Sufficient Statistics

Mingwei Fei

January 1, 2023

We learn sufficient statistics from a specific example, then extend to generalized case. So the first sufficient statistics that easily to look at is the sufficient statistics for mean. For every distribution which we can write in exponential family, the sufficient statistics for mean is the statistics with the parameter.

Similarly, we can extend this finding to other statistics(ie. variance) by factorization.

The denotation of a distribution, $f(x|\theta)$ write in this form as the sample data X is known, while the θ is unknown. So the distribution is only available when the θ is known.

1 Factorization Theorem

Let $f(x|\theta)$ denote the joint pdf or pmf of a sample X. A statistic $T(X)$ is a sufficient statistic for θ if and only if there exist functions

$$f(x|\theta) = g(T(x)|\theta)h(x)$$

such that, for all sample points X and all parameter points θ,

1.1 Relate the theorem

To understand this theorem, we can see that in the exponential family distribution

$$f(x|\theta) = \phi exp(\theta - b(\theta) - c(y))$$

So we can write the distribution in factorization form, one part involves θ and the other part does not.

1.2 Proof

The proof needs to use the definition of the sufficient statistics that the distribution based on sufficient statistics does not depend on θ. So we will construct a conditional distribution, which the nominator is the joint distribution of the sample X, and denominator is the distribution of the sufficient statistics.

When we prove the theorem using the specific distribution (ie. normal distribution), we can easily write out the joint distribution and distribution of $T(x)$. But now we are using the general form of distribution.

So the proof needs to find the distribution of $T(x)$. We need to see if there is any relationship between $g(T(x)|\theta)$ and $P_\theta(T(x)=t)$. Need to pay attention that, the distribution of $T(X)$ is actually the sum of the distributions with any $y = T(X)$. We use the concept in the nuisance parameter, which need to find the distribution of statistics.

First, suppose $T(x)$ is sufficient statistics.

As it is the general case, so we just choose $g(t|\theta) = P_\theta(T(x) = t)$.

Then $h(x) = P(X = x|T(X))$, Because $T(X)$ is sufficient, the conditional probability defining $h(x)$ dose not depend on θ. Thus, the choice of $h(x)$ and $g(t|\theta)$ is legitimate.

For this choice,

$$
\begin{aligned}
f(x|\theta) &= P_\theta(X = x) \\
&= P_\theta(X = x, T(X) = T(x)) \\
&= P_\theta(X = x|T(X) = T(x))P_\theta(T(X) = T(x)) \\
&= g(T(x)|\theta)h(x)
\end{aligned}
$$

Now assume factorization exists. Let $q(t|\theta)$ be the pmf of $T(X)$. To show that $T(X)$ is sufficient we exam the ratio $f(x|\theta)/q(T(X)|\theta)$. Define $A_{T(x)} = y : T(y) = T(x)$. Then

$$
\begin{aligned}
\frac{f(x|\theta)}{q(T(X)|\theta)} &= \frac{g(T(x)|\theta)h(x)}{q(T(X)|\theta)} \\
&= \frac{g(T(x)|\theta)h(x)}{\sum_{A_{T(x)}} g(T(y)|\theta)h(y)} \\
&= \frac{g(T(x)|\theta)h(x)}{g(T(x)|\theta)\sum_{A_{T(x)}} h(y)} \\
&= \frac{h(x)}{\sum_{A_{T(x)}} h(y)}
\end{aligned}
$$

Since the ratio does not depend on θ, by Theorem, $T(X)$ is a sufficient statistic for θ. We factor the joint pdf into two parts, one part not depending on θ, which is $h(x)$ function. The other part depends on θ, depends on the sample x only through some function $T(x)$ and this function is a sufficient statistic for θ.

1.3 Example

Uniform Sufficient Statistic: Let $X_1, X_2, ..., X_n$ be i.i.d. observations from the discrete uniform distribution on $1, ..\theta$. The pmf of X_i is

$$f(x|\theta) = \begin{cases} \frac{1}{\theta} & x = 1, 2, ..\theta \\ 0 & \text{otherwise} \end{cases}$$

Thus the joint pmf of $X_1, ...X_n$ is

$$f(x|\theta) = \begin{cases} \frac{1}{\theta^n} & x_i \in \{1, 2, .., \theta\}\theta \\ 0 & \text{otherwise} \end{cases}$$

Let $N = \{1, 2..\}$ be the set of positive integers and let $N_\theta = \{1, 2, ..\theta\}$. Then the joint pmf of $X_1, ...X_n$ is

$$f(x|\theta) = \prod_{i=1}^{n} \theta^{-1} I_{N_\theta}(x_i) = \theta^{-n} \prod_{i=1}^{n} I_{N_\theta}(x_i)$$

Defining $T(x) = max_i x_i$ then

$$\prod_{i=1}^{n} I_{N_\theta}(x_i) = (\prod_{i=1}^{n} I_N(x_i)) I_{N_\theta}(T(x))$$

The equation here is similar as the $I_{x_1<x_2<..x_{max}}(x_m < \theta)$.

2 Minimum Sufficient Statistics

Let $X \sim Gamma(\alpha, 1)$ and $Y \sim Gamma(\beta, 1)$ where the paramaterization is such that α is the shape parameter. Then

$$\frac{X}{X+Y} \sim Beta(\alpha, \beta)$$

(i) How do we create the transformed variables $U = \frac{X}{X+Y}$, V, such that it is easy to get the distribution of U and V, and easy to integrate out V to get the margin distribution of U? This requires familiarity of Beta and Gamma distribution.

One proposal is $V = Y$, and the other is $V = X + Y$. Compare between the two and see which one is better.

Note: The way to choose new parameter is easier to get the original parameter by simple arithmetic calculation or just itself. If we are using the first set of new parameter, it would get the distribution very complicated, the Beta distribution has the form of $U, 1 - U$, need to keep in mind of the achieving the product of two distribution form.

(ii) We could see that $V = X + Y$ is better as U and V are independent from each other.

$$f(X) = \frac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-x}$$
$$f(Y) = \frac{1}{\Gamma(\beta)} y^{\beta-1} e^{-y}$$

Let

$$U = \frac{X}{X+Y}, \qquad V = X + Y$$

Then

$$X = UV, \qquad Y = V - UV$$

The Jacobian transformation matrix

$$J = \begin{pmatrix} \frac{\partial X}{\partial U} & \frac{\partial X}{\partial V} \\ \frac{\partial Y}{\partial U} & \frac{\partial Y}{\partial V} \end{pmatrix} = \begin{pmatrix} V & U \\ -V & 1-U \end{pmatrix}$$
$$|J| = V$$

X and Y are independent, so the joint distribution of (X, Y)

$$f(X,Y) = \frac{1}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} e^{-x} y^{\beta-1} e^{-y}$$
$$f(U,V) = \frac{1}{\Gamma(\alpha)\Gamma(\beta)} U^{\alpha-1} (1-U)^{\beta-1} V^{\alpha+\beta-1} e^{-V}$$

We don't always need to integrate out the other parameter, if we can write the distribution in the form that we can recognize, then will directly get the distribution.

$V \sim Gamma(\alpha+\beta, 1)$ and $U \sim Beta(\alpha, \beta)$.

2.1 Gamma distribution and 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.

2.1.1 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 $(U, V)^T \to (Z_i, W)^T$, 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

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

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

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

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

2.1.2 Prove Uniform distribution

We need to note that, we always construct a joint distribution as we are using the $Y_1, ..., Y_{n+1}$ in joint distribution. So the transformed variables are also joint distribution.

Then we need to find the relationship between $Y_1, ..., Y_{n+1}$ and $(Z_1, ...Z_n)$. We can set the total sum as $S_{n+1} = Y_1 + .. + Y_{n+1}$.

$$
\begin{aligned}
Y_1 &= Z_1 \times S_{n+1}\\
Y_2 &= Z_2 \times S_{n+1} - Y_1 = (Z_2 - Z_1) \times S_{n+1}\\
Y_3 &= Z_3 \times S_{n+1} - Y_1 - Y_2 = (Z_3 - Z_2) \times S_{n+1}\\
&....\\
Y_n &= Z_n \times S_{n+1} - Y_1 - ... - Y_{n-1} = (Z_n - Z_{n-1}) \times S_{n+1}\\
Y_{n+1} &= S_{n+1} - Y_1 - ... - Y_n = (1 - Z_n) \times S_{n+1}
\end{aligned}
$$

Then we have the Jacobian transform distribution

$$
J = \begin{pmatrix}
\frac{\partial Y_1}{\partial Z_1} & \frac{\partial Y_1}{\partial Z_2}... & \frac{\partial Y_1}{\partial Z_n} & \frac{\partial Y_1}{\partial S_{n+1}}\\
.. & .. & .. & ..\\
\frac{\partial Y_n}{\partial Z_1} & \frac{\partial Y_n}{\partial Z_2}... & \frac{\partial Y_n}{\partial Z_n} & \frac{\partial Y_n}{\partial S_{n+1}}\\
\frac{\partial Y_{n+1}}{\partial Z_1} & \frac{\partial Y_{n+1}}{\partial Z_2} & ... & \frac{\partial Y_{n+1}}{\partial S_{n+1}}
\end{pmatrix}
= \begin{pmatrix}
S_{n+1} & 0 & 0 & .. & .. & Z_1\\
-S_{n+1} & S_{n+1} & 0 & .. & .. & Z_2 - Z_1\\
.. & .. & .. & & & \\
0 & 0 & .. - S_{n+1} & S_{n+1} & 0 & Z_n - Z_{n-1}\\
0 & 0 & .. & 0 & S_{n+1} & 1 - Z_n
\end{pmatrix}
$$

$|J| = S_{n+1}^n$

The joint distribution of $Y_1, ..., Y_{n+1}$

$$
\begin{aligned}
f(Y_1, Y_2, ..Y_{n+1}) &= \prod_{i=1}^{n+1} \frac{1}{\theta} exp(-\frac{y_i}{\theta}) I(0 < y_i < \infty)\\
&= \frac{1}{\theta^{n+1}} exp(-\frac{\sum_{i=1}^{n+1} y_i}{\theta}) I(0 < y_i < \infty)\\
&= \frac{1}{\theta^{n+1}} exp(-\frac{S_{n+1}}{\theta}) I(0 < y_i < \infty)
\end{aligned}
$$

Then the transformed variables

$$
f(Z_1, Z_2, ..Z_n, S_{n+1}) = \frac{1}{\theta^{n+1}} exp(-\frac{S_{n+1}}{\theta}) S_{n+1}^n I(0 < z_1 < z_2 < .. < z_n < 1) I(S_{n+1} > 0)
$$

We obtain the joint distribution of $(Z_1, ...Z_n)$

$$
\begin{aligned}
f(Z_1, Z_2, ..Z_n) &= \int_s f(Z_1, Z_2, ..Z_n, S_{n+1}) dS\\
&= \int_s \frac{1}{\theta^{n+1}} exp(-\frac{S}{\theta}) S^n I(0 < z_1 < z_2 < .. < z_n < 1) I(S > 0) dS\\
&= n! I(0 < z_1 < z_2 < .. < z_n < 1), \qquad \text{Gamma integral}
\end{aligned}
$$

which is the joint density of order statistics of n uniform (0,1) random variables.

(b) Let

$$W_1 = Y_1, W_2 = Y_1 + Y_2, ..., W_n = Y_1 + ... + Y_n, S = Y_1 + ... + Y_{n+1}.$$

Consider the transformation $(Y_1, ..., Y_{n+1})' \mapsto (W_1, ..., W_n, S)'$. Note that the transformation is one-to-one with the Jacobian

$$|det\left(\frac{\partial(Y_1, ..., Y_{n+1})}{\partial(W_1, ..., W_n, S)}\right)| = |det\begin{pmatrix} 1 & 0 & \dots & 0 & 0 \\ -1 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & \dots & -1 & 1 \end{pmatrix}| = 1$$

Figure 1: Alt method

3 Construct Distribution

3.1 Order Statistics

The joint distribution of minimum and maximum:

Let's go for the joint cdf of the minimum and maximum

$$F_{X_1,X_n}(x, y) = P(X_{(1)} \le x, X_{(n)} \le y)$$

Why do we start from cdf? It is much easier to get cdf than pdf, as the pdf need to take derivative, while cdf only needs to get the integral.

And if the observations are independent, the pdf and cdf also follows the same rule.

We will need to write this in terms of the individual X_i as the minimum and maximum are statistics of the individuals. Consider instead the relationship

$$P(X_n) \le y = P(X_{(1)} \le x, X_{(n)} \le y) + P(X_{(1)} > x, X_{(n)} \le y)$$

This is the integral of x, which is a common in getting the marginal distribution from joint distribution.

We know how to write out the term on the left-hand side. The first term on the right-hand side is what we want to compute. As for the final term, $P(X_{(1)} > x, X_{(n)} \le y)$,

note that this is 0 if $x > y$. So, we consider $x < y$

$$\begin{aligned} P(X_{(1)} > x, X_{(n)} \le y) &= P(x < X_1 \le y, x < X_2 \le y, .., x < X_n \le y) \\ &= [P(x < X_1 \le y)]^n, \qquad \text{i.i.d} \\ &= [F(y) - F(x)]^n \end{aligned}$$

So, we have

From the joint density of $(Y_1, ..., Y_{n+1})'$,

$$\prod_{i=1}^{n+1} \frac{1}{\theta} \exp\{-\frac{1}{\theta} y_i\} I(0 < y_i < \infty) = (\frac{1}{\theta})^{n+1} \exp\{-\frac{\sum_{i=1}^{n+1} y_i}{\theta}\} I(0 < y_i < \infty)$$

we obtain the joint density of $(W_1, ..., W_n, S)$ as

$$(\frac{1}{\theta})^{n+1} \exp\{-\frac{s}{\theta}\} I(0 < w_1 < w_2 < ... < w_n < s < \infty),$$

Let

$$Z_1 = W_1/S, Z_2 = W_2/S, ..., Z_n = W_n/S, S.$$

Consider the transformation $(W_1, ..., W_n, S)' \mapsto (Z_1, ..., Z_n, S)'$. Note that the formation is one-to-one with the Jacobian

$$|det \left(\frac{\partial(W_1, ..., W_n, S)}{\partial(Z_1, ..., Z_n, S)} \right) | = |det \begin{pmatrix} s & 0 & \dots & 0 & z_1 \\ 0 & s & \dots & 0 & z_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & s & z_n \\ 0 & 0 & \dots & 0 & 1 \end{pmatrix} | = s^n$$

From the joint density of $(Z_1, ..., Z_n, S)'$,

$$= (\frac{1}{\theta})^{n+1} \exp\{-\frac{s}{\theta}\} s^n I(0 < z_1 < z_2 < ... < z_n < 1) I(s > 0)$$

Figure 2: Alt method 2

$$
\begin{aligned}
F_{(X_{(1)},X_{(n)})}(x,y) &= P(X_{(1)} \le x, X_{(n)} \le y) \\
&= P(X_n \le y) - P(X_1 > x, X_n \le y) \\
&= [F(y)]^n - [F(y) - F(x)]^n
\end{aligned}
$$

Now the joint pdf is

$$
\begin{aligned}
f_{(X_{(1)},X_{(n)})}(x,y) &= \frac{d}{dx}\frac{d}{dy}\{[F(y)]^n - [F(y) - F(x)]^n\} \\
&= \frac{d}{dx} nF(y)^{n-1}f(y) - n(F(y) - F(x))^{n-1}f(y) \\
&= n(n-1)(F(y) - F(x))^{n-2}f(x)f(y)
\end{aligned}
$$

This hold for $x < y$ and for x and y both in the support of the original distribution.

4 Moment Generating Function

4.1 Chi-square MGF

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:

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

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.

4.2 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.

4.3 Exponential Distribution Family

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

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

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.

5 Fisher Information

The Fisher Information always comes with the asymptotic normal distribution of the estimator, and hypothesis testing.

5.1 Multinomial Distribution

Multinomial distribution is a very typical distribution to demonstrate the relationship between the covariance matrix and Fisher Information.

If the observations from multinomial distribution are independent, so we can construct the variance and covariance between two observations, and then get the covariance matrix. This step we can use the MGF or by definition.

But the Fisher Information don't use the inverse of Covariance, use the definition of Fisher Information, which is the variance of score function.

The log-likelihood function of Multinomial distribution

$$p(x,\theta) = \binom{n}{x_1, x_2..x_k} \theta_1^{x_1}\theta_2^{x_2}\theta_3^{x_3}..\theta_k^{x_k}$$

$$l_n(\theta) = const + \sum_{i=1}^{k} x_i log(\theta_i)$$

Then the score function

$$S(x,\theta) = \left(\frac{x_1}{\theta_1}, \frac{x_2}{\theta_2}, ..\frac{x_k}{\theta_k}\right) = diag(\frac{1}{\theta})x$$

Consequently, the Fisher Information

$$\begin{aligned}
I_n(\theta) &= Var(S(\theta)) = diag(\frac{1}{\theta})Var(x)diag(\frac{1}{\theta}) \\
&= ndiag(1/\theta)[diag(\theta) - \theta\theta^T]diag(1/\theta) \\
&= n\left[diag(1/\theta) - (diag(1/\theta)\theta)(diag(1/\theta)\theta)^T\right] \\
&= n\left[diag(1/\theta) - 11^T\right] \\
&= n\begin{pmatrix} \frac{1-p_1}{p_1} & -1 & -1 & .. & -1 \\ -1 & \frac{1-p_2}{p_2} & -1 & .. & -1 \\ .. & .. & .. & .. & \\ -1 & -1 & -1... & & \frac{1-p_k}{p_k} \end{pmatrix}
\end{aligned}$$

6 Weighted Statistics

Let $X_1, ..X_n$ be i.i.d $N(0, \sigma^2)$. $w_1, ..., w_n$ is a constant vector such that $w_1, .., w_n > 0$ and $w_1 + .. + w_n = 1$. Define $\bar{X}_{nw} = \sqrt{w_1}X_1 + .. + \sqrt{w_n}X_n$. Show that $Y_n = \bar{X}_{nw}/\sigma \sim N(0, 1)$.

6.1 Question

Note that $\bar{X}_{nw}$ is a linear combination of $X_1, ..X_n$, we need to use the vector/matrix to show the distribution, while not single one variable.

If $X_i \sim N(\mu_i, \sigma_i^2)$, which we can have a MVN distribution, which each X_i has its own normal distribution. Then the transformation matrix, orthogonal matrix, etc could be applied. Here all the X_i follows the same distribution, and we also can use the similar concept by applying orthogonal matrix.

Also we have the slutsky's theorem, delta method for the asymptotic distribution, however that is under the $n \to \infty$. In this problem, we can't use that.

So this is the exact distribution using the transformation (just the transform is by orthogonal matrix). The MGF or characteristic distribution is always the method when doing transformation.

6.2 MGF

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

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

6.3 Orthogonal Matrix

Consider an orthogonal matrix Σ such that the first row is $(\sqrt{w_1}, \sqrt{w_2}, ..\sqrt{w_n})$. Let

$$(Z_1, Z_2, ..Z_n)^T = \Sigma(X_1, X_2, ..X_n)^T$$

We have $Z^T Z = (\Sigma X)^T(\Sigma X) = X^T\Sigma^T\Sigma X = X^T X$. The characteristic function of Z is

$$\phi_Z(t) = E[exp(it'Z)] = E[exp(i(\Sigma' t)'X)] = exp(-\sigma^2 t't/2)$$

Need to get familiar with the vector form in MGF/characteristic function
Therefore, we have $Z_1, ..Z_n \sim N(0, \sigma^2)$

$$\begin{aligned} Y_n &= \bar{X}_{nw}/\sigma = (\sqrt{w_1}X_1 + \sqrt{w_2}X_2 + .. + \sqrt{w_n}X_n)/\sigma \\ &= Z_1/\sigma \sim N(0,1) \end{aligned}$$

Also,

$$\begin{aligned} (n-1)S_n^2/\sigma^2 &= \sum_{i=1}^{n}(X_i^2 - \bar{X}_{nw}^2)/\sigma^2 \\ &= (X^T X - Z_1^2)/\sigma^2 = (Z_2^2 + ... + Z_n^2)/\sigma^2 \sim \chi^2_{n-1} \end{aligned}$$

Since Y_n and S_n^2 are functions of Z_1 and $(Z_2, ..Z_n)$ respectively, and from the independence of $Z_i, (i = 1, ..n)$, we have Y_n and S_n^2 are independent. It follows that, by the definition of t-distribution, $T_n \sim t_{n-1}/\sigma$. When $w_1 = w_2 = .. = w_n = 1/n, Y_n = \sum_{i=1}^{n} X_i/(\sigma\sqrt{n})$, which is the standardized sample mean. Also,

$$
\begin{aligned}
S_n^2/\sigma^2 &= \frac{\sum_{i=1}^n X_i^2 - \bar{X}_{nw}^2}{n-1} \\
&= \frac{\sum_{i=1}^n X_i^2 - \sum_{i=1}^n {X_i}^2/n}{n-1} \\
&= \frac{\sum_{i=1}^n X_i^2 - n\bar{X}_i^2}{n-1}
\end{aligned}
$$

which is the sample variance.

If there are quadratic forms, we can consider the orthogonal matrix that transform to standard normal distribution.

7 Sufficient and Complete Statistics

7.1 Minimum Sufficient Statistics

7.2 Complete Statistics

7.3 Ancillary Statistics

8 Bivariate Normal Distribution / Partition Matrix

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

Definition: 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

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

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.

8.1 Conditional Distribution

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

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

8.1.1 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

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

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.

8.1.2 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 $V_{ab} = 0$.

8.2 Schur Complement

In order to derive the conditional distributions, we are going to reply on Schur complements.

In linear algebra and the theory of matrices, the Schur complement of a block matrix is defined as follows.

Suppose p, q are nonnegative integers, and suppose A, B, C, D are respectively $p \times p, p \times q, q \times p$, and $q \times q$ matrices of complex numbers. Let

$$M = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

So that M is a $(p+q) \times (q+p)$ matrix. If D is invertible, then the **Schuler complement** of the block D of the matrix M is the $p \times p$ matrix defined by

$$M/D := A - BD^{-1}C$$

If A is invertible, the Schur complement of the block A of the matrix M is the q $\times$ q matrix defined by

$$M/A := D - CA^{-1}B$$

In the case that A or D is singular, substituting a generalized inverse for the inverses on M/A and M/D yields the generalized Schur complement.

8.2.1 Background

The Schur complement arises when performing a block Gaussian elimination on the matrix M. In order to eliminate the elements below the block diagonal, one multiplies the matrix M by a block lower triangular matrix on the right as follows:

$$M = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \to \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} I_p & 0 \\ -D^{-1}C & I_q \end{bmatrix} = \begin{bmatrix} A - BD^{-1}C & B \\ 0 & D \end{bmatrix}$$

where Ip denotes a $p \times p$ identity matrix. As a result, the Schur complement $M/D = A - BD^{-1}C$ appears in the upper-left $p \times p$ block.

Continuing the elimination process beyond this point (i.e., performing a block Gauss–Jordan elimination),

$$\begin{bmatrix} A - BD^{-1}C & B \\ 0 & D \end{bmatrix} \to \begin{bmatrix} I_p & -BD^{-1} \\ 0 & I_q \end{bmatrix} \begin{bmatrix} A - BD^{-1}C & B \\ 0 & D \end{bmatrix} = \begin{bmatrix} A - BD^{-1}C & 0 \\ 0 & D \end{bmatrix}$$

leads to an LDU decomposition of M, which reads

$$M = \begin{bmatrix} I_p & -BD^{-1} \\ 0 & I_q \end{bmatrix} \begin{bmatrix} A - BD^{-1}C & 0 \\ 0 & D \end{bmatrix} \begin{bmatrix} I_p & 0 \\ -D^{-1}C & I_q \end{bmatrix}$$

Thus, the inverse of M may be expressed involving D^{-1} and the inverse of Schur's complement, assuming it exists, as

$$M^{-1} = \begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \left(\begin{bmatrix} I_p & BD^{-1} \\ 0 & I_q \end{bmatrix} \begin{bmatrix} A - BD^{-1}C & 0 \\ 0 & D \end{bmatrix} \begin{bmatrix} I_p & 0 \\ D^{-1}C & I_q \end{bmatrix} \right)^{-1}$$

Here I need to separate the inverse of 2×2 matrix and this partitioned matrix.

$$\begin{bmatrix} I_p & -BD^{-1} \\ 0 & I_q \end{bmatrix} \begin{bmatrix} I_p & BD^{-1} \\ 0 & I_q \end{bmatrix} = \begin{bmatrix} I_p & 0 \\ 0 & I_q \end{bmatrix}$$

$$\begin{bmatrix} I_p & -BD^{-1} \\ 0 & I_q \end{bmatrix}^{-1} = \begin{bmatrix} I_p & BD^{-1} \\ 0 & I_q \end{bmatrix}$$

$$\begin{bmatrix} I_p & 0 \\ -D^{-1}C & I_q \end{bmatrix}^{-1} = \begin{bmatrix} I_p & 0 \\ D^{-1}C & I_q \end{bmatrix}$$

So, we have

$$M^{-1} = \begin{bmatrix} I_p & 0 \\ -D^{-1}C & I_q \end{bmatrix} \begin{bmatrix} [A - BD^{-1}C]^{-1} & 0 \\ 0 & D^{-1} \end{bmatrix} \begin{bmatrix} I_p & -BD^{-1} \\ 0 & I_q \end{bmatrix}$$

If p and q are both 1 (i.e., A, B, C and D are all scalars), we get the familiar formula for the inverse of a 2-by-2 matrix:

$$M^{-1} = \frac{1}{AD - BC} \begin{bmatrix} D & -B \\ -C & A \end{bmatrix}$$

8.2.2 Applications to probability theory and statistics

Suppose the random column vectors X, Y live in R_n and R_m respectively, and the vector (X, Y) in R_{n+m} has a multivariate normal distribution whose covariance is the symmetric positive-definite matrix

$$\Sigma = \begin{bmatrix} A & B \\ B^T & C \end{bmatrix}$$

where $\mathbf{R}^{n\times n}$ is the covariance matrix of X, $C \in \mathbf{R}^{m\times m}$ is the covariance matrix of Y and $B \in \mathbf{R}^{n\times m}$ is the covariance matrix between X and Y.

Then the conditional covariance of X given Y is the Schur complement of C in $\Sigma\Sigma$

$$Cov(X|Y) = A - BC^{-1}B^T$$
$$E(X|Y) = E(X) + BC^{-1}(Y - E(Y))$$

Let V_a be invertible. Let V/V_a be the Schur complement of V_a in V, defined as

$$V/V_a = V_b - V_{ab}V_a^{-1}V_{ab}^T$$

If V/V_a is invertible, then V is invertible

8.3 b

Consider the following

(a) For an arbitrary model, consider the conditional score statistic

$$U_\psi(\xi) = \frac{\partial l_c(\xi, \psi_0)}{\partial \psi}|_{\psi_0=\psi}$$

Show that the conditional score statistic for any model can be written as

$$U_\psi(\xi) = \partial_\psi logp(Y|\xi) - E[\partial_\psi logp(Y|\xi)|s_\lambda(\psi_0)]|_{\psi_0=\psi}$$

The conditional score statistic is the derivative of the conditional distribution

$$U_\psi(\xi) = \frac{\partial l_c(\xi, \psi_0)}{\partial \psi}|_{\psi_0=\psi}$$
$$p(\mathbf{Y}|\xi) = p(\mathbf{Y}|s_\lambda(\psi_0), \xi)p(s_\lambda(\psi_0)|\xi), \qquad p(\mathbf{Y}|s_\lambda(\psi_0), \xi) = \frac{p(\mathbf{Y}|\xi)}{p(s_\lambda(\psi_0)|\xi)}$$
$$l_c(\xi, \psi_0) = logp(\mathbf{Y}|s_\lambda(\psi_0), \xi) = logp(\mathbf{Y}|\xi) - logp(s_\lambda(\psi_0)|\xi)$$

Then we need to prove

$$U_\psi(\xi) = \frac{\partial l_c(\xi, \psi_0)}{\partial \psi}|_{\psi_0=\psi} = \partial_\psi logp(\mathbf{Y}|\xi) - \partial_\psi logp(s_\lambda(\psi_0)|\xi)$$
$$\partial_\psi logp(s_\lambda(\psi_0)|\xi) = E[\partial_\psi logp(Y|\xi)|s_\lambda(\psi_0)]|_{\psi_0=\psi}$$

We can write

$$logp(\mathbf{Y}|\xi) = logp(\mathbf{Y}|s_\lambda(\psi_0), \xi) + logp(s_\lambda(\psi_0)|\xi)$$
$$E\left(\partial_\psi[logp(\mathbf{Y}|\xi)|s_\lambda]\right) = E\left(\partial_\psi[logp(\mathbf{Y}|s_\lambda(\psi_0), \xi)|s_\lambda]\right) + E\left(\partial_\psi[logp(s_\lambda(\psi_0), \xi)|s_\lambda]\right)$$

in which, the integral and expectation can switch, then we have

$$E\left(\partial_\psi[logp(\mathbf{Y}|s_\lambda(\psi_0), \xi)|s_\lambda]\right) = \partial_\psi E\left([logp(\mathbf{Y}|s_\lambda(\psi_0), \xi)|s_\lambda]\right) = \partial_\psi E\left([logp(\mathbf{Y}|\xi)]\right) = 0$$

So,

$$E\left(\partial_\psi[logp(\mathbf{Y}|\xi)|s_\lambda]\right) = \partial_\psi logp(s_\lambda(\psi_0), \xi)$$

Then we show

$$U_\psi(\xi) = \partial_\psi logp(Y|\xi) - E[\partial_\psi logp(Y|\xi)|s_\lambda(\psi_0)]|_{\psi_0=\psi}$$

(b) Suppose that $y_1; ..y_n$ are independent and y_i follows a Poisson distribution with mean $exp(\lambda_0 + \lambda_1 x_{i1} + \psi x_{i2})$, where $(x_{i1}; x_{i2})$ are covariates, $\lambda = (\lambda_0; \lambda_1)$ is the

nuisance parameter vector and ψ is the parameter of interest. Derive the conditional likelihood of ψ and show that this conditional likelihood is free of λ.
The joint distribution of $(y_1, , y_n)$ is given by

$$P(Y|\lambda, \psi) = exp\left(\sum_{i=1}^{n} y_i(\lambda_0 + \lambda_1 x_{i1} + \psi x_{i2}) - \sum_{i=1}^{n} exp(\lambda_0 + \lambda_1 x_{i1} + \psi x_{i2}) - logy_i!\right)$$

Thus, $S_0 = \sum_{i=1}^{n} y_i$ is the sufficient and complete statistics for λ_0, and $S_1 = \sum_{i=1}^{n} y_i x_{i1}$ is the sufficient and complete statistics for λ_1.
The conditional distribution of ψ given S_0, S_1 is given by

$$\begin{aligned} p(\mathbf{Y}, \psi | S = (S_0, S_1)) &= \frac{exp\left(\sum_{i=1}^{n} y_i(\lambda_0 + \lambda_1 x_{i1} + \psi x_{i2}) - \sum_{i=1}^{n} exp(\lambda_0 + \lambda_1 x_{i1} + \psi x_{i2}) - logy_i!\right)}{\sum_{y' \in S} exp\left(\sum_{i=1}^{n} y_i'(\lambda_0 + \lambda_1 x_{i1} + \psi x_{i2}) - \sum_{i=1}^{n} exp(\lambda_0 + \lambda_1 x_{i1} + \psi x_{i2}) - logy\right.} \\ &= \frac{exp\left(S_1\lambda_0 + S_2\lambda_1 + S_3\psi\right) - \sum_{i=1}^{n} exp(\lambda_0 + \lambda_1 x_{i1} + \psi x_{i2}) - logy_i!)}{\sum_{y' \in S} exp\left(S_1'\lambda_0 + S_2'\lambda_1 + S_3'\psi\right) - \sum_{i=1}^{n} exp(\lambda_0 + \lambda_1 x_{i1} + \psi x_{i2}) - logy_i'!)} \\ &= \frac{exp\left(S_3\psi - logy_i!\right)}{\sum_{y' \in S} exp\left(S_3'\psi - logy_i'!\right)}, \qquad S_3 = \sum_{i=1}^{n} y_i x_{i2}, S_3' = \sum_{i=1}^{n} y_i' x_{i2} \end{aligned}$$

which is independent of λ.

(c) Derive the conditional score statistic for part (b) and write out a Newton-Raphson algorithm for obtaining the conditional maximum likelihood estimate of ψ based on $U_\psi(\xi)$.
The log likelihood of the conditional distribution is

$$l_c(\psi) = S_3\psi - logy_i! - log\left[\sum_{y' \in S} exp\left(S_3'\psi - logy_i'!\right)\right], \qquad S_3 = \sum_{i=1}^{n} y_i x_{i2}, S_3' = \sum_{i=1}^{n} y_i' x_{i2}$$

The score function and observed fisher information is

$$\begin{aligned} U_\psi(\xi) &= \frac{\partial l_c(\xi, \psi_0)}{\partial \psi}|_{\psi_0 = \psi} \\ &= \psi - \frac{\sum_{y' \in S} S_3' exp\left(S_3'\psi - logy_i'!\right)}{\sum_{y' \in S} exp\left(S_3'\psi - logy_i'!\right)} \\ \frac{\partial^2 l_c(\xi, \psi_0)}{\partial \psi^2} &= \left[\frac{\sum_{y' \in S} S_3' exp\left(S_3'\psi - logy_i'!\right)}{\sum_{y' \in S} exp\left(S_3'\psi - logy_i'!\right)}\right]^2 - \frac{\sum_{y' \in S} S_3'^2 exp\left(S_3'\psi - logy_i'!\right)}{\sum_{y' \in S} exp\left(S_3'\psi - logy_i'!\right)} \end{aligned}$$

The newton-Raphson algorithm

$$\psi^{k+1} = \psi^k - \left[\frac{\partial^2 l_c(\psi^k)}{\partial \psi^2}\right]^{-1} U_\psi(\psi^k)$$

where $\frac{\partial^2 l_c(\psi^k)}{\partial \psi^2}, U_\psi(\psi^k)$ are from above equations.

(d) Now suppose that we only have two random variables $y_1 \sim Poisson(\mu_1)$ and $y_2 \sim Poisson(\mu_2)$, where y_1 and y_2 are independent. We are interested in making inferences on the ratio $\psi = \mu_1/\mu_2$. Let $\xi = (\psi, \lambda)$, where λ represents the nuisance parameter.

(i) Show that the log-likelihood function of ξ can be written as

$$l(\xi) = (y_1 + y_2)\lambda + y_1 log(\psi) - exp(\lambda)(1 + \psi)$$

where λ is a function of μ_2. Explicitly state what λ is.
Write the joint distribution of y_1, y_2

$$\begin{aligned}
P(y_1, y_2) &= \frac{\mu_1^{y_1} e^{-\mu_1}}{y_1!} \frac{\mu_2^{y_2} e^{-\mu_2}}{y_2!} \\
logP(y_1, y_2) &= y_1 log\mu_1 - \mu_1 + y_2 \log \mu_2 - \mu_2 - logy_1! - logy_2! \\
&= y_1 log\frac{\mu_1}{\mu_2} + y_1 log\mu_2 + y_2 log\mu_2 - \mu_1 - \mu_2 - logy_1! - logy_2! \\
&= y_1 log\frac{\mu_1}{\mu_2} + (y_1 + y_2)log\mu_2 - \mu_2(\mu_1/\mu_2 + 1) - logy_1! - logy_2!
\end{aligned}$$

where

$$\begin{aligned}
\psi &= log\frac{\mu_1}{\mu_2} \\
\lambda &= log\mu_2
\end{aligned}$$

(ii) Derive the conditional likelihood of ψ and write out a Newton-Raphson algorithm for obtaining the conditional maximum likelihood estimate of ψ .
From part (a), we see $y_1 + y_2$ is the sufficient statistics for λ, while $y_1 + y_2 \sim Poission(\mu_1 + \mu_2)$ then we have conditional distribution of ψ condition on $S = y_1 + y_2$.

$$\begin{aligned}
Y(\psi|S = y_1 + y_2, \lambda) &= \frac{exp\left[y_1\psi + (y_1 + y_2)\lambda - exp(\lambda)(\psi + 1) - logy_1! - logy_2!\right]}{exp\left[(y_1 + y_2)log(\mu_1 + \mu_2) - (\mu_1 + \mu_2) - log(y_1 + y_2)!\right]} \\
&= \frac{exp\left[y_1\psi + S\lambda - exp(\lambda)(\psi + 1) - logy_1! - logy_2!\right]}{exp\left[S(\lambda + log(\psi + 1)) - exp(\lambda)(\psi + 1) - logS!\right]} \\
&= \frac{exp\left[y_1\psi - logy_1! - logy_2!\right]}{exp\left[(y_1 + S - y_1)log(\psi + 1)) - logS!\right]} \\
&= \binom{S}{y_1} \left(\frac{\psi}{1 + \psi}\right)^{y_1} \left(\frac{1}{1 + \psi}\right)^{S - y_1}
\end{aligned}$$

The conditional distribution is a binomial, $B(S, \psi/(1 + \psi))$.

The score function and observed fisher information

$$
\begin{aligned}
logY(\psi|S,\lambda) &= y_1 log\psi - Slog(1+\psi) + log\binom{S}{y_1} \\
\partial_\psi logY(\psi|S,\lambda) &= \frac{y_1}{\psi} - \frac{S}{1+\psi} = 0, \qquad \hat{\psi} = y_1/(S-y_1) \\
\partial_\psi^2 logY(\psi|S,\lambda) &= -\frac{y_1}{\psi^2} + \frac{S}{(1+\psi)^2}
\end{aligned}
$$

The $CMLE = \hat{\psi} = y_1/(S-y_1)$. And the newton-Raphson equation

$$
\begin{aligned}
\psi^{k+1} &= \psi^k - \left[\frac{\partial^2 l_c(\psi^k)}{\partial \psi^2}\right]^{-1} U_\psi(\psi^k) \\
&= \psi^k - \left[-\frac{y_1}{\psi^2} + \frac{S}{(1+\psi)^2}\right]^{-1} \left[\frac{y_1}{\psi} - \frac{S}{1+\psi}\right]|_{\psi=\psi^k} \\
&= \psi^k + \frac{y_1/\psi^k - S/(1+\psi^k)}{y_1/{\psi^k}^2 - S/(1+\psi^k)^2}
\end{aligned}
$$