Mathematical Statistics

Donghyun Park

February 15, 2026

1 Distributions

1.1 Table of distributions

Distribution	**PDF**	**MGF** $M_X(t)$	**CGF** $K_X(t)$	**Mean**	**Variance**
Binomial	$\binom{n}{x}p^x(1-p)^{n-x}$	$(pe^t+1-p)^n$	$n\log(1-p+pe^t)$	np	$np(1-p)$
Negative Binomial	$\binom{x-1}{r-1}p^r(1-p)^{x-r}$	$(\frac{pe^t}{1-(1-p)e^t})^r$	$r\log(\frac{pe^t}{1-(1-p)e^t})$	$\frac{r}{p}$	$\frac{r(1-p)}{p^2}$
Geometric	$p(1-p)^{x-1}$	$\frac{pe^t}{1-(1-p)e^t}$	$\log p+t-\log(1-(1-p)e^t)$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Poisson	$\frac{e^{-\lambda}\lambda^x}{x!}$	$e^{\lambda(e^t-1)}$	$\lambda(e^t-1)$	λ	λ
Multinomial	$\frac{n!}{x_1!\cdots x_k!}p_1^{x_1}\cdots p_k^{x_k}$	$(\sum p_j e^{t_j})^n$	$n\log(\sum p_j e^{t_j})$	$n\mathbf{p}$	$[np_i(\delta_{ij}-p_j)]$
Gamma	$\frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}$	$(1-t/\beta)^{-\alpha}$	$-\alpha\log(1-t/\beta)$	$\frac{\alpha}{\beta}$	$\frac{\alpha}{\beta^2}$
Beta	$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$	$1+\sum_{k=1}^{\infty}\frac{t^k}{k!}\frac{(\alpha)_k}{(\alpha+\beta)_k}$	*No simple form*	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
Normal	$\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	$e^{\mu t+\frac{1}{2}\sigma^2t^2}$	$\mu t+\frac{1}{2}\sigma^2t^2$	μ	σ^2
Chi-square (χ_k^2)	$\frac{1}{2^{k/2}\Gamma(k/2)}x^{\frac{k}{2}-1}e^{-\frac{x}{2}}$	$(1-2t)^{-k/2}$	$-\frac{k}{2}\log(1-2t)$	k	$2k$
Student's t	$\frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}(1+\frac{x^2}{\nu})^{-\frac{\nu+1}{2}}$	*Does not exist*	*Does not exist*	0 $(\nu>1)$	$\frac{\nu}{\nu-2}$ $(\nu>2)$
F-distribution	$\frac{\Gamma((r_1+r_2)/2)}{\Gamma(r_1/2)\Gamma(r_2/2)}\left(\frac{r_1}{r_2}\right)^{r_1/2}(1+r_1x/r_2)^{-(r_1+r_2)/2}$	*Does not exist*	*Does not exist*	$\frac{r_2}{r_2-2}$	$\frac{2r_2^2(r_1+r_2-2)}{r_1(r_2-2)^2(r_2-4)}$
Dirichlet	$\frac{1}{B(\boldsymbol{\alpha})}\prod_{i=1}^{K}x_i^{\alpha_i-1}$	*Multivariate Form*	*Multivariate Form*	$\frac{\alpha_i}{\sum\alpha_k}$	*Covariance Matrix*

1.2 Multivariate Normal

We say $X \sim N_n(\mu, \Sigma)$ if it has pdf:

$$pdf(x) = \frac{1}{\sqrt{\det(2\pi\Sigma)}}\exp\{-(x-\mu)^t\Sigma^{-1}(x-\mu)/2\}$$

$X \overset{d}{=} \Sigma^{1/2}Z + \mu$ where $Z \sim N_n(0, I)$.

$$M_X(t) = \exp\left(i\mu^t t - \frac{1}{2}t^t\Sigma t\right)$$

Similar to the relation of chi-square distribution,

$$(X-\mu)^t\Sigma^{-1}(X-\mu) \sim \chi^2(n)$$

1.3 Motivations of distribution

Chi-square distribution : $Y \sim \chi^2(r)$ by $Y \overset{d}{=} X_1^2 + \cdots + X_r^2$, $X_i \overset{iid}{\sim} N(0,1)$.
Student's t distribution : $X \sim t(r)$ by $X \overset{d}{=} \frac{Z}{\sqrt{V/r}}$, $Z \sim N(0,1)$, $V \sim \chi^2(r)$.
F distribution : $X \sim F(r_1, r_2)$ by $X \overset{d}{=} \frac{V_1/r_1}{V_2/r_2}$, $V_i \sim \chi^2(r_i)$ independent.

1.4 Statistics following the distribution

1.4.1 $X_1, \cdots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$

Following holds : (1) $\bar{X} \sim N(\mu, \sigma^2/n)$, (2) $(n-1)S^2/\sigma^2 \sim \chi^2(n-1)$, (3) $\bar{X}, S^2$ are independent

Proof. (1) $mgf_{\bar{X}}(s) = mgf_{X_1}(s/n) \cdots mgf_{X_n}(s/n) = \exp\{\mu s + (\sigma^2/n)s^2/2\}$
(3) $(X_1 - \bar{X}, \cdots, X_n - \bar{X})$ and $\bar{X}$ are independent by calculation of mgf.
(2) Below equality,

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X} - \mu)^2$$

each follows $\chi^2(n)$ and $\chi^2(1)$, by mgf analysis $(n-1)S^2/\sigma^2 \sim \chi^2(n-1)$ □

(4) $\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$
From (4), we can get the following inequality.

$$\mathbb{P}\left\{\left|\frac{\bar{X}-\mu}{S/\sqrt{n}}\right| \leq t_{\alpha/2}(n-1)\right\} = 1 - \alpha$$

From (2), we can get the following inequality.

$$\mathbb{P}\left\{\chi^2_{1-\alpha/2}(n-1) \leq (n-1)S^2/\sigma^2 \leq \chi^2_{\alpha/2}(n-1)\right\} = 1 - \alpha$$

1.4.2 $X_{11}, \cdots, X_{1n_1} \sim N(\mu_1, \sigma_1^2)$, $X_{21}, \cdots, X_{2n_2} \sim N(\mu_2, \sigma_2^2)$

Then the statistic

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

follows the $F(n_1 - 1, n_2 - 1)$ distribution.

1.4.3 One way classification model

We assume $X_{ij} = \mu_i + e_{ij}$ where $e_{ij} \overset{iid}{\sim} N(0, \sigma^2)$
The goal is to compare means. Then following holds.

$$\sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X} - (\mu_i - \bar{\mu}))^2/\sigma^2 \sim \chi^2(k-1)$$

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2/\sigma^2 \sim \chi^2(n-k)$$

Above two terms are independent. Here, $\bar{X}_i$ be ith class mean, $\bar{X}$ a overall mean, $\bar{\mu}$ is average of μ_i with weight n_i.
Thus, defining the following statisics, it follows F-distribution.

$$\frac{\sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X} - (\mu_i - \bar{\mu}))^2/(k-1)}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2/(n-k)} \sim F(k-1, n-k)$$

1.5 Linear Regression Model

$Y_i = x_{i0}\beta_0 + x_{i1}\beta_1 + \cdots + x_{ip}\beta_p + e_i$, $e_i \overset{iid}{\sim} N(0, \sigma^2)$.
Or, $Y = X\beta + e$, $e \sim N_n(0, \sigma^2 I)$. Defining $\hat{\beta} = (X^tX)^{-1}X^tY$ and $\hat{\sigma^2} = (Y - X\hat{\beta})^t(Y - X\hat{\beta})/(n-p-1)$,
(1) $\hat{\beta} \sim N_{p+1}(\beta, \sigma^2(X^tX)^{-1})$
(2) $(n-p-1)\hat{\sigma^2}/\sigma^2 \sim \chi^2(n-p-1)$
(3) $\hat{\beta}, \hat{\sigma^2}$ are independent.

2 Matrix Equalities

$$(I+ab^t)^{-1} = I + \left(-\frac{1}{1+b^ta}\right)ab^t$$

$$\det\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} = \det(A_{11})\det(A_{22} - A_{21}A_{11}^{-1}A_{12})$$

(If A_{11}^{-1} exists. We notate $A_{22\cdot 1} = A_{22} - A_{21}A_{11}^{-1}A_{12}$)

Proof.

$$\begin{pmatrix} I & 0 \\ -A_{21}A_{11}^{-1} & I \end{pmatrix}\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}\begin{pmatrix} I & -A_{11}^{-1}A_{12} \\ 0 & I \end{pmatrix} = \begin{pmatrix} A_{11} & 0 \\ 0 & A_{22} - A_{21}A_{11}^{-1}A_{12} \end{pmatrix}$$

□

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}^{-1} = \begin{pmatrix} A_{11}^{-1} + A_{11}^{-1}A_{12}A_{22\cdot 1}^{-1}A_{21}A_{11}^{-1} & -A_{11}^{-1}A_{12}A_{22\cdot 1}^{-1} \\ -A_{22\cdot 1}^{-1}A_{21}A_{11}^{-1} & A_{22\cdot 1}^{-1} \end{pmatrix}$$

To look at more matrix equalities, see Multivariate Statistics.

3 Convergences

Converge in distribution : $X_n \xrightarrow{d} X$ if $\lim_{n\to\infty} \mathbb{P}(X_n \le x) = \mathbb{P}(X \le x)$ for all continuity points of X.
Converge in probability : $X_n \xrightarrow{P} X$ if $\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \epsilon) = 0$
Converges almost surely : $X_n \xrightarrow{a.s.} X$ if $\mathbb{P}(\lim_{n\to\infty} X_n = X) = 1$
Two theorems are useful to transfer one convergence result to other.

Theorem 1 (Slutsky theorem)**.** *If $X_n \xrightarrow{d} Z$ and $Y_n \xrightarrow{P} c$ then*
(1) $X_n + Y_n \xrightarrow{d} Z + c$
(2) $X_nY_n \xrightarrow{d} cZ$

Theorem 2 (Delta method)**.** *$X_n = (X_{n1}, \cdots, X_{nk})^t$ with $X = (X_1, \cdots, X_k)^t$ and $\theta = (\theta_1, \cdots, \theta_k)^t$. If*

$$\sqrt{n}(X_n - \theta) \xrightarrow{d} X$$

and $g : \Theta \to \mathbb{R}$ have continuous first order derivatives $(\partial g(\theta)/\partial\theta_1, \cdots, \partial g(\theta)/\partial\theta_k)^t$ then

$$\sqrt{n}(g(X_n) - g(\theta)) \xrightarrow{d} (\nabla(g))^tX$$

If X follows multivariate normal distribution $N(0, \Sigma)$ then

$$\sqrt{n}(g(X_n) - g(\theta)) \xrightarrow{d} N(0, \nabla(g)^t\Sigma\nabla(g))$$

4 Likelihood, MLE

MLE is important in statistics. First, I would like to give some examples.

4.1 Examples of MLE

4.1.1 Poisson Distribution

Since joint pdf is written as

$$\prod_{i=1}^n \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = e^{-n\lambda}\lambda^{x_1+\cdots+x_n}/(x_1!\cdots x_n!).$$

$$\hat{\lambda}^{MLE} = \bar{X}.$$

4.1.2 Bernoulli Distribution

Joint pdf is written as

$$\prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{x_1+\cdots+x_n}(1-p)^{n-x_1-\cdots-x_n}.$$

$$\hat{p}^{MLE} = \bar{X}.$$

4.1.3 Normal Distribution

The log likelihood function is given by

$$l(\mu,\sigma^2) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 - \frac{n}{2}\log\sigma^2 - \frac{n}{2}\log 2\pi.$$

Thus, regardless of σ, $\hat{\mu} = \bar{X}$ minimizes l among other μ's. Moreover, by differentiating, we get

$$(\hat{\mu}^{MLE}, \hat{\sigma^2}^{MLE}) = \Big(\bar{X}, \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2\Big).$$

This can be deduced by following MLE on Exponential family.

4.2 MLE on Exponential family

Theorem 3. *In (multi-parameter) exponential family, MLE is given by the solution satisfying following equation when it satisfies conditions (1), (2), (3).*

$$pdf = \exp\Big\{\sum_{j=1}^{k}\eta_j T_j(x) - A(\eta) + S(x)\Big\}$$

$$\mathbb{E}_\eta[(T_1(X_1),\cdots,T_k(X_1))^t] = \frac{1}{n}\Big(\sum_{j=1}^{n}T_1(X_j),\cdots,\sum_{j=1}^{n}T_k(X_j)\Big)^t$$

(1) The support of distributions do not depend on the parameter space Θ.
(2) Parameter space Θ is opened.
(3) For any vector $c \in \mathbb{R}^k$, $c\cdot(T_1(x),\cdots,T_k(x))$ is not constant.

4.2.1 Bi-variate Normal Distribution

By the theorem above, we get

$$\hat{\mu_1}^{MLE} = \bar{X_1},\ \hat{\mu_2}^{MLE} = \bar{X_2},$$
$$\hat{\sigma_1^2}^{MLE} = \frac{1}{n}\sum_{i=1}^{n}X_{1i}^2 - \bar{X_1}^2,\ \hat{\sigma_2^2}^{MLE} = \frac{1}{n}\sum_{i=1}^{n}X_{2i}^2 - \bar{X_2}^2,$$
$$\hat{\rho}^{MLE} = \frac{\sum_{i=1}^{n}(X_{1i}-\bar{X_1})(X_{2i}-\bar{X_2})}{\sqrt{\sum_{i=1}^{n}(X_{1i}-\bar{X_1})^2}\sqrt{\sum_{i=1}^{n}(X_{2i}-\bar{X_2})^2}}.$$

4.3 Asymptotic properties of MLE

We gain two results: **Consistency of MLE** and **Asymptotic Normal property of MLE**. For general distributions, we require stricter conditions to guaranty each theorem. However, in exponential family, MLE estimator is easily obtained by solving equations, thus more easier to gain.

Theorem 4 ([BD15] Theorem 5.2.2)**.** *Suppose $\mathcal{P}$ is a canonical exponential family of rank d generated by $\mathbf{T}$. Then if $X_1,\cdots,X_n$ are sampled from $\mathbb{P}_\eta \in \mathcal{P}$,*
(1) $\mathbb{P}_\eta[MLE\ exists] \to 1$
(2) $\hat{\eta}$ is consistent.

Theorem 5 ([BD15] Theorem 5.2.3)**.** *Suppose*
(1) For all compact $K \subset \Theta$,

$$\sup\left\{\left|\frac{1}{n}\sum_{i=1}^{n} l(\theta;X_i) - \mathbb{E}_{\theta_0}[\log f(\theta;X_1)]\right| : \theta \in K\right\} \xrightarrow[n\to\infty]{P_{\theta_0}} 0.$$

(2) For some compact $K \subset \Theta$,

$$\mathbb{P}_{\theta_0}\left[\inf\left\{\frac{1}{n}\sum_{i=1}^{n}(l(\theta;X_i) - l(\theta_0;X_i)) : \theta \in K^c\right\} > 0\right] \to 1.$$

(3) We have

$$\inf\{\mathbb{E}_{\theta_0}[\log f(\theta;X_1)] : |\theta-\theta_0| \geq \epsilon\} > \mathbb{E}_{\theta_0}[\log f(\theta_0;X_1)]$$

for every $\epsilon > 0$.
then consistency of MLE hold.

On the other hand, asymptotic normality holds when following conditions are satisfied.

Theorem 6 ([BD15] Theorem 5.3.5)**.** *Suppose $\mathcal{P}$ is a canonical exponential family of rank d generated by $\mathbf{T}$ with $\mathcal{E}$ open. Then if $X_1, \cdots, X_n$ are a sample from $\mathbb{P}_\eta \in \mathcal{P}$ and $\hat{\eta}$ is defined as the MLE (if MLE does not exist, let fixed value)*

$$\hat{\eta} = \eta + \frac{1}{n}\sum_{i=1}^{n}\ddot{A}^{-1}(\eta)(\mathbf{T}(X_i) - \dot{A}(\eta)) + o_{P_\eta}(n^{-1/2})$$

$$\sqrt{n}(\hat{\eta}^{MLE} - \eta) \xrightarrow[n\to\infty]{d} N_d(\mathbf{0}, \ddot{A}^{-1}(\eta))$$

Theorem 7 ([BD15] Theorem 5.4.3)**.** *Suppose $X_1, \cdots, X_n$ are iid sampled from P_θ for $\theta \in \Theta$ which is open set. Assume that*
(A0) $\dot{l}(\theta;x)$ is well-defined.
(A1) $t = \theta$ is the unique solution of

$$\int \dot{l}(t;x)dP_\theta = 0,$$

that is,

$$\int |\dot{l}(t;x)|dP_\theta(x) < \infty, \quad t, \theta \in \Theta$$

and $t = \theta$ is the unique solution.
(A2) $\mathbb{E}_\theta(\dot{l}(\theta;X_1))^2 < \infty$ for all $\theta \in \Theta$
(A3) $\ddot{l}(\theta;x)$ exist and have finite expectations,

$$\mathbb{E}_\theta(\ddot{l}(\theta;X_1)) \neq 0$$

(A4) If $\epsilon_n \to 0$,

$$\sup_t\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\left(\ddot{l}(t;X_i) - \ddot{l}(\theta;X_i)\right)\right| : |t-\theta| \leq \epsilon_n\right\} \xrightarrow{P_\theta} 0.$$

(A5) $\hat{\theta}^{MLE} \to \theta$.
(A6) The following information matrix exists.

$$I(\theta) = \mathrm{Var}_\theta(\dot{l}(\theta;X_1)) = -\mathbb{E}_\theta[\ddot{l}(\theta;X_1)]$$

If (A0) to (A6) holds,

$$\hat{\theta}^{MLE} = \theta + \frac{1}{n}\sum_{i=1}^{n} I(\theta)^{-1}\dot{l}(\theta;X_i) + o_P(n^{-1/2})$$

and

$$\sqrt{n}(\hat{\theta}^{MLE} - \theta) \xrightarrow[n\to\infty]{d} N_k(0, [I(\theta)]^{-1})$$

4.4 Information inequality

Theorem 8 (Information inequality, or Cramer-Rao inequality)**.** *Suppose $X_1, \cdots, X_n$ are iid sampled from P_θ for $\theta \in \Theta$ which is open set. Assume (A0) to (A6) holds. Then for the estimator $\hat{\eta}_n = \hat{\eta}_n(X_1, \cdots, X_n)$ of $\eta = \eta(\theta)$,*

$$\mathrm{Var}_\theta(\hat{\eta}_n) \succeq \left(\frac{\partial}{\partial\theta}\mathbb{E}_\theta(\hat{\eta}_n)\right)^t [nI(\theta)]^{-1} \left(\frac{\partial}{\partial\theta}\mathbb{E}_\theta(\hat{\eta}_n)\right).$$

where $A \succeq B$ means $A - B$ be positive semidefinite matrix.

Proof. Motivation comes from Cauchy-Schwartz inequality in one dimensional case.

$$\mathrm{Var}_\theta(\hat{\eta}_n)\mathrm{Var}_\theta(\dot{l}_n(\theta)) \geq \left(\mathrm{Cov}_\theta(\hat{\eta}_n, \dot{l}_n(\theta))\right)^2$$

note that $\mathrm{Var}_\theta(\dot{l}_n(\theta)) = nI(\theta)$. □

As a corollary,

Corollary 1. *Suppose $X_1, \cdots, X_n$ are iid sampled from P_θ for $\theta \in \Theta$ which is open set. Assume (A0) to (A6) holds. Then for the unbiased estimator $\hat{\theta}_n$,*

$$\mathrm{Var}_\theta(\hat{\theta}_n) \succeq \frac{1}{n}I^{-1}(\theta)$$

Fisher's conjecture conjectured any other estimator that have asymptotic distribution, then variant of information inequality holds.

$$\sqrt{n}(T_n - \theta) \xrightarrow[n\to\infty]{d} N(0, V(\theta)),$$
$$V(\theta) \succeq I(\theta)^{-1}.$$

However, Hodge's Superefficient Estimator gives counterexample. Define

$$T_n = \begin{cases} 0 & |\hat{\theta}_n| < n^{-1/4} \\ \hat{\theta}_n & |\hat{\theta}_n| \geq n^{-1/4} \end{cases}$$

then Hodge's estimator dominates at $\theta = 0$.

Theorem 9 (Le Cam's Theorem)**.** *$X_1, \cdots, X_n$ be iid random vectors from $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$. Assume the model satisfies **Local Asymptotic Normality** condition at every $\theta \in \Theta$. Then for any estimator of θ that*

$$\sqrt{n}(T_n - \theta) \xrightarrow[n\to\infty]{d} N(0, \Sigma(\theta))$$

then the set

$$S = \{\theta \in \Theta : I(\theta)^{-1} - \Sigma(\theta) \succeq 0\}$$

is measure zero.

4.5 Wald and Rao statistics

When testing $H_0 : \theta = \theta_0$, MLE testing statistics is

$$2\{l(\hat{\theta}^{MLE}) - l(\theta_0)\}$$

where

$$l(\theta) = \sum_{i=1}^n \log f(X_i; \theta) = \sum_{i=1}^n l(\theta; X_i).$$

By Theorem 7, we can approximate this statistic. Under H_0, by consistency and $\dot{l}(\hat{\theta}^{MLE}) = 0$,

$$l(\theta_0) \approx l(\hat{\theta}^{MLE}) + \dot{l}(\hat{\theta}^{MLE})^t(\theta_0 - \hat{\theta}^{MLE}) + \frac{1}{2}(\theta_0 - \hat{\theta}^{MLE})^t\ddot{l}(\hat{\theta}^{MLE})(\theta_0 - \hat{\theta}^{MLE}).$$

Thus,

$$2\{l(\hat{\theta}^{MLE}) - l(\theta_0)\} \approx \sqrt{n}(\theta_0 - \hat{\theta}^{MLE})^t I(\theta_0)\sqrt{n}(\theta_0 - \hat{\theta}^{MLE})$$

because by Law of large numbers, $-\ddot{l}(\hat{\theta}^{MLE})/n \approx I(\theta_0)$. Therefore,

$$2\{l(\hat{\theta}^{MLE}) - l(\theta_0)\} \xrightarrow[n\to\infty]{d} \chi^2(k).$$

In this computation, the intermediate terms appear, which we call Wald test statistics and Rao test statistics.

$$W_n(\theta_0) = n(\hat{\theta}^{MLE} - \theta_0)^t I(\theta_0)(\hat{\theta}^{MLE} - \theta_0),\ R_n(\theta_0) = \frac{1}{n}\dot{l}(\theta_0)^t[I(\theta_0)]^{-1}\dot{l}(\theta_0)$$

4.5.1 Multibinomial distribution testing

In the distribution $\mathrm{Multi}(1, (p_1, \cdots, p_k)^t)$ and sampled $Z_1, \cdots, Z_n$, we want to test $H_0 : p = p_0$ vs $H_1 : p \neq p_0$. Let $X_i = \sum_{j=1}^n Z_{ij}$ and $\theta = (p_1, \cdots, p_r)^t$ where $r = k - 1$. Moreover, write $\theta. = \theta_1 + \cdots + \theta_r$.
We get

$$l(\theta) = \sum_{i=1}^r x_i \log \theta_i + x_k \log(1 - \theta.)$$

so

$$\hat{\theta}^{MLE} = (x_1/n, \cdots, x_r/n)^t$$

and

$$-\ddot{l}(\theta) = \mathrm{diag}(x_i/\theta_i^2) + \frac{x_k}{(1-\theta.)^2} 11^t,$$

$$I(\theta) = \mathrm{diag}(1/\theta_i) + p_k^{-1} 11^t.$$

Thus, the MLE testing statistics, Rao and Wald statistics are

$$2\{l(\hat{\theta}^{MLE}) - l(\theta_0)\} = 2 \sum_{i=1}^k X_i \log \left(\frac{X_i}{np_{0i}} \right),$$

$$W_n(\theta_0) = \sum_{i=1}^k \frac{(X_i - np_{0i})^2}{np_{0i}},$$

$$R_n(\theta_0) = \sum_{i=1}^k \frac{(X_i - np_{0i})^2}{np_{0i}}.$$

and under the H_0, as $n \to \infty$, these will converge to $\chi^2(k-1)$ distribution.

4.5.2 Independence test

We want to test $H_0 : p_{ij} = p_{i\cdot} p_{\cdot j}$ when $X = (X_{ij}) \sim \mathrm{Multi}(n, (p_{ij}))$.
Log likelihood is

$$l(p) = \sum_{i=1}^r \sum_{j=1}^c x_{ij} \log p_{ij} + (Const)$$

If the null hypothesis is true, then the log likelihood function will be (in this case, not p_{ij} are parameters but $(p_{i\cdot}^0, p_{\cdot j}^0)$ are parameters.

$$l(p^0) = \sum_{i=1}^r x_{i\cdot} \log p_{i\cdot}^0 + \sum_{j=1}^c x_{\cdot j} \log p_{\cdot j}^0 + (Const)$$

We get

$$\hat{p_{ij}} = \frac{x_{ij}}{n},\ \hat{p_{ij}}^0 = \hat{p_{i\cdot}}^0 \times \hat{p_{\cdot j}}^0 = \frac{x_{i\cdot}}{n} \frac{x_{\cdot j}}{n}.$$

Thus the MLE testing statistics is given by

$$2\{l(\hat{p}) - l(\hat{p}^0)\} = 2 \sum_{i=1}^r \sum_{j=1}^c X_{ij} \log \left(\frac{X_{ij}}{n \hat{p_{i\cdot}}^0 \hat{p_{\cdot j}}^0} \right),$$

and Wald and Rao's statistics are given by

$$W_n(\hat{p}^0) = R_n(\hat{p}^0) = \sum_{i=1}^r \sum_{j=1}^c \frac{(X_{ij} - n\hat{p_{i\cdot}}^0 \hat{p_{\cdot j}}^0)^2}{n\hat{p_{i\cdot}}^0 \hat{p_{\cdot j}}^0}.$$

The limit distribution of these statistics under the H_0 will be $\chi^2((r-1)(c-1))$.

5 Comparison of estimator

First and foremost, we use the MSE.

$$MSE(\hat{\eta},\theta)=\mathbb{E}_\theta[(\hat{\eta}-\eta(\theta))^2].$$

However, there exists an situation that (and most commonly happen) one estimator dominates other in all parameter spaces Θ. Therefore, we could compare it by maximum of MSE throughout the parameter space or take an Bayesian MSE to compare those.

$$\min_{\hat{\eta}}\max_{\theta\in\Theta}\mathbb{E}_\theta[(\hat{\eta}-\eta(\theta))^2].$$

$$r(\pi,\hat{\eta})=\int_\Theta \mathbb{E}_\theta[(\hat{\eta}-\eta(\theta))^2]\pi(\theta)d\theta.$$

Secondly, we could compare between narrow categories of estimators. The most common choice is unbiased estimators. In this case, the unbiased estimator that minimizes MSE (thus, the variance) is called uniformly minimum variance unbiased estimator(**UMVUE**).

Finally, we would like to compare the estimators when samples grows to infinity. Assume that two estimators $\hat{\theta}_n^1,\hat{\theta}_n^2$ satisfies asymptotic normality.

$$\sqrt{n}(\hat{\theta}_n^j-\theta)\xrightarrow[n\to\infty]{d}N(0,\sigma_j^2(\theta)).$$

Then ${\sigma_1}^{-2}(\theta)/{\sigma_2}^{-2}(\theta)$ is called asymptotic relative efficiency. The more variance be small, the better the statistic is.

6 Sufficient statistic, Complete statistic

If statistic Y satisfies

$$\mathbb{P}_{\theta_1}\left\{(X_1,\cdots,X_n)^t\in A|Y=y\right\}=\mathbb{P}_{\theta_2}\left\{(X_1,\cdots,X_n)^t\in A|Y=y\right\}$$

for every θ_1,θ_2 and A, we call Y a sufficient statistic. The famous factorization theorem states necessary and sufficient for being sufficient statistic.

Theorem 10 (Factorization theorem)**.** *Suppose the distribution of random sample $X_1,\cdots,X_n$ came from some of the $f(x;\theta)$, $\theta\in\Theta$. The statistics $Y=u(X_1,\cdots,X_n)$ is sufficient statistic if and only if there exists k_1,k_2 such that*

$$\prod_{i=1}^n f(x_i;\theta)=k_1(u(x),\theta)k_2(x)$$

Using factorization theorem, we could easily get
- In Possion distribution, $X_1+\cdots+X_n$ is sufficient statistic.
- In Gamma distribution, $(X_1+\cdots+X_n,X_1\cdots X_n)^t$ is sufficient statistic.
- In Exponential family, $\left(\sum_{i=1}^n T_1(X_i),\cdots,\sum_{i=1}^n T_k(X_i)\right)^t$ is sufficient statistic.
- In normal distribution, $(\bar{X},S^2)^t$ is sufficient statistic.

Rao-Blackwell theorem states that, **for any estimator, we can reduce the MSE of the estimator by having conditional expectation over sufficient statistic.**

Theorem 11 (Rao-Blackwell Theorem)**.** *Suppose $X_1,\cdots,X_n$ sampled from $f(x;\theta)$, and $Y=u(X_1,\cdots,X_n)$ is sufficient statistic of θ. For any estimator $\hat{\eta}$ of $\eta=\eta(\theta)$,*

$$\hat{\eta}^{RB}(Y):=\mathbb{E}[\hat{\eta}(X_1,\cdots,X_n)|Y]$$

$$MSE(\hat{\eta}^{RB},\theta)\le MSE(\hat{\eta},\theta).$$

Now, complete statistic is the statistic $Y=u(X_1,\cdots,X_n)$ satisfying following condition: If $\mathbb{E}_\theta g(Y)=0$ for all $\theta\in\Theta$, then $\mathbb{P}_\theta(g(Y)=0)=1$ for all $\theta\in\Theta$.

If complete sufficient statistic exists, we can guarantee the existence of **UMVUE** estimator.

Theorem 12. *Suppose $X_1, \cdots, X_n$ sampled from $f(x;\theta)$, and $Y = u(X_1, \cdots, X_n)$ is complete sufficient statistic of θ. Then*
(1) For any unbiased estimator $\hat{\eta}$ of η,

$$\hat{\eta}^{RB}(Y) = \mathbb{E}[\hat{\eta}(X_1, \cdots, X_n)|Y]$$

is uniformly minimum variance unbiased estimator.
(2) If $\delta(Y)$ is unbiased estimator of η then $\delta(Y)$ is uniformly minimum variance unbiased estimator.

There are several examples of complete sufficient statistics,
- Poisson distribution, $(X_1 + \cdots + X_n)$ is complete sufficient statistic.
- Exponential family, $\left(\sum_{i=1}^n T_1(X_i), \cdots, \sum_{i=1}^n T_k(X_i)\right)^t$ is complete sufficient statistic.
- Poisson distribution, Theorem 12 states that $(X_1 + \cdots + X_n)/n$ is UMVUE of $\theta = \lambda$.
- Normal distribution, Theorem 12 states that $(\bar{X}, S^2)$ are UMVUE of μ, σ^2.

Finally, I remark the Basu theorem.

Theorem 13 (Basu theorem)**.** *Let $Z = v(X_1, \cdots, X_n)$ be an ancillary statistic (that distribution does not depend on $\theta \in \Theta$) and $Y = u(X_1, \cdots, X_n)$ a complete sufficient statistic. Then Y and Z are independent.*

References

[BD15] Peter J. Bickel and Kjell A. Doksum. *Mathematical Statistics: Basic Ideas and Selected Topics, Volume I.* 2nd. Chapman and Hall/CRC, 2015. ISBN: 9781498723800.