UNC BIOS 760

Summary

Creating my own project

Author:
Mingwei Fei

Special colaborator:
Bonnie

Jan, 2023

Contents

Chapter 1

Distributions

1.1 Exponential Distribution Family

1.1.1 Moment Generating Function

KGF could be used to directly get the expectation and variance, more common than MGF. To get KGF, we will need to write distribution in exponential distribution.

Suppose the exponential distribution family is

$$f(Y, \theta) = exp\left(\phi(y\theta - b(\theta) - c(y)) - 0.5s(y, \phi)\right)$$

The MGF of exponential family

$$\begin{aligned}
M_Y(t) = E[exp(yt)] &= \int exp(yt)exp\left(\phi(y\theta - b(\theta) - c(y)) - 0.5s(y, \phi)\right) \\
&= \int exp\left(\phi(y(\theta + t/\phi) - b(\theta) - c(y)) - 0.5s(y, \phi)\right) dy \\
&= exp(\phi[b(\theta + t/\phi) - b(\theta)]) \int exp\left(\phi(y(\theta + t/\phi) - b(\theta + t/\phi) - c(y)) - 0.5s(y, \phi)\right) dy \\
&= exp(\phi[b(\theta + t/\phi) - b(\theta)])
\end{aligned}$$

The KGF is $\phi[b(\theta + t/\phi) - b(\theta)]$, then we can get the expectation and variance

$$E(y) = \left.\frac{\partial K(t)}{\partial t}\right|_{t=0} = \dot{b}(\theta)$$

$$Var(y) = \left.\frac{\partial^2 K(t)}{\partial t\, \partial t}\right|_{t=0} = \phi^{-1}\ddot{b}(\theta)$$

The MGF/KGF has shown that we can use the derivative function to get expectation or variance other than using the integral. The efficiency of computation also could be shown in the getting the covariance matrix using Fisher Information.

1.2 Chi-Square Distribution

The $\chi^2(n)$ distribution is defined as the distribution that results from summing the squares of n independent random variables $N(0,1)$, so

$\chi^2(n)$ will denote both a Chi squared distribution with n degrees of freedom and a random variable with such distribution. Now, the pdf of the $\chi^2(n)$ distribution is

$$f_{\chi^2}(x,n) = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}, \qquad x \le 0$$

So, indeed the $\chi^2(n)$ distribution is a particular case of the $\Gamma(a,b)$ distribution with pdf

$$f_{\Gamma}(x,a,b) = \frac{1}{b^a\Gamma(a)} x^{a-1} e^{-\frac{x}{b}}, \qquad x \le 0$$

$\chi^2(n) \sim \Gamma(\frac{n}{2}, \frac{1}{2})$.

1.2.1 Non-Central Chi-Square

1.2.2 Moment Generating Function

We can get MGF for chi-square from $E[x^2t]$ and $E[(\mu+Z)^2t]$, where $Z \sim N(0,1)$. Let's prove it in two methods:

(i) Method 1:

$$\begin{aligned}
M_i(t) = E[x^2t] &= \frac{1}{\sqrt{2\pi}} \int exp(x^2t) exp\left(-\frac{(x-\mu)^2}{2}\right) dx \\
&= \frac{1}{\sqrt{2\pi}} \int exp\left((t-\frac{1}{2})x^2 + \mu x - \frac{\mu^2}{2}\right) dx \\
&= \frac{1}{\sqrt{2\pi}} \int exp\left(-\frac{1}{2}(1-2t)\{x^2 - \frac{2\mu x}{(1-2t)} + \frac{\mu^2}{(1-2t)^2}\} + \frac{\mu^2}{2(1-2t)} - \frac{\mu^2}{2}\right) dx \\
&= \frac{1}{\sqrt{(1-2t)}} \int \frac{(1-2t)}{\sqrt{2\pi}} exp\left(-\frac{(x-\frac{\mu}{1-2t})^2}{2(1-2t)^{-1}}\right) dx \left[exp\left(\frac{\mu^2 t}{1-2t}\right)\right] \\
&= \frac{1}{\sqrt{(1-2t)}} exp\left(\frac{\mu^2 t}{1-2t}\right), \qquad \lambda = \mu^2 \\
&= \frac{1}{\sqrt{(1-2t)}} exp\left(\frac{\lambda t}{1-2t}\right)
\end{aligned}$$

Then the MGF for $Q_i \sim \chi^2_{k_i}(\lambda_i)$

$$\begin{aligned}
M(t) &= E[\sum_{i=1}^{k} x_i^2 t] = \prod_{i=1}^{k} M_i(t) \\
&= \left(\frac{1}{\sqrt{(1-2t)}}\right)^k exp\left(\frac{\sum_{i=1}^{k}\lambda_i t}{1-2t}\right) \\
&= \left(\frac{1}{\sqrt{(1-2t)}}\right)^k exp\left(\frac{\lambda t}{1-2t}\right) \\
&= (1-2t)^{-k/2} exp\left(\frac{\lambda t}{1-2t}\right), \qquad \text{i.i.d}
\end{aligned}$$

(ii) Method 2:

$$\begin{aligned}
M(t) = E[(\mu+Z)^2 t] &= \frac{1}{\sqrt{2\pi}}\int exp\left((\mu+Z)^2 t\right) exp\left(-\frac{Z^2}{2}\right) dz \\
&= \frac{1}{\sqrt{2\pi}}\int exp\left((t-\frac{1}{2})z^2 + 2\mu t z + \mu^2 t\right) dz \\
&= \frac{1}{\sqrt{2\pi}}\int exp\left(-\frac{(1-2t)}{2}\{z^2 - \frac{4\mu t z}{(1-2t)} + \frac{2\mu^2 t^2}{(1-2t)^2}\} + \frac{2\mu^2 t^2}{(1-2t)} + \mu^2 t\right) dz \\
&= \frac{1}{\sqrt{(1-2t)}}\left[exp\left(\frac{\mu^2 t}{1-2t}\right)\right]\int \frac{1}{\sqrt{2\pi/(1-2t)}} exp\left(-\frac{(z-\frac{2\mu t}{1-2t})^2}{2(1-2t)^{-1}}\right) dz \\
&= \frac{1}{\sqrt{(1-2t)}} exp\left(\frac{\mu^2 t}{1-2t}\right), \qquad t < 1/2
\end{aligned}$$

The general case of a linear combination of independent $\chi^2_{k_i}(\lambda_i)$

$$Q = \sum_{i=1}^{k} a_i Q_i$$

We also can prove using MGF.

1.2.3 Linear Combination of Chi-Square Distribution

The linear combination of chi-square distribution Y_j. Let us denote by $X \sim \Gamma(r,\lambda)$ the fact that the r.v. X has a Gamma distribution with shape parameter r and rate parameter λ

$$f_X(x) = \frac{\lambda^x}{\Gamma(r)} exp(-\lambda x) x^{r-1}, \qquad (r, \lambda > 0, x > 0)$$

Then we have, for $j = 1, ..p$,

$$Y_j \sim \Gamma(\frac{k_j}{2}, \frac{1}{2}) \rightarrow Z_j = w_j Y_j \sim \Gamma(\frac{k_j}{2}, \frac{1}{2w_j})$$

The MGF for linear combinations $Z_j = w_j Y_j$

$$\begin{aligned}
M(t) &= E[exp(Y_j t)] = (1-2t)^{-k/2} exp\left(\frac{\lambda t}{1-2t}\right) \\
M_{Z_j}(t) &= E[exp(w_j Y_j t)] = E[exp(Y_j(w_j t))] \\
&= (1-2w_j t)^{-1/2} exp\left(\frac{\lambda w_j t}{1-2w_j t}\right)
\end{aligned}$$

$$\begin{aligned}
M_Y(t) &= E[exp(Yt)] = E[exp(t[w_1 Y_1 + w_2 Y_2 + w_3 Y_3 + ..w_n Y_n])] \\
&= E[exp(w_1 t Y_1)]E[exp(w_2 t Y_2)]...E[exp(w_n t Y_n)] \\
&= M_{X_1}(w_1 t) M_{X_2}(w_2 t) M_{X_3}(w_3 t)..M_{X_n}(w_n t) \\
&= \prod_{i=1}^{n} M_{X_i}(w_i t)
\end{aligned}$$

The third equation comes from the properties of exponents, as wells as from the expectation of the product of functions of independent random variables.

I need to pay attention that, only under independent and identical situation, we can write

$$M_Y(t) = M_X(t)^n$$

Other than that, we can not further simplify that. So back to the non-central chi-square distribution, we have the MGF of Y

$$\begin{aligned}
M_Y(t) &= \prod_{i=1}^{n} M_{X_i}(w_i t) \\
&= \prod_{i=1}^{n} (1-2w_j t)^{-1/2} exp\left(\frac{\lambda w_j t}{1-2w_j t}\right)
\end{aligned}$$

Then we can see that the shape parameter is $\frac{1}{2w_i}$. If we want to have a non-central chi-square distribution for Y, then all w_j need to be the same.

If $Z_1, ..., Z_k$ are independent, standard normal random variables, then the sum of their squares,

$$\begin{aligned}
Q &= Z_i^2 \sim \chi^2(k) \\
p(k) &= \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} exp(-\frac{x}{2})
\end{aligned}$$

(a) Non-Central Chi-square distribution

Definition 1.2.1. Let $(X_1, X_2, \ldots, X_i, \ldots, X_k)$ be k independent, normally distributed random variables with means μ_i and unit variances. Then the random variable

$$Q = \sum_{i=1}^{k} X_i^2 \sim \chi^2(k, \lambda), \qquad \lambda = \sum_{i=1}^{k} \mu_i^2$$

where the degrees of freedom is k.

The sample mean of n i.i.d. chi-squared variables of degree k is distributed according to a gamma distribution with shape α and scale θ parameters:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim Gamma(\alpha = nk/2, \theta = 2/n)$$

Lemma 1.2.1. Let $Q_i \sim \chi^2_{k_i}(\lambda_i)$ for $i = 1, \ldots, n$, be independent. Then, $Q = \sum_{i=1}^{n} Q_i$ is a noncentral $\chi^2_k(\lambda)$, where $k = \sum_{i=1}^{n} k_i$ and $\lambda = \sum_{i=1}^{n} \lambda_i$.

The distribution transformation use moment generating function.

We can get MGF from $E[x^2 t]$

$$\begin{aligned}
M_i(t) = E[x^2 t] &= \frac{1}{\sqrt{2\pi}} \int \exp(x^2 t) \exp\left(-\frac{(x-\mu)^2}{2}\right) dx \\
&= \frac{1}{\sqrt{2\pi}} \int \exp\left((t - \frac{1}{2})x^2 + \mu x - \frac{\mu^2}{2}\right) dx \\
&= \frac{1}{\sqrt{2\pi}} \int \exp\left(-\frac{1}{2}(1-2t)\{x^2 - \frac{2\mu x}{(1-2t)} + \frac{\mu^2}{(1-2t)^2}\} + \frac{\mu^2}{2(1-2t)} - \frac{\mu^2}{2}\right) dx \\
&= \frac{1}{\sqrt{(1-2t)}} \int \frac{(1-2t)}{\sqrt{2\pi}} \exp\left(-\frac{(x - \frac{\mu}{1-2t})^2}{2(1-2t)^{-1}}\right) dx \left[\exp\left(\frac{\mu^2 t}{1-2t}\right)\right] \\
&= \frac{1}{\sqrt{(1-2t)}} \exp\left(\frac{\mu^2 t}{1-2t}\right), \qquad \lambda = \mu^2 \\
&= \frac{1}{\sqrt{(1-2t)}} \exp\left(\frac{\lambda t}{1-2t}\right)
\end{aligned}$$

Then the MGF for $Q_i \sim \chi^2_{k_i}(\lambda_i)$

$$\begin{aligned}
M(t) &= E[\sum_{i=1}^{k} x_i^2 t] = \prod_{i=1}^{k} M_i(t) \\
&= \left(\frac{1}{\sqrt{(1-2t)}}\right)^k \exp\left(\frac{\sum_{i=1}^{k}\lambda_i t}{1-2t}\right) \\
&= \left(\frac{1}{\sqrt{(1-2t)}}\right)^k \exp\left(\frac{\lambda t}{1-2t}\right) \\
&= (1-2t)^{-k/2} \exp\left(\frac{\lambda t}{1-2t}\right), \qquad \text{i.i.d}
\end{aligned}$$

The general case of a linear combination of independent $\chi^2_{k_i}(\lambda_i)$

$$Q = \sum_{i=1}^{k} a_i Q_i$$

We also can prove using MGF.

1.2.4 Chi-square MGF

We can get MGF for chi-square from $E[x^2 t]$ and $E[(\mu + Z)^2 t]$, where $Z \sim N(0,1)$. Let's prove it in two methods:

(i) Method 1:

$$\begin{aligned}
M_i(t) = E[x^2 t] &= \frac{1}{\sqrt{2\pi}} \int exp(x^2 t) exp\left(-\frac{(x-\mu)^2}{2}\right) dx \\
&= \frac{1}{\sqrt{2\pi}} \int exp\left((t - \frac{1}{2})x^2 + \mu x - \frac{\mu^2}{2}\right) dx \\
&= \frac{1}{\sqrt{2\pi}} \int exp\left(-\frac{1}{2}(1-2t)\{x^2 - \frac{2\mu x}{(1-2t)} + \frac{\mu^2}{(1-2t)^2}\} + \frac{\mu^2}{2(1-2t)} - \frac{\mu^2}{2}\right) dx \\
&= \frac{1}{\sqrt{(1-2t)}} \int \frac{(1-2t)}{\sqrt{2\pi}} exp\left(-\frac{(x - \frac{\mu}{1-2t})^2}{2(1-2t)^{-1}}\right) dx \left[exp\left(\frac{\mu^2 t}{1-2t}\right)\right] \\
&= \frac{1}{\sqrt{(1-2t)}} exp\left(\frac{\mu^2 t}{1-2t}\right), \qquad \lambda = \mu^2 \\
&= \frac{1}{\sqrt{(1-2t)}} exp\left(\frac{\lambda t}{1-2t}\right)
\end{aligned}$$

Then the MGF for $Q_i \sim \chi^2_{k_i}(\lambda_i)$

$$\begin{aligned}
M(t) &= E[\sum_{i=1}^{k} x_i^2 t] = \prod_{i=1}^{k} M_i(t) \\
&= \left(\frac{1}{\sqrt{(1-2t)}}\right)^k exp\left(\frac{\sum_{i=1}^{k}\lambda_i t}{1-2t}\right) \\
&= \left(\frac{1}{\sqrt{(1-2t)}}\right)^k exp\left(\frac{\lambda t}{1-2t}\right) \\
&= (1-2t)^{-k/2} exp\left(\frac{\lambda t}{1-2t}\right), \qquad \text{i.i.d}
\end{aligned}$$

(ii) Method 2:

$$\begin{aligned}
M(t) = E[(\mu+Z)^2 t] &= \frac{1}{\sqrt{2\pi}} \int exp\left((\mu+Z)^2 t\right) exp\left(-\frac{Z^2}{2}\right) dz \\
&= \frac{1}{\sqrt{2\pi}} \int exp\left((t-\frac{1}{2})z^2 + 2\mu t z + \mu^2 t\right) dz \\
&= \frac{1}{\sqrt{2\pi}} \int exp\left(-\frac{(1-2t)}{2}\{z^2 - \frac{4\mu t z}{(1-2t)} + \frac{2\mu^2 t^2}{(1-2t)^2}\} + \frac{2\mu^2 t^2}{(1-2t)} + \mu^2 t\right) dz \\
&= \frac{1}{\sqrt{(1-2t)}} \left[exp\left(\frac{\mu^2 t}{1-2t}\right)\right] \int \frac{1}{\sqrt{2\pi/(1-2t)}} exp\left(-\frac{(z-\frac{2\mu t}{1-2t})^2}{2(1-2t)^{-1}}\right) dz \\
&= \frac{1}{\sqrt{(1-2t)}} exp\left(\frac{\mu^2 t}{1-2t}\right), \qquad t < 1/2
\end{aligned}$$

The general case of a linear combination of independent $\chi^2_{k_i}(\lambda_i)$

$$Q = \sum_{i=1}^{k} a_i Q_i$$

We also can prove using MGF.

(b) Linear Combination of Chi-Square Distribution The linear combination of chi-square distribution Y_j. Let us denote by $X \sim \Gamma(r, \lambda)$ the fact that the r.v. X has a Gamma distribution with shape parameter r and rate parameter λ

$$f_X(x) = \frac{\lambda^x}{\Gamma(r)} exp(-\lambda x) x^{r-1}, \qquad (r, \lambda > 0, x > 0)$$

Then we have, for $j = 1, ..p$,

$$Y_j \sim \Gamma(\frac{k_j}{2}, \frac{1}{2}) \rightarrow Z_j = w_j Y_j \sim \Gamma(\frac{k_j}{2}, \frac{1}{2w_j})$$

The MGF for linear combinations $Z_j = w_j Y_j$

$$\begin{aligned}
M(t) &= E[exp(Y_j t)] = (1-2t)^{-k/2} exp\left(\frac{\lambda t}{1-2t}\right) \\
M_{Z_j}(t) &= E[exp(w_j Y_j t)] = E[exp(Y_j(w_j t))] \\
&= (1-2w_j t)^{-1/2} exp\left(\frac{\lambda w_j t}{1-2w_j t}\right)
\end{aligned}$$

$$\begin{aligned}
M_Y(t) &= E[exp(Yt)] = E[exp(t[w_1 Y_1 + w_2 Y_2 + w_3 Y_3 + ..w_n Y_n])] \\
&= E[exp(w_1 t Y_1)]E[exp(w_2 t Y_2)]...E[exp(w_n t Y_n)] \\
&= M_{X_1}(w_1 t) M_{X_2}(w_2 t) M_{X_3}(w_3 t)..M_{X_n}(w_n t) \\
&= \prod_{i=1}^{n} M_{X_i}(w_i t)
\end{aligned}$$

The third equation comes from the properties of exponents, as wells as from the expectation of the product of functions of independent random variables.

I need to pay attention that, only under independent and identical situation, we can write

$$M_Y(t) = M_X(t)^n$$

Other than that, we can not further simplify that. So back to the non-central chi-square distribution, we have the MGF of Y

$$\begin{aligned}
M_Y(t) &= \prod_{i=1}^{n} M_{X_i}(w_i t) \\
&= \prod_{i=1}^{n} (1-2w_j t)^{-1/2} exp\left(\frac{\lambda w_j t}{1-2w_j t}\right)
\end{aligned}$$

Then we can see that the shape parameter is $\frac{1}{2w_i}$. If we want to have a non-central chi-square distribution for Y, then all w_j need to be the same.

(c) Independence

If $y \sim N(0, \sigma^2 I)$, M is a symmetric idempotent matrix of order n, and L is a $k \times n$ matrix, then Ly and $y'My$ are independently distribution if $LM = 0$.

The proof will use the orthogonal matrix Q, which is $Q'Q = I$, we could add this term wherever we want to.

Define the matrix Q so that

$$Q^T MQ = \Lambda = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}$$

Let r denote the dimension of the identity matrix which is equal to the rank of M. Thus $r = trM$.

Let $v = Q'y$ and partition v as follows

$$v = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} v_1 \\ v_2 \\ .. \\ v_r \\ v_{r+1} \\ .. \\ v_n \end{bmatrix}$$

The number of elements of v_1 is r, while v_2 contains $n - r$ elements. Clearly v_1 and v_2 are independent of each other since they are independent standard normals. What we will show now is that $y'My$ depends only on v_1 and Ly depends only on v_2. Given that the v_i are independent, $y'My$ and Ly will be independent.

$$\begin{aligned} y'My &= v'Q'MQv \\ &= v' \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} v \\ &= v_1^T v_1 \end{aligned}$$

Now consider the product of L and Q which we denote C. Partition C as (C_1, C_2). C_1 has k rows and r columns. C_2 has k rows and $n - r$ columns. Now consider the following product

$$\begin{aligned} C(Q'MQ) &= LQQ'MQ, \qquad C = LQ \\ &= LMQ = 0, \qquad LM = 0 \end{aligned}$$

Now consider the product of C and matrix $Q'MQ$

$$
\begin{aligned}
C(Q'MQ) &= (C_1, C_2)\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \\
&= 0
\end{aligned}
$$

This implies that $C_1 = 0$, then implies that $LQ = C = (0, C_2)$

Now consider Ly. It can be written

$$Ly = LQQ'y = Cv = C_2 v_2$$

Now note that Ly depends only on v_2, and $y'My$ depends only on v_1. But since v_1 and v_2 are independent, so are Ly and $y'My$.

1.3 Normal Distribution

1.3.1 MGF for Univariate Normal

$$
\begin{aligned}
M(t) &= exp(\mu t + \sigma^2 t^2/2), \qquad \text{MGF for } N(\mu, \sigma^2) \\
M_{\sqrt{w_i}X_i}(t) &= E[exp(\sqrt{w_i} t X_i)] = exp(\mu\sqrt{w_i}t + \sigma^2[\sqrt{w_i}t]^2/2), \qquad \mu = 0 \\
&= exp(\sigma^2 w_i t^2/2)
\end{aligned}
$$

Then the linear combination y_n

$$
\begin{aligned}
M_{Y_n}(t) &= E[exp\left((\sqrt{w_1}X_1 + \sqrt{w_2}X_2 + .. + \sqrt{w_n}X_n)t\right)] \\
&= E[exp(\sqrt{w_1}X_1 t)]E[exp(\sqrt{w_2}X_2 t)]E[exp(\sqrt{w_3}X_3 t)]..E[exp(\sqrt{w_n}X_n t)] \\
&= exp(\sigma^2 w_1 t^2/2)exp(\sigma^2 w_2 t^2/2)exp(\sigma^2 w_3 t^2/2)..exp(\sigma^2 w_n t^2/2) \\
&= exp(\sigma^2[w_1 + w_2 + ..w_n]t^2/2) = exp(\sigma^2 t^2/2)
\end{aligned}
$$

So $Y_n \sim N(0, \sigma^2)$.

1.3.2 Bivariate Normal Distribution

The Bivariate Normal Distribution is always connected with partitioned covariance matrix. Assume vector (X, Y) is Gaussian.

Theorem 1.3.1. Two random variables X and Y are said to be bivariate normal, or jointly normal, if $aX + bY$ has a normal distribution for all $a, b \in R$.

If $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$ are jointly normal, then $X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2 + 2\rho(X,Y)\sigma_X\sigma_Y)$.

We consider $X + Y$ is a also normal distribution, then the covariance

$$Cov(X+Y) = Cov(X) + Cov(Y) + 2Cov(XY) = \sigma_X^2 + \sigma_Y^2 + 2\rho(X,Y)\sigma_X\sigma_Y$$

How to provide a simple way to generate jointly normal random variables? The basic idea is that we can start from several independent random variables and by considering their linear combinations, we can obtain bivariate normal random variables.

Let Z_1 and Z_2 be two independent N(0,1) random variables. Define

$$X = Z_1, \qquad Y = \rho Z_1 + \sqrt{1-\rho^2} Z_2$$

where ρ is a real number in (-1, 1). Show that X and Y are bivariate normal.

First, note that since Z_1 and Z_2 are normal and independent, they are jointly normal, with the joint PDF

$$\begin{aligned} f_{Z_1,Z_2}(z_1, z_2) &= f_{Z_1}(z_1) f_{Z_2}(z_2) \\ &= \frac{1}{2\pi} exp\left(-\frac{1}{2}[z_1^2 + z_2^2]\right) \end{aligned}$$

We need to show $aX + bY$ is normal for all $a, b \in R$. We have

$$\begin{aligned} aX + bY &= aZ_1 + b(\rho Z_1 + \sqrt{1-\rho^2} Z_2) \\ &= (a + b\rho) Z_1 + b\sqrt{1-\rho^2} Z_2 \end{aligned}$$

which is a linear combination of Z_1 and Z_2, and thus it is normal.

We can use the method of transformation to find the joint PDF of X and Y. The inverse transformation is given by

$$\begin{aligned} Z_1 &= X = h_1(X,Y) \\ Z_2 &= -\frac{\rho}{\sqrt{1-\rho^2}} X + \frac{1}{\sqrt{1-\rho^2}} Y = h_2(X,Y) \end{aligned}$$

We have

$$\begin{aligned} f_{XY}(z_1, z_2) &= f_{Z_1,Z_2}(h_1(X,Y), h_2(X,Y))|J| \\ &= f_{Z_1,Z_2}(x, -\frac{\rho}{\sqrt{1-\rho^2}} x + \frac{1}{\sqrt{1-\rho^2}} y)|J| \end{aligned}$$

where

$$J = det \begin{bmatrix} \dfrac{\partial h_1}{\partial x} & \dfrac{\partial h_1}{\partial y} \\ \dfrac{\partial h_2}{\partial x} & \dfrac{\partial h_2}{\partial y} \end{bmatrix} = det \begin{bmatrix} 1 & 0 \\ -\frac{\rho}{\sqrt{1-\rho^2}} & \frac{1}{\sqrt{1-\rho^2}} \end{bmatrix} = \frac{1}{\sqrt{1-\rho^2}}$$

Thus, we conclude that,

$$f_{XY}(z_1, z_2) = \frac{1}{2\pi\sqrt{1-\rho^2}} exp\left(-\frac{1}{2(1-\rho^2)}[x^2 - 2\rho xy + y^2]\right)$$

To find the ρ

$$\begin{aligned} Var(X) &= Var(Z_1) = 1 \\ Var(Y) &= \rho^2 Var(Z_1) + (1-\rho^2)Var(Z_2) = 1 \\ \rho(X, Y) &= Cov(X, Y) = Cov(Z_1, \rho Z_1 + \sqrt{1-\rho^2} Z_2) \\ &= \rho Cov(Z_1, Z_2) + \sqrt{1-\rho^2} Cov(Z_1, Z_2) \\ &= \rho \end{aligned}$$

Now, if you want two jointly normal random variables X and Y such that $X \sim N(\mu_X, \sigma_X^2), Y \sim N(\mu_Y, \sigma_Y^2)$, and $\rho(X, Y) = \rho$, you can start with two independent N(0,1) random variables, Z_1 and Z_2, and define

$$\begin{aligned} X &= \sigma_X Z_1 + \mu_X \\ Y &= \sigma_Y \left(\rho Z_1 + \sqrt{1-\rho^2} Z_2\right) + \mu_Y \end{aligned}$$

construction using Z1 and Z2 can be used to solve problems regarding bivariate normal distributions. Third, this method gives us a way to generate samples from the bivariate normal distribution using a computer program.

1.3.3 Conditional Distribution

Theorem 1.3.2. Suppose X and Y are jointly normal random variables with parameters $\mu_X, \sigma_X^2, \mu_Y, \sigma_Y^2$, and ρ. Then, given $X = x, Y$ is normally distributed with

$$\begin{aligned} E[Y|X = x] &= \mu_Y + \rho\sigma_Y \frac{x - \mu_X}{\sigma_X} \\ Var(Y|X = x) &= (1-\rho^2)\sigma_Y^2 \end{aligned}$$

One way to solve this problem is by using the joint PDF formula. In particular, since $X \sim N(\mu_X, \sigma_X^2), Y \sim N(\mu_Y, \sigma_Y^2)$, we can use

$$f_{Y|X=x}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)}$$

or we use

$$\begin{aligned}
X &= \sigma_X Z_1 + \mu_X \\
Y &= \sigma_Y \left(\rho Z_1 + \sqrt{1-\rho^2} Z_2\right) + \mu_Y
\end{aligned}$$

Thus, given $X = x$,

$$\begin{aligned}
Z_1 &= \frac{x - \mu_X}{\sigma_X} \\
Y &= \sigma_Y \rho \frac{x - \mu_X}{\sigma_X} + \sigma_Y \sqrt{1-\rho^2} Z_2 + \mu_Y
\end{aligned}$$

Since Z_1 and Z_2 are independent, knowing Z_1 does not provide any information on Z_2. We have shown that given $X = x, Y$ is a linear function of Z_2, thus it is normal. In particular

$$\begin{aligned}
E[Y|X=x] &= \sigma_Y \rho \frac{x - \mu_X}{\sigma_X} + \sigma_Y \sqrt{1-\rho^2} E[Z_2] + \mu_Y \\
&= \mu_Y + \rho \sigma_Y \frac{x - \mu_X}{\sigma_X} \\
Var[Y|X=x] &= \sigma_Y^2 (1-\rho^2) Var(Z_2) = (1-\rho^2)\sigma_Y^2
\end{aligned}$$

1.4 Multivariate Normal Distribution

1.4.1 Moment Generating Function

Suppose $Y \sim MVN(\mu, \Sigma)$, Σ is positive definite matrix.

Then we can decompose $\Sigma = BB^T$ for some nonsingular matrix B since Σ is positive definite.

$$X = B^{-1}(Y - \mu), \qquad Y = \mu + BX$$

So we have $X_1, \ldots X_n$ are independent standard normal, $X = (X_1, ..X_n)^T \sim MVN(0, I_n)$.

$$
\begin{aligned}
M_Y(t) &= E\Big[e^{t^TY}\Big] = E\Big[e^{t^T(\mu+BX)}\Big] = e^{t^T\mu}E\Big[e^{l^TX}\Big], \qquad l^T = t^TB \\
&= e^{t^T\mu}E\Big[e^{\sum_{i=1}^n l_iY_i}\Big], \qquad l = (l_1, ..l_n) \\
&= e^{t^T\mu}\prod_{i=1}^n E\Big[e^{l_iY_i}\Big] \\
&= e^{t^T\mu}\prod_{i=1}^n e^{l_i^2/2} \\
&= \exp\Big(\mu^Tt + \frac{1}{2}l^Tl\Big) \\
&= \exp\Big(\mu^Tt + \frac{1}{2}t^T\Sigma t\Big)
\end{aligned}
$$

1.4.2 Marginal and conditional distributions of a multivariate normal vector

A $K \times 1$ random vector X is multivariate normal if its joint probability density function is

$$
f_X(x) = (2\pi)^{-K/2}|det(V)|^{-1/2}exp(-\frac{1}{2}(x-\mu)^TV^{-1}(x-\mu))
$$

where μ is a $K \times 1$ mean vector, V is a $K \times K$ covariance matrix.
Partition of the vector:
We partition X into two sub-vectors X_a and X_b such that

$$
X = \begin{pmatrix} X_a \\ X_b \end{pmatrix}
$$

The sub-vectors X_a and X_b have dimensions $K_a \times 1$ and $K_b \times 1$ respectively. Moreover, $K_a + K_b = K$.
Partition of the parameters
We partition the mean vector and covariance matrix as follows:

$$
\mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}
$$

and

$$
V = \begin{pmatrix} V_a & V_{ab}^T \\ V_{ab} & V_b \end{pmatrix}
$$

Normality of the sub-vectors
The marginal distributions of the two sub-vectors are also multivariate normal.

(i) *Proof.* The random vector X_a can be written as a linear transformation of X:

$$X_a = AX$$

Where A is a $K_a \times K$ matrix whose entries are either zero or one. Thus, X_a has a multivariate normal distribution because it is a linear transformation of the multivariate normal random vector X and multivariate normality is preserved by linear transformations. Same as $X_b = BX$ where B is a $K_b \times K$ matrix whose entries are either zero or one.

Independence of the sub-vectors

X_a and X_b are independent if and only if $V_{ab} = 0$.

X_a and X_b are independent if and only if their joint moment generating function is equal to the product of their individual moment generating functions. Since X_a is multivariate normal, its joint moment generating function is

$$\begin{aligned}
M_{X_a}(t_a) &= exp(t_a^T \mu_a + \frac{1}{2} t_a^T V_a t_a) \\
M_{X_b}(t_b) &= exp(t_b^T \mu_b + \frac{1}{2} t_b^T V_b t_b)
\end{aligned}$$

The joint moment generating function of X_a and X_b, which is just the joint moment generating function of X, is

$$\begin{aligned}
M_{X_a,X_b}(t_a, t_b) &= M_X(t) \\
&= exp(t^T \mu + \frac{1}{2} t^T V t) \\
&= exp\left([t_a^T t_b^T] \begin{bmatrix} \mu_a \\ \mu_b \end{bmatrix} + [t_a^T t_b^T] \begin{bmatrix} V_a & V_{ab}^T \\ V_{ab} & V_b \end{bmatrix} [t_a t_b] \right) \\
&= exp\left(t_a^T \mu_a + t_b^T \mu_b + \frac{1}{2} t_a^T V_a t_a + \frac{1}{2} t_b^T V_b t_b + \frac{1}{2} t_b^T V_{ab} t_a + \frac{1}{2} t_a^T V_{ab}^T t_b \right) \\
&= exp\left(t_a^T \mu_a + t_b^T \mu_b + \frac{1}{2} t_a^T V_a t_a + \frac{1}{2} t_b^T V_b t_b + t_b^T V_{ab} t_a \right) \\
&= exp\left(t_a^T \mu_a + \frac{1}{2} t_a^T V_a t_a \right) exp\left(t_b^T \mu_b + \frac{1}{2} t_b^T V_b t_b \right) exp(t_b^T V_{ab} t_a)
\end{aligned}$$

from which it is obvious that $M_{X_a,X_b}(t_a, t_b) = M_{X_a}(t_a) M_{X_b}(t_b)$ if and only if $V_{ab} = 0$.

□

(ii) Conditional distributions of MVN are MVN. Suppose $X \sim N_n(\mu, \Sigma)$. Using the partition above, we have

$$X_1|X_2 = x_2 \sim N_r(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \Sigma_{11.2})$$
$$\Sigma_{11.2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = (1 - \rho^2)\sigma_1^2$$

1.4.3 Beta distribution

If $Y_1, ..., Y_{n+1}$ are i.i.d $Exp(\theta)$, then

$$Z_i = \frac{Y_1 + ... + Y_i}{Y_1 + .. + Y_{n+1}} \sim Beta(i, n - i + 1)$$

Particularly, $(Z_1, ...Z_n)$ has the same distribution as the order statistics $(\xi_{n:1}, ..\xi_{n:n})$ of n Uniform(0,1) random variables.

(i) Prove Beta distribution The $Z_{n+1} = 1$, so we don't need to calculate this distribution. We can use the method in above problem to get the distribution of Z_i, and we need to get the joint distribution of $(Z_1, ...Z_n)$.

Let

$$U = Y_1 + ... + Y_i, \qquad V = Y_{i+1} + .. + Y_{n+1}$$

Then $U \sim Gamma(i, \theta), V \sim Gamma(n + 1 - i, \theta)$, Let

$$Z_i = U/(U + V), \qquad W = U + V$$

Consider the transformation $(U, V)^T \rightarrow (Z_i, W)^T$, note that the transform is one-to-one with the Jacobian

$$\left| \frac{\partial(U, V)}{\partial(Z_i, W)} \right| = |W|$$

For joint distribution of $(U, V)^T$,

$$\frac{1}{\Gamma(i)}\theta exp(-\theta u)(\theta u)^{i-1}I(u > 0) \times \frac{1}{\Gamma(n + 1 - i)}\theta exp(-\theta v)(\theta v)^{n-i}I(v > 0)$$

We obtain the joint density of (Z_i, W) as

$$\frac{1}{\Gamma(i)}\theta exp(-\theta z_i w)(\theta z_i w)^{i-1}$$
$$\times \frac{1}{\Gamma(n + 1 - i)}\theta exp(-\theta(1 - z_i)w)(\theta(1 - z_i)w)^{n-i}w \times I(0 < z_i < 1)I(w > 0)$$

Thus, the marginal density of $Z_i = X/(X+Y)$ is equal to

$$(1-z_i)^{n-i} z_i^{i-1} I(0 < z_i < 1) \frac{\Gamma(n+1)}{\Gamma(i)\Gamma(n+1-i)} \int_w \theta exp(-\theta w)(\theta w)^n I(w > 0) dw$$
$$= \frac{\Gamma(n+1)}{\Gamma(i)\Gamma(n+1-i)} (1-z_i)^{n-i} z_i^{i-1} I(0 < z_i < 1)$$

That is $Z_i \sim Beta(i, n+1-i)$

1.5 Gamma distribution

1.5.1 Gamma and Chi-square distribution

We know that Gamma and Chi-square distribution relationship is $\chi^2(x,n) = \Gamma(\frac{n}{2}, \frac{1}{2})$.

In hypothesis testing, we only have the chi-square table, so we will need to transform the gamma distribution to chi-square distribution.

Suppose $X \sim \Gamma(\alpha, \beta)$, define a scaled version $W \equiv \frac{2X}{\beta}$ such that the inverse transform is $X = h(w) = \frac{\beta w}{2}$.

$$\begin{aligned}
f_w(\alpha,\beta) &= f_X(\frac{\beta w}{2}, n)\Big|\frac{dX}{dw}\Big| \\
&= \frac{1}{\beta^\alpha \Gamma(\alpha)} \Big(\frac{\beta w}{2}\Big)^{\alpha-1} \exp\Big(-\frac{\beta w}{2\beta}\Big)\frac{\beta}{2} \\
&= \frac{1}{\Gamma(\alpha) 2^\alpha} w^{\alpha-1} \exp(-w/2) \sim \chi^2(2\alpha)
\end{aligned}$$

$\alpha = \frac{n}{2}$, so the degrees fo freedom is 2α.

Chapter 2

Distribution Transformation

2.1 Conditional Distribution

Conditional distribution is used in sufficient statistics (ie. show T(X) is sufficient which the distribution based on sufficient statistics does not depend on θ), UMVUE $E[\theta|T(X)]$, nuisance parameter $p(\theta_1|\theta_2, ..\theta_n, X)$, Bayesian statistics.

Basically we can write the distribution of the based on statistics, if not, we will write the integral

Chapter 3

Parameter Estimates

3.1 The Standard Exponential Distribution

The standard exponential distribution family

$$p(y|\theta) = \phi\Big[\exp\Big(y\theta - b(\theta)\Big) - c(y)\Big] - \frac{1}{2}s(y,\phi)$$

We will explore the fun characteristics of the exponential family

(i) Mean and Variance by derivatives

$$log \int p(y|\theta) = log \int \phi\Big[\exp\Big(y\theta - b(\theta)\Big) - c(y)\Big] - \frac{1}{2}s(y,\phi)dv = 0$$
$$log \int \exp\{(y\theta)\}h(y)v(dy) = b(\theta)$$
$$\partial_\theta log \int \exp\{(y\theta)\}h(y)v(dy) = \partial_\theta b(\theta)$$

To proceed we need to move the gradient past the integral sign. In general derivatives can not be moved past integral signs (both are certain kinds of limits, and sequences of limits can differ depending on the order in which the limits are taken). However it turns out that the move is justified in this case by an appeal to the dominated convergence theorem.

$$
\begin{aligned}
\partial_\theta b(\theta) &= \partial_\theta log \int \exp\{(y\theta)\} h(y) v(dy) \\
&= \frac{\int y \exp\{(y\theta)\} h(y) v(dy)}{\int \exp\{(y\theta)\} h(y) v(dy)} \\
&= \int y \exp\{y\theta - b(\theta)\} h(x) v(dy) \\
&= E[y]
\end{aligned}
$$

Also we can see that the first derivative of $b(\theta)$ is equal to the mean of the sufficient statistics. Similar for the variance.

Another proof is to use the Bartlett's identities

Suppose that differentiation and integration are exchangeable and all the necessary expectations are finite. We have the following results:

$$
\begin{aligned}
E\_\xi\Big(\partial_j l_n\Big) &= 0, \\
E_\xi\Big(\partial^2_{j,k} l_n\Big) + E_\xi\Big(\partial_j l_n \partial_k l_n\Big) &= 0
\end{aligned}
$$

By the above two equations, we can get the expectation and variance.

3.2 The Bernoulli Distribution

The standard exponential distribution family

$$
p(y|\theta) = \phi\Big[\exp\Big(y\theta - b(\theta)\Big) - c(y)\Big] - \frac{1}{2}s(y, \phi)
$$

For Bernoulli distribution,

$$
\begin{aligned}
p(x|\pi) &= \pi^x(1-\pi)^{1-x} \\
&= \exp\{\log\Big(\frac{\pi}{1-\pi}\Big)x + \log(1-\pi)\}
\end{aligned}
$$

We see that Bernoulli distribution is an exponential family distribution with

$$
\begin{aligned}
\theta &= \log\Big(\frac{\pi}{1-\pi}\Big) \\
b(\theta) &= -\log(1-\pi) = \log\Big(1 + \exp(\theta)\Big)x \\
\phi &= 1
\end{aligned}
$$

3.2.1 Mean and Variance

For a univariate random variable Y, in this case, all the Y_i have the same π

$$\begin{aligned}
\frac{\partial b(\theta)}{\partial \theta} &= \frac{\exp(\theta)}{1+\exp(\theta)} = \frac{1}{1+\exp(-\theta)} = \mu = E(Y) \\
\frac{\partial^2 b(\theta)}{\partial \theta\, \partial \theta} &= \frac{\exp(\theta)}{\Big[1+\exp(\theta)\Big]^2} = \mu(1-\mu) = Var(Y)
\end{aligned}$$

In regression model, $logit(\pi) = X\beta$, which β is a vector, then we will use the chain rule. And each individual y_i has its own equation that π_i is different.

$$\begin{aligned}
\theta &= X\beta, \qquad \theta_i = x_i^T \beta \\
\partial_\beta b(\theta_i) &= \partial_{\theta_i} b(\theta_i) \partial_\beta \theta_i \\
&= \frac{\exp(\theta_i)}{1+\exp(\theta_i)} x_i = \frac{1}{1+\exp(-\theta)} x_i = \mu_i x_i \\
\partial_\beta^2 b(\theta_i) &= \frac{\exp(\theta_i)}{\Big[1+\exp(\theta_i)\Big]^2} x_i^{\otimes 2} = \mu_i(1-\mu_i) x_i^{\otimes 2}
\end{aligned}$$

And we will need to connect this with the Fisher Information or Newton-Raphson algorithm

$$\begin{aligned}
\theta_i &= k\Big(x_i^T\beta\Big) = x_i^T\beta \\
\xi &= (\beta, \phi) \\
ln(\xi) &= \sum_{i=1}^n \phi \Big[y_i k\Big(x_i^T\beta\Big) - b\Big(k\Big(x_i^T\beta\Big)\Big) - c(y_i)\Big] - \frac{1}{2} s(y_i, \phi) \\
\dot{ln}(\beta) &= \frac{\partial ln(\beta)}{\partial \beta} = \phi \sum_{i=1}^n \Big[y_i - \dot{b}\Big(k\Big(x_i^T\beta\Big)\Big)\Big] \dot{k}\Big(x_i^T\beta\Big) x_i \\
&= \sum_{i=1}^n \Big[y_i - \mu_i\Big] x_i \\
\ddot{ln}(\beta) &= \frac{\partial^2 ln(\beta)}{\partial \beta\, \partial \beta} = -\phi \sum_{i=1}^n \ddot{b}\Big(k(x_i^T\beta)\Big) \dot{k}(x_i^T\beta)^2 x_i x_i^T + \phi \sum_{i=1}^n \Big[y_i - \dot{b}(k(x_i^T\beta))\Big] \ddot{k}(x_i^T\beta) x_i x_i^T \\
&= -\sum_{i=1}^n \ddot{b}\Big(\theta_i\Big) x_i x_i^T = -\sum_{i=1}^n V(\theta_i) x_i x_i^T, \qquad \partial_\beta^2 b(\theta_i) = V(\theta_i)
\end{aligned}$$

let

$$
\begin{aligned}
V(\theta) &= diag\{V(\theta_i)\}, \qquad e_i = y_i - \mu_i \\
\sum_{i=1}^{n} V(\theta_i) x_i x_i^T &= XV(\theta)V^T \\
\mu_i &= \dot{b}(\theta_i), \qquad v_i = \ddot{b}(\theta_i) \\
\dot{\theta}_i &= \partial_\beta \theta_i = \dot{k}(x_i^T\beta)x_i, \qquad \ddot{\theta}_i = \partial_\beta^2 \theta_i = \ddot{k}(x_i^T\beta)x_i x_i^T \\
\dot{b}(\theta_i) &= \partial_\theta b(\theta)\Big|_{\theta=\theta_i}, \dot{k}(\eta) = \partial_\eta k(\eta), \ddot{k}(\eta) = \partial_\eta^2(\eta)
\end{aligned}
$$

So

$$
E\Big[-\ddot{l}n(\beta)\Big] = \phi \sum_{i=1}^{n} v_i \dot{\theta}_i^{\otimes 2}
$$

Another set is to use $E(y_i), Var(y_i)$ which is also used commonly as that are the information we generally get. It is used a lot in GEE.

$$
\begin{aligned}
\partial_\mu \theta &= \partial_\theta \mu^{-1}, \qquad \partial_\mu \mu = \partial_\theta \mu \partial_\mu \theta = 1 \\
\partial_\theta \mu &= \partial_\theta b(\theta) = \ddot{b}(\theta) \\
\partial_\mu \theta &= \Big(\partial_\theta \mu\Big)^{-1} = \ddot{b}(\theta)^{-1}
\end{aligned}
$$

Then we have the connection between the two system

$$
\begin{aligned}
\partial_\beta \theta &= \partial_\beta \mu_i \partial_{\mu_i} \theta_i = \partial_\beta \mu_i \Big[\ddot{b}(\theta_i)\Big]^{-1} \\
\partial_\beta^2 \theta_i &= \Big(\partial_{\mu_i}^2 \theta_i\Big)\Big(\partial_\beta \mu_i\Big)^{\otimes 2} + \partial_{\mu_i}\theta_i \Big(\partial_\beta^2 \mu_i\Big) \\
&= -\dddot{b}(\theta_i)\ddot{b}(\theta_i)^{-3}\Big(\partial_\beta \mu_i\Big)^{\otimes 2} + \Big[\ddot{b}(\theta_i)\Big]^{-1}\Big(\partial_\beta^2 \mu_i\Big)
\end{aligned}
$$

The generalized estimation model

$$
\begin{aligned}
V(\beta) &= \text{diag}\Big(v_1(\beta), \dots, v_n(\beta)\Big) \\
e(\beta) &= (y_1 - \mu_1(\beta), \dots, y_n - \mu_n(\beta))' \\
D_\theta(\beta)' &= \Big(\partial_\beta \beta_1(\beta), \dots, \partial_\beta \beta_n(\beta)\Big)_{p\times n} \\
D(\beta)^T &= \Big(\partial_\beta \mu_1(\beta), \dots, \partial_\beta \mu_n(\beta)\Big)_{p\times n} \\
\dot{l}_n(\beta) &= \phi D_\theta(\beta)^T e(\beta) = \phi D(\beta)' V(\beta)^{-1} e(\beta) \\
E\Big[-\ddot{l}_n(\beta)\Big] &= \phi D_\theta(\beta)' V D_\theta(\beta) = \phi D(\beta)' V(\beta)^{-1} D(\beta)
\end{aligned}
$$

Chapter 4

Convergence Theorem

4.1 Measurement Theorem and Integral

4.1.1 Continuous Convergence

Definition 4.1.1. f_n converges continuously to f, written $f_n \xrightarrow{c} f$ if for any convergent sequence $x_n \to x$ we have $f_n(x_n) \to f(x)$.

We can show by triangle inequality

$$|f_n(x_n) - f(x)| \leq |f_n(x_n) - f(x_n)| + |f(x_n) - f(x)| \leq \|f_n - f\|_K + |f(x_n) - f(x)|$$

the first term on the right-hand side converges to zero by uniform convergence on compact sets and the second term on the right-hand side converges to zeros by continuity of f.

4.1.2 Convergence Mode

It is very important to understand the definition and the notation for each definition.

(i) Converge almost everywhere

A sequence X_n converges almost everywhere (a.e) to X, denoted $X_n \xrightarrow{a.e.} X$, if $X_n(w) \to X(w)$ for all $w \in \Omega - N$ where $\mu(N) = 0$. If μ is a probability, we write a.e. as a.s. (almost surely).

$$\lim_{n\to\infty} X_n = X$$

$$P(\sup_{m\geq n}\Big|X_m - X\Big| > \epsilon) \to 0$$

Remarks: Pay attention to the notation, it says that among all the observations that after X_n, the biggest difference is less than a certain value. When the $\sup\limits_{m\geq n}$ come up, it has listed almost all the observations, which is the same as almost sure.

(ii) Converges in probability A sequence X_n converges in measure to a measurable function X, denoted $X_n \xrightarrow{\mu} X$, if $\mu(|X_n - X| \geq \epsilon) \to 0$ for all $\epsilon > 0$. If μ is a probability measure, we say X_n converges in probability to X.

$$lim_{n\to\infty} P(\|X_n - X\| > \epsilon) = 0$$

(iii) Converges in L_r-distance (rth moment)

Notation: $c = (c_1, ..., c_k) \in R^k, \|c\|_r = \left(\sum_{j=1}^k |c_j|^r\right)^{1/r}, r > 0$. If $r \geq 1$, then $\|c\|_r$ is the L_r- distance between 0 and c. When $r = 2, \|c\| = \|c\|_2 = \sqrt{c^t c}$.

$$X_n \xrightarrow{L_r} X$$

$$lim_{n\to\infty} E\|X_n - X\|_r^r = 0$$

(iv) Converges in distribution Let $F, F_n, n = 1, 2, ..,$ be c.d.f.'s on R^k and $P, P_n, n = 1, 2, ..$ be their corresponding probability measures. We say that $\{F_n\}$ converges to F weakly and write $F_n \xrightarrow{w} F$ iff, for each continuity point x of F,

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

We say that $\{X_n\}$ converges to X in distribution and write $X_n \xrightarrow{d} X$ iff $F_{X_n} \xrightarrow{w} F_X$.

Note: converges in distribution is the cumulative distribution is the same.

$\xrightarrow{a.s.}, \xrightarrow{p}, \xrightarrow{L_r}$: measures how close is between X_n and X as $n \to \infty$.

$F_{X_n} \xrightarrow{w} F_X$: F_{X_n} is close to F_X. but X_n and X may not be close, they may be on different spaces.

Example: Let $\theta_n = 1 + n^{-1}$ and X_n be an random variable having the exponential distribution $E(0, \theta_n), n = 1, 2..$ Let X be a random variable having the exponential distribution $E(0, 1)$.

For any $x > 0$, as $n \to \infty$,

$$F_{X_n}(x) = 1 - e^{-x/\theta_n} \to 1 - e^{-x} = F_X(x)$$

Since $F_{X_n}(x) = 0 = F_X(x)$ for $x \le 0$, we have shown that $X_n \xrightarrow{d} X$.

How about $X_n \xrightarrow{p} X$?

We will need the distribution of $X_n - X$ as we need to get the probability $P(|X_n - X| > \epsilon)$.

The distribution has two cases depends on whether X_n and X are independent or not.

(i) Suppose that X_n and X are not independent, and $X_n \equiv \theta_n X$ (then X_n has the given c.d.f.).
$X_n - X = (\theta_n - 1)X = n^{-1}X$, which has the c.d.f. $(1 - e^{-nx})I_{[0,\infty)}(x)$.
Then $X_n \xrightarrow{p} X$ because, for any $\epsilon > 0$,

$$P(|X_n - X| \ge \epsilon) = e^{-n\epsilon} \to 0$$

Also, $X_n \xrightarrow{L_p} X$ for any $p > 0$, because

$$E(|X_n - X|^p) = n^{-p} E X^p \to 0$$

(ii) Suppose that X_n and X are independent random variables. Since p.d.f.'s for X_n and $-X$ are $\theta_n^{-1}e^{-x/\theta_n}I_{(0,\infty)(x)}$ and $e^x I_{(-\infty,0)}(x)$, respectively, we have let $y = X_n - X, x = X_n$, then $-X = y - X_n < 0$. In the below range, $y \in (-\infty, x)$

$$P(|X_n - X| \leq \epsilon) = \int_{-\epsilon}^{\epsilon}\int_0^{\infty} \theta_n^{-1}e^{-x/\theta_n}e^{y-x}I_{(0,\infty)}(x)I_{(-\infty,x)}(y)dxdy$$

which converges to (by the dominated convergence theorem)

$$\begin{aligned}\int_{-\epsilon}^{\epsilon}\int_0^{\infty} e^{-x}e^{y-x}I_{(0,\infty)}(x)I_{(-\infty,-x)}(y)dxdy &= 1 - e^{-\epsilon} \\ &= \int_0^{\epsilon} e^{-2x}\int_{-\epsilon}^{x} e^y dydx \\ &= \int_0^{\epsilon} e^{-x}dx \\ &= 1 - e^{-\epsilon}\end{aligned}$$

Thus, $P(|X_n - X| \leq \epsilon) \to e^{-\epsilon} > 0$ for any $\epsilon > 0$ and, therefore, $X_n \xrightarrow{p} X$ does not hold.

4.1.3 Relationship between convergence modes

(i) If $X_n \xrightarrow{a.s.} X$, then $X_n \xrightarrow{p} X$.

Proof:

$$P(\Big|X_i - X\Big| > \epsilon) \leq P(\sup_{m\geq n}\Big|X_m - X\Big| > \epsilon) \to 0$$

(ii) If $X_n \xrightarrow{L_r} X$ for an $r > 0$, then $X_n \xrightarrow{p} X$. Consider the definition of moment convergence and probability convergence, the link that connect Expectation and Probability with inequality is Markov Inequality.

For any positive and increasing function $g(\dot{})$ and random variable Y,

$$P(\Big|Y\Big| > \epsilon) \leq E\Big[\frac{g(|Y|)}{g(\epsilon)}\Big]$$

In particular, we choose $Y = \Big|X_n - X\Big|$ and $g(y) = |y|^r$. It gives that

$$P(\Big|X_n - X\Big| > \epsilon) \leq E\Big[\frac{\Big|X_n - X\Big|^r}{\epsilon^r}\Big] \to 0$$

(iii) If $X_n \xrightarrow{p} X$, then $X_n \xrightarrow{d} X$.

Prove: need to use the definition of convergence in probability, and construct the cumulative probability $F_X(x)$.

The purpose is to induce $F_{X_n}(x)$, so that we can compare $F_{X_n}(x)$ and $F(x)$. So the $F(x)$ will be rewritten as $F_{X_n}(x)$ and a probability involves $X_n - X$.

Assume $k = 1$, let x be a continuity point of F_X and $\epsilon > 0$ be given. Then

$$\begin{aligned} F_X(x-\epsilon) &= P(X \le x-\epsilon, X_n \le x) + P(X \le x-\epsilon, X_n > x) \\ &\le P(X_n \le x) + P(X \le x-\epsilon, X_n > x), \qquad P(X_n \le x) > P(X \le x-\epsilon, X_n \le x) \\ &\le F_{X_n}(x) + P(|X_n - X| > \epsilon), \qquad X_n - X > x - (x-\epsilon) = \epsilon \end{aligned}$$

Letting $n \to \infty$, we obtain that

$$F_X(x-\epsilon) \le liminf_n F_{X_n}(x)$$

Switching X_n and X in the previous argument,

$$\begin{aligned} F_X(x+\epsilon) &= P(X \le x+\epsilon, X_n \le x) + P(X \le x+\epsilon, X_n > x) \\ &\ge P(X_n \le x) + P(X \le x+\epsilon, X_n > x) \\ &\ge F_{X_n}(x) + P(|X_n - X| > \epsilon) \end{aligned}$$

Letting $n \to \infty$, we obtain that

$$\begin{aligned} F_X(x-\epsilon) &\le liminf_n F_{X_n}(x) \\ F_X(x+\epsilon) &\ge limsup_n F_{X_n}(x) \end{aligned}$$

Since ϵ is arbitrary and F_X is continuous at x,

$$F_X(x) = lim_{n\to\infty} F_{X_n}(x)$$

(iv) Skorohod's theorem: a conditional converse of (i)-(iii). If $X_n \xrightarrow{d} X$, then there are random vectors $Y_n, Y_n \xrightarrow{a.s.} Y$.

(v) If, for every $\epsilon > 0, \sum_{n=1}^{\infty} P(\|X_n - X\| \ge \epsilon) < \infty$, then $X_n \xrightarrow{a.s.} X$.

(vi) If $X_n \xrightarrow{p} X$, then there is a subsequence $\{X_{n_j}, j = 1, 2..\}$ such that $X_{n_j} \xrightarrow{a.s.} X$ as $j \to \infty$.

We need to show that such a sequence exists, and prove by the almost surely definition. Such a sequence generally use the 2^{-k}. Because 2^{-k} is almost surely convergence, so any sequence that is smaller than this sequence, will definitely be almost surely convergence as well.

For any $\epsilon > 0, P(\Big|X_n - X\Big| > \epsilon) \to 0$, we choose $\epsilon = 2^{-m}$ then there exists a X_{n_m} such that

$$P(\Big|X_{n_m} - X\Big| > 2^{-m}) < 2^{-m}$$

Particularly, we can choose n_m to be increasing. For the sequence $\{X_{n_m}\}$, we note that for any $\epsilon > 0$, when n_m is large,

$$P(\sup_{k \geq m}\Big|X_{n_k} - X\Big| > \epsilon) \leq \sum_{k \geq m} P(\Big|X_{n_k} - X\Big| > 2^{-k}) \leq \sum_{k \geq m} 2^{-k} \to 0$$

Thus, $X_{n_m} \xrightarrow{a.s.} X$.

Remarks: Need to pay attention to the SUP and sum of probability, it is similar to the max of the sequence. So we need to think about the all sequence observations probability.

(vii) If $X_n \xrightarrow{d} X$, and $P(X \equiv c) \equiv 1$, where $c \in R^k$ is a constant vector, then $X_n \xrightarrow{p} c$. Let $X \equiv c$.

Prove by Polya's theorem:

$$P(\Big|X - n - c\Big| > \epsilon) \leq 1 - F_n(c + \epsilon) + F_n(c - \epsilon) \to 1 - F_X(c + \epsilon) + F_X(c - \epsilon) = 0$$

Remarks: Polya's theorem is very useful when dealing with the F_n change to F.

(viii) Moment convergence: Suppose that $X_n \xrightarrow{d} X$, then for any $r > 0$,

$$\lim_{n \to \infty} E\|X_n\|_r^r = E\|X\|_r^r < \infty$$

iff $\{\|X_n\|_r^r\}$ is uniformly integrable (UI) in the sense that

$$\lim_{t \to \infty} sup E(\|X_n\|_r^r I_{\|X_n\|_r > t}) = 0$$

In particular, $X_n \xrightarrow{L_r} X$ if and only if $\{\|X_n - X\|_r^r\}$ is UI

(viii) If $X_n \xrightarrow{p} X$ and $|X_n|^r$ is uniformly integrable, then $X_n \xrightarrow{r} X$.

4.1.4 Polya's theorem

If $F_n \xrightarrow{w} F$ and F is continuous on R^k, then

$$\underset{n\to\infty}{lim}\, sup_{X\in R^k}|F_n(x) - F(x)| = 0.$$

This proposition implies the following useful result: If $c_n \in R^k$ with $C_n \to C$, then

$$F_n(C_n) \to F(C)$$

4.1.5 Fatou's lemma

Given a measure space $(\Omega, \mathbf{F}, \mu)$, and a set $X \in F$, let $\{f_n\}$ be a sequence of $(F, B_{R\geq 0})$ - measurable non-negative functions: $f_n : X \to [0, +\infty]$. Define the function $f : X \to [0, +\infty]$ by setting $f(x) = \underset{n\to\infty}{liminf} f_n(x)$, for every $x \in X$. Then f is $(F, B_{R\geq 0})$ - measurable, and also

$$\int_X f d\mu \leq \underset{n\to\infty}{liminf} \int_X f_n d\mu,$$

where the integral may be infinite.

Remarks: this lemma is used a lot in expectation of sequence.

4.1.6 Big O and Little o

In calculus, two sequences of real numbers, $\{a_n\}$ and $\{b_n\}$, satisfy

(i) $a_n = O(b_n)$ iff $\|a_n\| \leq M|b_n|$ for all n and a constant $M < \infty$. Note that the equal sign does not mean equality.

(ii) $a_n = o(b_n)$ iff $a_n/b_n \to 0$ as $n \to \infty$.

Definition 4.1.2. Let $X_1, X_2, ..$ be random vectors and $Y_1, Y_2, ..$ be random variables defined on a common probability space.

(i) $X_n = O(Y_n)$ a.s. iff $P(\|X_n\| = O(|Y_n|)) = 1$

Since $a_n = O(1)$ means that $\{a_n\}$ is bounded, $\{X_n\}$ is said to be bounded in probability if $X_n = O_p(1)$. ie, O(1) - as $x \to 0$ if it is bounded on a neighborhood of zero. And we say it is o(1) as $x \to 0$ if $f(x) \to 0, x \to 0$.

$X_n = O(Y_n)$ and $Y_n = O(Z_n)$ implies $X_n = O(Z_n)$.

$X_n = O(Y_n)$ does not imply $Y_n = O_p(X_n)$.

If $X_n = O(Z_n)$, then $X_n Y_n = O_p(Y_n Z_n)$.

If $X_n = O(Z_n)$ and $Y_n = O(Z_n)$, then $X_n + Y_n = O_p(Z_n)$.

If $X_n \xrightarrow{d} X$ for a random variable X, then $X_n = O_p(1)$.

If $E(|X_n|) = O(a_n)$, then $X_n = O_p(a_n)$, where $a_n \in (0, \infty)$.

If $X_n \xrightarrow{a.s.} X$, then $sup_n|X_n| = O_p(1)$.

(ii) $X_n = o(Y_n)$ a.s. iff $X_n/Y_n \xrightarrow{a.s.} 0$

$X_n = o(Y_n)$ implies $X_n = O_p(Y_n)$.

(iii) $X_n = O_p(Y_n)$ iff, for any $\epsilon > 0$, there is a constant $C_\epsilon > 0$ such that

$$sup_n P(\|X_n\| \geq C_\epsilon(|Y_n|)) < \epsilon$$

(iv) $X_n = o_p(Y_n)$ iff $X_n/Y_n \xrightarrow{p} 0$.

4.1.7 Big O_p and Little o_p

A sequence X_n of random vectors is said to be $O_p(1)$ if it is bounded in probability (tight) and $o_p(1)$ if it converges in probability to zero. Suppose X_n and Y_n are random sequences taking values in any normed vector space, then

$$X_n = O_p(Y_n)$$
$$Pr(\|X_n\| \leq M\|Y\|) \geq 1 - \epsilon$$

Means $X_n/\|Y_n\|$ is bounded in probability

and

$$X_n = o_p(Y_n)$$
$$\frac{X_n}{\|Y_n\|} \xrightarrow{P} 0, \qquad n \to 0$$
$$Pr(\|X_n\| \geq \epsilon\|Y_n\|) \to 0$$

Means $X_n/\|Y_n\|$ converges in probability to zero.

These notations are often used when the sequence Y_n is deterministic, for example, $X_n = O_p(n^{-1/2})$.

they are also often used when the sequence Y_n is random, for example, we say two estimators $\hat{\theta}_n$ and $\tilde{\theta}_n$ of a parameter θ are asymptotically equivalent if

$$\hat{\theta}_n - \tilde{\theta}_n = o_p(\hat{\theta}_n - \theta)$$
$$\hat{\theta}_n - \tilde{\theta}_n = o_p(\tilde{\theta}_n - \theta)$$

We also use O(1), o(1) and O_p, o_p for terms in equations. For example, a function f is differentiable at x if

$$f(x+h) = f(x) + f^{'}(x)h + o(h)$$

one case of Slutsky's theorem says

$$X_n \xrightarrow{w} X \qquad \to X_n + o_p(1) \xrightarrow{w} X$$

4.2 Sample Variance Distribution

Suppose that $X_1, ..X_n$ are i.i.d. with $E(X_i) = \mu, Var(X_i) = \sigma^2$, and $Var\Big[(X_i - \mu)^2\Big] = \tau < \infty$. Define $S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$. Find the asymptotic distribution of S_n^2.

We have several methods to get the asymptotic distribution:

1. Direct method: Get the mean and variance, then use CLT;

2. Delta method: when there is a function of the known distribution, ie. We know X distribution, then we would like to know $f(X)$ distribution. The key is to get the mean and variance of the new function, which we times the derivative

In this problem, we could use the given information to use CLT method:

WRONG METHOD:

$$
\begin{aligned}
S_n^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - E(X_i) + E(X_i) - \bar{X}_n)^2, && \text{construct the variance component} \\
&= \frac{1}{n}\sum_{i=1}^{n}\Big(X_i - E(X_i)\Big)^2 + \frac{1}{n}\sum_{i=1}^{n}\Big(E(X_i) - \bar{X}_n\Big)^2 \\
&= \sigma^2 + \frac{1}{n}\sum_{i=1}^{n}\Big(E(X_i) - \bar{X}_n\Big)^2 && \text{THIS IS WRONG}
\end{aligned}
$$

CAN'T SAY $\frac{1}{n}\sum_{i=1}^{n}\Big(X_i - E(X_i)\Big)^2 = \sigma^2$, as this is not expectation, it is sample mean/variance.

CORRECT METHOD:

We need to use the information $Var\Big[(X_i - \mu)^2\Big] = \tau < \infty$, to construct a CLT with variance component of $(X_i - \mu)$. To simplify, let $Y_i = X_i - \mu, \bar{Y}_n = \bar{X}_n - \mu$,

$$
\begin{aligned}
S_n^2 &= \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \bar{Y}_n^2 \\
Var\Big(Y_i^2\Big) &= \tau^2, \qquad E\Big(Y_i^2\Big) = E\Big((X_i - \mu)^2\Big) = \sigma^2 \\
\bar{Y}_n^2 &\xrightarrow{d} 0
\end{aligned}
$$

It is very tricky to notice this, we only see the mean has asymptotic distribution based on CLT, while other series does not have asymptotic distribution. $\bar{Y}_n^2$ is only a number. By slutsky theorem,

$$
\begin{aligned}
\sqrt{n}\Big(S_n^2 - \sigma^2\Big) &= \sqrt{n}\Big(\frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \sigma^2\Big) + \sqrt{n}\bar{Y}_n^2 \\
\sqrt{n}\Big(S_n^2 - \sigma^2\Big) &\xrightarrow{d} N(0, \tau^2)
\end{aligned}
$$

Bibliography

[1] Armstrong, M.A. *Basic Topology.* England: Editorial Board, 2000.

[2] Coxeter, H.S.M. *Introduction to Geometry.* Toronto: John Wiley and Sons Inc, 1969.

[3] Chronister, James. *Blender Basics.* Estados Unidos: 2004.

[4] Engelking, Ryszard. *General Topology.* Berlin: Hildermann, 1989.

[5] Engelking, Ryszard y Karol Sieklucki. *Topology: A Geometric Approach.* Berlin: Hildermann, 1992.

[6] Hatcher, Allen. *Algebraic Topology.* Cambridge: Cambridge University Press, 2002.

[7] Hoffman, Kenneth y Ray Kunze. *Linear Algebra, second edition.* New Jersey: Prentice Hall. Inc, 1961.

[8] Kackzynski, T., K. Mischainkow y M. Mrozen. *Algebraic Topology: A Computational Approach.* 2000.

[9] Kosniowski, Czes. *Topología Algebraica.* Barcelona: Editorial Revert, 1986.

[10] Lefschetz, Solomon. *Algebraic Topology.* Rhode Island: American Mathematical Society, 1991.

[11] May, J. P. *A Concise Course in Algebraic Topology.* Chicago.

[12] Mill, J. Van. *Infinite-Dimensional Topology.* Amsterdam: Elsevier Science Publications B. V, 1989.

[13] Munkres, James R. *Topology.* New York: Prentice Hall Inc, 2000.

[14] Roseman, Dennis. *Elementary Topology.* New Jersey: Prentice Hall Inc, 1999.

[15] Sato, Hajime. *Algebraic Topology: An Intuituve Approach.* Rhode Island: American Mathematical Society, 1996.

[16] Zomorodian, Afra J. *Topology for Computing.* United States: Cambridge University Press, 2005.