let:

$$g(t, \theta) = P(T = t|\theta) \qquad h(\boldsymbol{x}) = P(\boldsymbol{X} = \boldsymbol{x}|T = t)$$

We then have:

$$\begin{aligned} P(\boldsymbol{X} = \boldsymbol{x}|\theta) &= P(T = t|\theta)P(\boldsymbol{X} = \boldsymbol{x}|T = t) \\ &= g(t, \theta)h(\boldsymbol{x}) \end{aligned}$$

□

Corollary 2.1. *If T is sufficient for θ, the MLE is a function of T.*

Proof. The likelihood is $g(T, \theta)h(\boldsymbol{x})$, which depends on θ only through T. To maximize this quantity, we need to maximize $g(T, \theta)$ □

Theorem 2.5. ***(Rao-Blackwell Theorem)*** *Let $\hat{\theta}$ be an estimator of θ with $\mathbb{E}[\hat{\theta}^2] < \infty$ for all θ. Suppose that T is sufficient statistics for θ, and let $\tilde{\theta} = \mathbb{E}[\hat{\theta}|T]$, then for all θ:*

$$\mathbb{E}[\tilde{\theta} - \theta]^2 \le \mathbb{E}[\hat{\theta} - \theta]^2$$

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

Proof. First note that from the property of iterated condition expectation, we have:

$$\mathbb{E}[\tilde{\theta}] = \mathbb{E}[\mathbb{E}[\hat{\theta}|T]] = \mathbb{E}[\hat{\theta}]$$

To compare the square-error, we will have to only consider their varince:

$$\begin{aligned} \text{var}(\hat{\theta}) &= \text{var}[\mathbb{E}[\hat{\theta}|T]] + \mathbb{E}[\text{var}[\hat{\theta}|T]] \\ &= \text{var}(\tilde{\theta}) + \mathbb{E}[\text{var}(\hat{\theta}|T)] \end{aligned}$$

Thus $\text{var}(\hat{\theta}) > \text{var}(\tilde{\theta})$ unless $\text{var}(\hat{\theta}|T) = 0$, which is when $\hat{\theta}$ is a function of T, which implies $\hat{\theta} = \tilde{\theta}$. □

3 Testing Hypothesis and Goodness of Fit

3.1 Introduction

Definition 3.1. (Likelihood Ratio) Consider the two hypothesis to be H_0 and H_1, we have the following, posterior:

$$P(H_0|x) = \frac{P(x|H_0)P(H_0)}{P(x)} \qquad P(H_1|x) = \frac{P(x|H_1)P(H_1)}{P(x)}$$

The ratio is given as:

$$\frac{P(H_0|x)}{P(H_1|x)} = \frac{P(H_0)}{P(H_1)}\frac{P(x|H_0)}{P(x|H_1)}$$

This is the product of the ratio of prior probability and the likelihood ratio. Now, we would like to choose the hypothesis H_0 if, we have:

$$\frac{P(H_0|x)}{P(H_1|x)} = \frac{P(H_0)}{P(H_1)}\frac{P(x|H_0)}{P(x|H_1)} > 1 \quad \iff \quad \frac{P(x|H_0)}{P(x|H_1)} > c$$

where value of c depends upon your prior probability.

Definition 3.2. (Neyman-Pearson Paradigm) One hypothesis is singled out as *null hypothesis* H_0 and other as *alternative hypothesis* H_1. We have the following terminology as:

- Rejecting H_0 when it is true is called *type I error*.
- Probability of a type I error is called *significance level* and it is denoted as α.
- Accepting the null hypothesis when it is false is called *type II error*, and it is denoted by β.
- The probability that the null hypothesis is rejected when it is false is called *power* of the test, which is equal to $1-\beta$.
- The likelihood ratio is called the *test statistics*.
- Set of values of the test statistics that leads to rejection of the null hypothesis is called *rejection region*, and set of values that leads to acceptance is called *acceptance region*
- The probability distribution of test statistics when the null hypothesis is true is called *null distribution*.

Definition 3.3. (Simple Hypothesis) If the null and alternative hypothesis each completely specify the probability distribution. This kind of setting is called simple hypothesis.

Lemma 3.1. ***(Neyman-Peason)*** *Suppose that H_0 and H_1 are simple hypothesis:*

- *The test that rejects H_0 whenever the likelihood ratio is less than c and significance level α.*
- *Then any other test for which significance level is less than or equal to α has power less than or equal to that of the likelihood ratio test.*

Proof. Let $p(x)$ denote the pdf or frequency function of the observation.

- A test of $H_0 : p(x) = p_0(x)$ and $H_1 : p(x) = p_1(x)$ amounts to using a decision function:

$$d(x) = \begin{cases} 0 & \text{if } H_0 \text{ is accepted} \\ 1 & \text{if } H_1 \text{ is rejected} \end{cases}$$

- Since $d(X)$ is a Bernoulli random varaible, where we have:
  - Significance Level: $\mathbb{E}_0[d(X)] = P_0(d(X) = 1)$
  - Power: $\mathbb{E}_1[d(X)] = P_1(d(X) = 0)$
- If we consider the likelihood ratio test as the decision function:

$$d(x) = \begin{cases} 1 & \text{if } p_0(X) < cp_1(X) \\ 0 & \text{otherwise} \end{cases}$$

  Please note that $\mathbb{E}_{X\sim p_0(X)}[X] = \alpha$.
- Let $d^*(X)$ be the decision function of another test satisfying $\mathbb{E}_0[d^*(X)] \le \mathbb{E}_0[d^*(x)] = \alpha$.
- Consider the following inequalities:

$$d^*(x)[cp_1(x) - p_0(x)] \le d(x)[cp_1(x) - p_0]$$

  This follows from the $d(x) = 1$, where $cf_1(x) - f_0(x) > 0$ and if $d(x) = 0$. where $cf_1(x) - f_0(x) \le 0$

- Integrating the both sides of the inequality above with respected to x as:

$$c\mathbb{E}_1[d^*(X)] - \mathbb{E}_0[d^*(X)] \le c\mathbb{E}_1[d(X)] - \mathbb{E}_0[d(X)]$$

  and, so we have:

$$\mathbb{E}_0[d(X)] - \mathbb{E}_0[d^*(X)] \le c\left[\mathbb{E}_1[d(X)] - \mathbb{E}_1[d^*(X)]\right]$$

  Since the LHS of this inequality is non-negative, we have: $\mathbb{E}[d^*(X)] \le \mathbb{E}_A[d(X)]$

□

Example 3.1. ***(First Test)*** *Consider $X_1, \dots, X_n$ be random sample from normal distribution, with unknown mean and variance σ^2. Given 2 hypothesis:*

$$H_0 : \mu = \mu_0 \qquad H_1 : \mu = \mu_1$$

where μ_1 and μ_0 are constant. Consider a significance level of α. Then consider likelihood ratio:

$$\frac{f_0(\boldsymbol{X})}{f_1(\boldsymbol{X})} = \frac{\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^n (X_i - \mu_0)^2\right]}{\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^n (X_i - \mu_1)^2\right]}$$

To consider the ratio, we consider the value of $\sum_{i=1}^n (X_i - \mu_1)^2 - \sum_{i=1}^n (X_i - \mu_0)^2$. Expanding the squares:

$$2n\bar{X}(\mu_0 - \mu_1) + n\mu_1^2 - n\mu_0^2$$

There are 2 conditions, so that the likelihood is small:

- *If $\mu_0 - \mu_1 > 0$, the likelihood ratio is small if $\bar{X}$ is small.*
- *If $\mu_0 - \mu_1 < 0$, the likelihood ratio is small if $\bar{X}$ is large.*

Let's consider the later case. Likelihood-ratio rejects for $\bar{X} > x_0$ for some x_0, which we will choose it to give a test of desired level α. This means choosing $\mathbb{P}(\bar{X} > x_0) = \alpha$ if H_0 is true:

$$\mathbb{P}(\bar{X} > x_0) = \mathbb{P}\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} > \frac{x_0 - \mu_0}{\sigma/\sqrt{n}}\right)$$

The null distribution of $\bar{X}$ is a normal distribution with mean μ_0 and variance σ^2/n, then, we can solve:

$$\frac{x_0 - \mu_0}{\sigma/\sqrt{n}} = z(\alpha)$$

for x_0 in order to find the rejection region for level α test.

Definition 3.4. (P-Value) As we can see, the testing requires only the null distribution, and we are required to consider the significance level α (which should be 0.01 and 0.05). P-value is the smallest significance level at which the null hypothesis would be rejected.

3.2 More Complex Hypothesis Testing

Definition 3.5. (Uniformly Most Powerful) If the alternative hypothesis H_1 is composite, a test is most powerful for every simple alternative in H_1 is said to be uniformly most powerful.

Example 3.2. ***(2-Sided Test)*** *Consider $X_1, \dots, X_n$ be random sample from normal distribution, with unknown mean and variance σ^2. Given 2 hypothesis:*

$$H_0 : \mu = \mu_0 \qquad H_1 : \mu \neq \mu_0$$

Please note that in this example, this kind of hypothesis is called two-sided alternative. Consider the test at a specific level α that reject for $\left|\bar{X} - \mu_0\right| > x_0$, where x_0 is determined such that $\mathbb{P}(\left|\bar{X} - \mu_0\right| > x_0) = \alpha$ if H_0 is true. We can see that $x_0 = z(\alpha/2)\sigma/\sqrt{n}$:

$$\begin{aligned} &\left|\bar{X} - \mu_0\right| < \frac{z(\alpha/2)\sigma}{\sqrt{n}} \\ \iff &\bar{X} - \frac{z(\alpha/2)\sigma}{\sqrt{n}} \le \mu_0 < \bar{X} + \frac{z(\alpha/2)\sigma}{\sqrt{n}} \end{aligned}$$

A $100(1-\alpha)\%$ interval for μ is give, and so if μ_0 is in the interval, then we accept the null hypothesis.

Theorem 3.1. *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 $A(\theta_0)$. Then set:*

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

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

Remark 4. This means that a $100(1-\alpha)\%$ confidence region for θ consists of all those values of θ_0 for which the hypothesis that θ equals θ_0 will not be rejected at level α.

Proof. Because A is the acceptance region of a test at level α:

$$\mathbb{P}[\boldsymbol{X} \in A(\theta_0) | \theta = \theta_0] = 1 - \alpha$$

Now, we have:

$$\mathbb{P}[\theta_0 \in C(\boldsymbol{X}) | \theta = \theta_0] = \mathbb{P}[\boldsymbol{X} \in A(\theta_0) | \theta = \theta_0] = 1 - \alpha$$

by the definition of $C(\boldsymbol{X})$ □

Definition 3.6. (Generalized Likelihood Ratio Test) Suppose that the observation: $\boldsymbol{X} = (X_1, \dots, X_n)$ have a joint density $p(\boldsymbol{x}|\theta)$:

- Then H_0 may specify that $\theta \in \omega_0$ where ω_0 is subset of all possible values of θ
- For H_1 we consider ω_1 is disjoint from ω_0.

Let $\Omega = \omega_0 \cup \omega_1$. The generalized likelihood ratio is Λ^* or with the truncated version Λ as the small value of Λ^* tends to discredit H_0:

$$\Lambda^* = \frac{\max_{\theta \in \omega_0} l(\theta)}{\max_{\theta \in \omega_1} l(\theta)} \qquad \Lambda = \frac{\max_{\theta \in \omega_0} l(\theta)}{\max_{\theta \in \Omega} l(\theta)}$$

Note that $\Lambda = \min(\Lambda^*, 1)$. The rejection region is given as $\Lambda \le \lambda_0$, where the threshold λ_0 is choosen so that

$$\mathbb{P}(\Lambda \le \lambda_0 | H_0) = \alpha$$

Example 3.3. ***(Testing Normal Mean)*** *Consider $X_1, \dots, X_n$ be random sample from normal distribution, with unknown mean and variance σ^2. Given 2 hypothesis:*

$$H_0 : \mu = \mu_0 \qquad H_1 : \mu \neq \mu_0$$

We have the following specification:

$$\omega_0 = \{\mu_0\} \qquad \omega_1 = \{\mu | \mu \neq \mu_0\} \qquad \Omega = \{-\infty < \mu < \infty\}$$

If we maximize over ω_0, as it has only one point, the numerator. For the denominator, we it is clear that the MLE is $\bar{X}$ and so:

$$\max_{\theta\omega_1} l(\theta) = \frac{1}{(2\sigma\pi)^n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu_0)^2\right) \qquad \max_{\theta\Omega} l(\theta) = \frac{1}{(2\sigma\pi)^n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\bar{X})^2\right)$$

The ratio is given as:

$$\begin{aligned}\Lambda &= \exp\left(-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(X_i-\mu_0)^2 - \sum_{i=1}^{n}(X_i-\bar{X})^2\right]\right) \\ \iff -2\log\Lambda &= \frac{1}{\sigma^2}\left(\sum_{i=1}^{n}(X_i-\mu_0)^2 - \sum_{i=1}^{n}(X_i-\bar{X})^2\right) \\ &= \frac{n(\bar{X}-\mu_0)^2}{\sigma^2}\end{aligned}$$

Rejecting for small value of Λ is equivalent to reject the large value of $-2\log\Lambda$. Together with the identity that $\sum_{i=1}^{n}(X_i-\mu_0)^2 = \sum_{i=1}^{n}(X_i-\bar{X})^2 + n(\bar{X}-\mu_0)^2$. It follows that, under H_0:

- $\bar{X} \sim \mathcal{N}(\mu_0, \sigma^2/n)$*, which implies that* $\sqrt{n}(\bar{X}-\mu_0)/\sigma \sim \mathcal{N}(0,1)$
- $-2\log\Lambda \sim \chi_1^2$ *is implied from above.*

We can now construct the rejection region for any significance level to be:

$$\frac{n}{\sigma^2}(\bar{X}-\mu_0)^2 > \chi_1^2(\alpha)$$

where $P(Z > \chi_1^2(\alpha)) = \alpha$, recall that we are rejecting the large value of $-2\log\Lambda$. This links back to the original consideration as, we this inequality is equivalent to:

$$\left|\bar{X}-\mu_0\right| \geq \frac{\sigma}{\sqrt{n}} z(\alpha/2)$$

Theorem 3.2. *Under smoothness condition on the probability density, the null distribution of $-2\log\Lambda$ tends to a chi-square distribution with degree of freedom of $\dim\Omega - \dim\omega_0$ as the sample size tends to infinity.*

Example 3.4. ***(Tests for Multinomial Distribution/Goodness of Fit)*** *We consider the following testing scenario*

- H_0*: The cell probabilities $p = p(\theta)$ for $\theta \in \omega_0$ (maybe unknown, with dimension of k) is constrained on some way.*
- H_1*: Cell probabilities are free except the constriants such that they are non-negative and sum to 1.*

We have Ω to be set of m non-negative numbers that sum to one. We have:

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

where x_i are observed counts in m cells. We will denote $\hat{\theta}$ as the MLE of θ. For the denominator, with unrestricted MLE, we have $\hat{p}_i = x_i/n$, and so, the ratio is:

$$\begin{aligned}\Lambda &= \frac{\frac{n!}{x_1!\cdots x_m!}p_1(\hat{\theta})^{x_1}\cdot\hat{\theta})^{x_m}}{\frac{n!}{x_1!\cdots x_m!}\hat{p}_1^{x_1}\cdots\hat{p}_m^{x_m}} = \prod_{i=1}^{m}\left(\frac{p_i(\hat{\theta})}{\hat{p}_i}\right)^{x_i} \\ \implies -2\log\Lambda &= -2n\sum_{i=1}^{m}\hat{p}_i\log\left(\frac{p_i(\hat{\theta})}{\hat{p}_i}\right) = 2\sum_{i=1}^{m}O_i\log\left(\frac{O_i}{E_i}\right)\end{aligned}$$

As we have $x_i = n\hat{p}_i$, $O_i = n\hat{p}_i$ and $E_i = np_i(\hat{\theta})$. Let's consider the test statistics:

- *Ω allows cell probability to be free (but have to be sum to 1) so* $\dim \Omega = m - 1$.

- *$p_i(\hat{\theta})$ denpends on k-dimensional parameter θ so* $\dim \omega_0 = k$

The large sample theory, tells us that the distribution of $-2 \log \Lambda$ is χ^2_{m-k-1}.

Definition 3.7. (Peason's Chi-Square Statistics) It is a commonly used to test for goodness of fit, where:

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

Proposition 3.1. *Peason's statistics and likelihood tests are asymptoticall equivalent under H_0*

Proof. (Sketch) Starting with the value:

$$-2 \log \Lambda = 2n \sum_{i=1}^{m} \hat{p}_i \log \left(\frac{\hat{p}_i}{p_i(\hat{\theta})} \right)$$

If H_0 is true and n is large, then $\hat{p}_i \approx p_i(\hat{\theta})$. Consider the following Taylor series expansion of:

$$\begin{aligned} f(x) &= x \log \left(\frac{x}{x_0} \right) \\ &= (x - x_0) + \frac{1}{2}(x - x_0)^2 \frac{1}{x_0} + \cdots \end{aligned}$$

Thus, we have:

$$-2 \log \Lambda \approx 2n \sum_{i=1}^{m} [\hat{p}_i - p_i(\hat{\theta})] + n \sum_{i=1}^{m} \frac{[\hat{p}_i - p_i(\hat{\theta})]^2}{p_i(\hat{\theta})}$$

The first term is zero due to the fact that probabilities sum to 1, while the second terms is equal to Peason's statistics. Note that Peason's statistics is easier to calculate than the likelihood ratio test. □

Example 3.5. ***(Poisson Dispersion Test)*** *Gives counts $x_1, \ldots, x_n$, we consider:*

- *H_0: The counts are poisson with common parameter λ. Under ω_0 the MLE of λ is $\hat{\lambda} = \bar{X}$.*

- *H_1: The counts have difference rates $\lambda_1, \ldots, \lambda_n$. Under Ω we have $\tilde{\lambda_i} = x_i$*

Please note that $\omega_0 \subset \Omega$ is the speical case that they are all equal, so the likelihood ratio is:

$$\begin{aligned} \Lambda &= \frac{\prod_{i=1}^{n} \hat{\lambda}^{x_i} \dfrac{\exp(-\hat{\lambda})}{x_i!}}{\prod_{i=1}^{n} \tilde{\lambda}^{x_i} \dfrac{\exp(-\tilde{\lambda})}{x_i!}} = \prod_{i=1}^{n} \left(\frac{\bar{x}}{x_i} \right)^{x_i} \exp(x_i - \bar{x}) \\ \iff -2 \log \Lambda &= -2 \sum_{i=1}^{n} \left[x_i \log \left(\frac{\bar{x}}{x_i} \right) + (x_i - \bar{x}) \right] \\ &= 2 \sum_{i=1}^{n} x_i \log \left(\frac{x_i}{\bar{x}} \right) \end{aligned}$$

We have the following dimensions for the parameter spaces:

- *Ω, there are n independent parameter $\lambda_1, \ldots, \lambda_n$ so* $\dim \Omega = n$.

- *ω_1, there is only one parameter so* $\dim \omega = 1$

Thus, the test statistics distribution is χ^2_{n-1}. We can interprete the test statistics as the ratio of n times the estimated variance to estimated mean.

Remark 5. We can use Taylor series argument to approximate the test statistics for poisson dispersion test:

$$-2\log\Lambda \approx \frac{1}{\bar{x}}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

3.3 Testing via Plotting

Definition 3.8. (Hanging Rootograms) Graphical display of the differences between observed and fitted values in historgram. There are multiple sections of rootograms:

- **Compare Observed Quatities:** We want to compare the observed frequencies with the frequencies fit by the normal distribution. Given the parameters are approximated as $\mu \approx \bar{x}$ and $\sigma \approx \hat{\sigma}$. If j-th interval has the left boundary x_{j-1} and right boundary x_j. The probability falls in that interval is:

  $$\hat{p}_j = \Phi\left(\frac{x_j-\bar{x}}{\hat{\sigma}}\right) - \Phi\left(\frac{x_{j-1}-\bar{x}}{\hat{\sigma}}\right)$$

  we can predict the count on j-th interval as $\hat{n}_j = n\hat{p}_j$, which can be compared to observed counts. Now, we can find the differences between the expected count and observed out. However, we neglet the variability in the estimated expected counts.

- **Variability:** If we neglect the variability in the estimated expected counts as we have:

  $$\text{var}(n_j-\hat{n}_j) = \text{var}(n_j) = np_j - np_j^2$$

  if p_j are small, we have $\text{var}(n_j-\hat{n}_j) \approx np_j$. For a large values of p_j have more varaible differences $n_j-\hat{n}_j$. And, so we expect larger fluctuation in the center than in the tails.

- **Variance-Stabilizing Transfromation:** Suppose that a random variable X has mean μ and variance $\sigma^2(\mu)$. If $Y = f(X)$, the method of propagation of error shows that:

  $$\text{Var}(Y) \approx \sigma^2(\mu)[f'(\mu)]^2$$

  If f is chosen so that $\sigma^2(\mu)[f'(\mu)]^2$ is constant, the variance of Y will not depends on μ. Thus the transformation accomplishes variance-stabilizing transformation.

- **Variability-Stabilizing:** Apply this to the case, and we have:

  $$\mathbb{E}[n_j] = np_j = \mu \qquad \text{var}(n_j) \approx np_j = \sigma^2(\mu)$$

  That is when $\sigma^2(\mu) = \mu$. The variance stabilizing transformation $\mu[f'(\mu)]^2$ should be $f(x) = \sqrt{x}$ does the job so:

  $$\mathbb{E}[\sqrt{n_j}] \approx \sqrt{np_j} \qquad \text{var}(\sqrt{n_j}) \approx \frac{1}{4}$$

  If the method is correct, and so we compare the differences as $\sqrt{n_j} - \sqrt{\hat{n}_j}$.

- **Interpretation:** We use the deviation of more than 2 and 3 standard deviations is large. The run of positive deviations followed by the run of negative deviations and then the large positive deviation in the extreme right tail. This indicates some asymmetry in the distribution.

- **Hanging Chi-Gram:** The plot of the components of Pearson's chi-square statistics:

  $$\frac{n_j-\hat{n}_j}{\sqrt{\hat{n}_j}} \implies \text{var}\left(\frac{n_j-\hat{n}_j}{\sqrt{\hat{n}_j}}\right) \approx 1$$

  Neglecting the variability in the expected counts, $\text{var}(n_j-\hat{n}_j) \approx np_j = \hat{n}_j$, while the it is stabilizes the variance. This leads to the hanging χ^2-gram.

Definition 3.9. (Order Statistics) Consider the sample of size n from a uniform distribution $[0, 1]$. The ordered sample values by $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$. These values are called order statistics.

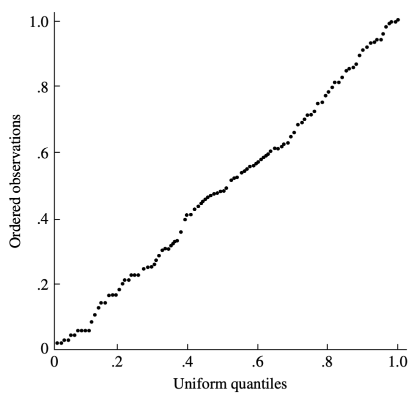
Remark 6. **(Understanding the Plots)** We can show that:

$$\mathbb{E}[X_{(j)}] = \frac{j}{n+1}$$

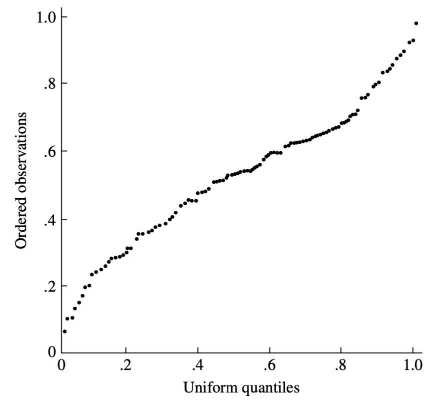
If the underlying distribution is uniform, the plot is shown in figure below, it is plotted for sample of size 100 from a uniform distribution. Now, we consider the triangular distribution as we have:

$$f(y) = \begin{cases} 4y & 0 \leq y \leq \frac{1}{2} \\ 4 - 4y & \frac{1}{2} \leq y \leq 1 \end{cases}$$

THe ordered observation $Y_1, \ldots, Y_{100}$ are plotted against the points $1/(n+1), \ldots, n/(n+1)$:



(a) Uniform-Uniform Probability Plot



(b) Uniform-Triangular Probability Plot

We can see that there is a clear deviation from the linearity and allow us to describe qualitatively the deviation of the distribution of Y's from the uniform distribution:

- The left tail of the plotted distribution are larger than the expected for a uniform distribution
- The right tail is smaller, which tells us that the distribution of Y decreases more quickly than the tails of the uniform distribution.

Definition 3.10. (Probability Integral Transform) The technique can be extended to other continuous probability. If X is a continuous random variable with a strictly increasing cumulative distribution function, and if $Y = F_X(X)$, then Y has a uniform distribution on $[0, 1]$, as:

$$P(Y \leq y) = P(F_X(X) \leq y) = P(X \leq F^{-1}(y)) = F(F^{-1}(y)) = y$$

This is the uniform of cdf. This transformation is known as probability integral transform.

Remark 7. **(Probability Plot)** Suppose that it is hypothesized that X follows a certain distribution F. Given a sample $X_1, \ldots, X_n$, we plot:

$$F(X_{(k)}) \quad \text{vs} \quad \frac{k}{n+1} \qquad \Longrightarrow \qquad X_{(k)} \quad \text{vs} \quad F^{-1}\left(\frac{k}{n+1}\right)$$

In some cases, F is of the form $F(X) = G\left(\frac{x-\mu}{\sigma}\right)$, where μ and σ are location and scale parameter. The normal distribution is of this form, we could plot:

$$\frac{X_{(k)} - \mu}{\sigma} \quad \text{vs} \quad G^{-1}\left(\frac{k}{n+1}\right)$$

or if we plot $X_{(k)}$ vs $G^{-1}\left(\frac{k}{n+1}\right)$. The result would be approixmately a straight line if the model were correct:

$$X_{(k)} \approx \sigma G^{-1}\left(\frac{k}{n+1}\right) + \mu$$

Remark 8. **(Slight Modification)** Slight modification of this procedure are sometimes used. For example $\mathbb{E}[X_{(k)}]$ is used instead, as we have:

$$\mathbb{E}[X_{(k)}] \approx F^{-1}\left(\frac{k}{n+1}\right) = \sigma G^{-1}\left(\frac{k}{n+1}\right) + \mu$$

The modification yields similar result to the original procedure.

Remark 9. **(Another Interpretation)** Recall that $F^{-1}[k/(n+1)]$ is the $k/(n+1)$ quantile of the distribution F, the point such that the probability that a random variable with distribution function F is less than it is $k/(n+1)$. We are plotting the ordered observations versus the quantile of the theoretical distribution.

3.4 Testing for Normality

Definition 3.11. (Coefficient of Skewness) The skewness is usually characterized by the third central moments as:

$$\int_{i\infty}^{\infty} (x-\mu)^2 \varphi(x) \, \mathrm{d}x$$

which is equal to 0 given the normal distribution. Now, coefficient of skewness is:

$$b_1 = \frac{1}{ns^3}\sum_{i=1}^{n}(X_i - \bar{X})^3$$

Definition 3.12. (Coefficient of Kurtosis) Symmetric distribution can depart from normality by being heavy tailed or light-tailed. This is characterized by coefficient of Kurtosis as:

$$b_2 = \frac{1}{ns^4}\sum_{i=1}^{n}(X_i - \bar{X})^4$$

Remark 10. **(Test for Normality)** We can use both coefficient for skewness and kurtosis to access the normality of the data. Otherwise, we can use the hypothesis test, but is are difficult to evaluate in closed form but can be approximated by simulation.

4 Summarizing Data

4.1 Methods Based on CDF

Definition 4.1. (Empirical CDF) Suppose we have $x_1, \dots, x_n$ be a batch of numbers. The empirical cumulative distribution function is defined as:

$$F_n(x) = \frac{1}{n}(\# x_i \le x)$$

Or, we have an ordered number of $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$. We have: if $x_{(k)} \le x < x_{(k+1)}$, then $F_n(x) = k/n$.

Remark 11. **(Comments on Empirical CDF)** In the analysis, it is better to express F_n in the following way, given random variables $X_1, \dots, X_n$:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x]}(X_i) \qquad \text{where} \qquad I_{(-\infty,x]}(X_i) = \begin{cases} 1 & \text{if } X_i \le x \\ 0 & \text{otherwise} \end{cases}$$

The random variable $I_{(-\infty,x]}(X_i)$ are independent Bernoulli random variables, where we have:

$$I_{(-\infty,x]}(X_i) = \begin{cases} 1 & \text{with probability } F(x) \\ 0 & \text{with probability } 1 - F(x) \end{cases}$$

Thus, $nF_n(x)$ is a binomial random variable (n trials with probability of $F(x)$ of success), as we have:

$$\mathbb{E}[F_n(x)] = F(x) \qquad \text{var}(F_n(x)) = \frac{1}{n}F(x)[1 - F(x)]$$

An estimate of $F_n(x)$ is unbiased and has a maximum variacne at the value of x such that $F(x) = 0.5$, which is at median.

Remark 12. **(Behavior of $\boldsymbol{F_n}$)** If we consider the stochastic behavior of $F(x)$, then we can show that:

$$\max_{-\infty<x<\infty} |F_n(x) - F(x)|$$

doesns't depend on F if F is continuous. This allow us to construct a simultaneous confidence band about F_n, which can be used to test goodness-of-fit. Please note that this isn't the same compared to the confidence interval of binomial distribution.

Definition 4.2. (Survival Function) It is equivalent to CDF and is defined as:

$$S(t) = \mathbb{P}(T > t) = 1 - F(t)$$

where T is a random variable with CDF of F. We use it where the data consists of times until failure or death and so non-negative. $S(t)$ denotes the lifetime will be longer than t, and so we can have empirical version to be $S_n(t) = 1 - F_n(t)$.

Definition 4.3. (Hazard Function) It is interpreted as the instantaneous death rate for individual who have survived up to a given time. If an individual is alive at time t, the probability that the individual will die at time interval $(t, t + \delta)$ is (assuming density function f is continuous at t):

$$\begin{aligned} P(t \le T \le t + \delta | T \ge t) &= \frac{P(t \le T \le t + \delta)}{P(T \ge t)} \\ &= \frac{F(t + \delta) - F(t)}{1 - F(t)} \approx \frac{\delta f(t)}{1 - F(t)} \end{aligned}$$

The hazard function is defined as:

$$h(t) = \frac{f(t)}{1 - F(t)}$$

If T is the lifetime of a manufactured component, it may be natural to think of $h(t)$ as the instantaneous or age-specific failure rate.

Remark 13. **(Interpretation of Hazard Function)** It can be expressed as:

$$h(t) = -\frac{d}{dt}\log[1 - F(t)] = -\frac{d}{dt}\log S(t)$$

Which is the negative of the log of survival funcion. With the method of propagation of error:

$$\text{var}\Big(1 - F_n(t)\Big) \approx \frac{\text{var}[1 - F_n(t)]}{(1 - F(t))^2} = \frac{1}{n}\left(\frac{F(t)}{1 - F(t)}\right)$$

For large value of t, the empirical log survial function is unrealiable, because $1 - F(t)$ is very small, and so in practice, last few data are disregarded.

Remark 14. **(Empirical Survial Function)** Suppose that there are no ties and the ordered failure times are: $T_{(1)} < T_{(2)} < \cdots < T_{(n)}$. If $t = T_{(i)}$, $F_n(t) = i/n$ and $S_n(t) = 1 - i/n$. But since $\log S_n(t)$ is undefined for $t \geq T_{(n)}$, it is ofen defined as:

$$S_n(t) = 1 - \frac{i}{n+1}$$

for $T_{(i)} \leq t < T_{(i+1)}$

Definition 4.4. (Quantile-Qunatile Plot) If X is a continuous random variable with a strictly increasing distribution function F, the p-th quantile of the to be value of x such that: $F(x) = p$ or $x_p = F^{-1}(p)$. In Q-Q plot, the quantile of one distribution is plotted against another.

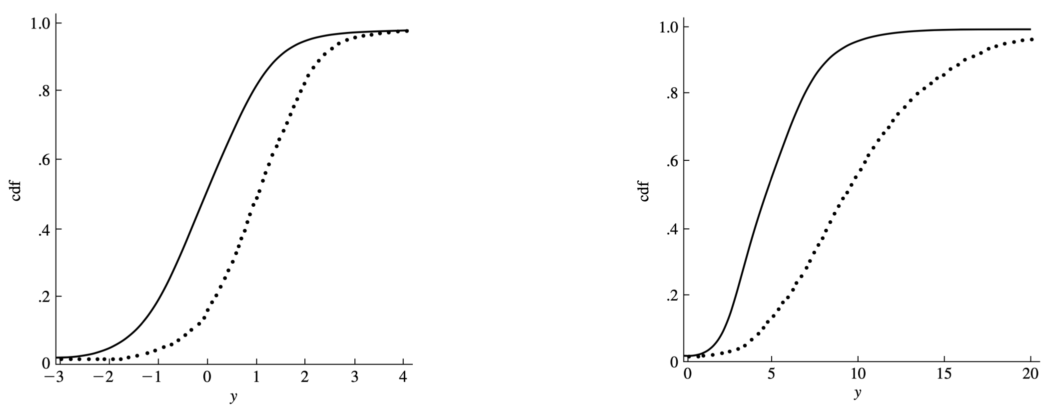
Remark 15. **(Usage of Q-Q)** Suppose we have 2 distributions:

- F is a model for observations of a control group.
- G is a model for observations of a group that has received some treatment.

Let's consider how difference update changes the plot:

- Suppose that there is an effect changned by h uniformly i.e $y_p = x_p + h$, where y_p is the group that received the treatment and vice versa. This gives us the relationship to be: $G(y) = F(y - h)$.
- Similarly, we have the effect with multiplicative differences i.e given $c \in \mathbb{R}$ where we have $y_p = cx_p$ with the relationship to be $G(y) = F(y/h)$

Given the number of samples, we have to use the empirical CDF to create thE Q-Q plot. Now, the results of the changes is shown in the following figure:



(a) Additive Treatment Effect

(b) Multiplicative Treatment Effect

Definition 4.5. (Kernel Probability Density Estimate) Let $w(x)$ be a non-negative, symmetric weight function, centered at zero and integrating to 1. It can be standard normal density, with the following rescaled version:

$$w_h(x) = \frac{1}{h} w\left(\frac{x}{h}\right)$$

is a rescaled version of w, as it approaches zero, w_h becomes more concentrated and peaked around zero. On the other hand, as h approaches infinity, w_h becomes flat. If $X_1, \ldots, X_n$ is a sample from a probability density function p, its esitmate is:

$$f_h(x) = \frac{1}{n} \sum_{i=1}^{n} w_h(x - X_i)$$

The parameter h represents bandwidth of estimating function as it controls the smoothness.

4.2 Meansure of Location

Definition 4.6. (Arithmetic Mean) The commonly used measure of location is the arithmetic mean, which is:

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

Remark 16. **(Problem with Arithmeic Mean)** By changing a single number, the arithmetic mean of a batch of numbers can be made arbitary large or smaller. Thus, when used blindly, without careful attention, the mean can produce a misleading results. Or, we need to have the measure of location that are robut or insensitive to outlier.

Remark 17. **(Why Sample Mean is Bad)** The sample mean minimizers the log-likelihood of:

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

This is the simpliest case of least square estimate. The outlier have a great effect on this estimate, as the deviation of μ from X_i is measured by square of their difference.

Definition 4.7. (Median) It is a middle value of the ordered observation; if the sample size is even, the median is the average of the 2 middle values.

Proposition 4.1. ***(Confidence Interval)*** *We can show that, given the population median η and the interval between the order statistics $(X_{(k)}, X_{(n-k+1)})$*

$$P(X_{(k)} \le \eta \le X_{(n-k+1)}) = 1 - \frac{1}{2^{n-1}}\sum_{j=0}^{k-1}$$

Proof. The coverage probability of this interval is:

$$\begin{aligned} P(X_{(k)} \le \eta \le X_{(n-k+1)}) &= 1 - P(\eta < X_{(k)} \text{ or } \eta > X_{n-k+1}) \\ &= 1 - P(\eta < X_{(k)}) - P(\eta > X_{(n-k+1)}) \end{aligned}$$

Since the event are mutually exclusive. To evaluate both terms, we note that:

$$P(\eta > X_{(n-k+1)}) = \sum_{j=0}^{k-1}\mathbb{P}(j \text{ observations} > \eta)$$

$$P(\eta < X_{(k)}) = \sum_{j=0}^{k-1}\mathbb{P}(j \text{ observations } < \eta)$$

The median satisfies $P(X_i > \eta) = P(X_i < \eta) = 1/2$, since n observations $X_1, \dots, X_n$ are independent and identically distributed, the distribution of the number of observation greater than median is binomial with n trials and probability $1/2$:

$$P(j \text{ observations } > \eta) = \frac{1}{2}\binom{n}{j}$$

and, so we have:

$$P(\eta > X_{(n-k+1)}) = \frac{1}{2^n}\sum_{j=0}^{k-1}\binom{n}{j}$$

This is the same for $P(\eta < X_{(k)})$ due to symmetry. Plugging it back to finish the proof ☐

Remark 18. Median can be seen as the minimizer of the following loss:

$$\sum_{i=1}^{n} \left| \frac{X_i - \mu}{\sigma} \right|$$

Here, large deviation are not weighted as heavily, making median robust. The proof follows from the fact that the dervative of absolute is sgn(·), and so the loss is zero when the positive $x - \mu$ (of the normalized data) is equal to the negative item $x - \mu$, which is where the median situates.

Definition 4.8. (Trimmed Mean) The $100\alpha\%$ trimmed mean consider the valuse that is between the lower $100\alpha\%$ and the higher $100\alpha\%$, as we can write it as:

$$\bar{x}_\alpha = \frac{x_{[n\alpha]+1} + \cdots + x_{(n-[n\alpha])}}{n - 2[n\alpha]}$$

where $[n\alpha]$ denotes the greatest integer less than or equal to $n\alpha$.

Definition 4.9. (M-Estimates) Consider the class of esitmates called M-estimates, where it is a minimizer:

$$\sum_{i=1}^{n} \Psi\left(\frac{X_i - \nu}{\sigma}\right)$$

where Ψ is the weight function that is a compromise between weight function for mean and median.

Remark 19. **(Measure of Dispersion)** The most commonly used measure is sample standard deviation, where it is given as:

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

Using $n - 1$ as divisor gives unbiased estimate. But like a sample mean standard deviation is sensitive to outlying observation. Two simple robust measures alternative are:

- Interquartile range (IQR): Differences between 2 sample quantiles.
- Median absolute deviation from the median (MAD): If data are $x_1, \dots, x_n$ with median $\tilde{x}$, then MAD is the median of number $|x_1, \dots, x_n|$.

5 Comparing Two Samples

5.1 Comparing Two Independent Samples

Remark 20. **(Setting)** We will assume the sample $X_1, \dots, X_n \sim \mathcal{N}(\mu_X, \sigma^2)$ as the control group and we have $Y_1, \dots, Y_m \sim \mathcal{N}(\mu_Y, \sigma^2)$ as the group after receives treatment. The effect of the treatment is characterized by the differences $\mu_X - \mu_Y$ with the natural estimate $\bar{X} - \bar{Y}$.

Remark 21. **(Confidence Interval)** As $\bar{X} - \bar{Y}$ is expressed as a linear combination of independent normally distributed random variable is:

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

If we know σ^2, where the confidence interval for $\mu_X - \mu_Y$ could be based on:

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

This leads to the confidence interval, which is of the form of $(\bar{X} - \bar{Y}) \pm z(\alpha/2)\sigma\sqrt{\frac{1}{m} + \frac{1}{n}}$.

Definition 5.1. **(Pooled Sample Variance)** Generally, σ^2 will not be known and must be estimated from the data by calculating pooled sample variance:

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

where $s_X^2 = 1/(n-1)\sum_{i=1}^n (X_i - \bar{X})^2$ and similarly for s_Y^2, and so s_p^2 is a weighted average of sample variance X and Y with weights propotional to degree of freedom.

Theorem 5.1. *Suppose that $X_1, \dots, X_n \sim \mathcal{N}(\mu_X, \sigma^2)$ and $Y_1, \dots, Y_m \sim \mathcal{N}(\mu_Y, \sigma^2)$, and that Y_i are independent of X_i. The statistics:*

$$t = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{s_p\sqrt{\dfrac{1}{n} + \dfrac{1}{m}}}$$

This follows a t-distribution with $m + n - 2$ degree of freedom.

Proof. We note that $(n-1)s_X^2/\sigma^2 \sim \chi^2_{n-1}$ and $(m-1)s_Y^2/\sigma^2 \sim \chi^2_{m-1}$. Both are independent as X_i and Y_i are. Their sum is χ^2_{m+n-2} degree of freedom. We express the statistics as the ratio U/V, where:

$$U = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sigma\sqrt{\dfrac{1}{n} + \dfrac{1}{m}}}$$

$$V = \sqrt{\left[\frac{(n-1)\sigma_X^2}{\sigma^2} + \frac{(m-1)s_Y^2}{\sigma^2}\right]\frac{1}{m+n-2}}$$

Please note that U follows the standard normal distribution and V has the distribution of square root of χ^2 divided by its degree of freedom. The independent of U and V follows from independent of $\bar{X}$ and s^2. □

Corollary 5.1. *Under the assumption of thoerem above, a $100(1-\alpha)\%$ confidence interval for $\mu_X - \mu_Y$ is:*

$$(\bar{X} - \bar{Y}) \pm t_{m+n-2}(\alpha/2) s_{\bar{X}-\bar{Y}} \qquad \textit{where} \qquad s_{\bar{X}-\bar{Y}} = s_p\sqrt{\frac{1}{n} + \frac{1}{m}}$$

Remark 22. **(Notes on One and Two-Sided Alternative)** In the current case, the null hypothesis to be tested is $H_0 : \mu_X = \mu_Y$, where there are 3 common alternatives, as we have:

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

The test statistics that will be used to make a decision to reject the null-hypothesis is:

$$t = \frac{\bar{X} - \bar{Y}}{s_{\bar{X}-\bar{Y}}}$$

The t-statistics equals the multiple of its estimate standard deviation differs from zero. This is the same role in the comparison of 2 samples as is played by χ^2-statistics. We will reject for extreme value of t. We have the following rejection region:

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

Proposition 5.1. ***(Two-Sided Alternative T-Statistics)*** *The test for H_1 (see above) rejects the large value of the following value:*

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

Which is t statistics apart from constant that don't depend on the data. Thus, the likelihood ratio test is equivalent to t-test as claimed.

Proof. Consider the set Ω is the set of all possible parameter values:

$$\Omega = \Big\{ -\infty < \mu_X < \infty,\ -\infty < \mu_Y < \infty,\ 0 < \sigma < \infty \Big\}$$

The unknown parameters are $\theta = (\mu_X, \mu_Y, \sigma)$. Under H_0 where $\theta \in \omega_0$ where:

$$\omega_0 = \Big\{ \mu_X = \mu_Y : 0 < \sigma < \infty \Big\}$$

The likelihood of 2 samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ is given as:

$$\begin{aligned} l(\mu_X, \mu_Y, \sigma^2) &= \log \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2} \left(\frac{X_i - \mu_X}{\sigma} \right)^2 \right] \prod_{j=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2} \left(\frac{Y_j - \mu_Y}{\sigma} \right)^2 \right] \\ &= -\frac{m+n}{2} \log 2\pi - \frac{m+n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \left[\sum_{i=1}^{n} (X_i - \mu_X)^2 + \sum_{j=1}^{m} (Y_j - \mu_Y)^2 \right] \end{aligned}$$

Let's consider the MLE and its log-likelihood are given as:

- Under ω_0, we have a sample of size $m+n$ from a normal distribution with unknown mean μ_0 and unknown variance σ_0^2. The MLE of μ_0 and σ_0^2 is:

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

- To find the MLE's $\hat{\mu}_X, \hat{\mu}_Y$ and σ_1^2 under Ω, we consider the log-likelihood is

$$\begin{aligned} \sum_{i=1}^{n} (X_i - \hat{\mu}_X) = 0 &\implies \hat{\mu}_X = \bar{X} \\ \sum_{j=1}^{m} (Y_j - \hat{\mu}_Y) = 0 &\implies \hat{\mu}_Y = \bar{Y} \end{aligned}$$

$$-\frac{m+n}{2\hat{\sigma}_1^2} + \frac{1}{2\hat{\sigma}_1^4} \left[\sum_{i=1}^{n} (X_i - \hat{\mu}_X)^2 + \sum_{j=1}^{m} (Y_j - \hat{\mu}_Y)^2 \right] = 0 \implies \hat{\sigma}_1^2 = \frac{1}{m+n} \left[\sum_{i=1}^{n} (X_i - \hat{\mu}_X)^2 + \sum_{j=1}^{m} (Y_j - \hat{\mu}_Y)^2 \right]$$

This implies that the log-likelihood, we obtain it as:

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

The log of likelihood ratio is given as:

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

Let's consider the alternatives expression for the numerator of this ratio:

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

Please note that:

$$\hat{\mu}_0 = \frac{1}{m+n} (n\bar{X} + m\bar{Y}) = \frac{n}{m+n} \bar{X} + \frac{m}{m+n} \bar{Y}$$

This implies that:

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

The alternatives expression for the numerator of the ratio is:

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

The test rejects for the large value of:

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

This is equivalent to the value above. Thus the proposition is proven. □

Remark 23. **(Difference Variance)** If 2 variances are not assumed to be equal, a natural estimate of $\text{var}(\bar{X} - \bar{Y})$ is given as:

$$\frac{s_X^2}{n} + \frac{s_Y^2}{m}$$

If this estimate is used in the denominator of t statistics, the distribution of that statistics is no longer the t-distribution. But it can be closely approximated by t-distribution with degree of freedom, where we round it to nearest integer.

$$\frac{[(s_X^2/n) + (s_Y^2/m)]^2}{\dfrac{(s_X^2/n)^2}{n-1} + \dfrac{(s_Y^2/m)^2}{m-1}}$$

Remark 24. **(Notes on Two Sample T-Test)** The power of 2-sample t-test depends on 4 factors:

- The real differences $\Delta = |\mu_X - \mu_Y|$. The larger the differences, the greater the power.
- The significant level α at which the test is done. The larger the more powerful the test.
- The population standard deviation σ, whichi s amplitude of the nose that hides the signal. The smaller the larger the power.
- The sample size n and m, The larger the sample size, and the greater the power.

The necessary sample sizes can be determined from the significant level of the test, the standard deviation, and the desired power agaisnt an alternatives hypothesis.

Remark 25. **(Finding Power of t Test)** To calculate the power of a t test exactly, we need special table of non-central t distribution. If the sample are reasonably large, one can perform approximation of it based on normal distribution.

Proposition 5.2. ***(Approximate Power of the Test)*** *The probability that the test statistics falls in rejection region is given as:*

$$1 - \Phi\left[z(\alpha/2) - \frac{\Delta}{\sigma}\sqrt{\frac{n}{2}}\right] + \Phi\left[-z(\alpha/2) - \frac{\Delta}{\sigma}\sqrt{\frac{n}{2}}\right]$$

where $\Delta = \mu_X - \mu_Y$ with test at level α. Now, Δ moves away from zero, one of these terms will be negligible with respect to others.

Proof. Consider the following variance:

$$\text{var}(\bar{X} - \bar{Y}) = \sigma^2\left(\frac{1}{n} + \frac{1}{n}\right) = \frac{2\sigma^2}{n}$$

The tes at level α of $H_0 : \mu_X = \mu_Y$ agaisnt the alternatives $H_1 : \mu_X \neq \mu_Y$ is based on test statistics:

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

The rejection region is given as:

$$\left|\bar{X} - \bar{Y}\right| > z(\alpha/2)\sigma\sqrt{\frac{2}{n}}$$

Let's consider the rejection region to be the following:

$$\begin{aligned}
&\mathbb{P}\left[\left|\bar{X} - \bar{Y}\right| > z(\alpha/2)\sigma\sqrt{\frac{2}{n}}\right] \\
&\quad = \mathbb{P}\left[\bar{X} - \bar{Y} > z(\alpha/2)\sigma\sqrt{\frac{2}{n}}\right] + \mathbb{P}\left[\bar{X} - \bar{Y} < -z(\alpha/2)\sigma\sqrt{\frac{2}{n}}\right]
\end{aligned}$$

As two of them are mutually exclusive. Both probability can be calculated by standardizing:

$$\begin{aligned}
\mathbb{P}\left[\bar{X} - \bar{Y} > z(\alpha/2)\sigma\sqrt{\frac{2}{n}}\right] &= \mathbb{P}\left[\frac{(\bar{X} - \bar{Y}) - \Delta}{\sigma\sqrt{2/n}} > \frac{z(\alpha/2)\sigma\sqrt{2/n} - \Delta}{\sigma\sqrt{2/n}}\right] \\
&= 1 - \Phi\left[z(\alpha/2) - \frac{\Delta}{\sigma}\sqrt{\frac{n}{2}}\right]
\end{aligned}$$

Similarly, we have the second probability is given as:

$$\Phi\left[-z(\alpha/2) - \frac{\Delta}{\sigma}\sqrt{\frac{n}{2}}\right]$$

Adding them together is given the above approximation of the test. □

5.2 Nonparametric Test

Remark 26. **(Setting for Mann-Whitney test)** Suppose we have $m + n$ experimental untis to assign to a treatment group and control group, as we have:

- n units are randomly chosen and assigned to the control.
- m units are assigned to the treatment.

We are interested in testing the null hypothesis that the treatment has not effect.

Definition 5.2. (Statistics for Mann-Whitney Test) We consider the following procedure:

- Group all $m + n$ observations together and rank them in order of increasing size.
- Calculate the sum of the ranks of those observations that came from the control group.

If the sum is too small or too large, we will reject the null hypothesis. Please note that this test doesn't depend on an assumption of normality. It is nearly as powerful as t-test and it is generally preferable (for small sample size).

Remark 27. **(Settings for Mann-Whitney Test)** Consider the control values as we have $X_1, \dots, X_n \sim F$ and the experimental values $Y_1, \dots, Y_m \sim G$. The Mann-Whitney test is a test of null hypothesis $H_0 : F = G$. We will denote T_Y to denotes the sum of ranks of $Y_1, Y_2, \dots, Y_m$.

Lemma 5.1. *From a simple random sampling without replacement, we have:*

$$\text{cov}(X_i, X_j) = -\sigma^2/(N-1)$$

where the $\text{var}(X_i) = \sigma^2$

Proof. Using the identities for covariance established:

$$\text{cov}(X_i, X_j) = \mathbb{E}[X_iX_j] - \mathbb{E}[X_i]\mathbb{E}[X_j]$$

And, we have the following:

$$\begin{aligned}
\mathbb{E}[X_iX_j] &= \sum_{k=1}^{m}\sum_{l=1}^{m} \xi_k\xi_l P(X_i = \xi_k \wedge X_j = \xi_l) \\
&= \sum_{k=1}^{m} \xi_k P(X_i = \xi_k) \sum_{l=1}^{k} \xi_l P(X_j = \xi_l | X_i = \xi_k)
\end{aligned}$$

from the multiplication law of conditional probability as we have:

$$P(X_j = \xi_l | X_i = \xi_k) = \begin{cases} n_l/(N-1) & \text{if } k \neq l \\ (n_l - 1)/(N-1) & \text{if } k = l \end{cases}$$

If we express is give as:

$$\begin{aligned}
\sum_{l=1}^{m} \xi_l P(X_j = \xi_l | X_i = \xi_k) &= \sum_{l \neq k} \xi_l \frac{n_l}{N-1} + \xi_k \frac{n_k - 1}{N-1} \\
&= \sum_{l=1}^{m} \xi_l \frac{n_l}{N-1} - \xi_k \frac{1}{N-1}
\end{aligned}$$

Now, we have the expression for $\mathbb{E}[X_iX_j]$ as we have:

$$\begin{aligned}
\sum_{k=1}^{m} \xi_k \frac{n_k}{N} \left(\sum_{l=1}^{m} \xi_l \frac{n_l}{N-1} - \frac{\xi_k}{N-1} \right) &= \frac{1}{N(N-1)} \left(\tau^2 - \sum_{k=1}^{m} \xi_k^2 n_k \right) \\
&= \frac{\tau^2}{N(N-1)} - \frac{1}{N(N-1)} \sum_{k=1}^{m} \xi_k^2 n_k \\
&= \frac{N\mu^2}{N-1} - \frac{1}{N-1}(\mu^2 + \sigma^2) \\
&= \mu^2 - \frac{\sigma^2}{N-1}
\end{aligned}$$

Finally, subtracting $\mathbb{E}[X_i]\mathbb{E}[X_j] = \mu^2$ fro mthe last equation, as we have:

$$\text{cov}(X_i, X_j) = -\frac{\sigma^2}{N-1}$$

for $i \neq j$. □

Corollary 5.2. *With simple random sampling, we can show that:*

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

Proof. We can see that:

$$\begin{aligned}
\text{var}(\bar{X}) &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \text{cov}(X_i, X_j) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \text{var}(X_i) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j \neq i} \text{cov}(X_i, X_j) \\
&= \frac{\sigma^2}{n} - \frac{1}{n^2} n(n-1) \frac{\sigma^2}{N-1}
\end{aligned}$$

This gives the desired result. □

Proposition 5.3. *If $F = G$ as we have:*

$$\mathbb{E}[T_Y] = \frac{m(m+n+1)}{2} \qquad \text{var}(T_Y) = \frac{mn(m+n+1)}{12}$$

Proof. Under the null hypothesis, T_Y is the sum of random sample of size m drawn without replacement from a population of integers $[m+n]$. T_Y thus equal to m times the average of such a sample as:

$$\mathbb{E}[T_Y] = m\mu \qquad \text{var}(T_Y) = m\sigma^2\left(\frac{N-m}{N-1}\right)$$

We can show that, where $N = m+n$ is the size of population. Using the identities (to calculate the values μ and σ^2) as we have (this follows from the thoerem above):

$$\sum_{k=1}^{N} k = \frac{N(N+1)}{2} \qquad \sum_{k=1}^{N} k^2 = \frac{N(N+1)(2N+1)}{6}$$

We find the population as we have $[m+n]$ as we have:

$$\mu = \frac{N+1}{2} \qquad \sigma^2 = \frac{N^2-1}{12}$$

The result follows from the algebraic simplification. □

Remark 28. **(Alternative Derivation of Mann-Whitney Test)** We consider the $X \sim F$ and $Y \sim G$ and

- Consider measuring of the effect of the treatment: $\pi = P(X < Y)$.
- The value of π is the probability that an observation from the distribution F is smaller than an independent observation from the distribution G.

The estimate of π can be obtained by comparing all n values of X to all m values of Y. Calculating the proposition of the comparison for which X was less than Y:

$$\hat{\pi} = \frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m} Z_{ij} \qquad Z_{ij} = \begin{cases} 1 & \text{if } X_i < Y_j \\ 0 & \text{otherwise} \end{cases}$$

Understand the relationship of $\hat{\pi}$ to the rank sum introduced earlier, we find the convenient to work with:

$$V_{ij} = \begin{cases} 1 & \text{if } X_{(i)} < Y_{(j)} \\ 0 & \text{otherwise} \end{cases}$$

Since V_{ij} are jusre reordering of Z_{ij}, also gives us:

$$\begin{aligned} \sum_{i=1}^{n}\sum_{j=1}^{m} V_{ij} = \#(X < Y_{(1)}) &+ \#(X < Y_{(2)}) \\ &+ \cdots + \#(X < Y_{(m)}) \end{aligned}$$

where $\#(X < Y_{(1)})$ is the number of X that are less than $Y_{(1)}$. If the rank of $Y_{(k)}$ in the combined sample is denoted by R_{yk}, then the number of X that is less than $Y_{(1)}$ is $R_{y1} - 1$ and number of X is less than $Y_{(2)}$ is $R_{y2} - 2$ and so on, thus we have:

$$\begin{aligned} \sum_{i=1}^{n}\sum_{j=1}^{m} V_{ij} &= (R_{y1} - 1) + (R_{y2} - 2) + \cdots + (R_{ym} - m) \\ &= \sum_{i=1}^{m} R_{yi} - \sum_{i=1}^{m} i \\ &= \sum_{i=1}^{m} R_{yi} - \frac{m(m+1)}{2} \\ &= T_y - \frac{m(m+1)}{2} \end{aligned}$$

Thus $\hat{\pi}$ may be expressed in terms of rank sum of Y.

Corollary 5.3. *Let $U_Y = \sum_{i=1}^n \sum_{j=1}^m Z_{ij}$. Under the null hypothesis $H_0 : F = G$ as we have:*

$$\mathbb{E}[U_Y] = \frac{mn}{2} \qquad \text{var}(U_Y) = \frac{mn(m+n+1)}{12}$$

Remark 29. For both n and m are both greater than 10, the null distribution U_Y is quite well approximated by a normal distribution as we have:

$$\frac{U_Y - \mathbb{E}[U_Y]}{\sqrt{\text{var}(U_Y)}} \sim \mathcal{N}(0, 1)$$

The distribution of the rank sum of the X and Y may be approximated by normal distribution as the rank sum differ from U_Y only by constant.

Remark 30. **(Mann-Whitney as CI)** Let's consider the shift model as we have $G(x) = F(x - \Delta)$. We will consider the confidence interval for Δ. To test $H_0 : F = G$, we use the statistics U_Y equal to number of $X_i - Y_j$ that are less than 0. We can use: (to test the hypothesis that the shift parameter is Δ)

$$U_Y(\Delta) = \#[X_i - (Y_j - \Delta) < 0] = \#(Y_j - X_i > \Delta)$$

The null distribution of $U_Y(\Delta)$ is symmetric about $mn/2$:

$$\mathbb{P}\left(U_Y(\Delta) = \frac{mn}{2} + k\right) = \mathbb{P}\left(U_Y(\Delta) = \frac{mn}{2} - k\right)$$

for all integer k. Suppose that $k = k(\alpha)$ is such that $\mathbb{P}(k \le U_Y(\Delta) \le mn - k) = 1 - \alpha$; the level α test then accepts for such $U_Y(\Delta)$. By the duality of CI and hypothesis tests, a $100(1-\alpha)\%$ confidence interval for Δ is thus:

$$C = \{\Delta : k \le U_Y(\Delta) \le mn - k\}$$

where C is the set of values for which the null hypothesis won't be rejected. Let's find the explicit form for this CI. Let $D_{(1)}, D_{(2)}, \dots, D_{(nm)}$ denote the ordered mn differences $Y_j - X_i$. We will show that:

$$C = [D_{(k)}, D_{(mn-k+1)})$$

To see this, first suppose that $\Delta = D_{(k)}$. Then:

$$\begin{aligned} U_Y(\Delta) &= \#(X_i - Y_j + \Delta < 0) \\ &= \#(Y_j - X_i > \Delta) \\ &= mn - k \end{aligned}$$

Similarly, if $\Delta = D_{(mn-k+1)}$, we have:

$$U_Y(\Delta) = \#(Y_j - X_i > \Delta) = k$$

5.3 Bayesian Approach

Remark 31. **(Setting For Bayesian Approach)** Consider the case where

- $X_i \sim \mathcal{N}(\mu_X, \xi^{-1})$
- $Y_j \sim \mathcal{N}(\mu_Y, \xi^{-1})$ and independent of X_i.
- The means μ_X and μ_Y are given improper prior that are constant on $(-\infty, \infty)$.
- ξ is given the improper prior $f_\Xi(\xi) = \xi^{-1}$.

This posterior is thus given by:

$$p(\mu_X, \mu_Y, \xi) \propto \xi^{(n+m)/2-1} \exp\left(-\frac{\xi^{m+n}}{2}\left[\sum_{i=1}^{n}(x_i-\mu_X)^2 + \sum_{j=1}^{m}(y_j-\mu_Y)^2\right]\right)$$

Using the identity that $\sum_{i=1}^{n}(x_i-\mu_X)^2 = (n-1)s_x^2 + n(\mu_X - \bar{x})^2$, and the analogous expression for y_j as:

$$\begin{aligned} p(\mu_X, \mu_Y, \xi) \propto \xi^{(n+m)/2-1} &\exp\left(-\frac{\xi}{2}\left[(n-1)s_x^2 + (m-1)s_y^2\right]\right) \\ &\times \exp\left(-\frac{n\xi}{2}(\mu_X - \bar{x})^2\right)\exp\left(-\frac{m\xi}{2}(\mu_Y - \bar{y})^2\right) \end{aligned}$$

For a fixed ξ, μ_X and μ_Y are independent normally distributed with means $\bar{x}$ and $\bar{y}$ and precisions $n\xi$ and $m\xi$, thus: the difference $\mu_X - \mu_Y$ is normally distributed with mean $\bar{x} - \bar{y}$ and variance $\xi^{-1}(n^{-1} + m^{-1})$. With further analysis, one can show that:

$$\frac{\Delta - (\bar{x} - \bar{y})}{s_p\sqrt{n^{-1}+m^{-1}}} \sim t_{n+m-2}$$

This may be similar to above result but has differences in interpretation, as $\bar{x} - \bar{y}$ and s_p are random in above result and Δ is fix. This is opposite in this case. The posterior probability that $\Delta > 0$ can be found using t distribution. Let T be random variable with t_{m+n-2} distribution, then:

$$\begin{aligned} P(\Delta > 0|X,Y) &= \mathbb{P}\left(\frac{\Delta - (\bar{x} - \bar{y})}{s_p\sqrt{n^{-1}+m^{-1}}} \ge \frac{-(\bar{x} - \bar{y})}{s_p\sqrt{n^{-1}+m^{-1}}} \middle| X, Y\right) \\ &= \mathbb{P}\left(T \ge \frac{\bar{y} - \bar{x}}{s_p\sqrt{n^{-1}+m^{-1}}}\right) \end{aligned}$$

As, we can use this as CI.

5.4 Compare Paired Samples

Remark 32. **(Conditions for Paired Samples Test)** Some of the experiments, the samples are paired. In medical experiment, the subjects might be matched by age or severity of the condition, while one of them are randomly assigned to treatment group and other control group.

Proposition 5.4. ***(Releative Efficiently)*** *We will denote the pair as (X_i, Y_i) where $i = 1, \dots, n$ adn assume X and Y have means μ_X and μ_Y and variances σ_X^2 and σ_Y^2. We will assume that different pairs are independently distributed that* $\text{cov}(X_i, Y_i) = \sigma_{XY}$. *Given the estimate of $D = X_i - Y_i$ (in the pair setting):*

$$\frac{\text{var}(\bar{D})}{\text{var}(\bar{X} - \bar{Y})} = 1 - \rho$$

where ρ is the correlation of members of a pair, and $\sigma_X = \sigma_Y = \sigma$. This menas that if the correlation coefficient is 0.5*, a paired design with n pairs of subjects yields same precision as an unpaired design with* $2n$ *subject per treatment.*

Proof. Starting with paired experiment, as we have:

- We will work with the differences: $D_i = X_i - Y_i$, which are independent with:

$$\mathbb{E}[D_i] = \mu_X - \mu_Y \qquad \begin{aligned} \text{var}(D_i) &= \sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY} \\ &= \sigma_X^2 + \sigma_Y^2 - 2\rho\sigma_X\sigma_Y \end{aligned}$$

  A natural estimate of $\mu_X - \mu_Y$ is $\bar{D} = \bar{X} - \bar{X}$, the average difference. From the properties of D_i:

$$\mathbb{E}[\bar{D}] = \mu_X - \mu_Y \qquad \text{var}(\bar{D}) = \frac{1}{n}(\sigma_X^2 + \sigma_Y^2 - 2\rho\sigma_X\sigma_Y)$$

- An experiment had been done by taking a sample of n X's and an independent sample of n Y's, then $\mu_X - \mu_Y$ would be estimated by $\bar{X} - \bar{Y}$ and:

$$\mathbb{E}[\bar{X} - \bar{Y}] = \mu_X - \mu_Y \qquad \text{var}(\bar{X} - \bar{Y}) = \frac{1}{n}(\sigma_X^2 + \sigma_Y^2)$$

We see that the variance of $\bar{D}$ is smaller if the correlation is positive. If X and Y are positively correlated. Consider the case where $\sigma_X = \sigma_Y = \sigma$, the 2 variances may be simply expressed as:

$$\text{var}(\bar{X}) = \frac{2\sigma^2(1-\rho)}{n} \qquad \text{var}(\bar{X} - \bar{Y}) = \frac{2\sigma^2}{n}$$

Thus the relative efficiently is give. □

Remark 33. **(Method Based on Normal Distribution)** Assume the differences that are sample of a normal distribution:

$$\mathbb{E}[D_i] = \mu_X - \mu_Y = \mu_D \qquad \text{var}(D_i) = \sigma_D^2$$

Generally, σ_D will be unknown, the inferences will be based on:

$$t = \frac{\bar{D} - \mu_D}{s_{\bar{D}}}$$

This follows a t distribution with $n-1$ degree of freedom. With similar reasoning $100(1-\alpha)\%$ confidence interval is given as:

$$\bar{D} \pm t_{n-1}(\alpha) s_{\bar{D}}$$

If sample size n is large, the approximate validity of the CI and hypothesis test follows from CLT.

Definition 5.3. (Signed Rank Test) We consider a paired sample (X_i, Y_i), we then find the absoluate differences $|X_i - Y_i|$ and rank them in order, denoted by $D_{(i)}$. The signed rank is calculated as:

$$S_{(i)} = \begin{cases} -D_{(i)} & \text{if } X_i > Y_i \\ D_{(i)} & \text{otherwise} \end{cases}$$

Now, we have $W_{(+)} = \sum_{S_{(i)}>0} S_{(i)}$. If there is no differences between the two paired conditions, as we expect about half of D_i to be positive and half negative.

Remark 34. **(Finding Rejection Region)** THe null distribution can be calculated this way. If H_0 is true, it makes no difference:

- The difference $X_i - Y_i = D_i$ has the same distribution as the difference $Y_i - X_i = -D_i$, so the distribution of D_i is symmetric about zero.
- The k-th largest value of D is thus equally likely to be positive or negative, and any particular assignment of signs to the integer $1, \ldots, n$ (the ranks) is equally likely.
- We obtain a list of 2^n value of W_+ each of which occurs with probability $1/2^n$. The probability of each distinct value of W_+ may be calculated, given the desired null distribution.

If the sample size is greater than 20, a noraml approximation to the null distribution can be used. We calculate the mean and variance of W_+

Proposition 5.5. *Under the null hypothesis that the D_i are independent and symmetrically distribution about zero:*

$$\mathbb{E}[W_+] = \frac{n(n+1)}{4} \qquad \text{var}[W_+] = \frac{n(n+1)(2n+1)}{24}$$

Proof. To facilitate the calculation, we represent W_+ in the following way:

$$W_+ = \sum_{i=1}^{n} kI_k \qquad I_k = \begin{cases} 1 & \text{if } k \text{ largest } |D_i| \text{ has } D_i > 0 \\ 0 & \text{otherwise} \end{cases}$$

Under H_0 and I_k are independent Bernoulli random variable $p = 1/2$, so we have:

$$\mathbb{E}[I_k] = \frac{1}{2} \qquad \text{var}(I_k) = \frac{1}{4}$$

We thus have:

$$\mathbb{E}[W_+] = \frac{1}{2}\sum_{k=1}^{n} k = \frac{n(n+1)}{4} \qquad \text{var}(W_+) = \frac{1}{4}\sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{24}$$

□

Remark 35. **(When Tie is Encountered)** If some of the differences are equal to zero, the most common way to discard those observation. If there are ties, each $|D_i|$ is assigned the average value of the ranks for which it is tied. If there are not too many ties, the significant level of the test isn't greatly affected.

6 Analysis of Variance

Definition 6.1. (One-Way Layout) The independent measurement are made under each of several treatments. It is the generalization of the above test. We will denote the I groups that contains J samples. We will denote, the following values:

$$Y_{ij} = \text{The } j\text{-th observation in the } i\text{-th treatments.}$$

6.1 Normal Theory: F Test

Remark 36. **(Statistical Model of One-Way Layout)** We have the statistical model model is given as $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$. Here μ is the overall mean and α_i is the differential effect of the i-th treatment. We will assume to be independent, normally distributed with mean 0 and variance σ^2. The α_i are normalized:

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

Remark 37. **(Defining Null-Distribution)** The expected response to the i-th treatment is $\mathbb{E}[Y_{ij}] = \mu + \alpha_i$. If $\alpha_i = 0$ for $i = 1, \dots, I$ all treatments have the same expected respoinse, and in general $\alpha_i - \alpha_j$ is the differeces between the expected values under treatments i and j.

Lemma 6.1. *We consider the following identity:*

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

where we have:

$$\bar{Y}_{i.} = \frac{1}{J}\sum_{j=1}^{J} Y_{ij} \qquad \bar{Y}_{..} = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} Y_{ij}$$

This means that the total sum of squares equals to the sum of square within groups plus the squares between groups, as we have $SS_{TOT} = SS_W + SS_B$

Proof. To establish the identity, we express the left-hand side as:

$$
\begin{aligned}
\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij}-\bar{Y}_{..})^2 &= \sum_{i=1}^{I}\sum_{j=1}^{J}[(Y_{ij}-\bar{Y}_{i.})+(\bar{Y}_{i.}-\bar{Y}_{..})]^2 \\
&= \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij}-\bar{Y}_{i.})^2+\sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{Y}_{i.}-\bar{Y}_{..})^2 \\
&\quad +2\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij}-\bar{Y}_{i.})(Y_{i.}-\bar{Y}_{..}) \\
&= \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij}-\bar{Y}_{i.})^2+\sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{Y}_{i.}-\bar{Y}_{..})^2 \\
&\quad +2\sum_{i=1}^{I}\left[(Y_{i.}-\bar{Y}_{..})\sum_{j=1}^{J}(Y_{ij}-\bar{Y}_{i.})\right]
\end{aligned}
$$

The last term of the final expression vanishes because some of deviation from a mean is zero. □

Lemma 6.2. *Let X_i where $i = 1, \vdots n$ be independent random variable with $\mathbb{E}[X_i] = \mu_i$ and $\operatorname{var}(X_i) = \sigma^2$. Then we have, the following identity:*

$$
\mathbb{E}[(X_i-\bar{X})^2] = (\mu_i-\bar{\mu})^2+\frac{n-1}{n}\sigma^2 \qquad \textit{where} \qquad \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\mu_i
$$

Proof. We used the fact that $\mathbb{E}[U^2] = \mathbb{E}[U]^2 + \operatorname{var}(U)$ for any random variable with finite variance. Let's consider the second term: $\operatorname{var}(X_i - \bar{X})$:

$$
\begin{aligned}
\operatorname{var}(X_i-\bar{X}) &= \operatorname{var}(X_i)+\operatorname{var}(\bar{X})-2\operatorname{cov}(X_i,\bar{X}) \\
&= \sigma^2+\frac{1}{n}\sigma^2+\operatorname{cov}\left(X_i,\frac{1}{n}\sum_{j=1}^{n}X_j\right) \\
&= \sigma^2+\frac{1}{n}\sigma^2-\frac{2}{n}\sigma^2
\end{aligned}
$$

This concludes the proof. □

Theorem 6.1. *We consider the expectation:*

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

Proof. Under the assumption for the model stated at the beginning of this section:

$$
\mathbb{E}[SS_{\mathrm{W}}] = \sum_{i=1}^{I}\sum_{j=1}^{J}\mathbb{E}[(Y_{ij}-\bar{Y}_{i.})^2] = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{J-1}{J}\sigma^2 = I(J-1)\sigma^2
$$

We have used lemma above with the role of X_i being played by Y_{ij}. The second equality follows since $\mathbb{E}[Y_{ij}] = \mathbb{E}[\bar{Y}_i] = \mu + \alpha_i$. Now, let's find the $\mathbb{E}[SS_{\mathrm{B}}]$, we use the lemma with $\hat{Y}_{i.}$ and $\hat{Y}_{..}$ as:

$$
\mathbb{E}[SS_{\mathrm{B}}] = J\sum_{i=1}^{I}\mathbb{E}(\bar{Y}_{i.}-\bar{Y}_{..})^2 = J\sum_{i=1}^{I}\left[\alpha_i^2+\frac{(I-1)\sigma^2}{IJ}\right] = J\sum_{i=1}^{I}\alpha_i^2+(I-1)\sigma^2
$$

□

Remark 38. **(Notes on the Sum of Squares)** SS_{W} may be used to estimate σ^2, where the estimate is:

$$s_p^2 = \frac{SS_{\text{W}}}{I(J-1)}$$

which is unbiased. The subscript p stands for pooled. Estimates of σ^2 from the I treatments are pooled together, since:

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

where s_i^2 is the sample variance in the i-th group.

Remark 39. **(Introduction to the Test)** If all the α_i are equal to zero, then the expectation of $SS_{\text{B}}/(I-1)$ is also σ^2. In this case, $SS_{\text{W}}/[I(J-1)]$ and $SS_{\text{B}}/(I-1)$ should be abount equal. If some of the α_i are non-zero, SS_{B} will be inflated.

Theorem 6.2. *If the errors are independent and normally distributed with means* 0 *and variance* σ^2*, then we have* $SS_W/\sigma^2 \sim \chi^2_{I(J-1)}$*. If additionally, the* α_i *are all equal to zero, then* $SS_B/\sigma^2 \sim \chi^2_{I-1}$ *and it is independent of* SS_W*.*

Proof. Let's consider the distribution function over random variable, as we have:

- We consider SS_{W}, where we have:

$$\frac{1}{\sigma^2}\sum_{j=1}^{J}(Y_{ij} - \bar{Y}_{i.})^2 \sim \chi^2_{J-1}$$

  There are I such sums in SS_{W}, they are independent of each other. The sum of I independent χ^2_{J-1} random variable give a $\chi^2_{I(J-1)}$. This also applied to SS_{B} noting that $\text{var}(\bar{Y}_{i.}) = \sigma^2/J$

- Now, we will show that 2 sums of square are independent of each other.
  - SS_{W} is a function of vector $\boldsymbol{U}$, which has the element $Y_{ij} - \bar{Y}_{i.}$, where $i = 1, \dots, I$ and $j = 1, \dots, J$
  - SS_{B} is a function of vector $\boldsymbol{V}$ whose element are $\bar{Y}_{i.}$ where $i = 1, \dots, I$, since $\bar{Y}_{..}$ can be obtained from $\bar{Y}_{i.}$

  It is sufficent to how that these 2 vectors are independent of each other, we consider:

  - If $i \neq i'$ then $Y_{ij} - \bar{Y}_{i.}$ and $\hat{Y}_{i'.}$ are independent since they are function of differeces observations.
  - On the other hand, $Y_{ij} - \bar{Y}_{i.}$ and $\bar{Y}_{i.}$ are independent by the previous result.

  This completes the proof of the thoerem.

□

Definition 6.2. (F Statistics) We use the following statistics:

$$F = \frac{SS_{\text{B}}/(I-1)}{SS_{\text{W}}/[I(J-1)]}$$

And it is used to the the following null hypothesis:

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_I = 0$$

If the null hypothesis is true, the F-statistics should be close to 1, and if it is false, the statistics should be larger. If the null hypothesis is false the numerator reflects variation between the different groups as well as variation within groups.

Theorem 6.3. *Under the assumption that the errors are noramlly distributed, the null distribution of F is F distribution with $I-1$ and $I(J-1)$ degree of freedom.*

Proof. The theorem follows from theorem above and for the definition of the F distribution. □

Remark 40. **(When number are not necessarily equal)** The analysis is the same as for the case of equal sample sizes. Suppose that there are J_i observation under treatment i, for $i = 1, \dots, I$. The basic identity still holds:

$$\sum_{i=1}^{I}\sum_{j=1}^{J_i}(Y_{ij} - \bar{Y}_{..})^2 = \sum_{i=1}^{I}\sum_{j=1}^{J_i}(Y_{ij} - \bar{Y}_{i.})^2 + \sum_{i=1}^{I} J_i(\bar{Y}_i - \bar{Y}_{..})^2$$

By reasoning similar to that used here for the simple case, as it can be shown that:

$$\mathbb{E}[SS_{\text{W}}] = \sigma^2 \sum_{i=1}^{I}(J_i - 1) \qquad \mathbb{E}[SS_{\text{B}}] = (I-1)\sigma^2 + \sum_{i=1}^{I} J_i\alpha_i^2$$

The degree of freedom for these sum of squares are $\sum_{i=1}^{I} J_i - I$ and $I-1$, respectively.

6.2 Problem of Multiple Comparisons

Remark 41. We are interested in comparing pairs or groups of treatments and estimating the treatment means and their differeces. The naive approach is to compare all pairs of treatment means using t test:

- Although each individual comparison would have a type I error rate of α
- The collection of all Comparisons considered simulataneous would not.

Definition 6.3. (Tukey's Method) It is used to construct confidence intervals for the differences of all pairs of means.

- If the sample sizes are all equal and the errors are normally distributed with a constant variance
- The centered sample means: $\bar{Y}_{i.} - \mu_i$ are independent and distributed with $\mathcal{N}(0, \sigma^2/J)$, where $\sigma^2 \approx s_p^2$.

Tukey's method is based on the probability distribution of the random variable:

$$\max_{i_1,i_2} \frac{\left|(\bar{Y}_{i_1.} - \mu_{i_1}) - (\bar{Y}_{i_2.} - \mu_{i_2})\right|}{s_p/\sqrt{J}}$$

where maximum is taken over all pairs. This distribution is called studentized range distribution with parameter I (number of samples being compared) and $I(J-1)$ (degree of freedom in s_p).

Remark 42. **(Confidence Bound for Turkey Method)** The upper 100α percentage point of the distribution is denoted by $q_{I,I(J-1)}(\alpha)$. Now, we have:

$$\begin{aligned}&\mathbb{P}\left[\left|(\bar{Y}_{i_1.} - \mu_{i_1}) - (\bar{Y}_{i_2.} - \mu_{i_2})\right| \le q_{I,I(J-1)}(\alpha)\frac{s_p}{\sqrt{J}}, \text{ for all } i_1, i_2\right] \\ &\quad= \mathbb{P}\left[\max_{i_1,i_2}\left|(\bar{Y}_{i_1.} - \mu_{i_1}) - (\bar{Y}_{i_2.} - \mu_{i_2})\right| \le q_{I,I(J-1)}(\alpha)\frac{s_p}{\sqrt{J}}\right] = 1-\alpha\end{aligned}$$

This can be converted to confidence interval as that holds for all differeces $\mu_{i_1} - \mu_{i_2}$ with confidence $100(1-\alpha)\%$. The interval are:

$$\bar{Y}_{i_1.} - \bar{Y}_{i_2.} \pm q_{I,I(J-1)}(\alpha)\frac{s_p}{\sqrt{J}}$$

Definition 6.4. (Bonferroni Method) If k null hypotheses are to be tested, a desired overally type I error rate of at most α can be guarantee by testing each null hypothesis at level α/k, and so if k confidence intervals are each formed to have a confidence level $100(1-\alpha/k)\%$, they hold simulataneously with confidence interval of at least $100(1-\alpha)\%$

Definition 6.5. (Kruskal-Wallis Test) The observations are assumed to be independent, but no particular distributional form. We consider:

$$R_{ij} = \text{ the rank of } Y_{ij} \text{ in the sample.}$$

Let's consider the following quantities:

$$\bar{R}_{i.} = \frac{1}{J_i}\sum_{j=1}^{J_i} R_{ij} \qquad \bar{R}_{..} = \frac{1}{N}\sum_{i=1}^{I}\sum_{j=1}^{J_i} R_{ij} = \frac{N+1}{2} \qquad SS_{\text{B}} = \sum_{i=1}^{I} J_i(\bar{R}_{i.} - \bar{R}_{..})^2$$

SS_{B} is the measure of dispersion of $\bar{R}_{i.}$ where the larger SS_{B} is the stronger is the evidence against the null hypotheses. The exact null distribution of this statistics for various combination of I and J_i can be enumerated. Or, we can use the statistics:

$$K = \frac{12}{N(N+1)} SS_{\text{B}}$$

is approximately distributed as χ^2_{I-1}. The value of K can be found by running the ranks through and analysis of variance program. It can be shown that:

$$K = \frac{12}{N(N+1)}\left(\sum_{i=1}^{I} J_i \bar{R}_i^2\right) - 3(N-1)$$

which is easier to compute by hand.

6.3 Two-Way Layout

Definition 6.6. (Two-Way Layout) Two-Way Layout is an experimental design involving 2 factors. The level of one factor might be various drugs and the level of the other factor might be genders. If there are I levels of one factor and J of the other, then there are $I \times J$ combinations. We will assume that K independent observations are taken for each of these combination. We will assume that there are $K > 1$ observations per cell.

Remark 43. **(Statistical Models)** This leads to the simple additive model as:

$$\hat{Y}_{ij} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j$$

We use the $\hat{Y}_{ij}$ to denote the fitted or predicted value of Y_{ij}. According to this additive model, we have:

$$\hat{Y}_{i1} - \hat{Y}_{i2} = (\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_1) - (\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_2) = \hat{\beta}_1 - \hat{\beta}_2$$

This may not always be the case as there can be *interaction* between each factors, and so this can be incorporated into the model to make it fit the data exactly. Consider the residual in cell ij to be:

$$Y_{ij} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \hat{\delta}_{ij}$$

Please note that the transformation can be used to stabilize the variance. Finally, to include the random error the model is given as:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \delta_{ij} + \varepsilon_{ijk}$$

where $\varepsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$, thus we have the following expected value: $\mathbb{E}[Y_{ijk}] = \mu + \alpha_i + \beta_j + \delta_{ij}$. The parameter will satisfy the following constraints to be:

$$\sum_{i=1}^{I} \alpha_i = 0 \qquad \sum_{j=1}^{J} \beta_j = 0 \qquad \sum_{i=1}^{I} \delta_{ij} = \sum_{j=1}^{J} \delta_{ij} = 0$$

Proposition 6.1. ***(MLE Estimate of Statistical Model)*** *The cell* ij *are normally distributed with mean* $\mu + \alpha_i + \beta_j + \delta_{ij}$ *and variance* σ^2*. The MLE, given the constraints, is*

$$\begin{aligned}
\hat{\mu} &= \bar{Y}_{...} \\
\hat{\alpha}_i &= \bar{Y}_{i..} - \bar{Y}_{...} \qquad i = 1, \dots, I \\
\hat{\beta}_j &= \bar{Y}_{.j.} - \bar{Y}_{...} \qquad j = 1, \dots, J \\
\hat{\delta}_{ij} &= \bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...}
\end{aligned}$$

Proof. We have the following log-likelihood:

$$l = -\frac{IJK}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{ijk} - \mu - \alpha_i - \beta_j - \delta_{ij})^2$$

Setting the derivative subjected to constraints gives us the MLE. □

Proposition 6.2. ***(Sum of Square Decomposition)*** *We can consider the sum of the square to be:*

$$\begin{aligned}
SS_A &= JK\sum_{i=1}^{I}(\bar{Y}_{i..} - \bar{Y}_{...})^2 \\
SS_B &= IK\sum_{j=1}^{J}(\bar{Y}_{.j.} - \bar{Y}_{...})^2 \\
SS_{AB} &= K\sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...})^2 \\
SS_E &= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{ijk} - \bar{Y}_{ij.})^2 \\
SS_{TOT} &= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{ijk} - \bar{Y}_{...})^2
\end{aligned}$$

The sum of square satisfy the algebraic identity:

$$SS_{TOT} = SS_A + SS_B + SS_{AB} + SS_E$$

Proof. This identity is proved by writing, follows:

$$Y_{ijk} - \bar{Y}_{...} = (Y_{ijk} - \bar{Y}_{ij.}) + (\bar{Y}_{i..} - \bar{Y}_{...}) + (\bar{Y}_{.j.} - \bar{Y}_{...}) + (\bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...})$$

Squaring both side, summing and verifying that the cross product vanishes. □

Proposition 6.3. *Under the assumption that the errors are independent with means* 0 *and variance* σ^2*:*

$$\begin{aligned}
\mathbb{E}[SS_A] &= (I-1)\sigma^2 + JK\sum_{i=1}^{I}\alpha_i^2 \\
\mathbb{E}[SS_B] &= (J-1)\sigma^2 + IK\sum_{j=1}^{J}\beta_j^2 \\
\mathbb{E}[SS_{AB}] &= (I-1)(J-1)\sigma^2 + K\sum_{i=1}^{I}\sum_{j=1}^{J}\delta_{ij}^2 \\
\mathbb{E}[SS_E] &= IJ(K-1)\sigma^2
\end{aligned}$$

Proof. The result of SS_{A}, SS_{B} and SS_{E}. Apply the lemma to SS_{TOT} as we have:

$$\begin{aligned}\mathbb{E}[SS_{\mathrm{TOT}}] &= \mathbb{E}\left[\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{ijk} - \bar{Y}_{...})^2\right] \\ &= (IJK-1)\sigma^2 + \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(\alpha_i + \beta_j + \delta_{ij})^2 \\ &= (IJK-1)\sigma^2 + JK\sum_{i=1}^{I}\alpha_i^2 + IK\sum_{j=1}^{J}\beta_j^2 + K\sum_{i=1}^{I}\sum_{j=1}^{J}\delta_{ij}^2\end{aligned}$$

The last step, we use the constraints on parameter. For example, we have:

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\alpha_i\beta_j = K\left(\sum_{i=1}^{I}\alpha_i\right)\left(\sum_{j=1}^{J}\beta_j\right) = 0$$

The values of expectation now follows. □

Theorem 6.4. *Assume that the error are independent and noramlly distributed with mean* 0 *and variance* σ^2*, then:*

- SS_E/σ^2 *follows a* χ^2*-distribution with* $IJ(K-1)$ *degree of freedom.*
- *Under null hypotheses:* $H_A : \alpha_i = 0, i = 1, \dots, I$ *where* SS_A/σ^2 *follows a* χ^2_{I-1}*-distribution*
- *Under null hypotheses:* $H_B : \beta_j = 0, j = 1, \dots, J$ *where* SS_B/σ^2 *follows a* χ^2_{J-1}*-distribution*
- *Under null hypotheses:* $H_{AB} : \beta_{ij} = 0, i = 1, \dots, I, j = 1, \dots, J$ *where* SS_{AB}/σ^2 *follows a* $\chi^2_{(I-1)(J-1)}$*-distribution*
- *Sums of squares are independently distributed*

Remark 44. **(On the use of F-Test)** The format of F-Test is the same. The mean squares are the sums of squares divided by their degree of freedom and F statistics are ratios of means squares. Let's consider the example:

- We have the following quantities $\mathbb{E}[MS_{\mathrm{A}}] = \sigma^2 + (JK/(I-1))\sum_i \alpha_i^2$ and $\mathbb{E}[MS_{\mathrm{E}}] = \sigma^2$
- If the ratio $MS_{\mathrm{A}}/MS_{\mathrm{E}}$ is large, it suggested that some α_i is non-zero.
- The null distribution of this F-statistics is $F_{(I-1),IJ(K-1)}$

6.4 Randomized Block Design

Definition 6.7. (Randomized Block Design) We want to study the effects of I different fertilizers, with J relatively homogeneous plots of land, each is divided into I plots. Within each block the assignment of fertilizer to plot is made at random, by comparing fertilizers within blocks, the variablity between blocks, which would contribute "noise" to the result is control.

Remark 45. **(Deriving the Null Distribution)** The null distribution of a test statistics can be dervied from the permuation argument just like null distribution of the Mann-Whitney test. The parameteric test can be a good approximation as we use the following model:

$$a$$

We will assume no interaction between the blocks and treatements.

Proposition 6.4. *We can show that, using the same calculation as above result, and consider no interaction:*

$$\mathbb{E}[MS_A] = \sigma^2 + \frac{J}{I-1}\sum_{i=1}^{I}\alpha_i^2$$
$$\mathbb{E}[MS_B] = \sigma^2 + \frac{I}{J-1}\sum_{j=1}^{J}\beta_j^2$$
$$\mathbb{E}[MS_{AB}] = \sigma^2$$

Remark 46. We can see that we can estimate σ^2 from MS_{AB}. The mean squares are independently distributed, F test can be performed to test, the hypotheses: $H_A : \forall i \in [I] : \alpha_i = 0$ uses the following statistics

$$F = \frac{MS_{\text{A}}}{MS_{\text{AB}}}$$

where under H_A, this statistics follows an F-distribution with $I-1$ and $(I-1)(J-1)$ degree of freedom. Contrary to the assumption, there is an interaction then:

$$\mathbb{E}[MS_{\text{AB}}] = \sigma^2 + \frac{1}{(I-1)(J-1)}\sum_{i=1}^{I}\sum_{j=1}^{J}\delta_{ij}^2$$

As MS_{AB} will tend to overestimate σ^2 making F statistics to be small that it should be.

Definition 6.8. (Friedman's Test) Like all non-parameteric methods, Friedman's test relies on ranks and doesn't assume normality. Within each of J blocks, the observation is ranked. To test the hypothesis that there is no effect due to factor corresponding to treatments I, we use the following statistics:

$$SS_{\text{A}} = J\sum_{i=1}^{I}(\bar{R}_{i.} - \bar{R}_{..})^2$$

Under null hypothesis there is no treatment effect, the permuation distribution of the statistics can be calculated.

Definition 6.9. (Approxiation of Friedman's Test) For the large sample sizes, we can use the approximation of friedman's test where the null distribution is, given as:

$$Q = \frac{12J}{I(I+1)}\sum_{i=1}^{I}(\bar{R}_{i.} - \bar{R}_{..})^2$$

is approximately χ^2_{I-1}.

7 The Analysis of Categorical Data

7.1 Fisher's Exact Test

Remark 47. **(Setting for the Tests)** Let's consider the data that we are given as: We want the see whether

	Variation 1	**Variation 2**	Total
Category 1	N_{11}	N_{12}	$n_{1.}$
Category 2	N_{21}	N_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	$n_{..}$

the count in each category is affected by the some variation of data or not (the null hypothesis is that thet are all randomly assigned). There are auxillary variables denoted (total).

Remark 48. **(Probability Under Null Hypothesis)** Under the null hypothesis (randomly generated), and so the probability that $N_{11} = n_{11}$ is given as:

$$p(n_{11}) = \frac{\binom{n_{1.}}{n_{11}}\binom{n_{2.}}{n_{21}}}{\binom{n_{..}}{n_{.1}}}$$

We can use N_{11} as the test statistics for testing the null hypothesis. We can generate the table to create 2 sided rejects for extreme value of N_{11}

7.2 χ^2-Test for Homogeneity

Remark 49. **(Settings for χ^2-Test)** We consider the larger setting compared to Fisher's exact test, where we comparing J multinomial distribution each having I categories. If the probability of i-th category of j-th multinomial is denoted as π_{ij}, the null hypothesis is:

$$H_0 : \pi_{i1} = \pi_{i2} = \cdots = \pi_{iJ} \qquad i = 1, \dots, J$$

Under H_0 each of the J multinomial has the same probability for the i-th category as π_i.

Proposition 7.1. *Under H_0, the MLE of the parameter $\pi_1, \pi_2, \dots, \pi_I$ are given as:*

$$\hat{\pi}_i = \frac{n_{i.}}{n_{..}} \qquad i = 1, \dots, I$$

where $n_{i.}$ is the total number of response in the i-th category and $n_{..}$ is the grand total number of response.

Proof. Since the multinomial distribution are independent:

$$\begin{aligned} \text{lik}(\pi_1, \pi_2, \dots, \pi_I) &= \prod_{j=1}^{J} \binom{n_{.j}}{n_{1j}n_{2j}\cdots n_{Ij}} \pi_1^{n_{1j}}\pi_2^{n_{2j}}\cdots\pi_I^{n_{Ij}} \\ &= \pi_1^{n_{1j}}\pi_2^{n_{2j}}\cdots\pi_I^{n_{Ij}} \prod_{j=1}^{J} \binom{n_{.j}}{n_{1j}n_{2j}\cdots n_{Ij}} \end{aligned}$$

Consider maximizing the log-likelihood subject to constraint $\sum_{i=1}^{I} \pi_i = 1$. Introducing multiplier, we have to maximizing:

$$\mathcal{L}(\pi, \lambda) = \sum_{j=1}^{J} \log\binom{n_{.j}}{n_{1j}n_{2j}\cdots n_{Ij}} + \sum_{i=1}^{I} n_{i.}\log\pi_i + \lambda\left(\sum_{i=1}^{I}\pi_i - 1\right)$$

Now, we have:

$$\begin{aligned} \frac{\partial l}{\partial \pi_i} &= \frac{n_{i.}}{\pi_i} + \lambda \qquad i = 1, \dots, I \\ \iff \hat{\pi}_i &= -\frac{n_i}{\lambda} \end{aligned}$$

Summing over both sides and applying the constraint, we find that $\lambda = -n_{..}$ and the theorem is proven. □

Definition 7.1. (Peason's χ^2-Test) For j-th multinomial, the expected count in the i-th category is the etimated probability of the cell times the total number of observation for j-th multinomial:

$$E_{ij} = \frac{n_{i.}}{n_{..}} n_{.j}$$

This gives us the Peason's χ^2-statistics as we have:

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - n_{i.}n_{.j}/n_{..})^2}{n_{i.}n_{.j}/n_{..}}$$

For large sample size, the approximate null distribution of this statistics is χ^2. We have the degree of freedom are number of independent counts minus the number of independent parameter:

- Each multinomial has $I-1$ independent counts, since the total are fixed.
- $I-1$ independent parameter have been estimated.

And so the degree of freedom are given as $J(I-1)-(I-1)=(I-1)(J-1)$.

7.3 χ^2-Test of Independent

Definition 7.2. (Contingency Table) We will discuss the statistical analysis of sample of size n cross-classifed in table with I rows and J columns. This configuration is called contingency table.

Remark 50. **(Settings for the Test)** We are interested in the relationship between factors on the table. The joint distribution of the counts n_{ij} where $i=1,\dots,I$ and $j=1,\dots,J$ is multinomial with cell probabilities denoted as:

$$\pi_{i.} = \sum_{j=1}^{J} \pi_{ij} \qquad \pi_{.j} = \sum_{i=1}^{I} \pi_{ij}$$

Both are the marginal probability that the observation will fall in i-th row or j-columns. If both row and columns are independent of each other then: $\pi_{ij} = \pi_{i.}\pi_{.j}$. This leads to the following null hypothesis:

$$H_0 : \pi_{ij} = \pi_{i.}\pi_{.j} \qquad i = 1,\dots,I \quad j = 1,\dots,J$$

Remark 51. **(Defining the χ^2-Test)** Let's consider the MLE estimate under each hypothesis

- Under H_0 is the MLE of π_{ij} is given as:

$$\hat{\pi}_{ij} = \hat{\pi}_{i.}\hat{\pi}_{.j} = \frac{n_{i.}}{n}\frac{n_{.j}}{n}$$

- Under alternative MLE of π_{ij} is given as:

$$\tilde{\pi}_{ij} = \frac{n_{ij}}{n}$$

Now we consider χ^2-test as we have:

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(O_{ij}-E_{ij})^2}{E_{ij}} = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij}-(n_{i.}n_{.j})/n)^2}{(n_{i.}n_{.j})/n}$$

where O_{ij} are the observation count as we have n_{ij}. The expected count is $E_{ij} = n\hat{\pi}_{ij} = (n_{i.}n_{.j})/n$.

- Let's consider the degree of freedom as under Ω, the cell probabilities sum to 1 as it has the dimension to be $IJ-1$.
- Under the null hypothesis, the marginal probabilities are estimated from the data are specified to $(I-1)+(J-1)$

We have the following degree of freedom:

$$\text{df} = IJ - 1 - (I-1) - (J-1) = (I-1)(J-1)$$

7.4 Matched-Pairs Designs

Remark 52. **(Setting for the test)** We consider the following table

	No Cure (Sibling)	**Cure (Sibling)**	Total
No Cure (Patient)	π_{11}	π_{12}	$\pi_{1.}$
Cure (Patient)	π_{21}	π_{22}	$\pi_{2.}$
Total	$\pi_{.1}$	$\pi_{.2}$	1

The appropriate null hypothesis is $\pi_{i.} = \pi_{.i}$, where $i = 1, 2$ (the probabilities of cure and no cure should be the same for patient and sibling), and so we have:

$$\pi_{11} + \pi_{12} = \pi_{11} + \pi_{21} \qquad \pi_{12} + \pi_{22} = \pi_{21} + \pi_{22}$$

The equation is simplified to $\pi_{12} = \pi_{21}$, where the null hypothesis is thus:

$$H_0 : \pi_{12} = \pi_{21}$$

Proposition 7.2. ***(MLE of Cell Probabilities)*** *Under the H_0, the MLE of the cell probabilities are:*

$$\hat{\pi}_{11} = \frac{n_{11}}{n} \qquad \hat{\pi}_{22} = \frac{n_{22}}{n} \qquad \hat{\pi}_{12} = \hat{\pi}_{21} = \frac{n_{12} + n_{21}}{2n}$$

Definition 7.3. (McNemar's Test) The contribution to the χ^2 statistics from n_{11} and n_{22} cells are equal to zero. The remainder of statistics is:

$$X^2 = \frac{[n_{12} - (n_{12} + n_{21})/2]^2}{(n_{12} + n_{21})/2} + \frac{[n_{21} - (n_{12} + n_{21})/2]^2}{(n_{12} + n_{21})/2} = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}$$

Let's consider the degree of freedom, as under Ω there are 3 free parameters (since there are 4 probability that are constrianted to one). On the null hypothesis, there are addiitonal constraint $\pi_{12} = \pi_{21}$ so there are 2 free parameter. Thus we have 1 degree of freedom.

7.5 Odd Ratios

Definition 7.4. (Odd) If an event A has probability $P(A)$ of occuring, the odds of A occuring are defined as (please note that this works with conditional probability):

$$\text{odds}(A) = \frac{P(A)}{1 - P(A)} \implies P(A) = \frac{\text{odds}(A)}{1 + \text{odds}(A)}$$

Definition 7.5. (Odds Ratio) We have the following:

$$\Delta = \frac{\text{odds}(D|X)}{\text{odds}(D|\bar{X})}$$

where $\bar{X}$ is the complementary element. This measures the influenced of some event X to the event D.

Remark 53. **(Setting for Test)** We consider how the odds and odds ratio could be estimated by sampling from a population with joint and marignal probability defined as:

	$\bar{D}$	D	Total
$\bar{X}$	π_{00}	π_{01}	$\pi_{0.}$
X	π_{10}	π_{11}	$\pi_{1.}$
Total	$\pi_{.0}$	$\pi_{.1}$	1

With this notation, as we have:

$$P(D|X) = \frac{\pi_{11}}{\pi_{10} + \pi_{11}} \qquad P(D|\bar{X}) = \frac{\pi_{01}}{\pi_{00} + \pi_{01}}$$

And, so we have:

$$\text{odds}(D|X) = \frac{\pi_{11}}{\pi_{10}} \qquad \text{odds}(D|\bar{X}) = \frac{\pi_{01}}{\pi_{00}} \qquad \Delta = \frac{\pi_{11}\pi_{00}}{\pi_{01}\pi_{10}}$$

The product of diagonal probabilities in the preceding table divided by the product of the off-diagonal probabilities.

Remark 54. **(Ways to Sample the Data)**

- *Naive Sample*: We can consider drawing a random sample from the entire population. But if the event D is rare, the total sample size would have to be quite large to guarantee that substantial number of D is included.
- *Prospective Study*: Fixed number of even X and $\bar{X}$ are sample, then incidence of D are compared. This allow use to compare $P(D|X)$ and $P(D|\bar{X})$ and the odd ratio. However π_{ij} can not be estiamte from the data.
- *Retrospective Study*: We fixed number of D and $\bar{D}$ and we compared the number of X and $\bar{X}$. We can estimate $P(X|D)$ and $P(X|\bar{D})$ by the proportion. But, we can't estimate $P(D|X)$ and $P(D|\bar{X})$ or the joint probability.

Proposition 7.3. *The odds ratio on the contingency table Δ can be expressed as:*

$$\Delta = \frac{\text{odds}(X|D)}{\text{odds}(X|\bar{D})}$$

Proof. This follows from the calculation of $P(X|D)$ and $1 - P(X|D)$ where we have:

$$P(X|D) = \frac{\pi_{11}}{\pi_{01} + \pi_{11}} \qquad 1 - P(X|D) = \frac{\pi_{01}}{\pi_{01} + \pi_{11}} \qquad \text{odds}(X|D) = \frac{\pi_{11}}{\pi_{01}} \qquad \text{odds}(X|\bar{D}) = \frac{\pi_{10}}{\pi_{00}}$$

We can see that the odds ratio Δ can be expressed as above, thus complete the proof. □

Remark 55. **(Retrospective Study - Odds Ratio)** We can't find the odds ratio of given the restrospective study but we can approximate it. Using the above result. where we replace π_{ij} with n_{ij} where n is the count of the observation.

Remark 56. **(Statistical Testing)** Since the value $\hat{\Delta}$ is non-linear function of the counts, we will have to use the boostrap to construct the approximation of the distribution $\hat{\Delta}$

8 Linear Least Squares

Remark 57. **(Vocabulary Used)** We consider the straight like is to fit the points (y_i, x_i) where $i = 1, \dots, n$ where we call the following components: y is called dependent/response variables. x is called independent/predictor variables.

Definition 8.1. (Objective) We are interested to minimize the following objective function:

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

where we consider to find the β_0 and β_1 that minimizes this value.

Proposition 8.1. ***(Solution of Simiple Linear Regression)*** *We can show that the expression of β_0 and β_1 can be found (given the dataset $\{(y_i, x_i)\}_{i=1}^n$) as:*

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Proof. We consider the derivative of the objective with respected to β_0 and β_1 as we have:

$$\frac{\partial S}{\partial \beta_0} = -2\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) \qquad \frac{\partial S}{\partial \beta_1} = -2\sum_{i=1}^n x_i(y_i - \beta_0 - \beta_1 x_i)$$

Setting the partial derivative to zero, we have the minimizer of $\hat{\beta}_0$ and $\hat{\beta}_1$ to be:

$$\sum_{i=1}^n y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^n x_i \qquad \sum_{i=1}^n x_i y_i = \hat{\beta}_0 \sum_{i=1}^n x_i + \hat{\beta}_1 \sum_{i=1}^n x_i^2$$

which we can solve for the $\hat{\beta}_0$ and $\hat{\beta}_1$ to obtain:

$$\hat{\beta}_0 = \frac{\left(\sum_{i=1}^n x_i^2\right)\left(\sum_{i=1}^n y_i\right) - \left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n x_i y_i\right)}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \qquad \hat{\beta}_1 = \frac{n\sum_{i=1}^n x_i y_i - \left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2}$$

With some rearrangement, and we have the required expression. □

Remark 58. **(Adding Non-Linearity)** We can consider the non-linear transformation of the input x_i before perform the linear Least square to increase the capacity of the model.

Definition 8.2. (Linear Least Square) It is a function of the form:

$$f(x_1, x_2, \dots, x_{p-1}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{p-1} x_{p-1}$$

This involves p unknown parameters $\beta_0, \beta_1, \dots, \beta_{p-1}$ as we fit the n data points:

$$\begin{gathered} y_1, x_{11}, x_{12}, \dots, x_{1,p-1} \\ y_2, x_{21}, x_{22}, \dots, x_{2,p-1} \\ \vdots \\ y_n, x_{n1}, x_{n2}, \dots, x_{n,p-1} \end{gathered}$$

The function $f(x)$ is called linear regression of y on x. We will always assume that $p < n$.

8.1 Simple Linear Regression

Definition 8.3. (Statistical Model) We consider the observed value of y is a linear function x plus the random noise:

$$y_i = \beta_0 + \beta_1 x_i + e_i \qquad i = 1, \dots, n$$

Here e_i is the independent random variable with $\mathbb{E}[e_i] = 0$ and $\text{var}(e_i) = \sigma^2$. Furthermore, x_i is assumed to be fixed. We will consider the statistics of β_0 and β_1, which are $\hat{\beta}_0$ and $\hat{\beta}_1$ respectively.

Proposition 8.2. *Under the assumption of the standard statistical model, the least square estimate are unbiased as $\mathbb{E}[\hat{\beta}_j] = \beta_j$ for $j = 0, 1$*

Proof. We will consider the proof for $\hat{\beta}_0$ only as the proof for β_1 is similar. Note that $\mathbb{E}[y_i] = \beta_0 + \beta_1 x_i$:

$$\begin{aligned}\mathbb{E}[\hat{\beta}_0] &= \frac{\left(\sum_{i=1}^n x_i^2\right)\left(\sum_{i=1}^n \mathbb{E}[y_i]\right) - \left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n x_i \mathbb{E}[y_i]\right)}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \\ &= \frac{\left(\sum_{i=1}^n x_i^2\right)\left(n\beta_0 + \beta_1 \sum_{i=1}^n x_i\right) - \left(\sum_{i=1}^n x_i\right)\left(\beta_0 \sum_{i=1}^n x_i + \beta_1 \sum_{i=1}^n x_i^2\right)}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \\ &= \beta_0\end{aligned}$$

Thus complete the proof. □

Theorem 8.1. *Under the assuption of the standard statistical model, we have:*

$$\text{var}(\hat{\beta}_0) = \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \qquad \text{var}(\hat{\beta}_1) = \frac{n\sigma^2}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2}$$

$$\text{cov}(\hat{\beta}_0, \hat{\beta}_1) = \frac{-\sigma^2 \sum_{i=1}^n x_i}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2}$$

Proof. We will consider the more general proof later. □

Definition 8.4. (Residual Sum of Squares) We define RSS to be:

$$\text{RSS} = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

Remark 59. **(Statistical Testing)** The value of σ^2 is used to find the variance $\hat{\beta}_0$ and $\hat{\beta}_1$. Replacing the σ^2 by s^1 yielding estimates that we will denote $s^2_{\hat{\beta}_0}$ and $s^2_{\hat{\beta}_1}$. We will show that:

$$s^2 = \frac{\text{RSS}}{n-2}$$

It is unbiased estimate of σ^2. If the error e_i are independent normal random variable, then the linear combination of them are normal distributed as well. Furthermore, we have:

- If e_i are independent and x_i satisfies certain assumption, a version of CLT implies that (for large n)m the estimated slow and intercept are approximately normally distributed.
- The normality assuption, makes possible to construct of confidence interval and hypothesis test, which can be shown that:

  $$\frac{\hat{\beta}_i - \beta_i}{s_{\hat{\beta}_i}} \sim t_{n-2}$$

  We allows the t distribution to be used for CI and hypothesis tests.

Remark 60. **(Correlation)** Let's start with finding the correlation coefficient, which is equal to:

$$r = \frac{s_{xy}}{\sqrt{s_{xx}s_{yy}}}$$

where we have:

$$s_{xx} = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2 \qquad s_{yy} = \frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^2 \qquad s_{xy} = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

Remark 61. **(On connection between Correlation)** We can show that the slope of the least square line is given by:

$$\hat{\beta}_1 = \frac{s_{xy}}{s_{xx}} \qquad r = \hat{\beta}_1\sqrt{\frac{s_{xx}}{s_{yy}}}$$

The correlation is zero iff the slope is zero. Furthermore, if $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, as we have the $\hat{\beta}_1$ is expressed the terms of r, then after some manipulation, we have:

$$\frac{\hat{y} - \bar{y}}{\sqrt{s_{yy}}} = r\frac{x - \bar{x}}{\sqrt{s_{xx}}}$$

We can interpret the following equation to be:

- Suppose that $r > 0$ and that x is one standard deviation greater than its average, the y is r standard deviation bigger than its average.
- The predicted value thus deviates from its average by few standard deviation than does the predictor. (as $r \leq 1$)
- In unit of standard deviations, it is closer to its average than is the predictor.

8.2 Matrix Approach

Remark 62. **(Matrix Formulation)** Consider the model of the form:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1}x_{p-1}$$

It is to be fit to data, which we denote as $y_i, x_{i1}, x_{i2}, \dots, x_{ip-1}$ as we have $i = 1, \dots, n$. We have:

- $\boldsymbol{Y}$ is a vector of observations y_i where $i = 1, \dots, n$.
- $\boldsymbol{\beta}$ is the unknown $\beta_0, \dots, \beta_{p-1}$.
- $\boldsymbol{X}_{n\times p}$ being the matrix:

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1,p-1} \\ 1 & x_{21} & x_{22} & \cdots & x_{2,p-1} \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{n,p-1} \end{bmatrix}$$

We have the predicted value to be given by $\hat{\boldsymbol{Y}} = \boldsymbol{X}\boldsymbol{\beta}$. We want to find $\boldsymbol{\beta}$ to minimize:

$$\begin{aligned} S(\boldsymbol{\beta}) &= \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_{p-1}x_{i,p-1})^2 \\ &= \left\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\right\|^2 = \left\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\right\|^2 \end{aligned}$$

Note that the residual can be find out as $\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is the solution to the optimization problem.

Proposition 8.3. ***(Solution of Least Square)*** *If $\boldsymbol{X}^T\boldsymbol{X}$ is non-singular, the formal solution is given as:*

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

Proof. If differetiate the S with respected to each β_k, then we see that the minimizer of $\hat{\beta}_0, \dots, \hat{\beta}_{p-1}$ satisfies the following equation:

$$\begin{aligned} n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^n x_{i1} + \cdots + \hat{\beta}_{p-1} \sum_{i=1}^n x_{i,p-1} &= \sum_{i=1}^n y_i \\ \hat{\beta}_0 \sum_{i=1}^n x_{ij} + \hat{\beta}_1 \sum_{i=1}^n x_{i1}x_{ik} + \cdots + \hat{\beta}_{p-1} \sum_{i=1}^n x_{ik}x_{i,p-1} &= \sum_{i=1}^n y_i x_{ik} \qquad k = 1, \dots, p-1 \end{aligned}$$

This can be written in the matrix form as $\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{Y}$ this is called normal equation, and the results above follows. □

Lemma 8.1. *If $\boldsymbol{X}^T\boldsymbol{X}$ is non-singular iff the rank of $\boldsymbol{X}$ equals to p.*

Proof. Suppose that $\boldsymbol{X}^T\boldsymbol{X}$ is singular. There exiss a non-zero vector $\boldsymbol{u}$ such that: $\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{u} = \boldsymbol{0}$. Multiply the left-handside of this equation by $\boldsymbol{u}^T$, we have:

$$\boldsymbol{0} = \boldsymbol{u}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{u} = (\boldsymbol{X}\boldsymbol{u})^T(\boldsymbol{X}\boldsymbol{u})$$

And so $\boldsymbol{X}\boldsymbol{u} = \boldsymbol{0}$, thu rank $\boldsymbol{X}$ is less than p. Now suppose that the rank of $\boldsymbol{X}$ is less than p, then there is a vector $\boldsymbol{u}$ such that $\boldsymbol{X}\boldsymbol{u} = \boldsymbol{0}$. Then $\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{u} = \boldsymbol{0}$ hence $\boldsymbol{X}^T\boldsymbol{X}$ is singular. □

Remark 63. **(On equivalent to eariler derivation)** Let's consider each matrices:

$$\begin{aligned} \boldsymbol{X}^T\boldsymbol{X} &= \begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix} \\ (\boldsymbol{X}^T\boldsymbol{X})^{-1} &= \frac{1}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \begin{bmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{bmatrix} \\ \boldsymbol{X}^T\boldsymbol{Y} &= \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{bmatrix} \end{aligned}$$

And so, we have:

$$\begin{aligned} \hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \\ &= \frac{1}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \begin{bmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{bmatrix} \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{bmatrix} \\ &= \frac{1}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \begin{bmatrix} \left(\sum_{i=1}^n x_i^2\right)\left(\sum_{i=1}^n y_i\right) - \left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n x_i y_i\right) \\ n\sum_{i=1}^n x_i y_i - \left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right) \end{bmatrix} \end{aligned}$$

Thus the equivalent is established.

8.3 Statistical Properties of Least Square

Definition 8.5. (Mean Vector and Covariance Matrix) Given the random vector, $\boldsymbol{Y}$, the element, which are jointly distributed random variables:

$$\mathbb{E}[Y_i] = \mu_i \qquad \operatorname{cov}(Y_i, Y_j) = \sigma_{ij}$$

The mean vector $\boldsymbol{\mu}_Y$ and the covariance matrix $\boldsymbol{\Sigma}$ of $\boldsymbol{Y}$, are defined as:

$$\mathbb{E}[\boldsymbol{Y}] = \boldsymbol{\mu}_Y = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix} \qquad \boldsymbol{Z} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{bmatrix}$$

Proposition 8.4. *If $\boldsymbol{Z} = \boldsymbol{c} + \boldsymbol{A}\boldsymbol{Y}$ where $\boldsymbol{Y}$ is a random variable and $\boldsymbol{A}$ a matrix with $\boldsymbol{c}$ a fixed vector, then:*

$$\mathbb{E}[\boldsymbol{Z}] = \boldsymbol{c} + \boldsymbol{A}\mathbb{E}[\boldsymbol{Y}]$$

Proof. The i-th components of $\boldsymbol{Z}$ is given as:

$$Z_i = c_i + \sum_{j=1}^{n} a_{ij} Y_j \implies \mathbb{E}[Z_i] = Z_i = c_i + \sum_{j=1}^{n} a_{ij}\mathbb{E}[Y_j]$$

The implication follows from the linearity of the expectation. As this can be written in matrix form, this completes the proof. □

Proposition 8.5. *Given the same setting as the above, if the covariance matrix of $\boldsymbol{Y}$ is $\boldsymbol{\Sigma}_{YY}$, then the covariance of $\boldsymbol{Z}$ is:*

$$\boldsymbol{\Sigma}_{ZZ} = \boldsymbol{A}\boldsymbol{\Sigma}_{YY}\boldsymbol{A}^T$$

Proof. The constant $\boldsymbol{c}$ doesn't affect the covariance:

$$\operatorname{cov}(Z_i, Z_j) = \operatorname{cov}\left(\sum_{k=1}^{n} a_{ik} Y_k, \sum_{l=1}^{n} a_{jl} Y_l\right) = \sum_{k=1}^{n}\sum_{l=1}^{n} a_{ik}a_{jl}\operatorname{cov}(Y_k, Y_l) = \sum_{k=1}^{n}\sum_{l=1}^{n} a_{ik}\sigma_{kl}a_{jl}$$

The last expression in ij element of the desired matrix. □

Proposition 8.6. *Let $\boldsymbol{X}$ be a random n vector with means $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ and let $\boldsymbol{A}$ be fixed matrix:*

$$\mathbb{E}[\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}] = \operatorname{tr}[\boldsymbol{A}\boldsymbol{\Sigma}] + \boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu}$$

Proof. The trace of square matrix is defined to be sum of diagonal terms, as we have:

$$\mathbb{E}[X_iX_j] = \sigma_{ij} + \mu_i\mu_j$$

We have the following:

$$\begin{aligned}\mathbb{E}\left(\sum_{i=1}^{n}\sum_{j=1}^{n} X_iX_ja_{ij}\right) &= \sum_{i=1}^{n}\sum_{j=1}^{n}\sigma_{ij}a_{ij} + \sum_{i=1}^{n}\sum_{j=1}^{n}\mu_i\mu_ja_{ij} \\ &= \operatorname{tr}[\boldsymbol{A}\boldsymbol{\Sigma}] + \boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu}\end{aligned}$$

□

Remark 64. **(Alternative Proof of Variance Estimator)** We are interested in finding $\mathbb{E}[\sum_{i=1}^{n}(X_i - \bar{X})^2]$ where X_i is uncorrelated random variable with common mean μ. Note that the vector $\bar{X}$ is given as:

$$\bar{X} = \frac{1}{n}\boldsymbol{1}^T\boldsymbol{X}$$

THe entries of check are $\bar{X}$ can be written as: $(1/n)\boldsymbol{1}\boldsymbol{1}^T\boldsymbol{X}$ and $\boldsymbol{A}$ can be written as:

$$\boldsymbol{A} = \boldsymbol{I} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^T$$

Thus, it is clear that:

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \|\boldsymbol{A}\boldsymbol{X}\|^2 = \boldsymbol{X}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{X} = \boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}$$

Note that the matrix $\boldsymbol{A}$ is symmetric and $\boldsymbol{A}^2 = \boldsymbol{A}$, where note that $\boldsymbol{1}^T\boldsymbol{1} = n$. Finally, consider the expectation of the summation, which we can use our results:

$$\mathbb{E}[\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}] = \sigma^2\operatorname{tr}[\boldsymbol{A}] + \boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu} = \sigma^2(n-1)$$

where $\boldsymbol{\mu} = \mu\boldsymbol{1}$, it can be verified that $\boldsymbol{A}\boldsymbol{\mu} = 0$, als trace $\boldsymbol{A} = n - 1$, so we have the value required.

Definition 8.6. (Cross-Covariance Matrix) Given the random vectors $\boldsymbol{Y} \in \mathbb{R}^{p\times 1}$ and $\boldsymbol{Z} \in \mathbb{R}^{m\times 1}$, then the cross-covariance of $\boldsymbol{Y}$ and $\boldsymbol{Z}$ is defined to be $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times m}$ with ij element $\sigma_{ij} = \text{cov}(Y_i, Z_j)$. The entries quanify the strengths of linear relationship between elements of $\boldsymbol{Y}$ and $\boldsymbol{Z}$.

Proposition 8.7. *Let $\boldsymbol{X}$ be a random vector with covariance matrix $\boldsymbol{\Sigma}_{XX}$ if $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X}$ and $\boldsymbol{Z} = \boldsymbol{B}\boldsymbol{X}$ where $\boldsymbol{A} \in \mathbb{R}^{p\times n}$ and $\boldsymbol{B} \in \mathbb{R}^{m\times n}$, where the cross-covariance matrix of $\boldsymbol{Y}$ and $\boldsymbol{Z}$ is:*

$$\boldsymbol{\Sigma}_{YZ} = \boldsymbol{A}\boldsymbol{\Sigma}_{XX}\boldsymbol{B}^T$$

Remark 65. **(Alternative Proof of Independence)** Consider a random vector $\boldsymbol{X}$ of size n with $\mathbb{E} = \mu\boldsymbol{I}$ and $\boldsymbol{\Sigma}_{XX} = \sigma^2\boldsymbol{I}$. Let $Y = \bar{X}$ and $\boldsymbol{Z}$ be vector with i-th element $\boldsymbol{X}_i - \bar{\boldsymbol{X}}$. Let's consider the $\boldsymbol{\Sigma}_{ZY} \in \mathbb{R}^{n\times 1}$ as we have:

$$\boldsymbol{Z} = \left(\boldsymbol{I} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^T\right)\boldsymbol{X} \qquad Y = \frac{1}{n}\boldsymbol{1}^T\boldsymbol{X}$$

From theorem above, we have:

$$\boldsymbol{\Sigma}_{ZY} = \left(\boldsymbol{I} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^T\right)(\sigma^2\boldsymbol{I})\left(\frac{1}{n}\boldsymbol{1}\right)$$

This comes $\mathbb{R}^{n\times 1}$ vector of zeros. Thus, the mean $\bar{X}$ is uncorrelated with each of $X_i - \bar{X}$ for $i = 1, \dots, n$. This implies that $\bar{X}$ and S^2 are independent of each other.

Remark 66. **(Least Squares Estimates)** We consider the following model to be:

$$Y_i = \beta_0 + \sum_{j=1}^{p-1}\beta_j x_{ij} + e_i \qquad i = 1, \dots, n$$

where e_i are the random error, as we have:

$$\mathbb{E}[\boldsymbol{e}_i] = 0 \qquad \text{var}(e_i) = \sigma^2 \qquad \text{cov}(e_i, e_j) = 0 \quad i \neq j$$

Given the matrix notation as we have $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, as we have:

$$\mathbb{E}[\boldsymbol{e}] = \boldsymbol{0} \qquad \boldsymbol{\Sigma}_{ee} = \sigma^2\boldsymbol{I}$$

Theorem 8.2. ***(Unbias)*** *Given the assumption that the error has mean* 0*, the least square estimate is unbiased.*

Proof. The least square estimate of $\boldsymbol{\beta}$ is given:

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}) \\ &= \boldsymbol{\beta} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{e}\end{aligned}$$

From the results aboue, we have the following expectation:

$$\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathbb{E}[\boldsymbol{e}] = \boldsymbol{\beta}$$

□

Theorem 8.3. ***(Covaraince Matrix of Least Square)*** *Under the assumption that the error have mean zero and uncorrelated with constant variance σ^2, the covariance matrix of the least square estimate $\hat{\boldsymbol{\beta}}$ is:*

$$\boldsymbol{\Sigma}_{\hat{\beta}\hat{\beta}} = \boldsymbol{\sigma}^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

Proof. From the results above, we have:

$$\begin{aligned}\boldsymbol{\Sigma}_{\hat{\beta}\hat{\beta}} &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}_{ee}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\ &= \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\end{aligned}$$

□

Remark 67. **(Recovery of Original Result)** We return to the case of fitting a straight like. From the computation of $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ as we have:

$$\boldsymbol{\Sigma}_{\hat{\beta}\hat{\beta}} = \frac{\sigma^2}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2}\begin{bmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{bmatrix}$$

And so we have the variance and covaraince results, which are the same as above.

Remark 68. **(Residual Vector)** Because σ^2 is the expected square value of an error e_i, its is natural to use the sample average squared the residual, as we have:

$$\hat{\boldsymbol{e}} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{Y} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{Y} - \boldsymbol{P}\boldsymbol{Y}$$

where $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ is an $n \times n$ matrix.

Lemma 8.2. *Let $\boldsymbol{P}$ be defined as before, then we have:*

$$\boldsymbol{P} = \boldsymbol{P}^T = \boldsymbol{P}^2 \qquad (\boldsymbol{I} - \boldsymbol{P}) = (\boldsymbol{I} - \boldsymbol{P})^T = (\boldsymbol{I} - \boldsymbol{P})^2$$

Remark 69. The $\boldsymbol{P}$ is the projection matrix that is $\boldsymbol{P}$ projects on the subspace of $\mathbb{R}^n$ spanned by the columns of $\boldsymbol{X}$. We may think geometrically of the fitted values, $\hat{\boldsymbol{Y}}$ as being the projection of $\boldsymbol{Y}$ onto subspace spanned by columns of $\boldsymbol{X}$.

Theorem 8.4. *Under the assumption that the error are uncorrelated with constant variance σ^2, an unbiased estimate of σ^2 is:*

$$s^2 = \frac{\left\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\right\|^2}{n-p}$$

The sum of squared residual, $\left\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\right\|^2$ is often denoted by RSS.

Proof. The sum of squared residual is, using the lemma:

$$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \|\boldsymbol{Y} - \boldsymbol{P}\boldsymbol{Y}\|^2 = \|(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}\|^2 = \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{P})^T(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}$$

We can compute the expected value of this quadratic form:

$$\mathbb{E}[\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}] = \mathbb{E}[\boldsymbol{Y}]^T(\boldsymbol{I} - \boldsymbol{P})\mathbb{E}[\boldsymbol{Y}] + \sigma^2\operatorname{tr}(\boldsymbol{I} - \boldsymbol{P})$$

Please note that $\mathbb{E}[\boldsymbol{Y}] = \boldsymbol{X}\boldsymbol{\beta}$ so we have:

$$(\boldsymbol{I} - \boldsymbol{P})\mathbb{E}[\boldsymbol{Y}] = \left[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\right]\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{0}$$

Furthermore, we consider the trace terms as we have:

$$\begin{aligned} \operatorname{tr}(\boldsymbol{I} - \boldsymbol{P}) &= \operatorname{tr}(\boldsymbol{I}) - \operatorname{tr}(\boldsymbol{P}) \\ &= n - \operatorname{tr}\left[\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\right] \\ &= n - \operatorname{tr}\left[\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right] \\ &= n - \operatorname{tr}[\boldsymbol{I}] = n - p \end{aligned}$$

adding them together given us the result. □

Proposition 8.8. *The covariance matrix of the residual is given by:*

$$\boldsymbol{\Sigma}_{\hat{\boldsymbol{e}}\hat{\boldsymbol{e}}} = (\boldsymbol{I} - \boldsymbol{P})(\sigma^2\boldsymbol{I})(\boldsymbol{I} - \boldsymbol{P})^T = \sigma^2(\boldsymbol{I} - \boldsymbol{P})$$

Definition 8.7. (Standardized Residual) To put residual in the familar scale corrsponding to the normal distribution with means 0 and variance is:

$$\frac{Y_i - \hat{Y}_i}{s\sqrt{1 - p_{ii}}}$$

where p_{ii} is the i-th diagonal element of $\boldsymbol{P}$.

Theorem 8.5. *If the error have the covariance matrix $\sigma^2 \boldsymbol{I}$, the residual are uncorrelated with the fitted values.*

Proof. The residual are $\hat{\boldsymbol{e}} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}$, and the fitted values are:

$$\hat{\boldsymbol{Y}} = \boldsymbol{P}\boldsymbol{Y}$$

from the theorem above, the cross-covaraince matrix of $\hat{\boldsymbol{e}}$ and $\hat{\boldsymbol{Y}}$ is given by:

$$\Sigma_{\hat{e}\hat{Y}} = (\boldsymbol{I} - \boldsymbol{P})(\sigma^2 \boldsymbol{I})\boldsymbol{P}^T = \sigma^2(\boldsymbol{P}^T - \boldsymbol{P}\boldsymbol{P}^T) = 0$$

Thus the theorem result is proven. □

Remark 70. **(Inference About $\boldsymbol{\beta}$)** We have the following observation of the result:

- Each components $\hat{\beta}_i$ of $\hat{\boldsymbol{\beta}}$ can be show that it sample $\mathcal{N}(\beta_i, \sigma^2 c_{ii})$, where $\boldsymbol{C} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}$
- The standard error of $\hat{\beta}_i$ may thus be estimated as $s_{\hat{\beta}_i} = s\sqrt{c_{ii}}$

We will use this result to construct the CI and hypothesis test. Under normality assumption is given as:

$$\frac{\hat{\beta}_i - \beta_i}{s_{\hat{\beta}_i}} \sim t_{n-p}$$

Now we have $100(1-\alpha)\%$ CI for β_i so that: $\hat{\beta}_i \pm t_{n-p}(\alpha/2)s_{\hat{\beta}_i}$

Remark 71. **(Test for Parameter)** To test the null hypothesis $H_0 : \beta_i = \beta_{i0}$ where β_{i0} is a fixed number, we can use the test statistics:

$$t = \frac{\hat{\beta}_i - \beta_{i0}}{s_{\hat{\beta}_i}}$$

Under the H_0 the test statistics follows the t_{n-p}. The most commonly tested null hypothesis is $H_0 : \beta_i = 0$, which states that x_i has no predicted value.

Remark 72. **(Test for Prediction)** We can see that the obvious estimate is given as $\hat{\mu}_0 = \boldsymbol{x}_0^T\hat{\boldsymbol{\beta}}$. The variance of this estimate is given as:

$$\text{var}(\hat{\mu}_0) = \boldsymbol{x}_0^T \boldsymbol{\Sigma}_{\hat{\beta}\hat{\beta}} \boldsymbol{x}_0 = \sigma^2 \boldsymbol{x}_0^T (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0$$

This variance can be estimated by substuiting s^2 for σ^2 as we have: $\hat{\mu}_0 \pm t_{n-p}(\alpha/2)s_{\hat{\mu}_0}$. Note that the variance depends on $\boldsymbol{x}_0$.

Definition 8.8. (Squared Multiple Correlated Coefficient) This coefficient is simply defined as the squared correlation of the dependent variable and fitted values. It can be shown that it is equal to:

$$R^2 = \frac{s_y^2 - s_{\hat{e}}^2}{s_y^2}$$

It is used as a crude measure of the strength of relationship that has been fitted by least squares.

8.4 Conditional Inferece, Unconditional Inference, and Bootstrap

Remark 73. **(Differece View)** Instead of consider $\boldsymbol{X}$ and $\boldsymbol{Y}$ to be constant like most of the analysis above, we consider both variable to be random and use the boostrap to quanify the uncertainty in parameter estimates.

Remark 74. **(Some notations)** We consider the design matrix $\boldsymbol{\Xi}$ and particular realization of this random matrix will be denoted as $\boldsymbol{X}$. The rows of $\boldsymbol{\Xi}$ will be denoted by $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \dots, \boldsymbol{\xi}_n$. In place of model $Y_i = \boldsymbol{x}_i\boldsymbol{\beta} + e_i$, where $\boldsymbol{x}_i$ is fixed and e_i is random with mean 0 and variance σ^2, where we have:

$$\mathbb{E}[Y|\boldsymbol{\xi} = \boldsymbol{x}] = \boldsymbol{X}\boldsymbol{\beta} \qquad \text{var}[Y|\boldsymbol{\xi} = \boldsymbol{x}] = \sigma^2$$

In the random $\boldsymbol{X}$ model, $\boldsymbol{Y}$ and $\boldsymbol{\xi}$ have a joint distribution and the data are modeled as $\boldsymbol{n}$ independent random vectors:

$$(Y_1, \boldsymbol{\xi}_1), (Y_n, \boldsymbol{\xi}_2), \dots, (Y_n, \boldsymbol{\xi}_n)$$

Let's consider how the mean and the variance of the parameter given the uncertainty in the data points:

- $\mathbb{E}[\hat{\boldsymbol{\beta}}|\boldsymbol{\Xi} = \boldsymbol{X}] = \boldsymbol{\beta}$. Using the nested expectation, we have:

$$\mathbb{E}[\hat{\boldsymbol{\beta}}] = \mathbb{E}[\mathbb{E}[\hat{\boldsymbol{\beta}}|\boldsymbol{\Xi}]] = \mathbb{E}[\hat{\boldsymbol{\beta}}] = \hat{\boldsymbol{\beta}}$$

- $\text{var}[\hat{\beta}_i|\boldsymbol{\Xi} = \boldsymbol{X}] = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{ii}$. Using the marginalized, we have:

$$\begin{aligned}\text{var}(\hat{\beta}_i) &= \text{var}[\mathbb{E}[\hat{\beta}_i|\boldsymbol{\Xi}]] + \mathbb{E}[\text{var}[\hat{\beta}_i|\boldsymbol{\Xi}]] \\ &= \text{var}(\beta_i) + \mathbb{E}[\sigma^2(\boldsymbol{\Xi}^T\boldsymbol{\Xi})^{-1}_{ii}] \\ &= \sigma^2\mathbb{E}[\boldsymbol{\Xi}^T\boldsymbol{\Xi}]^{-1}_{ii}\end{aligned}$$

This is highlt non-linear function of the random vectors $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \dots, \boldsymbol{\xi}_n$. This is hard to evaluate the analysically.

- Surprisingly, it turns out that the CI still holds at their nominal level of coverage. Let $C(\boldsymbol{X})$ denote the $100(1-\alpha)\%$ CI for β_j for the old model.
- Using the I_A denotes the indicator variable of the event A, we can express the fact that $100(1-\alpha)\%$ CI as: $\mathbb{E}[I\left\{\beta_j \in C(\boldsymbol{X})\right\}|\boldsymbol{\Xi} = \boldsymbol{X}] = 1-\alpha$
- Because the conditional probability of coverage is the same for every value of $\boldsymbol{\Xi}$, the unconditional probability of coverage $1-\alpha$:

$$\mathbb{E}[I\left\{\beta_j \in C(\boldsymbol{X})\right\}] = \mathbb{E}[\mathbb{E}[I\left\{\beta_j \in C(\boldsymbol{X})\right\}|\boldsymbol{\Xi} = \boldsymbol{X}]] = \mathbb{E}[1-\alpha] = 1-\alpha$$

  This is every useful result for forming the CI as we can use the old fixed-$\boldsymbol{X}$ model.

We can complete this section by discussing how the boostrap can be used to estimate the variability of the parameter estimate under the new model.