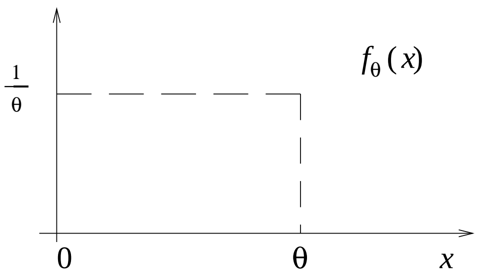
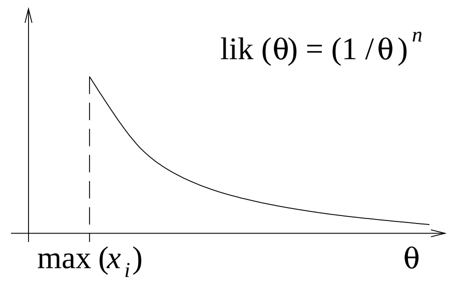
Lecture 18: MLE and Goodness-of-fit

Example of MLE

$X_1, \dots, X_n$ i.i.d. Uniform $[0, \theta]$. Find MLE of θ.





$$\text{lik}(\theta) = \prod_{i=1}^{n} f_\theta(x_i) = \begin{cases} \left(\frac{1}{\theta}\right)^n & x_i \le \theta \text{ all } i \\ 0 & \text{otherwise} \end{cases}$$
$$= \begin{cases} \left(\frac{1}{\theta}\right)^n & \max\,(x_i) \le \theta \\ 0 & \text{otherwise} \end{cases}$$

$\text{lik}(\theta)$ is maximized at $\hat{\theta} = \max(x_i)$ for $1 \le i \le n$. No differentiability for $\text{lik}(\theta)$ at $\hat{\theta}$. It can be shown that $\sqrt{n}(\hat{\theta} - \theta) \longrightarrow$ distribution is not Normal.

The uniform case is rather special. Most of the time, the MLE has an approximate Normal distribution at $\sqrt{n}$-rate when n is large. This generalizes the fact that the sample mean has an approximate Normal distribution at $\sqrt{n}$-rate (CLT), and we have seen a few cases where the MLE is the sample mean.

Result: Let

$$I(\theta) = E_\theta \left\{ \frac{\partial}{\partial\theta} \log f(x|\theta) \left[\frac{\partial}{\partial\theta} \log f(x|\theta) \right]^T \right\}$$

be the Fisher information Matrix. $\frac{\partial}{\partial\theta} \log f(x|\theta)$ is the partial derivative vector of the log-likelihood function. Assume $X_1, \dots, X_n$ i.i.d. $\sim f(x|\theta_0) \longleftarrow$ "**nice**". Then

$$\sqrt{n}(\hat{\theta} - \theta_0) \longrightarrow N\left(0, I^{-1}(\theta_0)\right)$$

(Rice. p. 264 Theorem B for 1-dimensional case)

In 1-dimensional case,

$$\hat{\theta} \sim N\left(\theta_0, \frac{1}{nI(\theta_0)}\right)$$

95% confidence interval for θ,

$$\hat{\theta} \pm 2\frac{1}{\sqrt{n}\sqrt{I(\theta_0)}}$$

θ_0 unknown, use

$$\hat{\theta} \pm 2\frac{1}{\sqrt{n}\sqrt{I(\hat{\theta})}}$$

Example:

$$\begin{aligned}
f(x|\theta) &= \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad \theta = (\mu, \sigma^2) \\
\log f(x|\theta) &= -\frac{1}{2}\log\sigma^2 - \frac{1}{2}\log 2\pi - \frac{(x-\mu)^2}{2\sigma^2} \\
\frac{\partial}{\partial\mu}\log f(x|\theta) &= \frac{(x-\mu)}{\sigma^2} \\
\frac{\partial}{\partial\sigma^2}\log f(x|\theta) &= -\frac{1}{2\sigma^2} + \frac{(x-\mu)^2}{2\sigma^4}
\end{aligned}$$

$$\left[\frac{\partial}{\partial\theta}\log f(x|\theta)\right]\left[\frac{\partial}{\partial\theta}\log f(x|\theta)\right]^T = \begin{bmatrix} \frac{(x-\mu)^2}{\sigma^4} & -\frac{(x-\mu)}{2\sigma^4} + \frac{(x-\mu)^3}{2\sigma^6} \\ -\frac{(x-\mu)}{2\sigma^4} + \frac{(x-\mu)^3}{2\sigma^6} & \left(-\frac{1}{2\sigma^2} + \frac{(x-\mu)^2}{2\sigma^4}\right)^2 \end{bmatrix}$$

Fact about Normal (μ, σ^2): If $X \sim N(\mu, \sigma^2)$ then

$$E\{[X-\mu]^k\} = \begin{cases} 0, & k \text{ odd} \\ (k-1)E[X-\mu]^{k-2}\sigma^2, & k \text{ even} \end{cases}$$

so $E[X-\mu]^3 = 0$ and $E[X-\mu]^4 = 3\sigma^4$

$$I(\theta) = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{4\sigma^4} - \frac{\sigma^2}{2\sigma^6} + \frac{3\sigma^4}{4\sigma^8} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix}$$

i.e. $(\bar{X}, \hat{\sigma}^2)$ independent. Therefore we have,

$$\begin{pmatrix} \bar{X} \\ s^2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}, \frac{1}{n}\begin{pmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{pmatrix}\right)$$

95% confidence intervals for μ and σ^2 are respectively:

$$\begin{aligned}
\bar{X} &\pm 2\frac{1}{\sqrt{n}}\hat{\sigma} \\
\hat{\sigma}^2 &\pm 2\frac{1}{\sqrt{n}}\sqrt{2}\hat{\sigma}^2
\end{aligned}$$

In 1-dimensional case, there is another way to calculate $I(\theta)$:

$$I(\theta) = E_\theta \left[-\frac{\partial^2}{\partial \theta^2} \log f(X|\theta) \right]$$

Asymptotic Optimality of MLE

In "nice" cases, it has been shown that there is no other estimator of θ which has a smaller asymptotic variance smaller than $n^{-1}I^{-1}(\theta_0)$. So MLE is the "best" when n is large.

Goodness of fit Test

We now have a good method to find an estimator of parameters, assuming the model $\{f(x|\theta)\}$ is good. How do we check whether the model fits the data or not?

Pearson's χ^2-test for goodness of fit of the model:

- Divide data into m "cells". The cells can be natural or aggregated if data are counts to start with; in the continuous case, the possible value domain should be divided into intervals.
- The counts of data in the m cells $(O_1, \ldots, O_m)$ is Multinomial. Estimate the expected cell counts using the model $f(x|\theta)$ and MLE, $\hat{\theta}$.

$$p_j(\hat{\theta}) = P(X \text{ in } j^{th} \text{ cell } |\hat{\theta}) \qquad E_j = np_j(\hat{\theta})$$

- Form χ^2 for each cell: $\dfrac{(O_j - E_j)^2}{E_j}$
- χ^2-statistic: $\chi^2 = \sum_{j=1}^{m} \dfrac{[O_j - nP_j(\hat{\theta})]^2}{nP_j(\hat{\theta})}$
- H_0: $X_1, \ldots, X_n$ come from $f(x|\theta)$
  H_1: not
- Under H_0, χ^2 has a χ^2 distribution with m − # independent parameters fitted − 1 degrees of freedom. The χ^2-distribution is approximate.

Rule of thumb: This approximation is good when each cell should have at least five expected counts. What is a χ^2 distribution with d-degrees of freedom?

Let Z_1, ..., Z_d i.i.d. $N(0,1)$ then

$$\sum_{i=1}^{d} Z_i^2 \sim \chi_d^2 \qquad \Longrightarrow \qquad E\chi_d^2 = d, \;\; \text{Var}(\chi_d^2) = 2d$$

Example: $X_1, \ldots, X_n$ i.i.d. $N(\mu, \sigma^2)$ then

$$\frac{\sum(X_i - \bar{X})^2}{\sigma^2} \sim \chi^2_{n-1}$$

because we can find $Z_1, \ldots, Z_{n-1}$ such that

$$\frac{\sum(X_i - \bar{X})^2}{\sigma^2} = \sum_{i=1}^{n-1} Z_i^2, \qquad Z_i \text{ i.i.d. } N(0,1)$$

In our case, asymptotically, we can find $Z_1, \ldots, Z_d$, $d = m -$ #independent $- 1$ such that

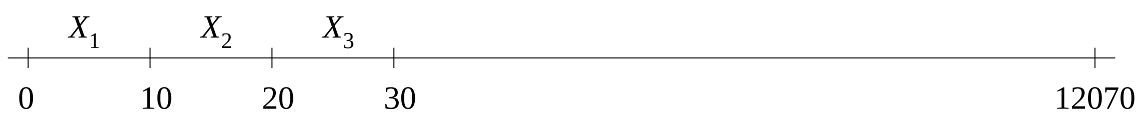
$$\sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i} \cong \sum_{i=1}^{d} Z_i^2$$

Example: Rice 8.2 p. 240-241

$X_1, \ldots, X_n$ i.i.d. Poisson(λ) where $X_i = \#$ counts in the i^{th} 10 second interval,
$m = 16$, $n = 1207$ intervals.

Cells:	1st	$0 \le X \le 2$
	2nd	$X = 3$
	$\vdots$	$\vdots$
	16th	$X \ge 17$

MLE $\hat{\lambda} = \frac{1}{n} \sum X_i = 8.392$ per 10 seconds.



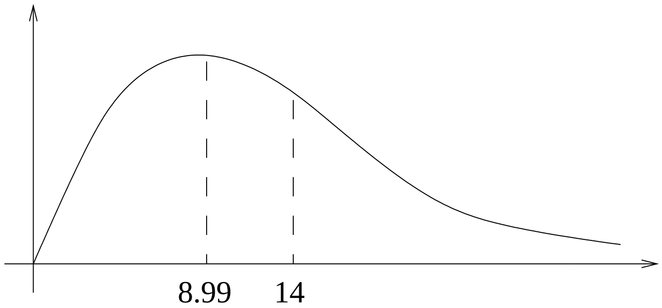
$n =$ 1207 intervals

Formulating the χ^2 statistic is tabular form gives

cell #	obs. count, O	exp. count, E	χ^2
1	18	$1207 \times e^{-\hat{\lambda}}[1 + \hat{\lambda} + \frac{\hat{\lambda}^2}{2}] = 12.2$	$\frac{(18-12.2)^2}{12.2} = 2.76$
2	28	$1207 \times e^{-\hat{\lambda}} \frac{\hat{\lambda}^3}{3!} = 27.0$	$\frac{(28-27.0)^2}{27.0} = 0.04$
3	58	$1207 \times e^{-\hat{\lambda}} \frac{\hat{\lambda}^4}{4!} = 56.5$	$\frac{(58-56.5)^2}{56.5} = 0.01$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
	1207	1207	8.99

$d = 16 - 1 - 1 = 14$ as there is 1 parameter in the Poisson model. Is 8.99 unusual for χ^2_{14}?

From χ^2 table,



70% < P−value < 90%

Accept H_0

i.e. Poisson model fits well