Likelihood Inference

Source: • Mathematical Statistics & Data Analysis - J. A. Rice (Section 8.5 & 9.4)
• Testing Statistical Hypotheses - Lehman & Romano

Maximum Likelihood Estimator

Suppose that random variables $X_1, \dots, X_n$ have a joint density $L(\theta, \underbrace{x}_{(x_1, \dots, x_n)})$, $\theta \in \Theta$. If the distribution is discrete, $L(\theta, x)$ represents the probability of observing the given data as a function of the parameter θ. The Maximum Likelihood Estimator (MLE) of θ is that value of θ that maximizes the likelihood $L(\theta, x)$. That, it makes the observed data "most probable" or "most likely".

Def: (MLE)

The Maximum Likelihood Estimator is the value $\hat{\theta}_{MLE}$ that maximizes the likelihood:

$$\hat{\theta}_{MLE} \in \underset{\theta \in \Theta}{\operatorname{argmax}}\, L(\theta, x)$$

Rk: • In full generality, a MLE might not be unique nor even exist.

• When the sample is i.i.d, we have

$$L(\theta, x) = f_\theta(x_1) \times f_\theta(x_2) \times \cdots \times f_\theta(x_n)$$

• In practice, we often consider the log-likelihood

$$\ell(\theta, x) = \log L(\theta, x) = \sum_{i=1}^{n} \log f_\theta(x_i),$$

which is easier to handle mathematically speaking.

Finding the MLE:

1) Find $\hat{\theta}_{MLE}$ having the "likelihood equations" (or "normal equations", or "score eq.") vanish

$$U_n(\hat{\theta}_{MLE}) := \nabla \ell(\hat{\theta}_{MLE}, x) = \left(\frac{\partial}{\partial \theta_k} \ell(\hat{\theta}_{MLE}, x) \right)_{1 \le k \le \dim \Theta} = 0$$

2) Check that $\hat{\theta}_{MLE}$ is indeed a maximum:

$$H_n(\hat{\theta}_{MLE}) = \nabla^2 \ell(\theta_{MLE}, x) \text{ is negative definite}$$

Ex 1: (Poisson distribution) $\mathbb{P}(X=x) = e^{-\lambda}\frac{\lambda^x}{x!}$, $x \geq 0$ integer.

If $X_1, \dots, X_n$ are iid Poisson(λ) for some $\lambda > 0$, then

$$\ell(\lambda, x) = \sum_{i=1}^{n} (X_i \log\lambda - \lambda - \log X_i!)$$
$$= \log\lambda \times \left(\sum_{i=1}^{n} X_i\right) - n\lambda - \sum_{i=1}^{n} \log X_i!$$

$$\frac{\partial \ell}{\partial \lambda}(\lambda, x) = \frac{1}{\lambda}\sum_{i=1}^{n} X_i - n, \text{ hence } \hat{\lambda}_{MLE} = \frac{\sum_{i=1}^{n} X_i}{n} = \bar{X}_n.$$

One easily check that $\bar{X}_n$ is a maximum by computing $\frac{\partial^2 \ell}{\partial \lambda^2}(\bar{X}_n, x) < 0$.

Ex 2: (Normal model) If $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$,

$\theta = (\mu, \sigma)$

$$L(\mu, \sigma, x) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{X_i - \mu}{\sigma}\right)^2\right)$$

$$\Rightarrow \ell(\mu, \sigma, x) = -n\log\sigma - \frac{n}{2}\log 2\pi - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2$$

The partials with respect to μ and σ are

$$\begin{cases} \dfrac{\partial \ell}{\partial \mu} = \dfrac{1}{\sigma^2} \displaystyle\sum_{i=1}^{n} (X_i - \mu) \\ \dfrac{\partial \ell}{\partial \sigma} = -\dfrac{n}{\sigma} + \dfrac{1}{\sigma^3} \displaystyle\sum_{i=1}^{n} (X_i - \mu)^2 \end{cases}$$

Setting $\frac{\partial \ell}{\partial \mu} = 0$ yields $\hat{\mu}_{MLE} = \overline{X}_n$. Then, setting $\frac{\partial \ell}{\partial \sigma} = 0$ yields

$$\hat{\sigma}_{MLE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2}.$$

Inspecting the hessian $\nabla^2 \ell = \begin{pmatrix} \frac{\partial^2 \ell}{\partial \mu^2} & \frac{\partial^2 \ell}{\partial \mu \partial \sigma} \\ \frac{\partial^2 \ell}{\partial \mu \partial \sigma} & \frac{\partial^2 \ell}{\partial \sigma^2} \end{pmatrix}$ shows we got a maximum.

Consistency of the MLE

Prop: The MLE obtained from a n-sample $X_1, \dots, X_n \overset{iid}{\sim} f_{\theta^*}$ is denoted $\hat{\theta}_{MLE}$.

Assume that:
- (A1) The model is identifiable (i.e. $\theta \mapsto f_\theta$ is one-to-one)
- (A2) Θ is compact, and $\forall x$, $\theta \mapsto f_\theta(x)$ is continuous
- (A3) Writing $g(x) = \sup_{\theta \in \Theta} |\log f_\theta(x)|$, for all $\theta \in \Theta$, $g \in L^1(\theta)$

Then, $\hat{\theta}_{MLE}$ is consistent: $\hat{\theta}_{MLE} \xrightarrow[n \to \infty]{\mathbb{P}} \theta^*$

Idea of the proof:

- (A3) $\log f_\theta(X_1)$ is integrable, so from the LLN,

$$\frac{1}{n}\log L(\theta, x) = \frac{1}{n}\sum_{i=1}^{n} \log f_\theta(X_i) \xrightarrow[n\to\infty]{a.s} \mathbb{E}_{\theta^*}\left[\log f_\theta(X_1)\right]$$

- (A1) Since the model is identifiable, θ^* is the unique maximum of

$$\theta \longmapsto K(\theta) := \mathbb{E}_{\theta^*}\left[\log f_\theta(X_1)\right]$$

- (A2 + A3) $\theta \mapsto f_\theta(x)$ continuous, Θ compact and $g \in L^1$, so the above convergence is uniform, and hence $\hat{\theta}_{MLE}$, which maximizes $L(\theta, x)$ converges towards the unique value θ^* maximizing the limiting function $K(\theta)$.

Rk: $\hat{\theta}_{MLE}$ can be shown to be asymptotically unbiased, but it may be biased for finite n.
(Exc: $\hat{\sigma}^2_{MLE}$ in the Gaussian model)

Notation:

$$\boxed{I_1(\theta) = -\mathbb{E}_\theta\left[\nabla^2 \log f_\theta(X_1)\right]}$$

Fisher information matrix

(6) Asymptotic Normality of the MLE

Prop: If $\hat{\theta}_{MLE}$ is consistent, that the model is regular (means the likelihood is "nice" analytically) and that $I_1(\theta^*)$ is invertible, then

$$\sqrt{n}\left(\hat{\theta}_{MLE} - \theta^*\right) \xrightarrow[n\to\infty]{\mathcal{D}} N\left(0, I_1(\theta^*)^{-1}\right)$$

Idea of the Proof:

- Recall that the score is $U_n(\theta^*) = \nabla \ell(\theta^*, x)$. Then we have

$$\frac{1}{n} U_n(\theta^*) = \frac{1}{n} \sum_{i=1}^{n} \nabla \log f_{\theta^*}(X_i) \xrightarrow[n\to\infty]{\text{a.s. (LLN)}} E_{\theta^*}\left[U_1(\theta^*)\right] = 0.$$

Furthermore, since $\mathrm{Var}_{\theta^*}\left(U_1(\theta^*)\right) = I_1(\theta^*)$, the CLT yields

$$\sqrt{n}\left(\frac{U_n(\theta^*)}{n}\right) = \frac{1}{\sqrt{n}} U_n(\theta^*) \xrightarrow[n\to\infty]{\mathcal{D}} N\left(0, I_1(\theta^*)\right)$$

- The Hessian matrix $H_n(\theta^*) = \nabla^2 \ell(\theta^*, x)$ satisfies

$$\frac{1}{n} H_n(\theta^*) \xrightarrow[n\to\infty]{\text{a.s. (LLN)}} -I_1(\theta^*)$$

- Taylor Expansion:

$$\sqrt{n}\,\underbrace{\frac{U_n(\theta^*)}{n}}_{\xrightarrow[n\to\infty]{\mathcal{D}} N(0, I_1(\theta^*))} \simeq \sqrt{n}\,\underbrace{\frac{U_n(\hat{\theta}_{MLE})}{n}}_{=0} + \underbrace{\frac{H_n(\bar{\theta}_n)}{n}}_{\to -I_1(\theta^*)} \times \sqrt{n}\left(\theta^* - \hat{\theta}_{MLE}\right)$$

Prop: (Another Formulation)

If one can find $\hat{V}_n$ such that $\hat{V}_n I_n^{-1}(\theta^*) \xrightarrow[n\to\infty]{P} I_d$, then

$$V_n^{-\frac{1}{2}}(\hat{\theta}_{MLE} - \theta^*) \xrightarrow[n\to\infty]{\mathcal{D}} N(0, I_d)$$

Proof: Follows straightforwardly from the previous result and Slutsky.

From this result, one can derive confidence regions for parameters of interest.

Corr: (Wald Statistic)

Let A be a $q \times p$ matrix with rank r. If

$$V_n^{-\frac{1}{2}}(\hat{\theta}_n - \theta^*) \xrightarrow[n\to\infty]{\mathcal{D}} N(0, I_{d_p}),$$

then

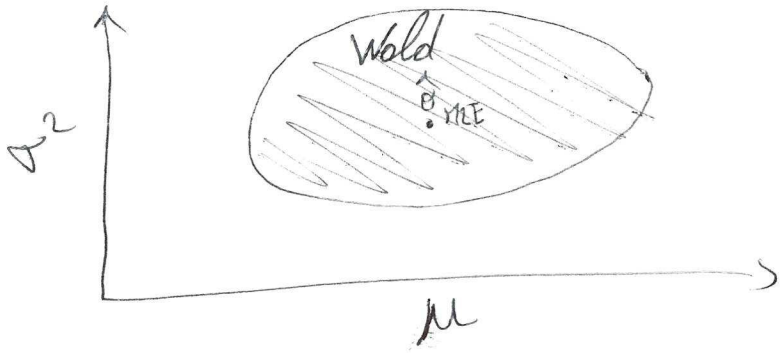
$$W = \left(A(\hat{\theta}_n - \theta^*)\right)^T \left(A V_n A^T\right)^{-1} A(\hat{\theta}_n - \theta^*) \xrightarrow[n\to\infty]{} \chi^2_r$$

Exc: In the Gaussian Model, one gets

$$\frac{n}{\sigma^2}(\bar{X}_n - \mu)^2 + \frac{n}{2\sigma^2}(S_n^2 - \sigma^2)^2 \xrightarrow[n\to\infty]{\mathcal{D}} \chi^2_2$$

from which we get the Asymptotic confidence region, at level $1-\alpha$, for $\theta = (\mu, \sigma^2)$,

$$\left\{ \frac{n}{\sigma^2}(\bar{X}_n - \mu)^2 + \frac{n}{2\sigma^2}(S_n^2 - \sigma^2)^2 \leq \chi^2_{1-\alpha,2} \right\}$$

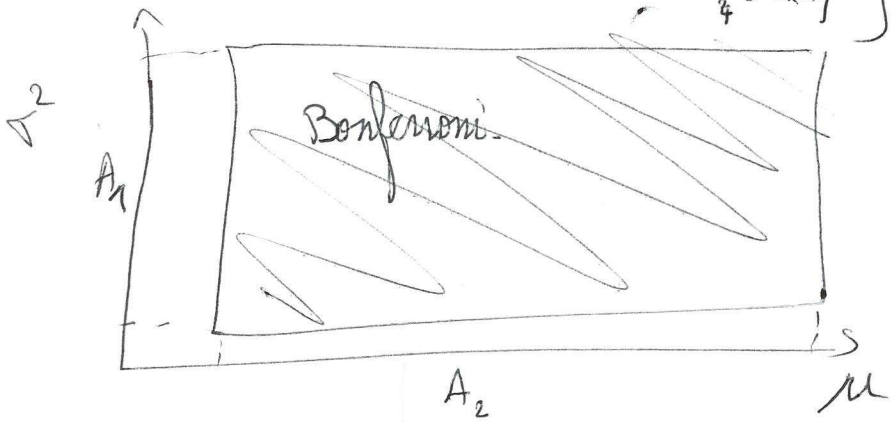


= Simultaneous region of confidence

This is to be compared to the Bonferroni confidence region $A_1 \times A_2$, where

$$\left\{ IC_{1-\frac{\alpha}{2}}(\sigma^2) = A_1 = \left\{ \frac{(n-1)\hat{\sigma}^2}{\chi_{1-\frac{\alpha}{4},n-1}} \leq \sigma^2 \leq \frac{(n-1)\hat{\sigma}^2}{\chi_{\frac{\alpha}{4},n-1}} \right\} \right.$$

$$\left. IC_{1-\frac{\alpha}{2}}(\mu) = A_2 = \left\{ (\bar{X}_n - \mu)^2 \leq \hat{\sigma}^2 \left| \right.^2_{1-\frac{\alpha}{4},n-1} \right\} \right.$$



Inferring functionals of θ

Sometimes, the MLE may help you estimate functionals of θ^*, such that $\theta^{*2}, e^{\theta^*}, \dots$

(Delta Method)

Prop: If h is a function of the variable θ that is differentiable, with differential $Dh(\theta)$, and that $\hat{\theta}_{MLE}$ is the MLE, then $h(\hat{\theta}_{MLE})$ is a MLE of $h(\theta)$, and

$$\sqrt{n}\left(h(\hat{\theta}_{MLE}) - h(\theta^*)\right) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}\left(0, Dh(\theta^*)\, I_1^{-1}(\theta^*)\, Dh(\theta^*)^T\right)$$

Ex: Asymptotic distribution of the estimator $\bar{X}_n(1-\bar{X}_n) = \hat{p}(1-\hat{p})$ of the variance $p(1-p)$ of a Bernoulli distribution.

Rk: The Delta-method is not limited to the MLE. It can be applied to any asymptotically normal estimator:

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}(0, V) \overset{\delta\text{-method}}{\Rightarrow} \sqrt{n}\left(h(\hat{\theta}_n) - h(\theta)\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, Dh(\theta) V Dh(\theta)^T)$$

(Generalized) Likelihood Ratio Tests

Review: building a test

- Define the model
- Define the null hypothesis H_0 and the alternative H_1
- Choose a test statistic $T(X_1, \dots, X_n)$, and determine its distribution under H_0
- Specify the decision rule by calibrating a rejection region $\mathcal{R}$, depending on the risk α.
- (Potential) computation of the power $1-\beta$.
- Calculation of the observed statistic and decision: rejection of H_0 or not.

General Setting: If $X_1, \dots, X_n \sim_{iid} f_\theta$ for some $\theta \in (H)$,

$$\boxed{H_0 : \theta \in (H)_0 \quad \text{(VS)} \quad H_1 : \theta \in (H)_1}$$

where $(H)_0, (H)_1 \subset (H)$,

Properties of a test:

- The power is (a function of θ) defined by

$$\pi(\theta_1) = 1 - \beta(\theta_1) = \mathbb{P}_{\theta_1}(\text{reject } H_0)$$
$$= \mathbb{P}_{\theta_1}(T \in \mathcal{R}) \quad , \text{ for } \theta_1 \in (H)_1$$

- The type I error is (a function of θ) defined by

$$\alpha(\theta_0) = \mathbb{P}_{\theta_0}(T \in \mathcal{R}) \quad , \text{ for } \theta_0 \in (H)_0$$

- The size of the test is $\sup_{\theta_0 \in (H)_0} \alpha(\theta_0)$. A test has level α if it has size $\leq \alpha$.

- A test is unbiased if $\inf_{\theta_1 \in (H)_1} (1 - \beta(\theta_1)) \geq \alpha$.

- A test is consistent if $1 - \beta(\theta_1) \xrightarrow[n \to \infty]{} 1$

Rk: Given two tests, one should pick the most powerful

Likelihood Ratio tests for Simple Hypotheses

We first focus on the test

$$H_0 : \theta = \theta_0 \quad \text{(VS)} \quad H_1 : \theta = \theta_1$$

Since $L(\theta, x)$ represents how likely θ is given x, it is natural to design a decision procedure based on the comparison of $L(\theta_0, x)$ and $L(\theta_1, x)$.

More precisely we want to design a rejection region R with:

- level α. This means that $\int_R L(\theta_0, x)\, dx = \alpha$

- maximal power. This means that

$$\pi = 1 - \beta = \int_R L(\theta_1, x)\, dx$$

$$= \int_R \frac{L(\theta_1, x)}{L(\theta_0, x)} L(\theta_0, x)\, dx$$

must be maximum.

This suggests to pick R to be the region where $\frac{L(\theta_1, x)}{L(\theta_0, x)}$ is large

optimal means it has best possible power.

⚠ We have not defined optimality properly. See Lehman & Romano if interested.

Theorem (Neyman-Pearson)

The optimal rejection region for the test $H_0: \theta = \theta_0$ (VS) $H_1: \theta = \theta_1$ is

$$\mathcal{R} = \left\{ x \;\middle|\; \frac{L(\theta_1, x)}{L(\theta_0, x)} > k_\alpha \right\},$$

where k_α is defined by $P_{\theta_0}\left(\frac{L(\theta_1, X)}{L(\theta_0, X)} > k_\alpha \right) = \alpha$

Likelihood Ratio Tests for Composite Hypotheses

We now move to the more general setting where

$$H_0: \theta \in \textcircled{H}_0 \qquad H_1: \theta \in \textcircled{H}_1$$

where $\textcircled{H}_0, \textcircled{H}_1 \subset \textcircled{H}$ are actual sets (with ≥ 1 elements).

Generalizing the idea above, we would reject H_0 if <u>there exists</u> some $\theta_1 \in \textcircled{H}_1$ that is more likely than <u>any</u> $\theta_0 \in \textcircled{H}$. In other words, we'd tend to reject if

$$\frac{\max_{\theta \in \textcircled{H}_0} L(\theta_0, x)}{\max_{\theta_1 \in \textcircled{H}_1} L(\theta_1, x)} \quad \text{is small}$$

It is preferable (for certain technical reasons) to use the test statistic $\max_{\theta_0 \in H_0} L(\theta_0, x)$ instead.

$$\frac{\max_{\theta_0 \in H_0} L(\theta_0, x)}{\max_{\theta \in H} L(\theta, x)}$$

Def: ((Generalized) Likelihood Ratio Statistic)

$$\Lambda = \frac{\max_{\theta_0 \in H_0} L(\theta_0, x)}{\max_{\theta \in H} L(\theta, x)}$$

$\Lambda = \Lambda(x)$ only depends on data x.

The likelihood ratio test then consists rejecting H_0 when the observed value of Λ is too small.

Def: ((Generalized) Likelihood Ratio Test)

The test is defined by the rejection region $R = \{x \mid \Lambda(x) < k_\alpha\}$, where k_α is such that $\sup_{\theta_0 \in H_0} P_{\theta_0}(\Lambda < k_\alpha) = \alpha$,

Rk: The GLRT does not benefit particular optimality properties (neither does the MLE), but in common settings we see it is unbiased.

Here comes a result that specifies how to pick k_α for the rejection region.

Thm: (Asymptotic distribution of Λ) (or affine)

Wilks Theorem

- Assume that Θ_0 is a linear subspace of Θ. Then under smoothness conditions on the MLE, under H_0 we have

$$-2\log(\Lambda) \xrightarrow[n\to\infty]{\mathcal{D}} \chi^2_r,$$

where $r = \dim \Theta - \dim \Theta_0$.

- Furthermore, the region region $R = \left\{ -2\log(\Lambda) \geq \chi^2_{1-\alpha, r} \right\}$ of the GLRT has asymptotic level α.

Exc 1: Testing a Normal Mean μ in the model $N(\mu, \sigma^2)$, μ unknown, σ^2 known

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

We have: $\theta = \mu$, $\Theta_0 = \{\mu_0\}$, $\Theta_1 = \mathbb{R} \setminus \{\mu_0\}$, $\Theta = \mathbb{R}$

(16)

• Since $\Theta_0 = \{\mu_0\}$ is a singleton, the numerator of Λ is

$$\frac{1}{(\sigma\sqrt{2\pi})^n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu_0)^2\right)$$

• For the denominator, one easily checks that $\hat{\mu}_{MLE} = \bar{X}_n$. Hence, we get

$$\frac{1}{(\sigma\sqrt{2\pi})^n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\bar{X}_n)^2\right),$$

yielding

$$\Lambda = \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(X_i-\mu_0)^2 + \sum_{i=1}^{n}(X_i-\bar{X}_n)^2\right)\right)$$

or equivalently

$$-2\log\Lambda = \frac{1}{\sigma^2}\left(\sum_{i=1}^{n}(X_i-\mu_0)^2 - \sum_{i=1}^{n}(X_i-\bar{X}_n)^2\right).$$

Here, $\dim \Theta_0 = 0$ and $\dim \Theta = 1 = \mathbb{R}$, so from Wilks,

Single point

$$-2\log\Lambda \xrightarrow[n\to\infty]{\mathcal{D}} \chi^2_{1-0},$$

from which we derive the rejection region $R = \left\{-2\log\Lambda \geq \chi^2_{1-\alpha,1}\right\}$

Rk: Using the identity $\sum (X_i - \mu_0)^2 = \sum (X_i - \bar{X}_n)^2 + n(\bar{X}_n - \mu_0)^2$, one can actually simply the expression to $-2\log\Lambda = \frac{n}{\sigma^2}(\bar{X}_n - \mu_0)^2$. Hence, in this case,

- We recover the usual Gaussian test
- $-2\log\Lambda$ has exact distribution χ^2_1 for all n.

Ex 2 GLRT for the Multinomial Distribution.

We'd like to test if data having multinomial distribution satisfies certain conditions, parametrized by $\theta_0 \in \Theta_0$. Write $p = (p_1, \dots, p_k)$ for the (unknown) parameter of the multinomial. (Θ_0 with $\dim \Theta_0 = s$; k bins)

$$H_0 : p = p(\theta_0) \text{ for some } \theta_0 \in \Theta_0 \quad \text{(VS)} \quad H_1 : p \notin \{p(\theta_0), \theta_0 \in \Theta_0\}$$

- The numerator of Λ is $\max_{\theta_0 \in \Theta_0} \frac{n!}{x_1! \cdots x_k!} p_1(\theta)^{x_1} \times \cdots \times p_k(\theta)^{x_k}$,

where the x_i's are the counts in the k bins. By definition, this likelihood is maximized for $\theta = \hat{\theta}_{MLE}$. The corresponding probabilities are $p_i(\hat{\theta}_{MLE})$

- Since for the denominator, $(P_1, \cdots, P_k)$ is unrestricted, the maximizer is found to be $\hat{P}_i = \frac{x_i}{n}$, $1 \le i \le k$.

Hence,

$$\Lambda = \frac{\frac{n!}{x_1! \cdots x_k!} P_1(\hat{\theta}_{MLE})^{x_1} \cdots P_k(\hat{\theta}_{MLE})^{x_k}}{\frac{n!}{x_1! \cdots x_k!} \hat{P}_1^{x_1} \cdots \hat{P}_k^{x_k}}$$

$$= \prod_{i=1}^{k} \left(\frac{P_i(\hat{\theta}_{MLE})}{\hat{P}_i} \right)^{x_i}.$$

Also, since $x_i = n\hat{P}_i$,

$$-2 \log \Lambda = -2n \sum_{i=1}^{k} \hat{P}_i \log \left(\frac{P_i(\hat{\theta}_{MLE})}{\hat{P}_i} \right)$$

$$= 2 \sum_{i=1}^{k} O_i \log \left(\frac{O_i}{E_i} \right),$$

where $O_i = n\hat{P}_i$ and $E_i = nP_i(\hat{\theta}_{MLE})$ denote the observed and expected counts respectively.

→ Since $\sum_{i=1}^{k} P_i = 1$, $\dim \Theta = k - 1$

→ By assumption, $\dim \Theta_0 = s$ ← Represents the number of parameters to estimate under H_0!

From Wilks, under H_0,

$$-2\log\Lambda = 2\sum O_i \log\left(\frac{O_i}{E_i}\right) \xrightarrow[n\to\infty]{\mathcal{D}} \chi^2_{t-1-s}$$

Rk: - We derive the region of rejection accordingly

- Doesn't it look like a result seen in Chapter 10? What about χ^2-test in this setting?

→ Let's find out!

Recall that $\chi^2 = \sum_{i=1}^{t} \frac{(O_i - E_i)^2}{E_i}$. Under the null, $\hat{p}_i \simeq p_i(\theta_{MLE})$, or equivalently $O_i \simeq E_i$.

But the Taylor expansion of $g(x) = x\log\left(\frac{x}{x_0}\right)$ about x_0 is

$$g(x) = (x-x_0) + \frac{1}{2}(x-x_0)^2 \cdot \frac{1}{x_0} + \dots$$

Thus,

$$-2\log\Lambda \simeq 2\sum_{i=1}^{t}(O_i - E_i) + \sum_{i=1}^{t}\frac{(O_i-E_i)^2}{E_i},$$

and since $\sum_{i=1}^{t} O_i = \sum_{i=1}^{t} E_i = n$, we get

$$-2\log\Lambda \approx \sum_{i=1}^{t} \frac{(O_i - E_i)^2}{E_i} = \chi^2.$$

We have argued for the approximate equivalence of the $\chi^2_{\text{Pearson's}}$ test and the GLRT in this example.

Rk: - Pearson's χ^2 has been more commonly used than the likelihood ratio test, because it is somewhat easier to calculate without the use of a computer.