Lecture Stat 461-561
Wald, Rao and Likelihood Ratio Tests

AD

February 2008

- Introduction
- Wald test
- Rao test
- Likelihood ratio test

Introduction

- We want to test $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ using the log-likelihood function.
- We denote $l(\theta)$ the loglikelihood and $\widehat{\theta}_n$ the consistent root of the likelihood equation.
- Intuitively, the farther $\widehat{\theta}_n$ is from θ_0, the stronger the evidence against the null hypothesis.
- How far is "far enough"?

Wald Test

- We use the fact that under regularity assumptions we have under H_0

$$\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) \xrightarrow{\text{D}} \mathcal{N}\left(0, I^{-1}\left(\theta_0\right)\right)$$

  where

$$I\left(\theta_0\right) = \mathbb{E}_{\theta_0}\left[\frac{\partial^2 \log f\left(X|\theta\right)}{\partial\theta^2}\right].$$

- This suggests defining the following *Wald statistic*

$$W_n = \sqrt{nI\left(\theta_0\right)}\left(\widehat{\theta}_n - \theta_0\right)$$

  or $W_n = \sqrt{n\widehat{I}\left(\theta_0\right)}\left(\widehat{\theta}_n - \theta_0\right)$ where $\widehat{I}\left(\theta_0\right)$ is a consistent estimate of $I\left(\theta_0\right)$, e.g. $I\left(\widehat{\theta}_n\right)$.

- Under H_0, we have

$$W_n = \sqrt{n\widehat{I}(\theta_0)}\left(\widehat{\theta}_n - \theta_0\right) \xrightarrow{D} \mathcal{N}(0,1)$$

- A **Wald test** is any test that rejects $H_0 : \theta = \theta_0$ in favor of $H_1 : \theta \neq \theta_0$ when $|W_n| \geq z_{\alpha/2}$ where $z_{\alpha/2}$ satisfies $P\left(Z \geq z_{\alpha/2}\right) = \alpha/2$.
- It follows that by construction Type I error probability is

$$P_{\theta_0}\left(|W_n| \geq z_{\alpha/2}\right) \rightarrow P_{\theta_0}\left(|Z| \geq z_{\alpha/2}\right) = \alpha$$

and this is an asymptotically size α test.
- Now consider $\theta \neq \theta_0$ then

$$W_n = \sqrt{n\widehat{I}(\theta_0)}\left(\widehat{\theta}_n - \theta_0\right) = \underbrace{\sqrt{n\widehat{I}(\theta_0)}\left(\widehat{\theta}_n - \theta\right)}_{\xrightarrow{D}\mathcal{N}(0,1)} + \underbrace{\sqrt{n\widehat{I}(\theta_0)}\left(\theta - \theta_0\right)}_{\xrightarrow{P}\pm\infty}$$

so

$$P_\theta\left(\text{reject } H_0\right) \rightarrow 1 \text{ as } n \rightarrow \infty.$$

- Note that the (approximate) size α Wald test rejects $H_0 : \theta = \theta_0$ in favor of $H_1 : \theta \neq \theta_0$ if and only if $\theta_0 \notin C$ where

$$C = \left(\widehat{\theta}_n - \frac{z_{\alpha/2}}{\sqrt{n\widehat{I}\left(\theta_0\right)}}, \widehat{\theta}_n + \frac{z_{\alpha/2}}{\sqrt{n\widehat{I}\left(\theta_0\right)}} \right).$$

- Thus testing the hypothesis is equivalent to checking whether the null value is in the confidence interval.

- Similarly we can test $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$.
- In this case we use $W_n = \sqrt{n\widehat{I}(\theta_0)}\left(\widehat{\theta}_n - \theta_0\right)$ and reject H_0 if $W_n \geq z_\alpha$.
- We have if $\theta = \theta_0$

$$P_\theta\left(W_n \geq z_\alpha\right) \rightarrow P\left(Z \geq z_\alpha\right) = \alpha$$

whereas if $\theta < \theta_0$

$$P_\theta\left(W_n \geq z_\alpha\right) \rightarrow 0$$

and if $\theta > \theta_0$

$$P_\theta\left(W_n \geq z_\alpha\right) \rightarrow 1.$$

- *Example*: Assume we have $X_i \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)$ and we want to test $H_0 : p = p_0$ against $H_1 : p \neq p_0$.
- We have for $S_n = \sum_{i=1}^n x_i$

$$\begin{aligned} f(x|\, p) &= p^x (1-p)^{1-x} \Rightarrow l(p) = S_n \log p + (n - S_n) \log (1-p) \\ l'(p) &= \frac{S_n}{p} - \frac{(n-S_n)}{1-p} \Rightarrow \widehat{p}_n = \frac{S_n}{n}. \end{aligned}$$

- It is easy to check that

$$\widehat{p}_n - p_0 \xrightarrow{\text{D}} \mathcal{N}\left(0, \frac{\sigma_0^2}{n}\right)$$

where $\sigma_0^2 = p_0 (1 - p_0)$. This quantity can be consistently estimated through $\widehat{\sigma_n^2} = \widehat{p}_n (1 - \widehat{p}_n)$ so the Wald test is based on the statistic

$$W_n = \frac{(\widehat{p}_n - p_0)}{\sqrt{\widehat{p}_n (1 - \widehat{p}_n) / n}}.$$

- Wald test is not limited to MLE estimate, you just need to know the asymptotic distribution of your test satistic.
- *Example*: Assume we have $X_1, ..., X_m$ and $Y_1, ..., Y_n$ be two independent samples from populations with mean μ_1 and v.
- We write $\delta = \mu_1 - \mu_2$ and we want to test $H_0 : \delta = 0$ versus $H_1 : \delta \neq 0$.
- We build

$$W = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{m}}}$$

  where S_1^2 and S_2^2 are the sample variances.
- Thanks to the CLT, we have $W \xrightarrow{\text{D}} \mathcal{N}(0, 1)$ as $m, n \to \infty$.

- *Example* (Comparing two prediction algorithms): We test a prediction algorithm on a test set of size m and the second prediction algorithm on a second test set of size n. Let X be the number of incorrect predictions for algorithm 1 and Y the number of incorrect prediction for algorithm 2. Then $X \sim$Binomial(m, p_1) and $Y \sim$Binomial(n, p_2).
- To test the null hypothesis $H_0 : \delta = p_1 - p_2 = 0$ versus $H_1 : \delta \neq 0$ we note that $\widehat{\delta} = \widehat{p}_1 - \widehat{p}_2$ with estimated standard error

$$\sqrt{\frac{\widehat{p}_1 \left(1 - \widehat{p}_1\right)}{m} + \frac{\widehat{p}_2 \left(1 - \widehat{p}_2\right)}{n}}.$$

- The (approximate) size α test is to reject H_0 when

$$|W| = \frac{\left|\widehat{p}_1 - \widehat{p}_2\right|}{\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{m} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n}}} \geq z_{\alpha/2}$$

as $W \xrightarrow{D} \mathcal{N}(0,1)$ as $m, n \to \infty$ thanks to the CLT.

- In practice, we often have mispecified models! That is there does not exist any θ_0 such that $X_i \sim f(x|\theta_0)$ but $X_i \sim g$ [g being obviously unknown].
- In this case θ_0 is not the 'true' parameter but the parameter maximizing

$$\int \log \frac{f(x|\theta)}{g(x)}.g(x)\,dx \Rightarrow \int \left.\frac{\partial \log f(x|\theta)}{\partial\theta}\right|_{\theta=\theta_0}.g(x)\,dx = 0$$

- The Wald test becomes incorrect as we do NOT have

$$\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) \xrightarrow{D} \mathcal{N}\left(0, I^{-1}(\theta_0)\right)$$

- To correct it, let's go back to the derivation of the asymptotic normality

$$0 = l'\left(\widehat{\theta}_n\right) \approx l'(\theta_0) + l''(\theta_0)\left(\widehat{\theta}_n - \theta_0\right)$$

so

$$\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) \approx \sqrt{n}\frac{-l'(\theta_0)}{l''(\theta_0)} = \frac{-l'(\theta_0)}{\sqrt{n}}\frac{n}{l''(\theta_0)}$$

- We have by the law of large numbers

$$\frac{l''(\theta)}{n} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \log f(X_i|\theta)}{\partial\theta^2} \xrightarrow{\text{P}} \int \frac{\partial^2 \log f(x|\theta)}{\partial\theta^2} .g(x)\, dx$$

- We also have

$$\frac{l'(\theta)}{\sqrt{n}} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \frac{\partial \log f(X_i|\theta)}{\partial\theta}$$

where

$$\begin{aligned} var\left[\frac{\partial \log f(X|\theta)}{\partial\theta}\right] &= \int \frac{\partial \log f(x|\theta)}{\partial\theta}^2 .g(x)\, dx \\ &\quad - \left(\int \frac{\partial \log f(x|\theta)}{\partial\theta} .g(x)\, dx\right)^2 \end{aligned}$$

So by the CLT

$$\frac{l'(\theta_0)}{\sqrt{n}} \xrightarrow{\text{D}} \mathcal{N}\left(0, \int \frac{\partial \log f(x|\theta)}{\partial\theta}^2 .g(x)\, dx\bigg|_{\theta=\theta_0}\right)$$

- By Slutzky's lemma

$$\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) \rightarrow \mathcal{N}\left(0, \frac{\int \left.\frac{\partial \log f(x|\theta)}{\partial \theta}\right|_{\theta_0}^2 .g(x)\, dx}{\left(\int \left.\frac{\partial^2 \log f(x|\theta)}{\partial \theta^2}\right|_{\theta_0} .g(x)\, dx\right)^2}\right).$$

- The asymptotic variance can be estimated consistently from the data.
- We can derive a Wald test based on this asymptotic variance.

- This suggests developing a test for misspecification as

$$\int \frac{\partial \log f\left(x\right|\theta)}{\partial\theta}^{2}.f\left(x\right|\theta)\,dx + \left(\int \frac{\partial^{2} \log f\left(x\right|\theta)}{\partial\theta^{2}}.f\left(x\right|\theta)\,dx\right) = 0$$

- So we can propose the following test statistic

$$T = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \log f\left(x_i\right|\theta)}{\partial\theta}^{2}\Bigg|_{\widehat{\theta}_n} + \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^{2} \log f\left(x_i\right|\theta)}{\partial\theta^{2}}\Bigg|_{\widehat{\theta}_n}$$

which under H_0: model specified is asymptotically Gaussian with zero-mean and variance which can be estimated consistently.

Rao Test

- We have under $H_0 : \theta = \theta_0$

$$\frac{1}{\sqrt{n}} l' (\theta_0) \xrightarrow{\text{D}} \mathcal{N} (0, I (\theta_0))$$

where $l' (\theta) = \frac{\partial \log L(\theta|\mathbf{x})}{\partial \theta}$.

- To prove this, remember that

$$0 = l' \left(\widehat{\theta}_n \right) \approx l' (\theta_0) + l'' (\theta_0) \left(\widehat{\theta}_n - \theta_0 \right)$$

thus

$$\frac{1}{\sqrt{n}} l' (\theta_0) \approx -\frac{l'' (\theta_0)}{\sqrt{n}} \left(\widehat{\theta}_n - \theta_0 \right) = -\frac{l'' (\theta_0)}{n} \sqrt{n} \left(\widehat{\theta}_n - \theta_0 \right)$$

where

$$-\frac{l'' (\theta_0)}{n} \xrightarrow{\text{P}} I (\theta_0) \text{ and } \sqrt{n} \left(\widehat{\theta}_n - \theta_0 \right) \xrightarrow{\text{D}} \mathcal{N} \left(0, I^{-1} (\theta_0) \right).$$

The result follows from Slutzky's lemma.

- This suggests defining the following Rao score statistic

$$R_n = \frac{l'(\theta_0)}{\sqrt{nI(\theta_0)}}$$

which converges in distribution to a standard normal under H_0.

- A major advantage of the score statistics is that it does not require computing $\widehat{\theta}_n$. We could also replace $I(\theta_0)$ by a consistent estimate $\widehat{I}(\theta_0)$. However using $I\left(\widehat{\theta}_n\right)$ would defeat the purpose of avoiding to compute $\widehat{\theta}_n$.
- A **score test** -also called **Rao score test**- is any test that rejects $H_0 : \theta = \theta_0$ in favor of $H_1 : \theta \neq \theta_0$ when $|R_n| \geq z_{\alpha/2}$.

- *Example*: Assume we have $X_i \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)$ and we want to test $H_0 : p = p_0$ against $H_1 : p \neq p_0$.
- We have for $S_n = \sum_{i=1}^{n} x_i$

$$
\begin{aligned}
f(x|p) &= p^x (1-p)^{1-x} \Rightarrow l(\lambda) = S_n \log p + (n - S_n) \log (1-p) \\
l'(p) &= \frac{S_n}{p} - \frac{(n - S_n)}{1-p} \Rightarrow \widehat{p}_n = \frac{S_n}{n}. \\
l''(p) &= -\frac{S_n}{p^2} - \frac{(n - S_n)}{(1-p)^2}
\end{aligned}
$$

- It follows that

$$
l'(p) = \frac{n(\widehat{p}_n - p)}{p(1-p)}, \; I(p) = \frac{1}{p(1-p)}
$$

and

$$
R_n = \frac{l'(p_0)}{\sqrt{nI(p_0)}} = \frac{\widehat{p}_n - p_0}{\sqrt{p_0(1-p_0)/n}}.
$$

- In practice, for complex models, one would use

$$R_n = \frac{l'(\theta_0)}{\sqrt{n\widehat{I}(\theta_0)}}$$

where

$$\widehat{I}(\theta_0) = -\frac{1}{n}\sum_{i=1}^{n} \left.\frac{\partial^2 \log f(x_i|\theta)}{\partial\theta^2}\right|_{\theta_0} \xrightarrow{P} I(\theta_0)$$

which is consistent thanks to Slutzky.

- However, once more one has to be careful for misspecified models

- Indeed we have

$$\frac{1}{\sqrt{n}} l'(\theta_0) \approx -\frac{l''(\theta_0)}{\sqrt{n}} \left(\widehat{\theta}_n - \theta_0\right) = -\frac{l''(\theta_0)}{n} \sqrt{n} \left(\widehat{\theta}_n - \theta_0\right)$$

where

$$-\frac{l''(\theta_0)}{n} \xrightarrow{\text{P}} \int \left.\frac{\partial^2 \log f(x|\theta)}{\partial \theta^2}\right|_{\theta=\theta_0} .g(x)\, dx$$

and

$$\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) \xrightarrow{\text{D}} \mathcal{N}\left(0, \frac{\left.\int \frac{\partial \log f(x|\theta)}{\partial \theta}^2 .g(x)\, dx\right|_{\theta=\theta_0}}{\left(\int \left.\frac{\partial^2 \log f(x|\theta)}{\partial \theta^2}\right|_{\theta=\theta_0} .g(x)\, dx\right)^2}\right)$$

so

$$\frac{1}{\sqrt{n}} l'(\theta_0) \xrightarrow{\text{D}} \mathcal{N}\left(0, \int \left.\frac{\partial \log f(x|\theta)}{\partial \theta}\right|_{\theta_0}^2 .g(x)\, dx\right)$$

- Thus if you use Rao test for misspecified model, you have to use the following consistent estimate of the asymptotic variance

$$\frac{1}{n}\sum_{i=1}^{n} \left.\frac{\partial \log f\left(\left. x_i \right| \theta\right)}{\partial \theta}\right|_{\theta_0}^{2} \xrightarrow{D} \int \left.\frac{\partial \log f\left(\left. x \right| \theta\right)}{\partial \theta}\right|_{\theta_0}^{2} .g\left(x\right) dx$$

- If you use the standard estimate $-\frac{1}{n}\sum_{i=1}^{n} \left.\frac{\partial^2 \log f\left(x_i|\theta\right)}{\partial \theta^2}\right|_{\theta_0}$, then you will get a wrong result.

Likelihood Ratio Test

- The LR test is based on

$$\Delta_n = l\left(\widehat{\theta}_n\right) - l(\theta_0) = \log\left(\frac{\sup_{\theta\in\Theta} L(\theta|\,\mathbf{x})}{L(\theta_0|\,\mathbf{x})}\right) \geq 0.$$

- If we perform a Taylor expansion of $l'\left(\widehat{\theta}_n\right) = 0$ around θ_0 then

$$l'\left(\widehat{\theta}_0\right) = \left(\widehat{\theta}_n - \theta_0\right)\left(-l''(\theta_0) + o(1)\right). \quad (1)$$

- By performing another Taylor expansion of $l\left(\widehat{\theta}_n\right)$ around θ_0 then

$$\Delta_n = \left(\widehat{\theta}_n - \theta_0\right) l'(\theta_0) + \frac{1}{2}\left(\widehat{\theta}_n - \theta_0\right)^2 \left[l''(\theta_0) + o(1)\right]. \quad (2)$$

- Hence, by substituting (1) in (2), then

$$\Delta_n = n\left(\widehat{\theta}_n - \theta_0\right)^2 \left\{-\frac{1}{n} l''(\theta_0) + \frac{1}{2n} l''(\theta_0) + o\left(\frac{1}{n}\right)\right\}.$$

- By Slutsky's theorem, this implies that under H_0

$$2\Delta_n \xrightarrow{\text{D}} \chi_1^2.$$

- Noting that the $1-\alpha$ quantile of a χ_1^2 distribution is $z_{\alpha/2}^2$, we can now define the LR test.
- A LR test is any test that rejects $H_0 : \theta = \theta_0$ in favor of $H_1 : \theta \neq \theta_0$ when $2\Delta_n \geq z_{\alpha/2}^2$ or, equivalently, when $\sqrt{2\Delta_n} \geq z_{\alpha/2}$; i.e. we reject for small values of the LR.

- *Example*: Assume we have $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{P}(\lambda)$ and we want to test $H_0 : \lambda = \lambda_0$ against $H_1 : \lambda \neq \lambda_0$.
- We have for $S_n = \sum_{i=1}^{n} x_i$

$$
\begin{aligned}
f(x|\lambda) &= \exp(-\lambda)\frac{\lambda^x}{x!} \Rightarrow l(\lambda) = -n\lambda + S_n \log\lambda - \left(\sum_{i=1}^{n} x_i!\right) \\
l'(\lambda) &= -n + \frac{S_n}{\lambda},\; l''(\lambda) = -\frac{S_n}{\lambda^2}.
\end{aligned}
$$

- It follows that $\widehat{\lambda}_n = \frac{S_n}{n}$ and

$$
\Delta_n = n\left(\lambda_0 - \widehat{\lambda}_n\right) + S_n \log\frac{\widehat{\lambda}_n}{\lambda_0}.
$$

Test equivalences

- We can show that (when there is no misspecification)

$$R_n \xrightarrow{P} W_n,$$
$$W_n^2 \xrightarrow{P} 2\Delta_n.$$

- The tests are thus asymptotically equivalent in the sense that under H_0 they reach the same decision with probability 1 as $n \longrightarrow \infty$.
- For a finite sample size n, they have some relative advantages and disadvantages with respect to one another.

- It is easy to create one-sided Wald and score tests.
- The score test does not require $\widehat{\theta}_n$ whereas the other two tests do.
- The Wald test is most easily interpretable and yields immediate confidence intervals.
- The score test and LR test are invariant under reparametrization, whereas the Wald test is not.

- *Example*: Assume $X_i \overset{\text{i.i.d.}}{\sim} f(x|\theta)$ where $f(x|\theta) = \theta \exp(-x\theta)\, \mathbb{I}\{x > 0\}$. Then

$$l(\theta) = n\left(\log \theta - \theta \overline{X}_n\right)$$

which yields

$$l'(\theta) = n\left(\frac{1}{\theta} - \overline{X}_n\right) \text{ and } l''(\theta) = -\frac{n}{\theta^2}.$$

- We also obtain

$$\widehat{\theta}_n = \frac{1}{\overline{X}_n},\ I(\theta) = \theta^{-2}.$$

- It follows that

$$\begin{aligned}
W_n &= \frac{\sqrt{n}}{\theta_0}\left(\frac{1}{\overline{X}_n} - \theta_0\right), \\
R_n &= \theta_0 \sqrt{n}\left(\frac{1}{\theta_0} - \overline{X}_n\right) = \frac{W_n}{\theta_0 \overline{X}_n}, \\
\Delta_n &= n\left\{\overline{X}_n\left(\overline{X}_n - \theta_0\right) - \log\left(\theta_0 \overline{X}_n\right)\right\}.
\end{aligned}$$

- *Example*: Assume $X_i \overset{\text{i.i.d.}}{\sim} f(x|\theta)$ where

$$f(x|\theta) = \theta c^{\theta} x^{-(\theta+1)} \mathbb{I}\{x > c\} \qquad \text{(Pareto distribution)}$$

where c is a known constant and θ is unknown.

- We have for $S_n = \sum_{i=1}^{n} \log(x_i)$

$$\begin{aligned} l(\theta) &= n(\log\theta + \theta\log c) - (\theta+1) S_n, \\ l'(\theta) &= n\left(\frac{1}{\theta} + \log c\right) - S_n, \;\; l''(\theta) = -\frac{n}{\theta^2}. \end{aligned}$$

- Thus we have $\widehat{\theta}_n = \frac{n}{S_n - n\log c}$, $I(\theta) = \theta^{-2}$ and it follows that

$$\begin{aligned} W_n &= \frac{\sqrt{n}}{\theta_0}\left(\frac{n}{S_n - n\log c} - \theta_0\right), \\ R_n &= \sqrt{n}\theta_0\left(\left(\frac{1}{\theta_0} + \log c\right) - S_n\right), \\ \Delta_n &= n\left(\log\frac{\widehat{\theta}_n}{\theta_0} + \left(\widehat{\theta}_n - \theta_0\right)\log c\right) - \left(\widehat{\theta}_n - \theta_0\right) S_n. \end{aligned}$$

Multivariate Generalizations

- When $\theta \in \mathbb{R}^d$, then the Wald, Rao and LR tests can be straightforwardly extended

$$
\begin{aligned}
W_n &: = n\left(\widehat{\theta}_n - \theta_0\right)^{\mathsf{T}} I\left(\theta_0\right)\left(\widehat{\theta}_n - \theta_0\right) \xrightarrow{\mathsf{D}} \chi_d^2, \\
R_n &: = \frac{1}{n} \nabla l\left(\theta_0\right)^{\mathsf{T}} I^{-1}\left(\theta_0\right) \nabla l\left(\theta_0\right) \xrightarrow{\mathsf{D}} \chi_d^2, \\
\Delta_n &: = l\left(\widehat{\theta}_n\right) - l\left(\theta_0\right) \xrightarrow{\mathsf{D}} \frac{1}{2}\chi_d^2
\end{aligned}
$$

- Therefore, if c_α^d denotes the $1-\alpha$ quantile of the χ_d^2 distribution, then we reject H_0 when $W_n \geq c_\alpha^d$, $R_n \geq c_\alpha^d$ and $2\Delta_n \geq c_\alpha^d$.
- As in the scalar case, in the Wald test, we can subsitute to $I\left(\theta_0\right)$ a consistent estimate.

- *Example*: Assume we observe a vector $X = (X_1, ..., X_k)$ where $X_j \in \{0, 1\}$, $\sum_{j=1}^{k} X_j = 1$ with

$$f(x| p_1, ..., p_{k-1}) = \left(\prod_{j=1}^{k-1} p_j^{x_j}\right)\left(1 - \sum_{j=1}^{k-1} p_j\right)^{x_k}$$

  where $p_j > 0$ and $p_k := 1 - \sum_{j=1}^{k-1} p_j < 1$. We have $\theta = (p_1, ..., p_{k-1})$.
- We have for n observations $X^1, X^2, ..., X^n$

$$\begin{aligned} l(\theta) &= \sum_{j=1}^{k} t_j \log p_j = \sum_{j=1}^{k-1} t_j \log p_j + t_k \log\left(1 - \sum_{j=1}^{k-1} p_j\right), \\ \frac{\partial l(\theta)}{\partial p_j} &= \frac{t_j}{p_j} - \frac{t_k}{p_k}, \ \frac{\partial^2 l(\theta)}{\partial p_j^2} = -\frac{t_j}{p_j^2} - \frac{t_k}{p_k^2}, \\ \frac{\partial^2 l(\theta)}{\partial p_j \partial p_l} &= -\frac{t_k}{p_k^2}, j \neq l < k. \end{aligned}$$

  where $t_j = \sum_{i=1}^{n} x_j^i$.

- Recall that X_j has a Bernoulli distribution with mean p_j so

$$I(\theta) = \begin{bmatrix} p_1^{-1} + p_k^{-1} & p_k^{-1} & p_k^{-1} \\ p_k^{-1} & p_2^{-1} + p_k^{-1} & \vdots \\ \vdots & p_k^{-1} & \vdots \\ \vdots & \vdots & p_{k-1}^{-1} + p_k^{-1} \end{bmatrix}$$

- It follows that

$$\begin{aligned} W_n &= n\left(\widehat{\theta}_n - \theta_0\right) I(\theta_0) \left(\widehat{\theta}_n - \theta_0\right), \\ R_n &= \frac{1}{n} \nabla l(\theta_0)^{\mathsf{T}} I^{-1}(\theta_0) \nabla l(\theta_0). \end{aligned}$$

After tiedous calculations, it can be shown that

$$W_n = R_n = \sum_{j=1}^{k} \frac{(t_j - np_i)^2}{np_i}$$

which is the usual Pearson chi-square test whereas $\Delta_n = n \sum_{j=1}^{k} t_j \log \frac{\widehat{p}_j}{p_j}$.