Notes - Statistical Models

June 6, 2020

1 Lecture 6:

1.1 Normalizing Constant

:
We know that

$$C(\theta) = \int h(y) \exp\{\theta^T t(y)\} dy$$

Where the canonical parameter space Θ is composed by all values of θ such that $C(\theta)$ is finite. There is no further restriction on the parameter space.

We now want to prove that the canonical statistic is minimal sufficient: ***Proof:*** Using the likelihood ration approach, consider two samples x, y from the same exponential family.

Consider:

$$\frac{L(\theta, x)}{L(\theta, y)} = \frac{a(\theta)h(\theta)\exp\{\theta^T t(x)\}}{a(\theta)h(\theta)\exp\{\theta^T t(y)\}} = \frac{h(x)}{h(y)} \exp\{\theta^T (t(x) - t(y))\}$$

In order to ensure that the likelihood ratio does not depend on θ $t(x) = t(y)$ meaning that $\frac{h(x)}{h(y)}$ will be independent of θ if $t(x) - t(y) = 0$ meaning that we have a minimal parametrization, i.e there is no linear dependency between the components of the canonical parameter θ.

1.2 Properties of θ

:

- **Regularity:** An exponential family of order k is caller regular if its canonical parameter space Θ is an open set in R^k. If we have regularity then it is easier to perform the optimization of the likelihood function, if the parameter space is not open then not for all points in the parameter space the derivative of the likelihood function exists. The finite support in t implies regularity (can be easily obtained by the properties of $C(\theta)$)

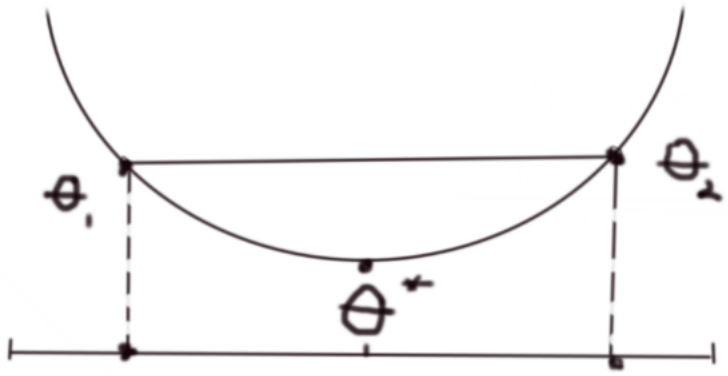
Proposition 3.7: *Finite support implies regularity*

$$C(\theta) = \sum_{t(y) \in T} h(y) \exp\{\theta^T t(y)\} \quad < \infty$$

Note that we are summing over $t(y)$ not over y. All possible values of θ are suitable for us. A finite sum can always be computed. It will provide us a finite number since $T = \{t_1, ..., t_N\} \quad \Rightarrow \Theta = R^{\dim(t)} \rightarrow$ open set
In here we consider such a y that belongs to an exponential family distribution, then the canonical statistic $t(y)$ also has an exponential family distribution, it s a different structural function but belongs to the exponential form, for the canonical statistic t we will have a finite number of possible values.

Proposition 3.8: *Properties of* $\log C(\theta)$

- (a) $\log C(\theta)$ **is a strictly convex function:**
  Let's define some convex function $d(\theta) = \log C(\theta)$. If we want to prove a function is convex we need to prove the following equation.



If we have some curve and 2 points θ_1, θ_2 if we draw a line connectivity between the two points, it should be always above the function. If we take θ^* which belongs in the interval (θ_1, θ_2). Then the value of the function at the point θ^* should be smaller than the value of the function that corresponds to the line. We aim

$$d(\theta^*) < wd(\theta_1) + (1-w)d(\theta_2) \quad \theta^* = w(\theta_1) + (1-w)\theta_2 \quad w \in (0,1)$$

Notice that the sign $<$ denotes strict inequality, and only in the limiting points we get equality.

From corollary we obtain:

$$\log C(\theta^*) \leq w \log C(\theta_1) + (1-w) \log C(\theta_2)$$

whenever $\theta_1 = \theta_2 = \theta^*$. Hence $\log C(\theta^*)$ is a strictly convex function.

- (b) $\log C(\theta)$ **is a continuously differentiable function**

$$d(\theta) = \log C(\theta) \in C_\infty$$

We can differenciate as many times as we want, and will be continuous.

Where

$$C(\theta) = \int_R \underbrace{h(y)\exp\{\theta^T t(y)\}}_{\text{non-negative function} \geq 0} \quad dy$$

Also, for each value of $y, \exp\{\theta^T t(y)\} \in C_\infty$ where C_∞ denotes a functional space, is the set of functions that are infinitely many times differentiable on some space, in our case is a canonical parameter space. then, if we compute the k-th order derivative:

$$D^k C(\theta) = \int_{R^{\dim(t)}} h(y) D^k \exp\{\theta^T t(y)\}$$

Mind that θ is a vector an not an univariate value. We will show the result for each element of this vector.

$$\theta = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_k \end{pmatrix}; \quad t(y) = \begin{pmatrix} t_1(y) = t_1 \\ \vdots \\ t_k(y) = t_k \end{pmatrix}$$

We want to prove that

$$\frac{\partial \log C(\theta)}{\partial \theta} = E[t_i] \quad \text{for} \quad i = 1, ..., k$$

In order to do this we will start our computations considering the norming constant $C(\theta)$ instead of $\log C(\theta)$. By definition we have that:

$$\begin{aligned}
C(\theta) &= \int h(y)\exp\{\theta^T t(y)\}dy \\
\Rightarrow \frac{\partial C(\theta)}{\partial \theta_i} &= \int \frac{\partial}{\partial \theta_i} h(y)\exp\{\theta^T t(y)\}dy \\
&= \int h(y)\exp\{\theta^T t(y)\}\frac{\partial \theta^T t(y)}{\partial \theta_i}dy \\
&= \int h(y)\exp\{\theta^T t(y)\}\Big(\frac{\partial(\theta_1 t_1(y) + \ldots + \theta_k t_k(y))}{\partial \theta_i}\Big)dy \qquad (1) \\
&= \int \underbrace{h(y)\exp\{\theta^T t(y)\}}_{\text{kernel expofam}} t_i(y) \underbrace{\frac{C(\theta)}{C(\theta)}}_{**} dy \\
&= C(\theta)\int t_i(y) f(y;\theta)dy = C(\theta)\cdot \mu_i(\theta) = C(\theta)\cdot E[t_i]
\end{aligned}$$

$**$ In order to get a true density we have to divide $h(y)\exp\{\theta^T t(y)\}$ by $C(\theta)$ and in order to make no change in the equality we should multiply by $C(\theta)$

Hence

$$\frac{\partial \log C(\theta)}{\partial \theta} = \frac{1}{C(\theta)} \cdot \frac{\partial C(\theta)}{\partial \theta_i} = \frac{1}{C(\theta)} \cdot C(\theta)\mu_i(\theta) = \mu_i(\theta)$$

$$\frac{\partial \log C(\theta)}{\partial \theta} = \begin{pmatrix} \frac{\partial \log C(\theta)}{\partial \theta_1} \\ \vdots \\ \frac{\partial \log C(\theta)}{\partial \theta_k} \end{pmatrix} = \begin{pmatrix} \mu_1(\theta) \\ \vdots \\ \mu_k(\theta) \end{pmatrix} = \mu(\theta)$$

- (c) **Matrix** $V_i(\theta)$

We know that

$$D^2 \log C(\theta) = \text{The Hessian of the transformation} \quad \log C(\theta)$$

$$\Rightarrow \Big(\frac{\partial^2}{\partial \theta_i \theta_j} \log C(\theta)\Big)_{i,j=1,\dots,k}$$

We want to compute this element-wise, hence

$$\begin{aligned} \frac{\partial}{\partial \theta_i \theta_j} C(\theta) &= \frac{\partial}{\partial \theta_j}\Big(\frac{\partial}{\partial \theta_i} C(\theta)\Big) \\ &= \frac{\partial}{\partial \theta_j}\Big(\int h(y)\exp\{\theta^T t(y)\} \cdot t_i(y) dy\Big) \\ &= \int \frac{\partial}{\partial \theta_j} h(y)\exp\{\theta^T t(y)\} \cdot t_i(y) dy \\ &= \int h(y)\exp\{\theta^T t(y)\} \cdot t_j(y) t_i(y) \frac{C(\theta)}{C(\theta)} dy \\ &= C(\theta) \int t_j(y) t_i(y) f(y;\theta) dy \\ &= C(\theta) E[t_j(y) t_i(y)] = C(\theta) E[t_j t_i] \end{aligned} \tag{2}$$

Hence,

$$\begin{aligned} \frac{\partial^2 \log C(\theta)}{\partial \theta_i \theta_j} &= \frac{\partial}{\partial \theta_j}\Big(\frac{\partial}{\partial \theta_i} \log C(\theta)\Big) \\ &= \frac{\partial}{\partial \theta_j}\Big(\frac{1}{C(\theta)} \frac{\partial C(\theta)}{\partial \theta_i}\Big) \\ &= \frac{\partial}{\partial \theta_j}\Big(\frac{1}{C(\theta)}\Big) \frac{\partial C(\theta)}{\partial \theta_i} + \frac{1}{C(\theta)} \cdot \frac{\partial}{\partial \theta_j}\Big(\frac{\partial C(\theta)}{\partial \theta_i}\Big) \\ &= -\frac{1}{C^2\theta} \cdot \underbrace{\frac{\partial C(\theta)}{\partial \theta_j}}_{C(\theta)E[t_j]} \cdot \underbrace{\frac{\partial C(\theta)}{\partial \theta_i}}_{C(\theta)E[t_i]} + \frac{1}{C(\theta)} \cdot \underbrace{\frac{\partial^2 C(\theta)}{\partial \theta_j \theta_i}}_{C(\theta)E[t_j t_i]} \end{aligned} \tag{3}$$

$$\frac{\partial^2 \log C(\theta)}{\partial \theta_i \theta_j} = E[t_j t_i] - E[t_i]E[t_j] = \mathrm{Cov}(t_i, t_j) = [V_t(\theta)]_{ij}$$

Where (ij) are components of the $V_t(\theta)$ matrix.
Remark 3.9: *Extension to higher order moments, mgf for t as a transformation of y*

$$\begin{aligned}
M_t(\Psi) = E[\exp\{\Psi^T t\}] = E[\exp\{\Psi^T t(y)\}] \\
&= \int \exp\{\Psi^T t(y)\} f(y;\theta) dy \\
&= \int \exp\{\Psi^T t(y)\} \frac{h(y)}{C(\theta)} \exp\{\theta^T t(y)\} dy \\
&= \frac{1}{C(\theta)} \int h(y) \exp\{(\Psi^T + \theta^T) + t(y)\} \cdot \underbrace{\frac{C(\Psi^T + \theta^T)}{C(\Psi^T + \theta^T}}_{**} dy \\
&= \frac{C(\Psi^T + \theta^T)}{C(\theta)} \int \frac{h(y)}{C(\Psi^T + \theta^T)} \exp\{(\Psi^T + \theta^T) + t(y)\} dy \\
&= \frac{C(\Psi^T + \theta^T)}{C(\theta)} \underbrace{\int f(y; \Psi + \theta) dy}_{1} = \frac{C(\Psi^T + \theta^T)}{C(\theta)}
\end{aligned} \tag{4}$$

$(**)$ So we can get the density. Remember that t is a function of y, i.e, we interpret t as a transformation of y, so if $t = y^2$ where $y \sim \mathcal{N}(0,1)$ then

$$E[t] = E[Y^2] \Rightarrow \int t f(t) dt = \int Y^2 f(y) dy$$

2 Lecture 7:

2.1 Likelihood & Maximum Likelihood

We have

$$\begin{aligned}
y \sim \text{expofam}(\theta) \quad f(y;\theta) = C(\theta)^{-1} h(y) \exp\{\theta^T t(y)\} \\
t(y) = t \sim \text{expofam}(\theta) \quad f(t;\theta) = C(\theta)^{-1} g(t) \exp\{\theta^T t\}
\end{aligned}$$

Notice that the two distributions are not the same.

Proposition 3.10: *Log-Likelihood properties in canonical parameter*

$$l(\theta, t) = \log f(t;\theta) = \log\left(\frac{1}{C(\theta)}\right) + \log(\exp\{\theta^T t\}) + \log(g(t)) = \theta^T t - \log(C(\theta)) + \underbrace{\log(g(t))}_{\text{constant}}$$

The log-likelihood function is strictly concave and $\log C(\theta)$ is also a strictly convex function, and so $-\log C(\theta)$ will be strictly concave.

$$\Rightarrow \theta^T t - \log(C(\theta)) \quad \text{strictly concave} \quad \Rightarrow l(\theta, t) \quad \text{strictly concave}$$

2.2 Score Function

$$U(\theta) = U(\theta; t) = \frac{\partial l(\theta, t)}{\partial \theta} = \frac{\partial \theta^T t}{\partial \theta} - \frac{\partial \log(C(\theta))}{\partial \theta} = t - E[t] = t - \mu_t(\theta)$$

2.3 Observed Fisher Information

In the case of the exponential family distribution the Observed Fisher Information is not dependent on t.

$$J(\theta) = J(\theta; t) = \frac{-\partial U(\theta)}{\partial \theta} = -\Big(\frac{\partial t}{\partial \theta} - \mu_t(\theta)\Big) = -\Big(-\frac{\partial}{\partial \theta}\mu_t(\theta)\Big) = V_t(\theta) = \text{Var}(\theta)$$

Where $V_t(\theta)$ is the variance of the canonical statistic θ while $\text{Var}(\theta)$ could be the variance in the univariate case or a matrix otherwise.

Proposition 3.11: *Maximum Likelihood*

$$U(\theta) = 0 \Rightarrow t = \mu_t(\theta) \Rightarrow \hat{\theta} = \mu_t^{-1}(t)$$

We have proved that $\mu = E[t]$ { under θ parametrization} $= (\theta) = \mu_t(\theta)$ The most important here is that it is a function of the parameter θ so the ML equation will have the following form $t = d(\theta)$. Since $l(\theta; t) = l(\theta)$ is strictly concave.

$$\Rightarrow \frac{\partial l(\theta)}{\partial \theta} = t - d(\theta) \quad \text{is monotonically decreasing in } \theta$$

$\iff$ $d(\theta)$ is a monotonical function $\iff$ there exists $d^{-1}(\cdot)$. We will now prove that the Expected Fisher information $I(\theta)$ equals the observed Fisher information $J(\theta)$

$$\Rightarrow I(\theta) = \underbrace{E[J(\theta)]}_{\text{deterministic quantity}} = J(\theta) = V_t(\theta)$$

Remark:

- (1)
  In case of the canonical parameter: The ML equation is given in a closed-form $t = \mu_t(\theta) = 0, \forall \theta \in \Theta : I(\theta) = J(\theta) = V_t(\theta)$ (directly obtained from the proposition). This equality id for all θ. Later we will consider for another parametrization of the exponential family distribution and then we will no longer have this equality. We will only have it for the point of the MLE but not other points. In the case of the **canonical parametrization** we have this equality everywhere regardless of the value of θ, $I(\theta) = J(\theta)$

- (2)
  $y \to t = t(y)$ in reality we have a sample of y observation, in practice $y_1, \ldots, y_n \overset{i.i.d}{\sim}$ expofam(θ) Then $Y = (y_1, \ldots, y_n) \sim$ expofam(θ)

$$f(Y) = C(\theta)^{-1} \cdot \prod h(y_i) \exp\{\theta^T \sum_{i=1}^{n} t(y_i)\}$$

on the other hand, define $t_i = t(y_i)$ We also have that $t_1, \ldots, t_n \overset{i.i.d}{\sim}$ expofam(θ)

$$\begin{aligned} \underline{\textbf{Canonical Parameter:}} \;& \theta \\ \underline{\textbf{Canonical Statistic:}} \;& t_n^* = \sum_{i=1}^{n} t(y_i) \\ \underline{\textbf{Norming Constant:}} \;& C_n(\theta) = C(\theta)^n \end{aligned} \tag{5}$$

Hence, the likelihood equation: $t_n^* - \mu_{t_n^*}(\theta) = 0$. Where

$$\mu_{t_n^*} = \frac{\partial \log C_n(\theta)}{\partial \theta} = \frac{\partial \log C(\theta)^n}{\partial \theta} = \frac{\partial n \log C(\theta)}{\partial \theta} = n\mu_t(\theta)$$

$$\Rightarrow = \sum_{i=1}^{n} t_i - \underbrace{n \cdot \mu_t(\theta)}_{\text{computed for a single observation}} = 0.$$

Why $C_n(\theta) = C(\theta)^n$?

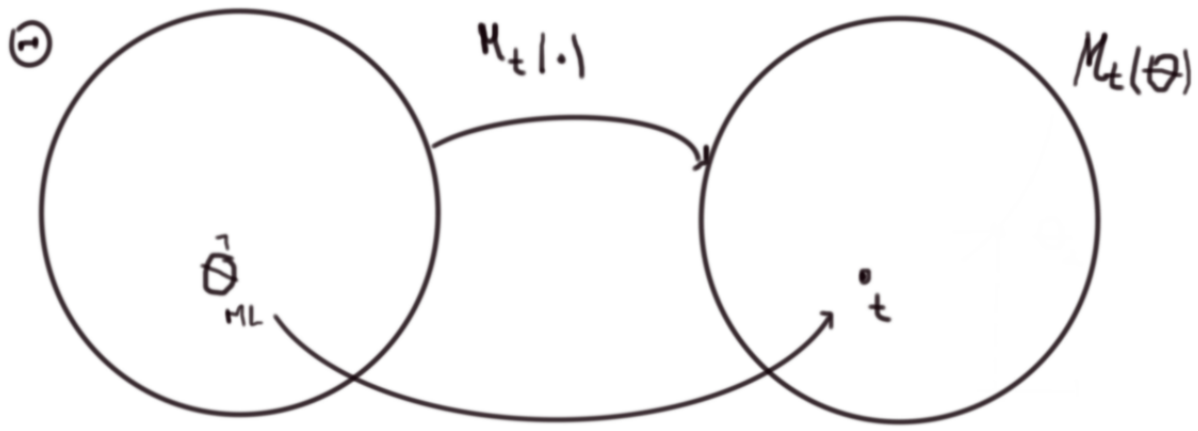
$$T = (t_1, \ldots, t_n) \sim \text{expofam}(\theta) : f(T;\theta) = \underbrace{C(\theta)^{-n}}_{C(\theta)^{-1} = a_n(\theta)} \prod_{i=1}^{n} g(t_i) \cdot \exp\{\theta^T \sum_{i=1}^{n} t_i\}$$

2.4 Information matrix in case of a sample

$$n \cdot I_n(\theta) = n \cdot J_n(\theta)$$

In the case of a sample the Fisher Information has an additional property

Proposition 3.13: *MLE Existence*



The RHS circle refers to the set we obtain by transforming the parameter space by using the mean value transformation μ_t. Notice there is a one-to-one transformation fro $\hat{\theta}_{ML} \to t$. Our ML equation is given by $\mu_t(\theta) = t$, we should be able to recover θ from $\mu_t(\theta) = t$. i.e, if we know t exists then we should know there exists $\hat{\theta}_{ML}$ which uses the transformation $\mu_t(\cdot)$ moving form one space into the other. In order to ensure this we should ensure that the canonical statistic t belongs to the space $\mu_t(\theta; \quad t \in \mu_t(\theta))$ which is enough to find a solution for the ML equation. Notice that $\mu_t(\theta)$ is an open set

$$\text{A is a open} \quad \Rightarrow \forall a \in AB_\epsilon(a) \in A, \quad \epsilon \to \{\tilde{a} : d(\tilde{a}, a) \leq \epsilon\}$$

2.5 Mean Value Parametrization

Instead of using θ for describing our model we will write it instead as a function of the parameter vector μ_t. We want to re-write the density function in terms of μ and then construct the likelihood equation.
Now: μ parametrization (or mean-value parametrization) In our case

$$\underbrace{\mu}_{\text{new parameter}} = E[t] = \mu_t(\underbrace{\theta}_{\text{old parameter}})$$

2.6 Reparametrization Lemma

Tells us how the score function and Fisher Information Matrix are related under different re-parametrization of the family of distributions.

Proposition 3.14: *Reparametrization*
Here

$$\frac{\partial \theta}{\partial \mu} = \Big(\frac{\partial \mu}{\partial \theta}\Big)^{-1} = \Big(\frac{\partial \mu_t(\theta)}{\partial \theta}\Big)^{-1} = V_t(\theta)^{-1}$$

$$
\begin{aligned}
\underline{\textbf{Score Function}}: \quad U_\mu(\mu) &= \Big(\frac{\partial\theta}{\partial\mu}\Big)^T U_\theta(\theta(\mu)) \\
&= (V_t^{-1})^T(t-\mu_t(\theta)) \\
&= V_t^{-1}(t-\mu_t^{-1}(\mu)) \\
&= V_t^{-1}(t-\mu) \\
\underline{\textbf{Likelihood Equation}}: \quad U_\mu(\mu) &= 0 \\
&\iff V_t^{-1}(t-\mu)=0 \Rightarrow t-\mu=0 \\
&\Rightarrow \hat{\mu}_{ML}=t \\
\underline{\textbf{Expected Fisher Information}}: I_\mu(\mu) &= \Big(\frac{\partial\theta}{\partial\mu}\Big)^T I_\theta(\theta(\mu))\Big(\frac{\partial\theta}{\partial\mu}\Big) \\
&= V_t^{-1}V_tV_t^{-1}=V_t^{-1} \\
\underline{\textbf{Observed Fisher Information at } \hat{\mu}}: J_\mu(\hat{\mu}_{ML}) &= \Big(\frac{\partial\theta}{\partial\mu}\Big)^T I_\theta(\theta(\hat{\mu}_{ML}))\frac{\partial\theta}{\partial\mu} \\
&= V_t^{-1}V_tV_t^{-1}=V_t^{-1}
\end{aligned}
\tag{6}
$$

3 Lecture 10:

3.0.1 Large Sample asymptotic

Let $t(y)$ be the canonical statistic for a single observation, while $t_n=\sum_i t(y_i)$ for the entire sample. From early in the course we have discussed that if $n=1$ then the ML equation $E[t]=\mu_t=t$ where t is the value of the canonical statistic, while μ_t is the vector of parameters in the mean value parametrization. Similarly, the ML equation for any n should be given by:

$$n\cdot\mu_t=E[t_n]=t_n=\sum_i t(y_i)$$

$$** \Rightarrow \mu_{t,ML}=\frac{1}{n}t_n=\frac{1}{n}\sum_i t(y_i) \Rightarrow \text{Sample mean}$$

In the case of the mean value parametrization, the ML estimator could be very easily obtained. In case of other parametrizations we can use the property of the ML estimator (one-one) we apply the transformation to our ML estimator. $\rightarrow$ $**$ is the MLE in any parametrization. From $**$ we can also derive asymptotic properties of this MLE (i) consistency, (ii) asymptotical normality.

If we want to derive the distribution of $\hat{\mu}$ we will have to use two classical results from probability theory (i) Law of Large Numbers, (ii) Central Limit Theorem.

These two results can be applied in $**$. In the case of the LLN we need to prove that $E[t]<\infty$, while in the case of the CLT we need to show the existence of the saddle point.

From $\hat{\mu}_{ML}=\frac{1}{n}t_n=\frac{1}{n}\sum_i t(y_i)$. Let $w_i=t(y_i)$ Then,

- $w_1, \ldots, w_n$ - i.i.d, since $y_1, \ldots, y_n$ - i.i.d and the same transformation $t(\cdot)$ is applied to the y-sequence.
- $E[W_i] = E[t(y_i)] = \mu_t$
- $\text{Var}(W_i) = \text{Var}(t(y_i)) = V_t$

Using these three results we can apply LLN, and CLT.

3.1 Regularity Condition:

3.1.1 Law of Large Numbers

Provides us answer to the question of what quantity the MLE converges to.

$$X_1, \ldots, X_n \quad \text{with} \;\; E[X_i] = \mu_t < \infty$$

Then

$$\bar{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i \rightarrow \;\; a.s \; \mu_t \;\; \text{as} \;\; n \rightarrow \infty$$

3.1.2 Central Limit Theorem

Tells us which asymptotic distribution this estimator will have.

$$X_1, \ldots, X_n \overset{i.i.d}{\sim} f(\cdot, \theta) \quad \text{with} \;\; \text{Var}(X_i) = \sigma^2 < \infty$$

Then

$$\sqrt{n}\frac{X_n - \mu}{\sigma} \xrightarrow{D} \mathcal{N}(0,1)$$

or

$$\iff \sqrt{n}\left(\frac{1/n\sum_{i=1}^{n} W_i - E[W_i]}{\sqrt{\text{Var}(W_i)}}\right) \xrightarrow{D} \mathcal{N}(0,1) \iff \sqrt{n}\left(\frac{\hat{\mu}_{ML} - \mu_t}{\sqrt{V_t}}\right) \xrightarrow{D} \mathcal{N}(0,1)$$

$$\sqrt{n}(1/n\sum_{i=1}^{n} W_i - E[W_i]) \xrightarrow{D} \mathcal{N}(0, \text{Var}(W_i)) \iff \sqrt{n}(\hat{\mu}_{ML} - \mu_t) \xrightarrow{D} \mathcal{N}(0, \text{V}_t) \quad \text{as } n \rightarrow \infty$$

3.2 Likelihood equation for a sample of size n

$$n\hat{\mu}_t = \hat{\mu}_{t_n} = t_n \quad \mu_t \text{ is the mean of a single observation}$$

$$\hat{\mu} = \frac{t_n}{n} = \frac{1}{n}\sum_{i=1}^{n} t(y_i) \xrightarrow{a.s} \mu_t \;\; \text{(By LLN)}$$

and

$$\sqrt{n}(\hat{\mu}_t - \mu_t) \xrightarrow{D} \mathcal{N}(0, V_t) \quad \text{(By CLT)}$$

3.2.1 Canonical Parametrization

$\theta \to \mu_t$ (one to one map), $\mu_t(\theta)$ is infinitely many times differentiable $\iff$ $\mu_t^{-1}(\cdot)$ exists. The MLE for θ is given by

$$\theta : \mu_t(\theta) = \frac{t_n}{n} \Rightarrow \hat{\theta} = \mu_t^{-1}(t_n/n) \to \hat{\theta} \overset{a.s}{\to} \theta$$

By the **Continuous mapping Theorem** that is,

$$\text{if} \epsilon_n \overset{a.s}{\to} \epsilon \quad \text{as} \quad n \to \infty \quad \text{and} \quad \Psi(\cdot) \quad \text{is continuous} \Rightarrow \Psi(\epsilon_n) \overset{a.s}{\to} \Psi(\epsilon) \quad \text{as} \quad n \to \infty$$

Remark: We cannot say that $\hat{\mu}_{t,ML} \overset{d}{\to} \mathcal{N}(\mu_t, V_t/n) \;\; n \to \infty$, the proper statement would be $\hat{\mu}_{t,ML} \approx \mathcal{N}(\mu_t, V_t/n)$

Proposition 4.3: *Asymptotic distribution of MLE*

Consider $\eta = \eta(\mu_t)$. If we want to prove **consistency** then we would need for η to be a continous function; for **asymptotic normality** we will require that η is a first order continuously differentiable function. We need to formulate the following results:

- 1) $\hat{\eta} \overset{a.s}{\to} \eta, \quad n \to \infty$ which is also a direct consequence of the continous mapping theorem.
- 2) $\sqrt{n}(\hat{\eta}_{ML} - \eta) \to \mathcal{N}\left(0, \left(\frac{\partial \eta}{\partial \mu}\right)^T \; V_t \;\; \frac{\partial \eta}{\partial \mu}\right))$ i.e, we apply the delta method.

Example: Canonical parametrization:

$$\sqrt{n}(\hat{\eta}_{ML} - \eta) \overset{D}{\to} \mathcal{N}(0, V_t^{-1})$$

3.3 Delta Method

If $\sqrt{n}(\eta_n - \eta) \sim \mathcal{N}(0, \Sigma) \quad \Psi(\cdot)$ is continous.

3.4 Speed of convergence

From the asymptotic distribution of the MLE we know that we can write any parametrization as

$$\hat{\theta}_{ML} - \theta = \underbrace{O(1/\sqrt{n}}_{\text{Bounded sequence}}$$

The upper and lower bound of our sequence depends on our parametrization and our family of distribution.

$$\hat{\theta}_{ML} \approx \theta + O(1/\sqrt{n}$$

The question now is, how big our sample should be for getting a good approximation? $1/\sqrt{n}$ is the rate of convergence, it tells us how many elements we should have in our sample in order to get a good approximation, if n is too large it could be expensive.

3.4.1 Saddle point approximation

Provides $1/n$ convergence. We could get a better approximation, better results with order $1/n$ instead of $1/\sqrt{n}$. The canonical statistic t has an exponential family distribution.

$$f(t;\theta) = a(\theta)g(\theta)\exp\{\theta^T t\} \qquad g(t) = \int_{t(y)=t} h(y) \;\; dy$$

Where $g(t)$ has a very complicated form; the analytical computation of $g(t)$ is unfeasible in most cases, hence for computing the density function $f(t;\theta)$ we would like to approximate $g(t)$. Some motivations for the saddle point approximation is that

- (i) improves the convergence, as now we will have $1/n$ instead of $1/\sqrt{n}$.
- (ii) We would like to get something that looks as an analytical expression for the density function of the canonical statistic for example. For this reason we will approximate the structural function $g(t)$.

The saddle point approximation is a general technique, but in this context we will apply it to the exponential family only. We aim to approximate the density function of the canonical statistic t for some parameter value θ_0. This proof consists on 3 parts (i) we write down what we know about the exact density of t (ii) we write down what we know about the approximate density of t. (iii) we write down the combination of this knowledge together.

If we have a sample / single observation y of the exponential distribution, then the canonical statistic will also have an exponential distribution. The difference is only in the structural function. We consider:

The density function of the canonical statistic t at point θ

$$** f(t;\theta) = \frac{1}{C(\theta)} g(t) \cdot \exp(\theta^T t)$$

The density function of t for the parameter value θ_0

$$f(t;\theta_0) = \frac{1}{C(\theta_0)} g(t) \cdot \exp(\theta_0^T t)$$

We use another parameter value θ_0 to get the saddle point. Using $f(t;\theta_0)$ we also can express the structural function $g(t)$ as:

$$g(t) = \frac{f(t;\theta_0) \cdot C(\theta_0)}{\exp\{\theta_0^T t\}}$$

We can now rewrite $**$ in terms of t parametrized by θ_0. Thus,

$$\begin{aligned} f(t;\theta) &= \frac{C(\theta_0)}{C(\theta)} \cdot f(t;\theta_0)\frac{\exp(\theta^T t)}{\exp(\theta_0^T t)} \\ &= \frac{C(\theta_0)}{C(\theta)} \cdot f(t;\theta_0)\exp(\theta^T t - \theta_0^T t) \end{aligned} \tag{7}$$

Where t is the MLE for μ in the mean value parametrization $t \sim \mathcal{N}(\mu_t, V_t^{-1})$ $t = \hat{\mu}_{t,ML}$, is asymptotically normal. We want to approximate the density function of t by using the asymptotic results for the MLE.

$$f(t;\theta) \approx \frac{1}{(2\pi)^{k/2}\sqrt{\det(V_t)}} \exp\Big\{ -\frac{1}{2}\underbrace{(t-\mu_t(\theta))}_{**\,\hat{\theta}_{MLE}}{}^T V_t^{-1}(t-\mu_t(\theta))\Big\} \quad k = \dim(t) = \dim(\theta)$$

The density function will hold true whenever θ is close to $\hat{\theta}_{MLE}$.

$$\mu_t(\hat{\theta}_{MLE}) = \hat{\mu}_{t.ML} = t \quad \text{then } ** \text{ will be } 0$$

$f(t;\theta)$ has the structure of an exponential distribution ($V_t \sim \hat{V}_t$)

The structural function $g(t)$ is given by:

$$g(t) \approx \frac{1}{(2\pi)^{k/2}\sqrt{\det(V_t(\hat{\theta}_{MLE}))}} C(\hat{\theta}_{MLE})\exp(\hat{\theta}_{MLE}^T t)$$

We now rewrite:

$$\begin{aligned} f(t;\theta_0) &\approx (2\pi)^{-k/2}|V_t(\hat{\theta}_{MLE})|^{-1/2}\frac{C(\theta_0)^{-1}\exp(\hat{\theta}_0^T t)g(t)}{C(\theta_{MLE})^{-1}\exp(\hat{\theta}_{MLE}^T t)g(t)} \\ &\approx (2\pi)^{-k/2}|V_t(\hat{\theta}_{MLE})|^{-1/2}\frac{L(\theta_0)}{L(\hat{\theta}_{ML})} \end{aligned} \tag{8}$$

$$\frac{1}{\sqrt{|\hat{V}_t|}} = \sqrt{|\hat{V}_t|^{-1}} = \sqrt{|V_t^{-1}|} = \sqrt{I_t}$$

where V_t^{-1} is the Fisher Information in case of the mean-value parametrization $\hat{I}_\mu$

4 Lecture 11:

To read: Section 5.1
We will construct the saddle point approximation, through the following example.

Example 4.4 *Time to n successes, continued from Example 4.2*
In this example we have a sample from the geometric distribution $y_1, ..., y_n \overset{i.i.d}{\sim} \text{Ge}(\pi_0)$ where

$\pi_0 = P(\text{success})$, while y_i refers to the number of trials until the first success, consequently $\sum_{i=1}^{n} y_i$ will be the number of trials until n successes . The aim of this example is to determine the number of trials until n success.

$$\Rightarrow \sum_{i=1}^{n} y_i = t_n$$

We will now write the MLE for the mean value parameter of each y_i and we will also like to derive the MLE for the probability of success π_0. In particular, in the example of saddle point approximation we would like to derive and improve the approximation for the density function of π_0. In order to do that we are going to present the sequence of random variables $y_1, ..., y_n$ as a member of the exponential family.

We will first consider a single observation y_i from our sample and then make some changes so we can consider the entire sample. The probability mass function of this random variable is parametrized by π_0. Note that if we have repeated our procedure y_i times then at the place y_i we will have our first success while the remaining positions will be zero's. $(\underbrace{0 \;\; 0 \;\; ... \;\; 0}_{y_i - 1} \;\; 1)$.

$$\begin{aligned} f(y_i; \pi_0) &= (1-\pi_0)^{y_i - 1}\pi_0 \\ &= (1-\pi_0)^{y_i}\frac{\pi_0}{1-\pi_0} \\ &= \frac{\pi_0}{1-\pi_0} \;\; \exp\{y_i \log(1-\pi_0)\} \end{aligned} \tag{9}$$

Where

$$\begin{aligned} \text{Canonical Statistic: } t(y_i) &= y_i \\ \text{Canonical parameter: } \theta &= \log(1-\pi_0) \\ \text{Norming Constant: } C(\theta) &= \frac{1-\pi_0}{\pi_0} \end{aligned} \tag{10}$$

We now want to present the previous result as a function of θ and also we would want to construct the norming constant as a function of θ and not π_0 in particular by our definition of the parameter θ

$$\theta = \log(1-\pi_0) \iff e^{\theta} = 1-\pi_0 \iff \pi_0 = 1 - e^{\theta}$$

also

$$C(\theta) = \frac{1-\pi_0}{\pi_0} = \frac{1-(1-e^{\theta})}{1-e^{\theta}} = \frac{e^{\theta}}{1-e^{\theta}}$$

We can rewrite the density as a function of θ

$$f(y_i; \theta) = \frac{e^{\theta}}{1-e^{\theta}} \exp\{y_i \theta\}$$

We now have everything in order to compute the mean parameter of the canonical statistic t, this is needed as the aim of this exercise is to construct the MLE for the mean value

parameter. (in case of the mean value parametrization) this is important because we know if we want to discuss asymptotic properties of the MLE in any parametrization, the starting point is to discuss these properties in the case of the mean value parametrization.

If we had another parametrization we would only need to use the delta method, as **Proposition 4.3**. Our next step is to compute the expected value of the canonical statistic in the sample parameter case because is where we get the mean value parameter μ and we would also like to compute the variance of the canonical statistic t because that is needed when we write down the asymptotic distribution for the MLE of μ.

Mean of $t(y_i)$:

$$\begin{aligned}\mu = \mu_t(\theta) &= \frac{\partial}{\partial\theta}\log C(\theta) = \frac{\partial}{\partial\theta}\log\Big(\frac{e^\theta}{1-e^\theta}\Big)\\ &= \frac{\partial}{\partial\theta}\log(e^\theta) - \log(1-e^\theta)\\ &= \frac{\partial}{\partial\theta}(\theta - \log(1-e^\theta))\\ &= 1 - \frac{1}{1-e^\theta}(-e^\theta) = 1 + \frac{e^\theta}{1-e^\theta}\\ &= \frac{1}{1-e^\theta}\end{aligned} \tag{11}$$

Since we want to approximate the density of the MLE for π_0 we would like to re-write this quantity in terms of π_0. Then

$$\mu = \frac{1}{1-e^\theta} = \frac{1}{\pi_0}$$

Next we compute the variance:

Variance of $t(y_i)$:

$$V_t = \frac{\partial\mu_t(\theta)}{\partial\theta} = \frac{\partial}{\partial\theta}\frac{1}{1-e^\theta} = \frac{e^\theta}{(1-e^\theta)^2} = \frac{1-\pi_0}{\pi_0^2} \tag{12}$$

This is what we basically need for deriving the asymptotic theorem of the MLE of the parameter μ given the sample $y_i, ..., y_n$.

Our next step will be the computation of the MLE for μ and then we will discuss it's properties.

Likelihood equation:

We will now use the whole sample, instead of the single observation y_i. This likelihood equation is derived in several steps. (i) if we know the canonical statistic for a single observation, then we know the canonical statistic for a whole sample.

- Canonical statistic :

$$t_n = \sum_{i=1}^{n} t(y_i) = \sum_{i=1}^{n} y_i$$

- Mean of the canonical statistic in case of the whole sample:

$$E[t_n] = \sum_{i=1}^{n} E[y_i] = \sum_{i=1}^{n} \mu = n\mu$$

$$\Rightarrow \quad \text{Likelihood equation: } n\mu = t_n \Rightarrow \hat{\mu}_{ML} = \frac{t_n}{n} = \frac{\sum_{i=1}^{n} y_i}{n} = \bar{y}$$

What else can we say about $\hat{\mu}_{ML}$ we could use **Proposition 4.3** where he have already asymptotic distribution of $\hat{\mu}_{ML}$ in case of the exponential family distribution.

Asymptotic distribution of $\hat{\mu}_{ML}$

$$\sqrt{n}(\hat{\mu}_{ML} - \mu) \overset{d}{\sim} \mathcal{N}(0, V_t) = \mathcal{N}(0, \frac{1-\pi_0}{\pi_0^2})$$

Consider for Exercise 4.2: In particular using the last statement we get the asymptotic variance for $\hat{\mu}$

$$\text{Var}(\hat{\mu}_{ML}) = \frac{1}{n}\frac{(1-\pi_0)}{\pi_0^2}$$

MLE for π_0

$$\hat{\pi}_{0,ML} = \frac{1}{\hat{\mu}_{ML}} = \frac{1}{t_n/n} = \frac{n}{t_n}$$

Asymptotic distribution for $\pi_{0,ML}$: Proposition 4.3

$$\sqrt{n}(\pi_{0,ML} - \pi_0) \overset{d}{\sim} \left(0, \left(\frac{\partial \pi_0}{\partial \mu}\right)^2 \cdot V_t\right)$$

$$\frac{\partial \pi_0}{\partial \mu} = \frac{\partial}{\partial \mu}\frac{1}{\mu} = -\frac{1}{\mu^2} = -\pi_0^2$$

$$\Rightarrow \sqrt{n}(\pi_{0,ML} - \pi_0) \overset{d}{\sim} \mathcal{N}\left(0, (\pi_0^2)^2 \ \ \frac{(1-\pi_0)}{\pi_0^2}\right) = \mathcal{N}(0, \pi_0^2(1-\pi_0))$$

Consider for Exercise 4.2: We also get $V(\hat{\pi}_{0,ML}) = \frac{1}{n}\pi_0^2(1-\pi_0)$

Saddle point Approximation:

$$f(\hat{\pi}_{0,ML;\pi_0}) \approx (2\pi)^{-k/2}\sqrt{\det(I_{\pi_0}(\hat{\pi}_{0,ML}))} \cdot \frac{L(\pi_0)}{L(\pi_{0,ML})}$$

Where k is the dimension of the parameter vector, in our case we have a single parameter, so $k = 1 = \dim(\pi_0)$.

Also, the determinant of a uni-variate quantity is just the quantity. So we only need to find $I_{\pi_0}(\hat{\pi}_{0,ML})$ which we already have, as we have previously computed the variance $V(\hat{\pi}_{0,ML})$, hence it should be the inverse of the latter. i.e

$$I_{\pi_0}(\hat{\pi}_{0,ML}) = (V(\hat{\pi}_{0,ML}))^{-1} = n/\hat{\pi}^2_{0,ML}(1 - \hat{\pi}_{0,ML})$$

Important remark: The Fisher information $I_{\pi_0}(\hat{\pi}_{0,ML})$ is computed in the whole sample.

Likelihood Function:

$$\begin{aligned} L(\pi_0) &= \prod_{i=1}^{n} f(y_i; \pi_0) = \prod_{i=1}^{n} (1-\pi_0)^{y_i - 1}\pi_0 = (1-\pi_0)^{\sum y_i - n}\pi_0^n = (1-\pi_0)^{(\frac{n}{\hat{\pi}_{0,ML}} - n)}\pi_0^n \\ L(\hat{\pi}_{0,ML}) &= (1 - \hat{\pi}_{0,ML})^{(\frac{n}{\hat{\pi}_{0,ML}} - n)}\hat{\pi}^n_{0,ML} \end{aligned} \tag{13}$$

Hence the saddle point approximation is given by,

$$\begin{aligned} f(\hat{\pi}_{0,ML}; \pi_0) &= \frac{1}{\sqrt{2\pi}}\sqrt{\frac{n}{\hat{\pi}^2_{0,ML}(1 - \hat{\pi}_{0,ML})}}\frac{(1-\pi_0)^{(\frac{n}{\hat{\pi}_{0,ML}} - n)}\pi_0^n}{(1 - \hat{\pi}_{0,ML})^{(\frac{n}{\hat{\pi}_{0,ML}} - n)}\hat{\pi}^n_{0,ML}} \\ &= \sqrt{\frac{n}{2\pi}}\frac{(1-\pi_0)^{(\frac{n}{\hat{\pi}_{0,ML}} - n)}\pi_0^n}{(1 - \hat{\pi}_{0,ML})^{(\frac{n}{\hat{\pi}_{0,ML}} - n + \frac{1}{2})}\hat{\pi}^{n+1}_{0,ML}} \end{aligned} \tag{14}$$

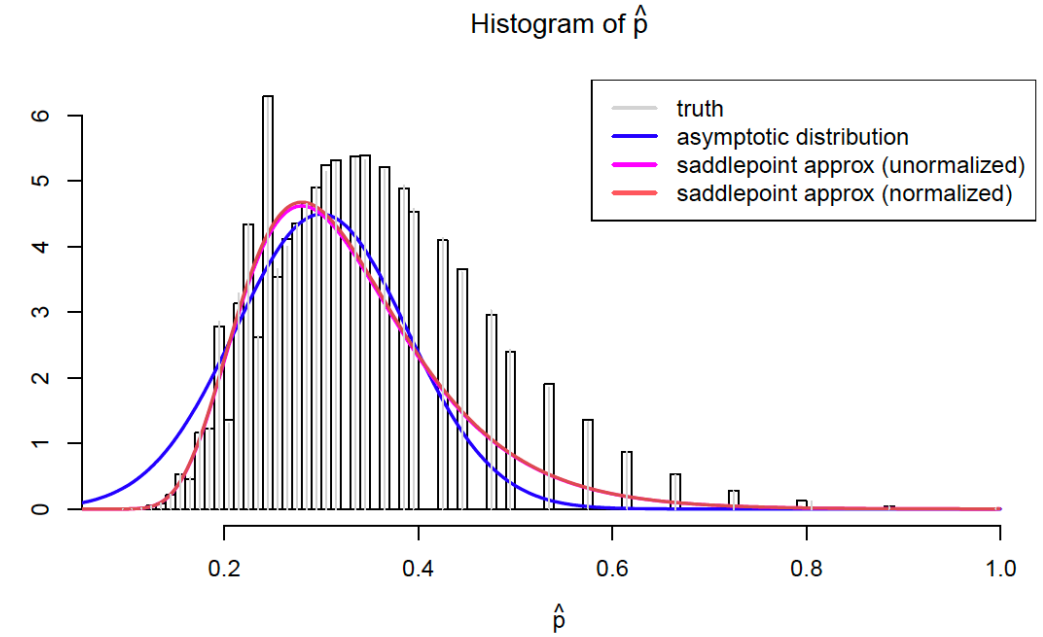
In the following example we can see the results of a simulation with a sample of 8 y_i variables where $\pi_0 = 0.3$ we generate a sample of 100 and computed a histogram for the values of $\hat{y}$ there are three approximations

(i) blue line: Normal approximation, asymptotic distribution $\sim \mathcal{N}(0, \pi_0^2(1 - \pi_0))$. It does not do a good job, prove of that is the left tail, it does not perform a good approximation, also it does not start from zero, hence we have some probability of $\hat{\pi}_0$ to be negative in case of the normal approximation which is expected but theoretically it should not be the case because by construction $\hat{\pi}_0$ is always positive.

(ii) pink line: the saddle point approximation with $f(\hat{\pi}_{0,ML})$, notice that this is not a true density function, we have to normalize it.

(iii) red line: normalized to the true density function saddle point approximation.
The left tail is very well approximated by (ii) and (iii) and the distance between (ii) and (iii) is very small, telling us that it is enough to use the unnormalized version.

Saddlepoint approximations (6) – Example with $n = 8$ and $p = 0.3$



Note: In both cases a discrete distribution is approximated by a continuous distribution.

We could also use the **Reparametrization lemma** for getting the same result.

4.1 Significance test:

Regular exponential family of order r describing data y We assume that the canonical statistic as well as the canonical parameter vector could be partitioned into two parts. $\theta = (\theta_u, \theta_v) = (\lambda, \psi)$ and $t = (u, v)$ where $\dim(\theta_u) = p$ and $\dim(\theta_v) = q$, i.e $r = p + q$. Our aim here is to do some inference about the parameter ψ. We will consider λ as the nuisance parameter in our model, and mainly we will deal with the so-called **model reduction hypothesis**, we will like to test $\psi = 0$.

Why is this important? Because if $\psi = 0$ this means that we can reduce the dimensionality of our model so the canonical statistic and also the minimal sufficient statistic in this case will have not r components but only p components.

This is why this hypothesis formulated as $H_0 : \psi = 0$ is nothing else as the model reduction hypothesis. Alternatively we can also perform the comparison of two models, (i) smaller model where $\psi = 0$ and (ii) larger model.

We have that:

$$t = \begin{pmatrix} u \\ v \end{pmatrix}; \quad \theta = \begin{pmatrix} \theta_u = \lambda \\ \theta_v = \psi \end{pmatrix}$$

The aim of this section is to test the model reduction hypothesis which tells us that is enough just to use the part of our canonical statistic t in our model which is u, this could be presented as:

$$H_0 : \psi = 0 \quad \text{against} \quad H_1 : \psi \neq 0$$

Where ψ is a vector. Let's now consider the density function of the canonical statistic t.

$$f(u, v; \lambda\psi) = \frac{1}{C(\lambda, \psi)} \cdot g(u, v) \cdot \exp(\lambda^T u + \psi^T v)$$

where $\lambda^T u + \psi^T v = \theta^T t$. If we have partitioned the vector t and θ in the similar way then $\dim(u) = \dim(\lambda)$ and $\dim(v) = \dim(\psi)$.

Assume our aim is to test:

$$H_0 : \psi = \psi_0 \quad \text{against} \quad H_1 : \psi \neq \psi_0$$

We compute:

$$\psi^T v = (\psi - \psi_0)^T + \psi_0^T v = \tilde{\psi}^T + \psi_0^T v$$

Then we have:

$$\begin{aligned} f(u, v; \lambda\tilde{\psi}) &= \frac{1}{C(\lambda, \tilde{\psi} + \psi_0)} \cdot g(u, v) \cdot \exp(\lambda^T u + \tilde{\psi}^T v) \cdot \exp(\psi_0^T v) \\ &= \frac{1}{C(\lambda, \tilde{\psi} + \psi_0)} \cdot \tilde{g}(u, v) \cdot \exp(\lambda^T u + \tilde{\psi}^T v) \end{aligned} \tag{15}$$

Where $\tilde{g}(u, v) = g(u, v) \cdot \exp(\psi_0^T v)$ as ψ_0 is a vector of constants which are known values and not a parameter, hence we join it together with the structural function $g(u, v)$.

We still have an expo-family, the only new difference is the canonical parameter vector which is now $(\lambda, \tilde{\psi})$

$$t = \begin{pmatrix} u \\ v \end{pmatrix}; \quad \tilde{\theta} = \begin{pmatrix} \lambda \\ \tilde{\psi} \end{pmatrix}$$

$$H_0 : \psi = \psi_0 \iff H_0 : \tilde{\psi} = 0$$

4.2 Exact Test:

- We will now consider a model reduction $t = (u, v) \Rightarrow u$ represented in **mixed parametrization** given by the parameter vector (μ_v, ψ). In the case of the mixed presentation we have a very nice expression for the likelihood function which can be written as a product of two likelihoods

$$L(\mu_u, \psi; u, v) = L_1(\psi; v|u)L_2(\mu_u\psi; u)$$

where v is the canonical statistic that corresponds to the main parameter ψ, and the conditional density function is a function of ψ only. It has no element related to λ. Then the marginal density of u contributes only to the estimation of μ_u (since u is the MLE for μ_u) hence all information we have of u is only used for μ_u. However $v|u$ is the only relevant part for us.

- We will use the **conditional principle** for testing $H_0 : \psi = 0$, where the conditional density function $f(v|u; \psi)$ depends only on ψ and under $H_0 : \psi = 0$ then $f_0(v|u) = f(v|u; 0)$ is a fully specified object, we have no other parameter in the expression of this density function.

- It is also important to notice that we can compute $f_0(v|u)$

$$f(v|u) = \frac{f(vu)}{f(u)} = \text{under the null hypothesis} \quad = \frac{f_0(vu)}{f_0(u)}$$

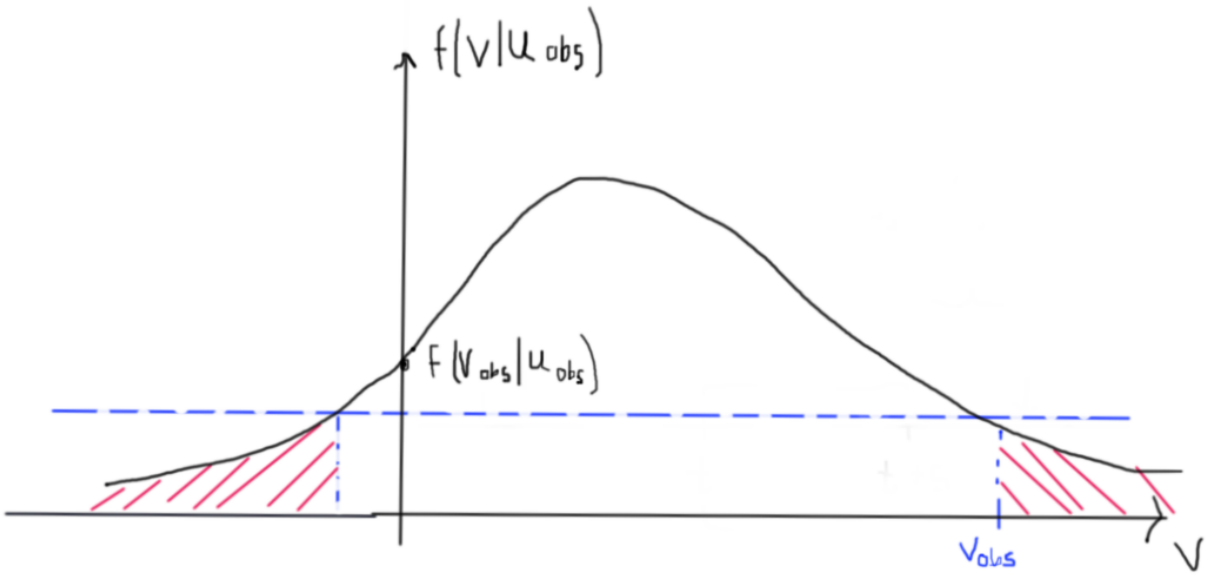
$$f(v|u) = \frac{c(\lambda, \psi = 0)^{-1} \cdot g(v, u) \cdot \exp(\lambda^T u + \overbrace{0^T}^{H_0 : \psi = 0} \cdot v)}{\int_v c(\lambda, \psi = 0)^{-1} \cdot g(v, u) \cdot \exp(\lambda^T u + 0^T \cdot v) \ dv} \tag{16}$$
$$= \frac{g(v, u)}{\int_v g(v, u) \ dv} = \frac{g(v, u)}{g_0(u)}$$

Where $g_0(u)$ is a function of u only, which is computed under the null hypothesis $\psi = 0$, hence the conditional density function is a structural function. We should be able to obtain a close form expression for this density function.

4.3 Testing procedure

Consider the data $y_1, ..., y_n \to t_{\text{obs}} = \begin{pmatrix} u_{\text{obs}} \\ v_{\text{obs}} \end{pmatrix}$ form which we compute the observed value of the canonical statistic t.

Question: Based on the conditional density we would like to answer how extreme is the observed value of the second part of the canonical statistic t, i.e, how extreme is v_{obs} We are not looking for how large/small is the value of v_{obs} because we have a two-sided test, hence by extreme we mean the values in both directions (lower and upper tail).
Answer: We have to draw the density function (pmf)



On the $x-axis$ we have the possible values of v while on the $y-axis$ we have the density of $v|u_{\text{obs}}$. Why do we want to analyze this object? Because we prefer to perform the conditional inference, as we do not need any knowledge of the nuisance parameter. We now assume we have an observed value v_{obs} having this value, we project it to the density function and then through this point we draw a parallel line to the $x-axis$, then we look for the second intersection of this parallel line and we will like to compute the red area, these areas are exactly the **p-value** of our test.

Then our **decision** is the following. If p-value $< \alpha \rightarrow H_0$ is rejected or H_1 is accepted (remember that by construction H_0 can not be "accepted", because H_0 always include type I error, hence we could do some mistake, whose probability will be included in H_0, because of the exact test this Type-I error probability will be close to α, in the case of H_1 we can say it both ways). If p-value $\geq \alpha \Rightarrow H_0$ cannot be rejected. In our formulation of the test it means that we have no support to use the bigger model. Here we are not doing it in the classical way as we would normally use a test statistic, here instead of the test statistic we determine the value of our density function $f(v_{\text{obs}}|u_{\text{obs}})$ then the p-value is nothing else than the probability that it is the area of the density function in some interval which consists in all points v_{obs} for which $f(v_{\text{obs}}|u_{\text{obs}})$ is no smaller than the density function of v at the point v_{obs}.

The p-value area can be represented by the following integral:

$$\int_{\{v:f(v|u_{\text{obs}})\leq f(v_{\text{obs}}|u_{\text{obs}})\}} f(v|u_{\text{obs}}) \quad dv$$

where $\rho = P(f(v|u_{\text{obs}}) \leq f(v_{\text{obs}}|u_{\text{obs}}))$ and reject if say $\rho < \alpha$. If v is discrete the integration is replaced by a summation.

It is under the null hypothesis that we have a smaller model, and then we look at the probability that we can observe some value of v more extreme than we already have, by extreme we mean those values of v for which $f(v|u_{\text{obs}}) \leq f(v_{\text{obs}}|u_{\text{obs}})$ holds true, Alternatively we can repeat the experiment several times and compute the average of these areas, which will be an approximate of the p-value. It is important to know that $f(v_{\text{obs}}|u_{\text{obs}})$ can be used not only in the uni-variate case, it can also be used in a multivariate case.

4.4 Chapter 5: Model reduction

Here, we only consider a two sided test, that under the null hypothesis we consider a single value 0, under H_0 we can reduce the dimension of the parameter space. This is why it's called the model reduction hypothesis.

$$H_0 : \psi = 0 \quad \text{against} \quad H_1 : \psi \neq 0$$

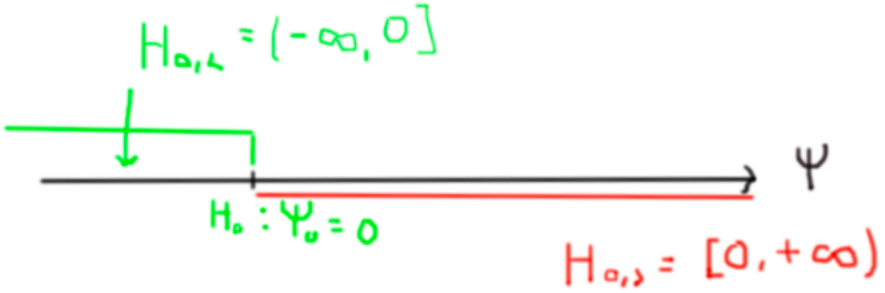
If we have a one-sided test, let's say that $\psi \in R$ (one-dimensional) and

$$H_0 : \psi \leq 0 \quad \text{against} \quad H_1 : \psi > 0$$

This is not a model reduction hypothesis, because under H_1 the dimension of the parameter space is 1.

$$H_0 : \psi \leq 0$$

Furthermore, generally we have, one side or two sided tests, so if we have a one sided test, we compute the p-value and divide it by 2. We can look at the theory of **multiple test** where we consider a set of the parameter space only,



the hypothesis $H_0 : \psi = 0$ can be presented as the intersection of two sets, we can consider $H_{0,<} = (-\infty, 0]$ and the set $H_{0,>} = [0, +\infty)$. i.e, $\{H_0 : \psi = 0\} = H_{0,<} \cap H_{0,>}$. Our

null hypothesis can be considered as the global hypothesis in two tests, (i) one sided test is smaller (ii) one sided test is larger. This relates to the **Bonferroni correction** in multiple test theory. (several tests performed at the same time) $H_{0,<}$ and $H_{0,>}$.

5 Lecture 12:

To read: Section 5.2, 5.4, and 5.5

5.1 Fisher's exact test in 2x2 tables

Depending on the experiment results we could have three possible models, related to the 2 x 2 models.

- Model 1: **Poisson model**
  We do not know the sample size before the experiment, e.g we would like to conduct a survey for a month, however we do not know how many people will participate in the study, meaning that n is an unknown quantity.

  $$y_{ij} \sim \text{Po}(\lambda_{ij})$$

  where y_{ij} are independent (not identically distributed) realizations of the Poisson distribution. Generally each Poisson distribution will have it's own parameter λ_{ij}

- Model 2: **Multinomial model**
  We fix the size of the experiment n (possibly caused by budget constraint). We can present t

  $$\mathbf{y} = \begin{pmatrix} y_{00} \\ y_{01} \\ y_{10} \\ y_{11} \end{pmatrix} \sim \text{Mult}\left(n, \pi = \begin{pmatrix} \pi_{00} \\ \pi_{01} \\ \pi_{10} \\ \pi_{11} \end{pmatrix}\right)$$

  where n how many "answers" we get in our experiment and the vector π which represent the probability of success in each group of observations.

- Model 3: **Binomial model**
  This is a more restricted model, which provides more information. We not only restrict the size of our sample n if not the rows size (e.g same number of female and male). i.e r_0, r_1 are fixed.

  $$y_{00} \sim \text{Bi}(r_0, \pi_0); \quad y_{01} \sim \text{Bi}(r_1, \pi_1)$$

  Also, as in the case of the Poisson distribution we need some assumption of independence which we will assume here, y_{00}, y_{10} are independent. We would now like to know the probability of "smoking" between woman and man. This probability is denoted by π_0, π_1, and probably we would perform some hypothesis testing to verify if these two probabilities are the same or not.

From a probabilistic point of view, these three models are completely different models, we have different distribution assumptions but now we would like to show that these three models are related and if we would like to do some hypothesis testing then it does not matter how we interpret the entries in our table. Under some considerations all three models are equivalent. And we can transform one model into another.

Example: Model 1 *Poisson model conditioned on the sample size n*

We will want to derive the following probability mass function:

$$\begin{aligned} f(\{y_{ij}\}|n;\{\lambda_{ij}\}\}) &= \frac{f(\{y_{ij},n;\{\lambda_{ij}\}\})}{f(n;\{\lambda_{ij}\}\})} = \frac{f(\{y_{ij};\{\lambda_{ij}\}\})}{f(n;\{\lambda_{ij}\}\})} \\ f(\{y_{ij},n;\{\lambda_{ij}\}\}) &= P(Y_{00}=y_{00},Y_{01}=y_{01},Y_{10}=y_{10},Y_{11}=y_{11},Y_{00}+Y_{01}+Y_{10}+Y_{11}=n) \\ &= P(Y_{00}=y_{00},Y_{01}=y_{01},Y_{10}=y_{10},Y_{11}=y_{11}) \\ &= f(\{y_{ij};\{\lambda_{ij}\}\}) \text{ the presence of } n \text{ is irrelevant.} \end{aligned} \tag{17}$$

As we have specified all four values $Y_{00}=y_{00},Y_{01}=y_{01},Y_{10}=y_{10},Y_{11}=y_{11}$, which uniquely specifies the last sum (as $n=Y_{00}+Y_{01}+Y_{10}+Y_{11}$; $n\sim \text{Po}(\lambda_{00}+\lambda_{01}+\lambda_{10}+\lambda_{11})$) thus, not relevant for the computation of this probability.

$$\begin{aligned} f(\{y_{ij}\}|n;\{\lambda_{ij}\}\}) &= \frac{\frac{\lambda_{00}^{y_{00}}e^{-\lambda_{00}}}{y_{00}!}\cdot\frac{\lambda_{01}^{y_{01}}e^{-\lambda_{01}}}{y_{01}!}\cdot\frac{\lambda_{10}^{y_{10}}e^{-\lambda_{10}}}{y_{10}!}\cdot\frac{\lambda_{11}^{y_{11}}e^{-\lambda_{11}}}{y_{11}!}}{\frac{(\lambda_{00}+\lambda_{01}+\lambda_{10}+\lambda_{11})^n e^{-(\lambda_{00}+\lambda_{01}+\lambda_{10}+\lambda_{11})}}{n!}} \\ &= \frac{n!}{y_{00}!y_{01}!y_{10}!y_{11}!}\cdot\Big(\frac{\lambda_{00}}{\lambda_{00}+\lambda_{01}+\lambda_{10}+\lambda_{11}}\Big)^{y_{00}}\cdot\Big(\frac{\lambda_{01}}{\lambda_{00}+\lambda_{01}+\lambda_{10}+\lambda_{11}}\Big)^{y_{01}} \\ &\cdot\Big(\frac{\lambda_{10}}{\lambda_{00}+\lambda_{01}+\lambda_{10}+\lambda_{11}}\Big)^{y_{10}}\cdot\Big(\frac{\lambda_{11}}{\lambda_{00}+\lambda_{01}+\lambda_{10}+\lambda_{11}}\Big)^{y_{11}} \\ &= \frac{n!}{y_{00}!y_{01}!y_{10}!y_{11}!}\cdot\pi_{00}^{y_{00}}\cdot\pi_{01}^{y_{01}}\cdot\pi_{10}^{y_{10}}\cdot\pi_{11}^{y_{11}} \end{aligned} \tag{18}$$

Which is just the pmf of a $\text{Mult}\left(n,\pi=\begin{pmatrix}\pi_{00}\\ \pi_{01}\\ \pi_{10}\\ \pi_{11}\end{pmatrix}\right)$ which is not surprising as in the multinomial model we assume n to be known. It does not matter if we consider the Poisson model, and make a condition on n, or if we consider the multinomial model an say that n is known. Both cases yield to a multinomial model.
Now let's assume that we have

$$x=\begin{pmatrix}x_1\\ \vdots\\ x_k\end{pmatrix}\sim \quad \text{Mult}\left(n,\pi=\begin{pmatrix}\pi_1\\ \vdots\\ \pi_k\end{pmatrix}\right)$$

Now, consider the following partition

$$x = \begin{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_{k_1} \end{pmatrix} \\ \begin{pmatrix} x_{k_1+1} \\ \vdots \\ x_{k_2} \end{pmatrix} \\ \vdots \\ \begin{pmatrix} x_{k_r+1} \\ \vdots \\ x_{k_r} \end{pmatrix} \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} Z_1 = \sum_{i=1}^{k_1} x_i \\ \vdots \\ Z_r = \sum_{i=x_{k_r+1}}^{k_r} x_i \end{pmatrix}$$

Where $\begin{pmatrix} x_1 \\ \vdots \\ x_{k_1} \end{pmatrix}$ is used to construct Z_1 and so on. Then $Z \sim \text{Mult}\left(n, \tilde{\pi} = \begin{pmatrix} \sum_{i=1}^{k_1} \pi_i \\ \vdots \\ \sum_{i=k_r+1}^{k_r} \pi_i \end{pmatrix}\right)$

We have the same results in case of the multinomial distribution.

Example: Model 2 *Multinomial model*

If we now consider the multinomial distribution as a model for the contingency table, then it is the same as if we assumed we know r_0 and r_1 precisely and to assume that y_{00} and y_{10} have binomial distribution. We have to do the same computations as before, show the equivalence between the conditional Poisson model and the multinomial model. considering

$$\begin{pmatrix} r_0 = y_{00} + y_{01} \\ r_1 = y_{10} + y_{11} \end{pmatrix} \sim \text{Mult}\left(n, \tilde{\pi} = \begin{pmatrix} \pi_{00} + \pi_{01} \\ \pi_{10} + \pi_{11} \end{pmatrix}\right)$$

$$
\begin{aligned}
f(\{y_{ij}\}|r_0, r_1; \{\pi_{ij}\}) &= \frac{f(\{y_{ij}\}r_0, r_1; \{\pi_{ij}\})}{f(r_0, r_1; \{\pi_{ij}\})} = \frac{f(\{y_{ij}\}; \{\pi_{ij}\})}{f(r_0, r_1; \{\pi_{ij}\})} \\
&= \frac{\frac{n!}{y_{00}!y_{01}!y_{10}!y_{11}!} \ \pi_{00}^{y_{00}} \ \pi_{01}^{y_{01}} \ \pi_{10}^{y_{10}} \ \pi_{11}^{y_{11}}}{\frac{n!}{r_0!r_1!} \quad (\pi_{00}+\pi_{01})^{r_0} \ (\pi_{10}+\pi_{11})^{r_1}} \\
&= \frac{r_0!}{y_{00}!y_{01}!} \cdot \frac{r_1!}{y_{10}!y_{11}!} \cdot \Big(\frac{\pi_{00}}{\pi_{00}+\pi_{01}}\Big)^{y_{00}} \cdot \Big(\frac{\pi_{01}}{\pi_{00}+\pi_{01}}\Big)^{y_{01}} \\
&\cdot \Big(\frac{\pi_{10}}{\pi_{10}+\pi_{11}}\Big)^{y_{10}} \cdot \Big(\frac{\pi_{11}}{\pi_{10}+\pi_{11}}\Big)^{y_{11}} \\
&= f(y_{00}|r_0, r_1; \pi_0) \cdot f(y_{10}|r_0, r_1; \pi_1) \\
&= f(y_{00}|r_0; \pi_0) \cdot f(y_{10}|r_1; \pi_1)
\end{aligned} \tag{19}
$$

Where$f(y_{00}|r_0; \pi_0)$ is the p.m.f of Bi($r_0\pi_0$), and $f(y_{10}|r_1; \pi_1)$ is the p.m.f of Bi($r_1\pi_1$). We have proved that y_{00} and y_{10} are independent random variables and conditioned to r_0 and r_1 follow a binomial distribution.

5.2 Test Theory in 2 x 2 theory:

- Model 1: **Poisson model**
  We want to test the hypothesis that the intensity λ_{ij} can be factorized and presented as $\alpha_i \cdot \beta_j$ which is also known as "multiplicativity". We can reduce the parameter space if we are under the null hypothesis. We want to test

$$H_{p,0} : \lambda_{ij} = \alpha_i \cdot \beta_j$$

- Model 2: **Multinomial model**
  We note for the multinomial model that we need to define another type of parameter which are π_{ij} and we will make this by using λ_{ij}. These quantities should be computed in the following way

$$
\begin{aligned}
\lambda_{00} + \lambda_{01} + \lambda_{10} + \lambda_{11} &= \alpha_0\beta_0 + \alpha_0\beta_1 + \alpha_1\beta_0 + \alpha_1\beta_1 \\
&= \alpha_0(\beta_0 + \beta_1) + \alpha_1(\beta_0 + \beta_1) \\
&= (\alpha_0 + \alpha_1)(\beta_0 + \beta_1)
\end{aligned} \tag{20}
$$

$$\Rightarrow \pi_{ij} = \frac{\alpha_i\beta_j}{(\alpha_0 + \alpha_1)(\beta_0 + \beta_1)} = \frac{\alpha_i}{\alpha_0 + \alpha_1} \cdot \frac{\beta_j}{\beta_0 + \beta_1} = \tilde{\alpha_i} \cdot \tilde{\beta_j}$$

Which can be represented in the contingency table:

	0	1	
1	π_{00}	π_{01}	$\pi_{00}+\pi_{01}=\tilde{\alpha_0}$
0	π_{10}	π_{11}	$\pi_{10}+\pi_{11}=\tilde{\alpha_1}$
	$\pi_{00}+\pi_{10}=\tilde{\beta_0}$	$\pi_{01}+\pi_{11}=\tilde{\beta_1}$	1

Hence $H_{P,0}$ is equivalent to

$$H_{M,0}: \pi_{00} = \tilde{\alpha_0} \cdot \tilde{\beta_0} - \text{independence}$$

- Model 3: **Binomial model**
  We would like to check what would happen to the hypothesis that we have for independence in the case of multinomial model of the hypothesis will be transformed in the case of the binomial model in order to do that we only have to consider the probability which he have, π_0, π_1.

  We get
  $$\pi_0 = \frac{\pi_{00}}{\pi_{00}+\pi_{01}} = \frac{\tilde{\alpha_0}\tilde{\beta_0}}{\tilde{\alpha_0}\tilde{\beta_0}+\tilde{\alpha_0}\tilde{\beta_1}} = \frac{\tilde{\beta_0}}{\tilde{\beta_0}+\tilde{\beta_1}}$$
  and
  $$\pi_1 = \frac{\pi_{10}}{\pi_{10}+\pi_{11}} = \frac{\tilde{\alpha_1}\tilde{\beta_0}}{\tilde{\alpha_1}\tilde{\beta_0}+\tilde{\alpha_1}\tilde{\beta_1}} = \frac{\tilde{\beta_0}}{\tilde{\beta_0}+\tilde{\beta_1}} = \pi_0$$
  We get these computations under the null hypothesis, we have formulated on the multinomial model $H_{M,0}$. Hence $H_{M,0}$ can be present as $\underbrace{H_{B,0}: \pi_0 = \pi_1}_{\text{Hypothesis of homogeneity}}$ under the binomial model.

  In conclusion, if we make the test for multiplicativity in the case of the Poisson model, is the same as if we test independence on the multinomial model or if we do a test for homogeneity in the case of the binomial model. All three hypothesis are the same.

Another important remark is that all three hypothesis are read in a complete different way, but nevertheless we can present them in the same way and redefine the variable ψ. Define

$$\begin{aligned}
\psi &= \log\left(\frac{\lambda_{00}\lambda_{11}}{\lambda_{01}\lambda_{10}}\right) \overset{H_{P,0}}{=} \log\left(\frac{\alpha_0\beta_0\alpha_1\beta_1}{\alpha_0\beta_1\alpha_1\beta_0}\right) = 0 \\
&= \log\left(\frac{\pi_{00}\pi_{11}}{\pi_{01}\pi_{10}}\right) \overset{H_{M,0}}{=} \log\left(\frac{\tilde{\alpha_0}\tilde{\beta_0}\tilde{\alpha_1}\tilde{\beta_1}}{\tilde{\alpha_0}\tilde{\beta_1}\tilde{\alpha_1}\tilde{\beta_0}}\right) = 0 \\
&= \log\left(\frac{\pi_0(1-\pi_1)}{(1-\pi_0)\pi_1}\right) = \text{logit}(\pi_0) - \text{logit}(\pi_1) \overset{H_{B,0}}{=} \log\left(\frac{\pi_0(1-\pi_0)}{(1-\pi_0)\pi_0}\right) = 0
\end{aligned} \tag{21}$$

As a result $H_{P,0}, H_{M,0}$ and $H_{B,0}$ can be formulated as $H_0: \psi = 0$

5.3 Exact Fisher test

We would like to start with two binomials and rewrite our model in the case of one component of the canonical parameter is just equal to ψ or to the difference of the logit functions. This should be possible to do because we know that for each binomial distribution the canonical parameter is the logit of the probability of success. We should be able to parametrized the joint model by using this difference.

$$\begin{aligned}
f(y_{00}, y_{10}; \pi_0, \pi_1) &= f(y_{00}; \pi_0) f(y_{10}; \pi_1) \\
&= \frac{r_0!}{y_{00}!(r_0 - y_{00})!} \cdot \frac{r_1!}{y_{10}!(r_1 - y_{10})!} \quad \pi_0^{y_{00}} (1-\pi_0)^{r_0 - y_{00}} \cdot \quad \pi_1^{y_{10}} (1-\pi_1)^{r_1 - y_{10}} \\
&= \frac{r_0!}{y_{00}!(r_0 - y_{00})!} \cdot \frac{r_1!}{y_{10}!(r_1 - y_{10})!} \quad (1-\pi_0)^{r_0} \exp\left\{ y_{00} \log\left(\frac{\pi_0}{1-\pi_0} \right) \right\} \\
&\cdot (1-\pi_1)^{r_1} \exp\left\{ y_{10} \log\left(\frac{\pi_1}{1-\pi_1} \right) \right\} \\
&= \frac{r_0!}{y_{00}!(r_0 - y_{00})!} \cdot \frac{r_1!}{y_{10}!(r_1 - y_{10})!} (1-\pi_0)^{r_0} \exp\{ y_{00} \text{logit}(\pi_0) \} \\
&\cdot (1-\pi_1)^{r_1} \exp\{ y_{10} \text{logit}(\pi_1) \} \\
&= (1-\pi_0)^{r_0} (1-\pi_1)^{r_1} \frac{r_0!}{y_{00}!(r_0 - y_{00})!} \cdot \frac{r_1!}{y_{10}!(r_1 - y_{10})!} \\
&\cdot \exp\{ \text{logit}(\pi_0) y_{00} + \text{logit}(\pi_1) y_{10} \}
\end{aligned} \tag{22}$$

We have already presented the two independent binomials as the p.m.f and member of the expo family. However, we have to remind that our aim is to get the canonical statistic whose one element is the value of ψ which should be $\text{logit}(\pi_0) - \text{logit}(\pi_1)$. At the moment we have no such difference as an argument of the exponential function. We would like to do some transformation in order to get this equality. The most easiest way would be to add and subtract $\text{logit}(\pi_0) - \text{logit}(\pi_1)$. Hence, the last expression will become:

$$\begin{aligned}
&= \exp\{ (\text{logit}(\pi_0) - \text{logit}(\pi_1) + \text{logit}(\pi_1)) y_{00} + (\text{logit}(\pi_1)) y_{10} \} \\
&= \exp\{ (\text{logit}(\pi_0) - \text{logit}(\pi_1)) y_{00} + (\text{logit}(\pi_1)) y_{00} + (\text{logit}(\pi_1)) y_{10} \} \\
&= \exp\{ \underbrace{(\text{logit}(\pi_0) - \text{logit}(\pi_1))}_{\psi} \underbrace{y_{00}}_{v} + \underbrace{\text{logit}(\pi_1)}_{\theta_u} \underbrace{(y_{00} + y_{10})}_{u} \}
\end{aligned} \tag{23}$$

$$t = \begin{pmatrix} u = & y_{00} + y_{10} \\ v = & y_{00} \end{pmatrix}; \quad \theta = \begin{pmatrix} \theta_u = & \text{logit}(\pi_1) \\ \psi = & \text{logit}(\pi_0) - \text{logit}(\pi_1) \end{pmatrix}$$

Next, we need to derive $f_{v|u}$

$$f_0(v|u) = \frac{f(v,u)}{f_0(u)} \tag{24}$$

Where

$$f(u,v) = |\text{Jacobian}| \cdot f(y_{00}(u,v), y_{10}(u,v); \pi_0, \pi_1)$$

If we do a linear transformation where

$$y_{00} = v, \Rightarrow \quad y_{10} = u - v \Rightarrow |\text{Jacobian}| = 1$$

In the case of the joint p.m.f we will have:

$$f(u,v) = (1-\pi_0)^{r_0}(1-\pi_1)^{r_1}\frac{r_0!}{v!(r_0-v)!}\cdot\frac{r_1!}{(u-v)!(r_1-u+v)!}\cdot\exp\{\psi\cdot v + \text{logit}(\pi_1)u\}$$

Under H_0: $(\pi_1 = \pi_0)$

$$f_0(u,v) = (1-\pi_0)^{r_0+r_1}\frac{r_0!}{v!(r_0-v)!}\cdot\frac{r_1!}{(u-v)!(r_1-u+v)!}\cdot\exp\{\text{logit}(\pi_1)u\}$$

Computation of $f_0(u)$:

$$\begin{gathered} y_{00} \sim Bi(r_0,\pi_0);\quad y_{10} \sim Bi(r_1,\pi_1),\quad y_{00}, y_{01} - \text{independent} \\ \Rightarrow u = y_{00} + y_{10} \sim Bi(r_0+r_1,\pi_1) \\ f_0(u) = \frac{(r_0+r_1)!}{u!(r_0+r_1-u)!}\cdot(1-\pi_1)^{r_0+r_1}\cdot\exp\{\text{logit}(\pi_1)u\} \end{gathered}$$

Computation of $f_0(v|u)$:

$$\begin{aligned} f_0(v|u) &\overset{H_0(\psi=0)}{=} \frac{(1-\pi_0)^{r_0+r_1}\frac{r_0!}{v!(r_0-v)!}\cdot\frac{r_1!}{(u-v)!(r_1-u+v)!}\cdot\exp\{\text{logit}(\pi_1)u\}}{\frac{(r_0+r_1)!}{u!(r_0+r_1-u)!}\cdot(1-\pi_1)^{r_0+r_1}\cdot\exp\{\text{logit}(\pi_1)u\}} \\ &= \frac{\frac{r_0!}{v!(r_0-v)!}\cdot\frac{r_1!}{(u-v)!(r_1-u+v)!}}{\frac{(r_0+r_1)!}{u!(r_0+r_1-u)!}} \end{aligned} \tag{25}$$

Which turns out to be the p.m.f of the hyper-geometric distribution.

$$f_0(v|u) = f_0(y_{00}|r_0,s_0,n) = \frac{\binom{r_0}{y_{00}}\binom{r_1}{y_{10}}}{\binom{n}{s_0}} = \frac{\binom{r_0}{y_{00}}\binom{r_1}{s_0-y_{00}}}{\binom{n}{s_0}}$$

Consider that whenever n is large, the computation of $\binom{n}{s_0}$ could be a difficult task.
Idea: find an approximation for $f_0(v|u)$

6 Lecture 13:

To read: Section 5.4-5.6

6.1 Asymptotic equivalent tests:

Consider the likelihood

$$L(\theta) = f((u,v);(\theta_u,\psi))$$
$$\hat{\theta} = \text{argmax} \;\; L(\theta) \to L(\hat{\theta})$$
$$\hat{\theta}_0 = \text{argmax} \;\; L(\theta_u, \psi = 0) \to L(\hat{\theta}_0)$$

If $H_0; \psi = 0$ is true,

$$\hat{\theta} \approx \hat{\theta}_0 \to L(\hat{\theta}) \approx L(\hat{\theta}_0)$$

- **Approach 1:** Likelihood ratio test:

$$W = 2\log\frac{L(\hat{\theta})}{L(\hat{\theta}_0)} \overset{H_0}{\to} \chi^2_{\dim\psi}$$

- **Approach 2:** Quadratic form:

$$\hat{\theta}_0 \; x \; \hat{\theta} \Rightarrow \hat{\theta}_0 - \hat{\theta} \approx 0$$

$$\hat{\theta}_0 \;-\; \hat{\theta} \overset{H_0}{\approx} \mathcal{N}(0, I(\hat{\theta}_0)^{-1})$$

$$\Rightarrow (\hat{\theta}_0 \;-\; \hat{\theta})^T I(\hat{\theta}_0)(\hat{\theta}_0 \;-\; \hat{\theta}) \overset{H_0}{\to} \chi^2_{\dim\psi}$$

6.2 Score test

From

$$2\log\frac{L(\hat{\theta})}{L(\hat{\theta}_0)} = -2\log\frac{L(\hat{\theta}_0)}{L(\hat{\theta})}$$

Now using Taylor expression:

$$\log L(\theta) \approx \log L(\hat{\theta}) + U(\hat{\theta})(\theta - \hat{\theta}) - \frac{1}{2}((\theta-\hat{\theta})^T I(\hat{\theta})(\theta-\hat{\theta})) + O((\theta-\hat{\theta})^T I(\hat{\theta})(\theta-\hat{\theta}))$$

considering $U(\hat{\theta}) = 0; \quad \theta \to \hat{\theta}_0$

$$\log L(\theta_0) \approx \log L(\hat{\theta}) - \frac{1}{2}((\hat{\theta}_0-\hat{\theta})^T I(\hat{\theta})(\hat{\theta}_0-\hat{\theta}))$$

$$W \approx (\hat{\theta}_0 - \hat{\theta})^T \underbrace{I(\hat{\theta})}_{\text{since } \hat{\theta}_0 \overset{H_0}{\approx} \hat{\theta}} (\hat{\theta}_0 - \hat{\theta})$$

Consider

$$t = \begin{pmatrix} u \\ v \end{pmatrix}; \quad \theta = \begin{pmatrix} \theta_u \\ \theta_v \;\; = \psi \end{pmatrix}$$

Where $\dim t = k, \;\; \dim u = p, \;\; \dim v = q$

Hypothesis:

$$H_0 : \psi = 0 \quad \text{against} \quad H_1 : \psi \neq 0$$

Under H_0 $\theta = \begin{pmatrix} \theta_u \\ 0 \end{pmatrix}$

Statistic of the Score test:

$$W_u = u(\hat{\theta}_0)^T I(\hat{\theta}_0)^{-1} u(\hat{\theta}_0)$$

We get $u(\hat{\theta}_0) = t - \mu_t(\hat{\theta}_0) = \begin{pmatrix} u - \mu_u(\hat{\theta}_0) \\ v - \mu_v(\hat{\theta}_0) \end{pmatrix} = \begin{pmatrix} 0 \\ v - \mu_v(\hat{\theta}_0) \end{pmatrix}$

Since using conditional inference

$$L(t; (\lambda, \psi)) = L(\psi; v|u) L(\psi, \mu; u)$$

u is the MLE of μ_u, thus, $u = \mu_u(\hat{\theta}_0)$ and does not depend on the value of ψ Using notation:

$$I(\hat{\theta}_0) = V_t(\hat{\theta}_0) \to I(\hat{\theta}_0)^{-1} = V_t^{-1}(\hat{\theta}_0) = \begin{pmatrix} V^{uu}(\hat{\theta}_0) & V^{uv}(\hat{\theta}_0) \\ V^{vu}(\hat{\theta}_0) & V^{vv}(\hat{\theta}_0) \end{pmatrix} = \begin{pmatrix} A_{uu} & A_{uv} \\ A_{vu} & A_{vv} \end{pmatrix}$$

Hence,

$$W_u = (v - \mu_v(\hat{\theta}_0))^T V^{vv}(\hat{\theta}_0)(v - \mu_v(\hat{\theta}_0))$$

Under H_0: $W_u \sim \chi^2_q, q = \dim(v)$

Decision: Reject $H_0 \iff W_u > \chi^2_{q,1-\alpha}$

Large sample approximation of the exact test:

Proposition 5.1: *The principle of the 'exact test' of* $H_0 : \psi = 0$
Based on this proposition we can prove that as $n \to \infty$ the exact test (based on $f_u(v|u)$) is equivalent to the score test.

Proof:

For the exact test we had that

$$f_0(v|u) = \frac{g(u, v)}{g_0(u)}$$

Idea: Approximate $g(u, v)$ and $g_0(u)$

Approximating $g_0(u)$:

$$g_0(u) \approx (2\pi)^{-p/2} \frac{C(\hat{\theta}_0)}{\sqrt{\det(V_u(\hat{\theta}_0))}} \exp(-\hat{\theta}_0^T u) \quad \text{Saddle point approximation}$$

To approximate $g(u,v)$ we follow the proof on the proposition for the saddle point approximation where:

$$f(t,\theta) = \frac{C(\theta_0)}{C(\theta)} f(t,\theta_0) e^{-(\theta-\theta_0)^T t} \approx \frac{1}{(2\pi)^{r/2}} \frac{1}{\sqrt{\det V_t(\theta)}} \exp\Big\{ -\frac{1}{2}(t-\mu_t(\theta))^T V_t^{-1}(t-\mu_t(\theta))\Big\}$$

Previously: $\theta \to \hat{\theta}$
Now: $\theta \to \hat{\theta}_0$ Where $\hat{\theta}_0$ is the MLE for θ under $H_0 : \psi = 0$

$$f(t,\hat{\theta}_0) = \frac{C(\theta_0)}{C(\hat{\theta}_0)} f(t,\theta_0) e^{-(\hat{\theta}_0-\theta_0)^T t} \approx \frac{1}{(2\pi)^{r/2}} \frac{1}{\sqrt{\det V_t(\hat{\theta}_0)}} \exp\Big\{ -\frac{1}{2}(t-\mu_t(\hat{\theta}_0))^T V_t^{-1}(t-\mu_t(\hat{\theta}_0))\Big\}$$

$$f(t,\theta_0) \approx \frac{C(\hat{\theta}_0)}{C(\theta_0)} e^{-\hat{\theta}_0^T t} \frac{(2\pi)^{-r/2}}{\sqrt{\det V_t(\hat{\theta}_0)}} \exp\Big\{ -\frac{1}{2}(t-\mu_t(\hat{\theta}_0))^T V_t^{-1}(t-\mu_t(\hat{\theta}_0))\Big\} e^{-\theta_0^T t} = \frac{1}{C(\theta_0)} g(u,v) e^{-\theta_0^T t} \tag{26}$$

Approximating $g_0(u,v)$:

$$g(t) = g(u,v) \approx \frac{1}{(2\pi)^{r/2}} \frac{1}{\sqrt{\det V_t(\hat{\theta}_0)}} C(\hat{\theta}_0) e^{-\hat{\theta}_0^T t} \exp\Big\{ -\frac{1}{2}(t-\mu_t(\hat{\theta}_0))^T V_t^{-1}(t-\mu_t(\hat{\theta}_0))\Big\}$$

In here $\hat{\theta}_0 = (\hat{\theta}_u, \psi) = (\hat{\theta}_u, 0)$
Score vector: $U(\theta) = t = \mu_t(\theta) \Rightarrow U(\hat{\theta}_0) = \begin{pmatrix} 0 \\ v - \mu_v(\hat{\theta}_0) \end{pmatrix}$

$$\begin{aligned}(t-\mu_t(\hat{\theta}_0))^T V_t^{-1}(t-\mu_t(\hat{\theta}_0)) &= \begin{pmatrix} 0 \\ v-\mu_v(\hat{\theta}_0) \end{pmatrix} V_t(\hat{\theta}_0)^{-1} \begin{pmatrix} 0 \\ v-\mu_v(\hat{\theta}_0) \end{pmatrix} \\ &= (v-\mu_v(\hat{\theta}_0))^T (V_t(\theta_0))^{vv} (v-\mu_v(\hat{\theta}_0))\end{aligned} \tag{27}$$

Sylvester's determinant theorem:
Knowing that

$$V_t(\hat{\theta}_0) = \begin{pmatrix} V_{uu}(\hat{\theta}_0) & V_{uv}(\hat{\theta}_0) \\ V_{vu}(\hat{\theta}_0) & V_{vv}(\hat{\theta}_0) \end{pmatrix} = \begin{pmatrix} A^{uu} & A^{uv} \\ A^{vu} & A^{vv} \end{pmatrix}$$

$$\det V_t(\hat{\theta}_0) = \det(V_{uu}(\hat{\theta}_0)) \cdot \det(\underbrace{V_{vv}(\hat{\theta}_0) - V_{vu}(\hat{\theta}_0) - V_{vv}(\hat{\theta}_0)^{-1} \cdot V_{uv}(\hat{\theta}_0)}_{(V^{vv}(\hat{\theta}_0))^{-1}})$$

Hence,

$$
\begin{aligned}
f_0(v|u) &\approx \frac{g_0(v,u)}{g_0(u)} \\
&= \frac{(2\pi)^{-r/2}\frac{1}{\sqrt{(\det V_t(\hat{\theta}_0))}}C(\hat{\theta}_0)e^{-\hat{\theta}_0^T u}\exp\Big\{-\tfrac{1}{2}(v-\mu_v(\hat{\theta}_0))^T V_t^{-1}(\hat{\theta}_0)(v-\mu_v(\hat{\theta}_0))\Big\}}{(2\pi)^{-p/2}\frac{C(\hat{\theta}_0)}{\sqrt{\det(V_u(\hat{\theta}_0))}}\exp(-\hat{\theta}_0^T u)} \\
&= (2\pi)^{-(r-p)/2}\frac{1}{\sqrt{\frac{\det(V_t(\hat{\theta}_0))}{\det(V_u(\hat{\theta}_0))}}}\exp\Big\{-\frac{1}{2}(v-\mu_v(\hat{\theta}_0))^T(V_t(\hat{\theta}_0))^{vv}(v-\mu_v(\hat{\theta}_0))\Big\} \\
&= (2\pi)^{-(r-p)/2}\frac{1}{\sqrt{\frac{\det(V_{uu}(\hat{\theta}_0))\det(V^{vv}(\hat{\theta}_0))^{-1}}{\det(V_u(\hat{\theta}_0))}}}\exp\Big\{-\frac{1}{2}(v-\mu_v(\hat{\theta}_0))^T(V_t(\hat{\theta}_0))^{vv}(v-\mu_v(\hat{\theta}_0))\Big\} \\
&= \frac{1}{(2\pi)^{q/2}}\sqrt{\det(V^{vv}(\hat{\theta}_0))}\exp\Big\{-\frac{1}{2}\underbrace{(v-\mu_v(\hat{\theta}_0))^T(V_t(\hat{\theta}_0))^{vv}(v-\mu_v(\hat{\theta}_0))}_{W_u}\Big\}
\end{aligned}
\tag{28}
$$

$$\Rightarrow v|u \approx \mathcal{N}(\mu_v(\hat{\theta}_0), (V(\hat{\theta}_0)^{vv})^{-1})$$

Transformation to χ^2 distribution:

$$
\begin{aligned}
W_u &= (v-\mu_v(\hat{\theta}_0))^T\underbrace{((V_t(\hat{\theta}_0))^{vv})^{-1})^{-1}}_{I_t(\hat{\theta}_0)_{vv}}(v-\mu_v(\hat{\theta}_0)) \overset{H_0}{\sim} \chi^2_q \\
&\text{as } n\to\infty \\
&= U(\hat{\theta}_0)^T I_t(\hat{\theta}_0)U(\hat{\theta}_0)
\end{aligned}
\tag{29}
$$

Where $I_t(\hat{\theta}_0)_{vv}$ is the Fisher information corresponding to the v part.

Example 5.4 *Score test for 2x2 Table*

y_{00}	y_{01}	r_0
y_{10}	y_{11}	r_1
s_0	s_1	n

We assume a Binomial model where: (classical example for a 2x2 table)

$$y_{00}\sim Bi(r_0,\pi_0);\quad y_{10}\sim Bi(r_1,\pi_1)\quad y_{00},y_{10}-\text{independent}$$

Where π_0,π_1 denote the probability of success. We will test :

$$H_0:\pi_0=\pi_1\quad\text{against}\quad H_1:\pi_0\neq\pi_1$$

We will start with an **Exponential Family** model, we know that the joint probability mass function of y_{00}, y_{10} could be written on the following way:

$$\begin{aligned} f(y_{00}, y_{10}; \pi_0\pi_1) &= \binom{r_0}{y_{00}}(1-\pi_0)^{r_0}\exp\{y_{00}\text{logit}(\pi_0)\}\binom{r_1}{y_{10}}(1-\pi_1)^{r_1}\exp\{y_{10}\text{logit}(\pi_1)\} \\ &= \binom{r_0}{y_{00}}\binom{r_1}{y_{10}}(1-\pi_0)^{r_0}(1-\pi_1)^{r_1}\exp\{y_{00}\text{logit}(\pi_0) + y_{10}\text{logit}(\pi_1)\} \end{aligned}$$

$$= \binom{r_0}{y_{00}}\binom{r_1}{y_{10}}(1-\pi_0)^{r_0}(1-\pi_1)^{r_1}\exp\{\underbrace{y_{00}}_{\text{v}}\underbrace{(\text{logit}(\pi_0) - \text{logit}(\pi_1))}_{\psi} + \underbrace{(y_{00}+y_{10})}_{\text{u}}\underbrace{\text{logit}(\pi_1)}_{\theta_u}\}$$

The last line of the previous equation tells us how we have to define the quantities of the components of the canonical statistic t (we have two components) it also tells us how we can define the parameters of our model with respect to the definition of the canonical statistic t.

$$t = \begin{pmatrix} u = & y_{00} + y_{10} = s_0 \\ v = & y_{00} \end{pmatrix}; \quad \theta = \begin{pmatrix} \theta_u = & \text{logit}(\pi_1) \\ \psi = & \text{logit}(\pi_0) - \text{logit}(\pi_1) \end{pmatrix}$$

Finally,

$$H_0 : \pi_0 = \pi_1 \quad \text{is equivalent} \quad H_0 : \psi = 0$$

Our aim here is the construction of the **score test statistic**. The choice of this test is not obvious, we could have chosen the deviance test, quadratic form test, or any other test. We will use this test as we have previously derived the equivalent which is the exact test and the t-statistic which is based on the score test, but we are not restricted to apply this test to the given problem.

$$W_u = (v - \mu_v(\hat{\theta}_0))^2 \; V_t^{vv}(\hat{\theta}_0)$$

We square the difference as we have an uni-variate quantity, furthermore V_t^{vv} corresponds to a single component of the inverse Fisher information matrix, it is a function of parameters which should be estimated under the null hypothesis.
Let's now consider

$$\hat{\theta}_0 = \begin{pmatrix} \hat{\theta}_u = \text{logit}(\pi_1) \\ 0 \end{pmatrix} \quad \text{Under the null hypothesis}$$

In our example $\hat{\theta}_u$ is a function of the probability of success π_1 only. We have to estimate the probability of success under the null hypothesis, considering $\pi_1 = \pi_0 = \pi$ (we use π for practical reasons)

Next step: MLE for π:

We would like to use all the information which is relevant for the estimation of π which we have in our sample, in the case of the two binomial distribution we will start by constructing the MLE and then estimate π

$$\hat{\pi} = \frac{y_{00} + y_{10}}{n} = \frac{s_0}{n}$$

We now compute the estimate $\Rightarrow \mu_v = E[v] = E[y_{00}] = r_0\pi \Rightarrow \hat{\mu}_v = \mu_v(\hat{\theta}_0) = r_0\hat{\pi} = r_0\frac{y_{00}+y_{10}}{n} = r_0\frac{s_0}{n}$.

Furthermore, the variance matrix is expressed as:

$$V_t = \text{Var}\begin{pmatrix} u \\ v \end{pmatrix}\begin{pmatrix} \text{Var}(u) & \text{Cov}(u,v) \\ \text{Cov}(u,v) & \text{Var}(v) \end{pmatrix} = \begin{pmatrix} \text{Var}(y_{00}+y_{10}) & \text{Cov}(y_{00}, y_{00}+y_{10}) \\ \text{Cov}(y_{00}, y_{10}+y_{00}) & \text{Var}(y_{00}) \end{pmatrix}$$

Where

$$v = y_{00} \overset{H_0}{\sim} Bi(r_0, \pi) \rightarrow \text{Var}(v) = r_0\pi(1-\pi)$$
$$u = y_{00} + y_{10} \overset{H_0}{\sim} Bi(n = r_0 + r_1, \pi) \rightarrow \text{Var}(u) = n\pi(1-\pi)$$

We obtain the covariance as:

$$\text{Cov}(u,v) = \text{Cov}(y_{00}, y_{00}+y_{10}) = \text{Cov}(y_{00}, y_{00}) + \underbrace{\text{Cov}(y_{00}+y_{10})}_{0 \text{ by independence}} = \text{Var}(y_{00}) = r_0\pi_0(1-\pi_0)$$

$$V_t = \begin{pmatrix} n\pi_0(1-\pi) & r_0\pi(1-\pi) \\ r_0\pi(1-\pi) & r_0\pi(1-\pi) \end{pmatrix} = \pi(1-\pi)\begin{pmatrix} n & r_0 \\ r_0 & r_0 \end{pmatrix}$$

$$I_t = (V_t)^{-1} = \frac{1}{\pi(1-\pi)}\begin{pmatrix} n & r_0 \\ r_0 & r_0 \end{pmatrix}^{-1} = \frac{1}{\pi(1-\pi)}\frac{1}{nr_0 - r_0^2}\begin{pmatrix} r_0 & -r_0 \\ -r_0 & n \end{pmatrix}$$

$$I_t(\hat{\theta}_0) = (V_t(\hat{\theta}_0))^{-1} = \frac{1}{\hat{\pi}(1-\hat{\pi})}\frac{1}{r_0(n-r_0)}\begin{pmatrix} r_0 & -r_0 \\ -r_0 & n \end{pmatrix} = \frac{1}{\hat{\pi}(1-\hat{\pi})}\frac{1}{r_0 s_1}\begin{pmatrix} r_0 & -r_0 \\ -r_0 & n \end{pmatrix}$$

We are only interested on the element

$$V_t^{vv}(\hat{\theta}_0) = \frac{n}{\hat{\pi}(1-\hat{\pi})r_0 s_1} = \frac{(s_0 s_1/n^2)^{-1}n}{r_0 r_1} = \frac{n^3}{r_0 r_1 s_0 s_1}$$

Score statistic W_u:

$$U(\hat{\theta}_0) = \begin{pmatrix} 0 \\ v - \mu_v(\hat{\theta}_0) \end{pmatrix} = \begin{pmatrix} 0 \\ y_{00} - r_0 \frac{s_0}{n} \end{pmatrix}$$

$$\begin{aligned} W_u &= (y_{00} - r_0 \frac{s_0}{n}) \frac{n^3}{r_0 r_1 s_0 s_1} (y_{00} - r_0 \frac{s_0}{n}) \\ &= (y_{00} - r_0 \frac{s_0}{n})^2 \frac{n^3}{r_0 r_1 s_0 s_1} \overset{H_0}{\sim} \chi^2_1 \end{aligned} \tag{30}$$

Null Distribution: (W_u under H_0)

$$W_u \overset{H_0}{\sim} \chi^2_1$$

Decision: Reject H_0 i.f.f $W_u > \chi^2_{1;1-\alpha}$ or p-value $= 1 - F_{\chi^2_1}(W_u) < \alpha \rightarrow$ Reject H_0

Another Approach: Poisson model

Assume, $y_{ij} \sim Po(\lambda_{ij})$, $\quad y_{ij}$_independent Using this assumption we can also derive a test statistic for our null hypothesis. Our null hypothesis in case of the Poisson model is the following (multiplicative hypothesis):

$$H_0 : \lambda_{ij} = \alpha_i \beta_j$$

Under the assumption of independent Poisson distributed random variables, we know that

$$E[y_{ij}] = \lambda_{ij}, \quad \text{Var}(y_{ij}) = \lambda_{ij}, \quad \hat{\lambda}_{ij} = \frac{r_i s_j}{n} \quad \text{Since} \quad v(y) = \begin{pmatrix} y_{00} \\ y_{01} \\ y_{10} \\ y_{11} \end{pmatrix} \quad t(y) = \begin{pmatrix} n = u \\ v(y) \end{pmatrix}$$

Note that V_t^{vv} corresponds to the lower of the Fisher information matrix computed under the full model, also when we consider n as a random quantity, it can be shown that this matrix specially in this case, is a diagonal matrix, as it has independent elements.

$$V_t^{vv} = \begin{pmatrix} \hat{\lambda}_{00}^{-1} & 0 & 0 & 0 \\ 0 & \hat{\lambda}_{10}^{-1} & 0 & 0 \\ 0 & 0 & \hat{\lambda}_{01}^{-1} & 0 \\ 0 & 0 & 0 & \hat{\lambda}_{11}^{-1} \end{pmatrix},$$

Score test statistic:

$$\tilde{W_u} = \left(v(y) - \begin{pmatrix} \hat{\lambda}_{00} \\ \hat{\lambda}_{01} \\ \hat{\lambda}_{10} \\ \hat{\lambda}_{11} \end{pmatrix} \right)^T V_t^{vv} \left(v(y) - \begin{pmatrix} \hat{\lambda}_{00} \\ \hat{\lambda}_{01} \\ \hat{\lambda}_{10} \\ \hat{\lambda}_{11} \end{pmatrix} \right) = \sum_{i,j=0}^{n} \frac{(y_{ij} - \frac{r_i s_j}{n})^n}{r_i s_j / n}$$

This expression is not the same as the one we obtain in the binomial distribution. However we can show that both of them coincide and we can do it by using the results of **problem 5.7**. We used two different models to get the same expression of the score test statistic, which is expected as this two interpretations of the 2x2 table are the same.

Another Approach: Multinomial model

The canonical statistic:

$$t(y) = \begin{pmatrix} y_{00} \\ y_{01} \\ y_{10} \end{pmatrix} \quad \text{where } y_{00}, y_{01}, y_{10} - \text{Independent}$$

$$V_t^{-1} \iff V_t^{vv}$$

Because of this independent structure, we could have some problems in the computation of V_t^{vv}. Probably this is not the case of 2x2 tables since in this case V_t is a 3-dimensional matrix so we already know how the inverse can be computed, but if we consider a more general case when we have an n x n table then we could have some problems with the computation of the inverse of the covariance matrix and V_t^{vv}, it could be done but it might be difficult and could take a lot of time in practice.

Poisson trick application to the Multinomial model:

If instead of the Multinomial model we consider a Poisson model then the canonical statistic will have the same components as the multinomial but with an extra component n, even if we increase the size of the canonical statistic, we will not make the computation more difficult. Under the Poisson model the components of $t(y)$ are independent, meaning that the covariance matrix of V_t is also a diagonal matrix. By adding some quantity we will make the computation easier. The use of the Poisson trick is mainly due to computational aspects.

$$\tilde{t}(y) \begin{pmatrix} t(y) \\ n \end{pmatrix} \quad \text{where} \quad n \sim Po(\lambda) \ \text{ and } \ V_t = \begin{pmatrix} V_t & 0 \\ 0 & V_n \end{pmatrix}$$

Summary:

We have proved that the exact test can be equivalently presented as an asymptotic test when n is large. In the prove we used that the conditional density/ probability mass function that we have in the exact test can be equivalently presented as a density function which depends

on the score statistic W_u. We also discussed that the different asymptotic tests are equivalent, which tells us that when n is large it does not matter which test we use, if we use the exact test or an asymptotic test. Surely, each test has disadvantages and advantages, and we will have to make considerations before applying any of them as they can all lead to different results.

7 Lecture 14:

To read: Section 5.4 and 9.1

The theory of generalized linear models is very close related to the theory of exponential family, it is some an extension of the models that we have discussed up to now which allows us more flexibility in modeling the realized data.

The generalized linear model consists consists usually in three blocks that determines these models.

- (i) **Linear predictor:** collection of covariates which we would like to use to model some parameters in our model.

$$\eta = \mathbf{x}^T \beta, \quad \dim(\beta) = k < n$$

- **(ii) Distribution type:** This part is related with the exponential family distribution because a GLM is determined with respect to the exponential family distribution.

- **(iii) Link function:** connect the mean value of the canonical statistic which we have in the exponential family to the new parameter which we denote by η which we will like to model by using the independent covariate, also known as some factors in our model.

$$\eta = g(\mu)$$

  If the function $g(\mu)$ is of an specific way, we will have a specific type of a GLM specially if it can be made a transformation $\eta = g(\mu) = g(\mu(\theta)) = \theta$. Where θ is the canonical parameter in the exponential family distribution. Then this function is called *canonical link*.

,

7.1 Generalized Linear Models:

We assume the following:

- We have an exponential family distribution parametrized by $\theta \in R$.

- The canonical statistic is **linear** in each observation y. i.e

$$t(y) = y$$

We have a sample form the exponential family. Let $y_1, y_2, ..., y_n$ be an independent sample. If we assume to have a one parameter exponential family and a large sample, if we would like to feed the model to the large sample, probably the model will not be very nice, the reason is that the hole sample that we would like to describe by only one parameter could have some heterogeneity. If we assume that $y_1, y_2, ..., y_n$ are also identically distributed then we would like to capture all heterogeneity that we have in our sample only by a single parameter, which could be not usually the case (as it could take us far from reality). All the knowledge of stochastic behavior could be contained in the single parameter θ or mean value μ.

Classical Solution: Drop the assumption of identically distribution. The realizations of each element in our sample we will like to assign to the same probabilistic model (same classical statistical distribution expofam) but we will like to model it with different parameter θ_i, since it will be the most flexible approach, but on the other hand it will not be very useful for us as well. In this case we will have to estimate n parameters which we will base only on n observations and that is surely not good, as among many reasons, the error estimation will be very large and the data will be over-parametrized (over-fitting).

Furthermore, we know the ML equation is based on the following relationship $\mu_i = y_i$ it means that y_i contributes only to estimate the mean value and nothing else. This is no good as one of the aims of any statistical modeling is to perform forecast of some realizations and if we have such a relationship between μ_i and y_i and we would like to take the next observation y_{n+1} then it also should have it's own mean μ_{n+1} and then we know it's mean can be estimated only by the new value y_{n+1} which is not observable by time n meaning that given a large sample, if we have always different means, we are not able to do any forecast.

We now have to decide what to do in these two extreme situations (i) only one parameter (ii) n parameters. The true some be somewhere in between. The idea is to model μ_i with some external variables which could be helpful for us.

Idea: We need a model for μ_i by using some other variables (factors, covariates). We assume we have some realizations and factors i.e, $(y_i, x_{i1}, ..., x_{iq})$ using this information we have in $x_{i1}, ..., x_{iq}$ we would like to model the mean value of y_i by $x_{i1}, ..., x_{iq}$.

For example, we want to make some predictions of the air pollution, and then we have some data with many covariates regarding of how the data was collection (season, area, moment of the day, etc). All of this information can be used to model y_i and it will help us to reduce the number of parameters in our model, so we do not have μ_i but we will have some coefficients which are used to feed y_i by using the covariates $x_{i1}, ..., x_{iq}$.

Classical approach: we could use here is the linear model we want to model by using a linear function with independent factors. Meaning that we will like to see the following model for each μ_i:

$$E[y_i] = \mu_i = x_{i1}\beta_1 + x_{i2}\beta_2 + ... + x_{iq}\beta_q \qquad \Rightarrow \text{q-parameters}$$

We have now reduced the number of parameters to $\beta_1, .., \beta_q$ hence, we will have q parameters

only in our model and it should be smaller or equal than n, because in all likelihood theory, the dimension of the parameter should be fixed, otherwise the classical asymptotic theory does not work.
This model works very well in the case of the Gaussian sample for example, or in the case of samples from the normal distribution with some given variance σ^2 and mean μ. In some cases just to model the mean function by a linear function is not a very good idea. For instance, if we consider a Poisson model

$$y_i \sim Po(\lambda_i) \quad \text{and} \quad y_i, ..., y_n \ -\text{independent} \quad E[y_i] = \underbrace{\lambda_i}_{>0} = \underbrace{x_{i1}\beta_1 + ... + X_{iq}\beta_q}_{\text{Could be negative}}$$

Theoretically the RHS could be negative for some values $x_{i1}, ..., x_{iq}$ unless we impose some restrictions on the x variables. Furthermore we know that λ_i cannot be negative, hence to model λ_i with a linear function is not very nice as sometimes we could get results that we cannot explain. A solution for this problem would be before using a linear predictor we would like to transform the mean value. For example the classical transformation, the logarithm. This value us from R and with this values we can model by a linear predictor without imposing any further restriction.

$$\eta_i = \log(\mu_i) = \log(\lambda_i) \in R$$

And then η_i is a linear predictor. Similar, in the case of a Bernoulli sample where we have the probability of success between (0,1) then we can think in transforming this probability of success in the hole R and then to fix a linear predictor to this transformed variable.
Summary: We have

- (i) **Linear predictor** which is the first part of the linear model. We would like to capture the behaviour in the mean value by using some additional covariates.

$$x_{i1}\beta_1 + ... + X_{iq}\beta_q$$

- (ii) **Link Function**: We have the transformation

$$\eta_i = \log(\mu_i) \Rightarrow \eta_i = \underbrace{g}_{\text{Link function}} (\mu_i)$$

  It defines the re-parametrization of our original model, expofam in terms of the new parameter η_i.

- (iii) **Distribution Assumption**: an specify expofam (Poisson distribution, Normal distribution, Negative Binomial distribution, Multinomial distribution, among others). The most important fact in here is that for these family of distribution the canonical statistic should be characterized by one parameter only, and also that the canonical statistic here is linear, i.e equal to y_i.

In general we are not restricted to some specific functions, for some families of exponential distributions, we have already some good choices for these linear functions, which are motivated by the way the canonical parameter is constructed. For instance, in the Bernoulli

sample the two classical models are the *logit model* and *covit model*, or in case of the Poisson model we much rather use the logarithm of λ instead of using a multiplicative model transformation.
In case of the likelihood theory we do not have any analytical solutions, hence we should use numerical procedures.

8 Lecture 15:

To read: Sections 9.2 and 9.3

8.1 Canonical link function

:

$$g = \mu_t^{-1}(\cdot) \Rightarrow \theta = X^T\beta$$

Where θ should be modeled with the help of the linear predictors. In most of the cases we assume X are deterministic quantities, they are changeable variables but involving no randomness variables, randomness is contained on y. θ is a canonical parameter vector, for instance if we have a sample of n observations we will have:

$$\theta = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_n \end{pmatrix}$$

We now will like to reduce the parameter space to a k-dimensional parameter space, this will done by using the linear predictors.Now, we will like to model each component of the vector θ as a linear combination of k factors such that $\theta_i = X_{i1}\beta_1 + ... + X_{k1}\beta_k$. This is very important to us because in the case of the canonical link function we could get a very nice representation of the exponential family distributions.

We will have that the probability mass function / density function of our observations $\mathbf{y} = (y_1, ..., y_n)$ now we will only have to assume that these observations are independent, we do not have any assumptions that they are identically distributed because for each observation we have it's own canonical parameter, or mean value. If we have the canonical link function then the probability mass function of the observation could be given by:

$$f(y;\theta) = f(y;\beta) = \prod_{i=1}^{n} C(\theta_i)^{-1} \underbrace{\prod_{i=1}^{n} h(y_i)}_{\text{structural function}} \exp(\theta^T y) \tag{31}$$

we now want to replace θ on the exponential function

Introducing the notation $X = (X_1, ..., X_n)$ where X_i is the i-th column of $X = (X_{ji})$ where $j \in \{1, .., k\}$ and $i \in \{1, .., n\}$

Then

$$\theta_i = X_i^T \beta \quad \text{and} \quad \theta^T = (X^T\beta)^T = \beta^T X$$

Yielding to a new family of exponential distribution:

$$f(y;\beta) = \prod_{i=1}^{n} C(X_i^T\beta)^{-1} \prod_{i=1}^{n} h(y_i) \exp(\beta^T X y)$$

where the canonical statistic is $t(y) = Xy$ where X is a matrix of some known constants, hence $t(y)$ is a function of y only, and the canonical parameter vector is β. In the case of the norming constant we can notice the presence of X_i however we should keep in mind that this is a known quantity, hence it is a function of β only. The most important here is that we do not have the presence of y_i.

Example: *Bernoulli distribution*

We assume to have a sequence of independent random variables $y_1, ..., y_n$ with $y_i \sim Be(\pi_i)$ we will like to model π_i by using some predictor variables in order to reduce the number of parameters in this model. In order to do that we should find a way to model the parameter π_i which is just the probability of success.

$$\mu_i = E[y_i] = \pi_i$$

In the case of the Bernoulli distribution we know that

$$\theta_i = \text{logit}(\pi_i) = \text{logit}(\mu_i)$$

This equality tells us that the canonical parameter θ_i is now related to the mean function μ_i.

$$\Rightarrow \text{Canonical link function} = \text{logit}(\cdot) \Rightarrow \text{logit regression}$$

In the case of the Bernoulli sample the canonical link function is the logic function and this basically introduces a new model in statistics which is the **logit regression**. There are some other link functions that can be used in the Bernoulli function, such as the **probit function** in this case $\eta_i = g(\mu_i)$ is given by $g = Q^{-1}(\cdot)$ where Q is the CDF of $\mathcal{N}(0,1)$.

Estimation:
Until now the only parameter that needs to be estimated is β. Surely, we will prefer to use the ML approach.

Application of MLE:

- (a) **Canonical link function:**
  We can consider, taking the likelihood function apply the logarithm and then the first derivative. However in this example we will obtain the maximum likelihood equation by

$$\mu_t(\beta) - t = 0$$

.

$$
\begin{aligned}
(\mu_t(\beta))_j &= \frac{\partial \log C_n(\beta)}{\partial \beta} \\
&= \frac{\partial \log \prod_{i=1}^{n} C(X_i^T \beta)}{\partial \beta_j} \\
&= \sum_{i=1}^{n} \frac{\partial \log C(X_i^T \beta)}{\partial \beta_j} \\
&= \sum_{i=1}^{n} \mu_i(\beta) \frac{\partial X_i^T \beta}{\partial \beta_j} = \sum_{i=1}^{n} \mu_i(\beta) x_{ji}
\end{aligned}
\tag{32}
$$

Where we have used that

$$
\frac{\partial \log C(\theta_i)}{\partial \theta_i} = \mu_i(\theta_i) = \mu_i(\beta) \quad \text{where} \;\; \mu_i = E[y_i]
$$

Notice that $\mu_i(\theta_i) = \mu_i(\beta)$ holds as $\theta_i = X_i^T \beta$ where X_i is a deterministic quantity, and this is why we can say that $\mu_i(\theta_i)$ is a function μ_i applied to to the vector β. We write the index i in μ_i, given its presence on X_i.

If we now get everything together our likelihood function should be written as:

$$
\Rightarrow \begin{pmatrix} \sum_{i=1}^{n} \mu_i(\beta) X_{1i} \\ \sum_{i=1}^{n} \mu_i(\beta) X_{2i} \\ \vdots \\ \sum_{i=1}^{n} \mu_i(\beta) X_{ki} \end{pmatrix} - Xy = 0 \qquad \Rightarrow \sum_{i=1}^{n} \mu_i(\beta) X_{1i} = \begin{pmatrix} X_{11} \\ X_{12} \\ \vdots \\ X_{1n} \end{pmatrix}^T \begin{pmatrix} \mu_1(\beta) \\ \mu_2(\beta) \\ \vdots \\ \mu_n(\beta) \end{pmatrix}
$$

Furthermore we have:

$$
\Rightarrow \underbrace{\begin{pmatrix} X_{11} & X_{12} & \dots & X_{1n} \\ X_{21} & X_{22} & \dots & X_{2n} \\ \vdots & & & \\ X_{k1} & X_{k2} & \dots & X_{kn} \end{pmatrix}}_{X} \underbrace{\begin{pmatrix} \mu_1(\beta) \\ \mu_2(\beta) \\ \vdots \\ \mu_n(\beta) \end{pmatrix}}_{\mu(\beta)} - Xy = 0 \Rightarrow X \cdot \mu(\beta) - Xy = 0
$$

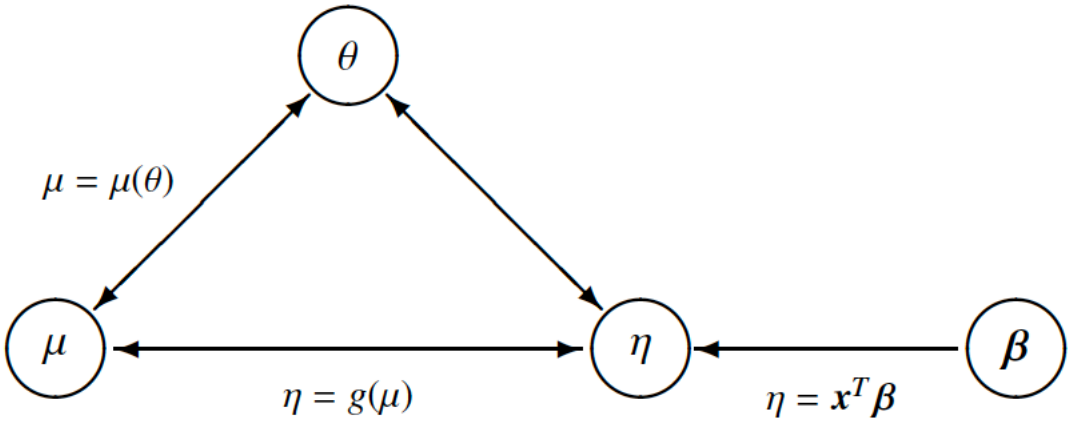
Then, the **Likelihood equation in the matrix form** is:

$$
\Rightarrow X(y - \mu(\beta)) = 0
$$

Let's now consider a sample from the exponential family distribution

$$
y_1, ..., y_n \overset{indp.}{\sim} \text{Expofam}(\theta_i)
$$

we now want to model some transformation of the mean vector with some linear predictor where we have a new vector of the parameter θ. We can draw a diagram of how the general linear model is design, we will have:



Where θ is the canonical parameter and μ the mean value parameter. Note that neither θ or μ are modeled with a linear predictor, and this is why we define a new parametrization $\eta = X^T \cdot \beta$ (as X is a k x n matrix). Hence, we have three parametrization which will be used when we would like to derive the likelihood equation of the parameter θ. Transformation should go in the following direction $\theta \to \mu \to \eta$.

Likelihood function:

$$L(\beta; y) = \prod_{i=1}^{n} f(y_i; \theta_i(\beta))$$

Log - likelihood function:

$$\begin{aligned} l(\beta; y) &= \sum_{i=1}^{n} \log f(y_i; \theta_i(\beta)) \\ &= \sum_{i=1}^{n} [\log C(\theta_i) + \log h(y_i) + \theta_i(\beta) y_i] \\ &= \sum_{i=1}^{n} \log h(y_i) + \sum_{i=1}^{n} [\theta_i(\beta) y_i - \log C(\theta_i)(\beta)] \end{aligned} \tag{33}$$

We will now like to optimize the log-likelihood function, with the inconvenient that we do not usually have the analytical expression of $\theta_i(\beta)$ which will make the computation of the derivatives much difficult. However, we do know that θ, μ and η are related. We

will apply **chain rule**

Score vector: We will take the derivative of the log likelihood function with respect to the j-th component of the vector θ.

$$U_j(\beta) = \frac{\partial l(\beta; y)}{\partial \beta_j} = \sum_{i=1}^{n} \Big[\frac{\partial \theta_i(\beta)}{\partial \beta_j} y_i - \frac{\partial \log C(\theta_i(\beta))}{\partial \beta_j} \Big]$$

Where keeping in mind the relation $\theta \to \mu \to \eta$ we apply chain rule:

$$\theta_i(\beta) = \theta_i(\mu_i(\eta_i(\beta))) \quad \Rightarrow \frac{\partial \theta_i(\beta)}{\partial \beta_j} = \frac{\partial \theta_i}{\partial \mu_i} \cdot \frac{\partial \mu_i}{\partial \eta_i} \cdot \frac{\partial \eta_i}{\partial \beta_j}$$

$$\begin{aligned} \frac{\partial \theta_i}{\partial \mu_i} &= \Big(\frac{\partial \mu_i}{\partial \theta_i} \Big)^{-1} = (\mathrm{Var}(y_i))^{-1} = (V_t(\theta_i))^{-1} \\ \frac{\partial \mu_i}{\partial \eta_i} &= \Big(\frac{\partial \eta_i}{\partial \mu_i} \Big)^{-1} = \Big(\frac{\partial g(\mu_i)}{\partial \mu_i} \Big)^{-1} = (g'(\mu_i))^{-1} \\ \frac{\partial \eta_i}{\partial \beta_j} &= \frac{\partial (X_i^t \beta)}{\partial \beta_j} = X_{ji} \end{aligned} \tag{34}$$

Furthermore

$$\frac{\partial \log C(\theta_i(\beta))}{\partial \beta_j} = \underbrace{\frac{\partial \log C(\theta_i(\beta))}{\partial \theta_i}}_{\mu_i} \cdot \underbrace{\frac{\partial \theta_i}{\partial \beta_j}}_{\frac{\partial \theta_i(\beta)}{\partial \beta_j}}$$

Hence, the score function is given by:

$$\begin{aligned} U_j(\beta) &= \sum_{i=1}^{n} \frac{\partial \theta_i}{\partial \beta_j} (y_i - \mu_i) \\ &= \sum_{i=1}^{n} (V_{y_i}(\theta_i))^{-1} \cdot g'(\mu_i)^{-1} \cdot X_{ji}(y_i - \mu_i) \\ &= (X_{j1}, \ldots, X_{jn})^T \cdot \begin{pmatrix} g'(\mu_1)^{-1} \cdot V_{y_1}^{-1}(\theta_1)(y_i - \mu_i) \\ \vdots \\ g'(\mu_n)^{-1} \cdot V_{y_n}^{-1}(\theta_n)(y_n - \mu_n) \end{pmatrix} \end{aligned} \tag{35}$$

In order to think about how such a vector can be obtained, we can use the following property. If we have the matrix

$$D = \begin{pmatrix} d_1 & & \\ & \ddots & \\ & & d_n \end{pmatrix}, \quad Z = \begin{pmatrix} z_1 \\ \vdots \\ z_n \end{pmatrix} \quad \Rightarrow \; D \cdot Z = \begin{pmatrix} d_1 Z_1 \\ \vdots \\ d_n Z_n \end{pmatrix}$$

We can rewrite the last matrix in (35)

$$\begin{pmatrix} g'(\mu_1)^{-1} \cdot V_{y_1}^{-1}(\theta_1)(y_i - \mu_i) \\ \vdots \\ g'(\mu_n)^{-1} \cdot V_{y_n}^{-1}(\theta_n)(y_n - \mu_n) \end{pmatrix} = \underbrace{\begin{pmatrix} g'(\mu_1)^{-1} & & \\ & \ddots & \\ & & g'(\mu_n)^{-1} \end{pmatrix}}_{g(\mu)} \cdot \underbrace{\begin{pmatrix} V_{y_1}(\theta_1)^{-1} \\ \vdots \\ V_{y_n}(\theta_n)^{-1} \end{pmatrix}}_{V_y(\theta)^{-1}} \cdot \underbrace{\begin{pmatrix} y_1 - \mu_1 \\ \vdots \\ y_n - \mu_n \end{pmatrix}}_{y - \mu(\beta)}$$

Since μ can be a difficult function of θ we do not have an analytical solutions of the equation

$$X^T g(\mu) V_y(\theta)^{-1}(y - \mu(\beta)) = 0$$

Hence, a numerical procedure will be needed for solving the previous equation. The **Newton - Raphson method** or **Fisher scoring** are two possible ways of solving this likelihood equation.

Fisher Information matrix: The general equation, depends in the score statistic.

$$I(\beta) = \mathrm{Cov}(U(\beta))$$

- (a) **Canonical link function**: we have that the score function is given by

$$U(\beta) = X(y - \mu(\beta))$$

$$\Rightarrow I(\beta) = J(\beta) = \mathrm{Cov}(X(y - \mu(\beta))) = X \underbrace{\mathrm{Cov}(y - \mu(\beta))}_{\mathrm{Cov}(y) = V_y} \cdot X^T = X V_y \cdot X^T$$

  $\mathrm{Cov}(X(y - \mu(\beta))) = \mathrm{Cov}(y)$ since $\mu(\beta)$ is a deterministic quantity.

- (b) **non-Canonical link function**:

$$\begin{aligned} I(\beta) &= \mathrm{Cov}(\underbrace{X G^{-1} V_y^{-1}}_{\text{deterministic}} \cdot (y - \mu(\beta))) \\ &= X G^{-1} V_y^{-1} \mathrm{Cov}(y) (X G^{-1} V_y^{-1})^T \\ &= X G^{-1} V_y^{-1} V_y V_y^{-1} G^{-1} X^T \\ &= X G^{-1} V_y^{-1} G^{-1} X^T \end{aligned} \tag{36}$$

8.2 Residuals:

Residuals are used in order to construct goodness of fit analysis of a model. In the case of the GLM we use the deviance, instead of the classical residuals which are different between observations y and each mean (which we model) $\hat{\mu}$ we define the residual in another way.

$$D = 2\{\log L(y; y) - \log L(\mu(\hat{\beta}); y)\}$$

Where the likelihood function $L(\mu(\hat{\beta}); y)$ can be written in terms of the mean value parametrization. If we have no assumption about the structure of the mean function, then we know that the MLE of μ_i should be the observation element y_i itself.

- (a) Saturated model: most general model where we have no restriction on it's parameters. The deviance will not be a very informative value for us, this value cannot be used in any asymptotic test as we have estimated too many parameters.
- (b) Null model:

The deviance can also be used in likelihood theory We can do comparison of two nested models M_1 and M_2 where $M_2 \subset M_1$ knowing that M_1 is a richer model, it has more parameters compared to M_2. When it comes to compare we would like to know if the richer model is enough for us or if we want to keep the simpler model M_2.

We have to note in the case of the model reduction hypothesis that we do not accept M_2, we can only say there is not evidence to use M_1 or if we reject the null hypothesis then we have evidence and then M_2 is not enough, but we never say M_2 is the best model, as the null hypothesis is based in whether we need to use a richer model or not. In this test we can rewrite the likelihood ratio test as the difference of two deviance

$$2 \log \left\{ \frac{L(\mu(\hat{\beta}_1)}{L(\mu(\hat{\beta}_2)} \right\} = D_2 - D_1$$

Which follows directly form the definition of D. Another application of the deviance is

Deviance Residuals:

$$D = 2[\log L(y, y) - \log L(\mu(\hat{\beta}); y)]$$

Considering that we have independence between the elements, we can rewrite this in the following way

$$2 \sum_{i=1}^{n} \log L(y_i; y_i) - \log L(\mu_i(\hat{\beta}); y) = \sum_{i=1}^{n} D_i^2$$

where D_i^2 is always positive. The deviance residuals are given by

$$\epsilon_i = \text{sign} \ \ (y_i - \mu_i(\beta)) \cdot D_i$$

Where "sign" is a sign function, we need this function only because we would like to have residuals of different signs as in the case of the classical linear regression model.

$$\text{sign}(x) = \begin{cases} 1 & x > 0 \\ 0 & x = 0 \\ -1 & x < 0 \end{cases}$$