Topics in statistics III/IV (MATH3361/4071) Michaelmas term, 2018-2019

Handout: Likelihood methods for large samples [a] , [b] , [c]

Lecturer: Georgios Karagiannis georgios.karagiannis@durham.ac.uk

[a] Basic reading list: [6, 7, 10, 11, 12, 9, 3, in the Section References]

[b] Author: Georgios P. Karagiannis.

[c] Acknowledgments to students in 2018 for spotting typos in the handouts.

1 Modes of convergence and relations

§ **Set-up and notation:**

Consider a probability triplet $(\Omega, \mathcal{F}, P)$.

Consider random variable $X : \Omega \to \mathbb{R}^d$, where for simplicity we will denote the d-dimensional random vector as $X := X(\omega)$, $\forall \omega \in \Omega$.

Likewise, we define a sequence of random variables $X_n : \Omega \to \mathbb{R}^d$, and for simplicity denote $X_n := X_n(\omega)$, for $n = 1, 2, ...$, and $\forall \omega \in \Omega$.

The distribution function of r.v. X is denoted as

$$F_X(x) = P(X \leq x) = P(X_1 \leq x_1, ..., X_d \leq x_d).$$

Hereafter, the norm $|\cdot|$ refers to the Euclidean norm; i.e. $|X| = \sqrt{\sum_{j=1}^{d} X_j^2}$, however the results can be generalized.

§ **Definitions of modes of convergence:**

Some modes of convergence are defined below.

Definition 1. X_n converges in distribution to X, symb. as $X_n \xrightarrow{D} X$, iff

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

for all points $x \in \mathbb{R}^d$ at which $F_X(x)$ is continuous.

- Other names: converges in law, and weak convergence

Definition 2. X_n converges in probability to X iff for every $\epsilon > 0$

$$\lim_{n\to\infty} P(|X_n - X| > \epsilon) = 0 \tag{1.1}$$

It is symbolized as $X_n \xrightarrow{P} X$.

- It means: for any $\epsilon > 0$, and for any $\delta > 0$, there exists $N_{\epsilon,\delta} > 0$, where $P(|X_n - X| < \epsilon) < \delta$

Definition 3. X_n converges in almost surely to X iff for every $\epsilon > 0$

$$P(\lim_{n\to\infty} X_n = X) = 1 \tag{1.2}$$

It is symbolized as $X_n \xrightarrow{a.s.} X$.

- Other names: converges with probability 1, and strong convergence

Definition 4. X_n converges in the r-th mean to X iff for every $\epsilon > 0$

$$\lim_{n\to\infty} \mathrm{E}|X_n - X|^r = 0$$

where $r \in \{1, 2, ...\}$. It is symbolized as $X_n \xrightarrow{r} X$.

Definition 5. X_n converges in quadratic mean to X iff

$$\lim_{n\to\infty} \mathrm{E}|X_n - X|^2 = 0 \tag{1.3}$$

It is symbolized as $X_n \xrightarrow{\mathrm{qm}} X$

§ **Convergence in probability versus almost surely:**

To better understand the difference/connection between the $\xrightarrow{P}$ and $\xrightarrow{a.s.}$, we restate the definitions in words.

convergence in probability $\xrightarrow{P}$: it requires that for every $\epsilon > 0$ the probability that X_n is within ϵ of X to tend to 1 n tends to infinity

convergence almost surely $\xrightarrow{a.s.}$: it requires that for every $\epsilon > 0$ the probability that X_k STAYS within ϵ of X for all $k \geq n$ to tend to 1 as n tends to infinity

The following Lemma shows the distinction between $\xrightarrow{P}$ and the $\xrightarrow{a.s.}$.

Lemma 6. $X_n \xrightarrow{a.s.} X$ *iff for every* $\epsilon > 0$

$$P(|X_k - X| < \epsilon\,,\, \forall k \geq n) \to 1, \qquad \text{as } n \to \infty$$

Proof. Let $A_{n,\epsilon} = \{|X_k - X| < \epsilon\,,\, \forall k \geq n\}$. Then

$$P(\lim_{n\to\infty} X_n = X) = P\{\forall \epsilon > 0,\, \exists n > 0,\, \text{s.t.}\, |X_k - X| < \epsilon,\, \forall k \geq n\} = P\{\cap_{\epsilon>0} \cup_{\forall n} A_{n,\epsilon}\}$$

So $X_n \xrightarrow{a.s.} X$ is equivalent to $P\{\cap_{\epsilon>0} \cup_{\forall n} A_{n,\epsilon}\} = 1$. Because sets $\cup_{\forall n} A_{n,\epsilon}$ decrease to $\cap_{\epsilon>0} \cup_{\forall n} A_{n,\epsilon}$ as $\epsilon \to 0$, it is

$$P\{\cap_{\epsilon>0} \cup_{\forall n} A_{n,\epsilon}\} = 1 \iff P\{\cup_{\forall n} A_{n,\epsilon}\} = 1,\, \forall \epsilon > 0$$

Because $A_{n,\epsilon}$ increases to $\cup_{\forall n} A_{n,\epsilon}$ as $n \to \infty$, it is

$$P\{\cup_{\forall n} A_{n,\epsilon}\} = 1 \iff P\{A_{n,\epsilon}\} = 1,\, \text{as } n \to \infty,\, \forall \epsilon > 0$$

□

§ **Relations between convergence modes:**

Theorem 7. *Relations between/among different modes of convergence*

1. $X_n \xrightarrow{a.s.} X \implies X_n \xrightarrow{P} X$
2. $X_n \xrightarrow{r} X$, *for some* $r > 0 \implies X_n \xrightarrow{P} X$
3. $X_n \xrightarrow{P} X \implies X_n \xrightarrow{D} X$

Proof.

1. For any $\epsilon > 0$, then

$$P(|X_n - X| > \epsilon) \geq P(|X_k - X| < \epsilon\,, \forall k \geq n) \to 1, \text{ as, } n \to \infty$$

   from Lemma 6.

2. It is

$$E|X_n - X|^r \geq E(|X_k - X|^r 1(|X_n - X| \geq \epsilon)) \geq \epsilon^r P(|X_n - X| \geq \epsilon) \to 0, \text{ as, } n \to \infty$$

   This is Markov inequality (Prob. I)

3. ($\star\star\star$) For any $\epsilon > 0$, $\{X > z + 1\epsilon\}$ and $|X_n - X| < \epsilon$ imply $\{X_n > z\}$. Hence, $\{X_n > z\} \supseteq \{X > z+\epsilon\} \cap \{|X_n - X| < \epsilon\}$. By taking complements, we get $\{X_n \leq z\} \subseteq \{X \leq z+1\epsilon\} \cup \{|X_n - X| > \epsilon\}$. So I get $P(X_n \leq z) \leq P(X \leq z+\epsilon) + P(|X_n - X| > \epsilon)$. In a similar way (by interchanging X and X_n), I get $P(X_n \leq z) \geq P(X \leq z - \epsilon) + P(|X_n - X| > \epsilon)$. [a]
   So as $n \to \infty$

$$P(X \leq z - 1\epsilon) \leq \liminf_{n\to\infty} P(X_n \leq z) \leq \limsup_{n\to\infty} P(X_n \leq z) \leq P(X \leq z + 1\epsilon)$$

   As $F_X(x) = P(X \leq x)$ is continuous at z, the two ends should converge to $F_X(z) = P(X \leq z)$ as $\epsilon \to 0$, which implies that $\lim_{n\to\infty} F_{X_n}(z) = F_X(z)$

[a] It is:

(a) $\limsup_{n\to\infty} f_n := \lim_{n\to\infty}(\sup_{\forall m \geq n} f_m)$ and $\liminf_{n\to\infty} f_n := \lim_{n\to\infty}(\inf_{\forall m \geq n} f_m)$

(b) It is $\liminf_{n\to\infty} f_n \leq \limsup_{n\to\infty} f_n$ if both exist.

(c) It is $\lim_{n\to\infty} f_n = \liminf_{n\to\infty} f_n = \limsup_{n\to\infty} f_n$ if $\lim_{n\to\infty} f_n$ exists

□

Example. ($\star$)Consider $Z \sim U(0,1)$, and $X_n = 2^n 1_{[0,1/n)}(Z)$. Check if $X_n \xrightarrow{r} 0$, $X_n \xrightarrow{a.s.} 0$, or $X_n \xrightarrow{P} 0$

Solution 8. It is $\mathrm{E}|X_n|^r = \frac{1}{n}2^{nr} \to \infty$, so $X_n \not\xrightarrow{r} 0$. It is $\mathrm{P}(\{\lim X_n = 0\}) = \mathrm{P}(\{Z > 0\}) = 1$, so $X_n \xrightarrow{as} 0$. It is $\mathrm{P}(\{|X_n| \geq \epsilon\}) = \mathrm{P}(\{X_n = 2^n\}) = \mathrm{P}(Z \in [0, 1/n)) = 1/n \to 0$, so $X_n \xrightarrow{P} 0$.

Consider a constant vector $c \in \mathbb{R}^d$. We say that X is a degenerate random variable/vector identically equal to $c \in \mathbb{R}^d$, iff $X(\omega) = c$, $\forall \omega \in \Omega$ (for every element of the sampling space). The distribution function of a degenerate random variable X equal to c is

$$F_X(x) = \begin{cases} 1 & , x \geq c \\ 0 & , \text{else} \end{cases}$$

Mostly, we will use the symbol $c \in \mathbb{R}^d$ to denote the constant point c, as well as the degenerate random vector identically equal to c.

The Theorem 9, together with Theorem 7, implies that $X_n \xrightarrow{D} c \iff X_n \xrightarrow{P} c$, if c is constant.

Theorem 9. *If $c \in \mathbb{R}^d$ is a constant, then $X_n \xrightarrow{D} c \implies X_n \xrightarrow{P} c$*

Proof. We will show the 2D case, just to understand how to use the degenerate random variable, however the generalization is obvious. It is

$$\begin{aligned} P(|X_n - c| \leq \epsilon\sqrt{2}) \geq & P(c - \epsilon \begin{bmatrix} 1 \\ 1 \end{bmatrix} < X_n \leq c + \epsilon \begin{bmatrix} 1 \\ 1 \end{bmatrix}) \\ = & \cancel{P(X_n \leq c + \epsilon \begin{bmatrix} 1 \\ 1 \end{bmatrix})}^{1} - \cancel{P(X_n \leq c + \epsilon \begin{bmatrix} 1 \\ -1 \end{bmatrix})}^{0} \\ & \cancel{- P(X_n \leq c + \epsilon \begin{bmatrix} -1 \\ 1 \end{bmatrix})}^{0} + \cancel{P(X_n \leq c - \epsilon \begin{bmatrix} 1 \\ 1 \end{bmatrix})}^{=1} = 1 \end{aligned}$$

□

§ Revision: Taylor expansion in many dimensions

We revise the 2nd order Taylor expansion in many dimensions. For more details see [2, 1] or any calculus book.

§ Derivative notation:

- If $f : \mathbb{R}^d \to \mathbb{R}^k$, then
$$\dot{f}(x) = \frac{\mathrm{d}}{\mathrm{d}x} f(x) = \nabla_x f(x)$$
is a $d \times k$ matrix whose (i,j)th element is $\frac{\mathrm{d}}{\mathrm{d}x_j} f_i(x)$.

- If $f : \mathbb{R}^d \to \mathbb{R}$, then

$$\ddot{f}(x) = \frac{\mathrm{d}}{\mathrm{d}x}\dot{f}(x)^T$$

is a $d \times d$ matrix whose (i, j)th element is

$$[\ddot{f}(x)]_{i,j} = \frac{\mathrm{d}^2}{\mathrm{d}x_i \mathrm{d}x_j} f(x)$$

§ **Consequences:**

Fact 10. *If $f : \mathbb{R}^d \to \mathbb{R}^s$, $g : \mathbb{R}^s \to \mathbb{R}^k$, and $h(x) = g(f(x))$ then*

$$\dot{h}(x) = \dot{g}(f(x))\dot{f}(x) \tag{1.4}$$

Fact 11. *If $f : \mathbb{R}^d \to \mathbb{R}^k$, $g : \mathbb{R}^s \to \mathbb{R}^k$, and $h(x) = f^T(x)g(x)$ then*

$$\dot{h}(x) = g(x)^T\dot{f}(x) + f(x)^T\dot{g}(x)$$

Theorem 12. *The Mean Value Theorem: If $f : \mathbb{R}^d \to \mathbb{R}^k$ and if $\dot{f}(x)$ is continuous in the ball $B_r(x_0) = \{x \in \mathbb{R}^d : |x - x_0| < r\}$, then for $|t| < r$,*

$$f(x_0 + t) = f(x_0) + \left(\int_0^1 \dot{f}(x_0 + ut)\,du\right) t$$

Proof. Let $h(u) = f(x_0 + ut)$, so that $\dot{h}(u) = \dot{f}(x_0 + ut)t$ (from (1.4)). Then,

$$\int_0^1 \dot{f}(x_0 + ut)t\mathrm{d}u = \int_0^1 h(u)\mathrm{d}u = h(1) - h(0) = f(x_0 + t) - f(x_0)$$

□

Theorem 13. *The Taylor's theorem: If $f : \mathbb{R}^d \to \mathbb{R}$ and if $\ddot{f}(x)$ is continuous in the ball $B_r(x_0) = \{x \in \mathbb{R}^d : |x - x_0| < r\}$, then for $|t| < r$,*

$$f(x_0 + t) = f(x_0) + \dot{f}(x_0)t + t^T\left(\int_0^1 \int_0^1 u\ddot{f}(x_0 + uvt)\,du\,dv\right) t$$

Proof. [FYI:] Same trick as above by using $g(v) = t^T\left(\int_0^1 \dot{f}(x_0 + uvt)\mathrm{d}u\right)$... □

Exercise sheet

Exercise #1
Exercise #2
Exercise #3
Exercise #5
Exercise #6

2 Characteristic functions

Characteristic functions provide an alternative way to the probability function for describing a random variable. In fact, it completely determines (see Theorem 15(9)) the behavior and properties of the probability distribution of the random variable X.

Definition 14. The characteristic function of a d dimensional random variable X is

$$\varphi_X(t) = \mathrm{E}(e^{it^T X})$$

for $t \in \mathbb{R}^d$, where $e^{it^T X} = \cos(t^T X) + i \sin(t^T X)$.

Theorem 15. *Some properties of characteristic functions*

1. $\varphi_X(t)$ *exists for all* $t \in \mathbb{R}^d$ *and is continuous*
2. $\varphi_X(0) = 1$ *and* $|\varphi_X(t)| \leq 1$ *for all* $t \in \mathbb{R}^d$
3. $\varphi_{A+BX}(t) = e^{it^T A}\varphi_X(B^T t)$ *if* $A \in \mathbb{R}^d$ *and* $B \in \mathbb{R}^{k\times d}$ *are constants*
4. $\varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t)$ *iff* X *and* Y *are independent*
5. *if* $E|X| < \infty$, *then* $\dot{\varphi}_X(t)$ *exists, it is continouous, and* $\dot{\varphi}_X(0) = iE(X)^T$
6. *if* $E|X|^2 < \infty$, *then* $\ddot{\varphi}_X(t)$ *exists, it is continouous, and* $\ddot{\varphi}_X(0) = -E(X^T X)$
7. *if* X *is degenerate at* $c \in \mathbb{R}^d$ *then* $\varphi_X(t) = e^{it^T c}$
8. *if* $M_X(t) = E(e^{t^T x})$ *is the moment generating function, then* $M_X(t) = \phi_X(-it)$
9. $F_Y(t) = F_X(t) \iff \varphi_Y(t) = \varphi_X(t)$, *for any* $t \in \mathbb{R}^d$
10. *if* $X \sim N(\mu, \Sigma)$ *then* $\varphi_X(t) = \exp(it^T\mu - \frac{1}{2}t^T\Sigma t)$

Proof. Straightforward from the Definition 14. □

Theorem 16. *[Continuity theorem] Let* $X, X_1, X_2, \ldots$ *random vectors*

$$X_n \xrightarrow{D} X \iff \varphi_{X_n}(t) \to \varphi_X(t), \text{ for any } t \in \mathbb{R}^d$$

Example 17. ($\star$) Show that if $X \sim \mathrm{Ex}(\lambda)$ then $\varphi_X(t) = \frac{\lambda}{\lambda - it}$.

Solution. It is

$$\varphi_X(t) = \int_{-\infty}^{\infty} e^{itX} \underbrace{\lambda e^{-\lambda x} 1(X > 0)}_{=f_{\mathrm{Ex}}(x|\lambda)} \mathrm{d}x = \lambda \int_{-\infty}^{\infty} e^{-x(\lambda - itX)} \mathrm{d}x = \frac{\lambda}{\lambda - it}$$

Example 18. ($\star$)

1. Find $\varphi_X(t)$ if $X \sim \text{Br}(p)$.
2. Find $\varphi_Y(t)$ if $Y \sim \text{Bin}(n, p)$

Solution.

1. It is
$$\varphi_X(t) = \sum_{x=0,1} e^{itX} P(X = x) = e^{it0}(1-p) + e^{it1}p = (1-p) + pe^{it}$$
2. Because Binomial r.v. results as a summation of n IID Bernoulli r.v., it is $Y = \sum_{i=1}^{n} X_i$, where $X_i \sim \text{Br}(p)$ $i = 1, ..., n$ and IID. Then
$$\varphi_Y(t) = \varphi_{\sum X_i}(t) = \prod_{i=1}^{n} \varphi_{X_i}(t) = ((1-p) + pe^{it})^n$$

Exercise sheet

Exercise #9
More exercise:

- `https://www.statlect.com/fundamentals-of-probability/characteristic-function`
- `https://www.statlect.com/fundamentals-of-probability/joint-characteristic-function`

3 Consistency

Assume that we have specified a parametric probabilistic model P_θ (aka sampling distribution) to model the data generating process of a sequence of random samples (i.e.; unseen/hypothetical observations) $X_{1:n} = (X_1, ..., X_n)$. This P_θ often depends on unknown parameter $\theta \in \Theta$ whose true value the statistician needs to learn from the observables. Often this is performed through the construction of estimators $\hat{\theta}_n := \hat{\theta}_n(X_{1:n})$ (statistic functions) which are functions of the observations $X_{1:n}$.

Idea. A desired property for $\hat{\theta}_n$ would be to be close to the true value of θ, in the limit $n \to \infty$; i.e. $\hat{\theta}_n$ has to be consistent to θ.

- There are different types of consistency of an estimator, each of them depending on the 'convergence mode'. Here, we extend the concept to the multivariate case.

Definition 19. [Weak consistency] We say that $\hat{\theta}_n$ is a weakly consistent sequence of estimators of θ iff for all $\theta \in \Theta$,
$$\hat{\theta}_n \xrightarrow{P} \theta$$

when the probability P in (1.1) is defined on the true parameter θ, i.e. $P = P_\theta$.

- Other names: consistency in probability

Definition 20. [Strong consistency] We say that $\hat{\theta}_n$ is a strongly consistent sequence of estimators of θ iff for all $\theta \in \Theta$,

$$\hat{\theta}_n \xrightarrow{a.s.} \theta$$

when the probability P in (1.2) is defined on the true parameter value θ, i.e. $P = P_\theta$.

Definition 21. [Consistency in quadratic mean] We say that $\hat{\theta}_n$ is a consistent in quadratic mean sequence of estimators of θ iff for all $\theta \in \Theta$,

$$\hat{\theta}_n \xrightarrow{\text{qm}} \theta$$

when the probability P in (1.3) is defined on the true parameter value θ, i.e. $P = P_\theta$

Question ($\star$) Which type of consistency implies the other?

Answer

Question ($\star$) Have you though of any good reason why most of the estimator are in the form of arithmetic average?

Answer

Idea. A useful tool to prove that our estimator is consistent (in some sense) to the parameter of interest are the Law of Large Numbers in Theorem 22!!!

Theorem 22. *Let $X, X_1, X_2, ...$ be i.i.d. random vectors, and let*

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

then

1. *(Weak law) If $E|X| < \infty$, then $\bar{X}_n \xrightarrow{P} E(X)$*
2. *(Strong law) $E|X| < \infty$, iff $\bar{X}_n \xrightarrow{as} E(X)$*
3. *(in qm) $E|X|^2 < \infty$, iff $\bar{X}_n \xrightarrow{qm} E(X)$*

Often it is easy to show that the 2nd moment is finite, and hence consistence in quadratic mean is feasible, and presents interest.

Proof. (of Thm 22)

1. Let $\varphi_X(t) = \mathrm{E}(e^{it^T X})$, and $\mu = \mathrm{E}(X)$. It is

$$\begin{aligned}\varphi_{\bar{X}_n}(t) =& \varphi_{X_1+\ldots+X_n}(\frac{t}{n}) = \prod_{i=1}^{n} \varphi_{X_j}(\frac{t}{n}) = \left(\varphi_X(\frac{t}{n})\right)^n \\ =& \left(\varphi_X(0) + \left(\int_0^1 \dot{\varphi}_X(u\frac{t}{n})\mathrm{d}u\right)\frac{t}{n}\right)^n\end{aligned}$$

   since by the Mean-Value theorem

$$\varphi_X(\frac{t}{n}) = \varphi_X(0) + \left(\int_0^1 \dot{\varphi}_X(u\frac{t}{n})\mathrm{d}u\right)\frac{t}{n}.$$

   Because $\varphi_X(0) = 1$, and $\lim_{\epsilon\to 0}\dot{\varphi}_X(\epsilon) = \dot{\varphi}_X(0) = i\mu^T$ it is

$$\lim_{n\to\infty} \varphi_{\bar{X}_n}(t) = \exp\left(\lim_{n\to\infty}\left(\int_0^1 \dot{\varphi}_X(u\frac{t}{n})\mathrm{d}u\right) t\right) = \exp(i\mu^T t) \tag{3.1}$$

   Here I used that $\lim_{n\to\infty}(1+a_n)^n = \exp(\lim_{n\to\infty} na_n)$ if $\lim_{n\to\infty} na_n$ exists (Exercise #27). So (3.1) says that the characteristic function of $\bar{X}_n$ converges to a characteristic function of the degenerate random variable μ

$$\varphi_{\bar{X}_n}(t) \to \varphi_\mu(t)$$

   From the continuity Theorem 16 it is $\bar{X}_n \xrightarrow{D} \mu$. Then from Theorem 7(3) it is $\bar{X}_n \xrightarrow{P} \mu$ because $\mu = \mathrm{E}(X)$ is just a constant point.

2. Proof is out of the scope; for more details see in[5].

3. It is

$$\begin{aligned}\mathrm{E}|\bar{X}_n - \mu|^2 =& \mathrm{E}(\bar{X}_n - \mu)^T(\bar{X}_n - \mu) \\ =& \frac{1}{n^2}\sum_i\sum_j \mathrm{E}(X_i - \mu)^T(X_j - \mu) \\ \overset{\text{simplify}}{=}& \frac{1}{n^2}\sum_i \mathrm{E}(X_i - \mu)^T(X_i - \mu) \overset{\text{iid}}{=} \frac{1}{n^2} n\mathrm{E}(X - \mu)^T(X - \mu) \\ =& \frac{1}{n}\mathrm{Var}(X) \to 0\end{aligned}$$

   as the 2nd mode is finite.

□

Exercise 23. (The regression model) Consider a regression model

$$Y_i = a + bX_i + Z_i$$

for (y_i, x_i) , $i = 1, ...,$ where $\mathrm{E}(Z_i) = 0$ and $\mathrm{Var}(z_i) = \sigma^2$. Consider Least squares estimator (not the MLE !!!), namely (Stats 2):

$$\begin{aligned}\hat{b}_n =& \frac{\sum_{i=1}^n Y_i(X_i - \bar{X}_n)}{\sum_{i=1}^n (X_i - \bar{X}_n)^2} \\ \hat{a}_n =& \bar{Y}_n - \hat{b}_n \bar{X}_n\end{aligned}$$

where $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$.

1. Show that $\hat{b}_n \xrightarrow{\text{qm}} b$, and impose any condition if necessary
2. Show that $\hat{a}_n \xrightarrow{\text{qm}} a$, and impose any condition if necessary

Solution. right... I know that $\mathrm{E}_\pi(Z-\theta)^2 = \mathrm{Var}_\pi(Z) + (\mathrm{E}_\pi(Z) - \theta)^2$

1. It is

$$\mathrm{E}(\hat{b}_n) = \frac{\sum_{i=1}^n \mathrm{E}(Y_i)(X_i - \bar{X}_n)}{\sum_{i=1}^n (X_i - \bar{X}_n)^2} = \frac{\sum_{i=1}^n (a + bX_i)(X_i - \bar{X}_n)}{\sum_{i=1}^n (X_i - \bar{X}_n)^2} = ... = b$$

$$\mathrm{Var}(\hat{b}_n) = \frac{\sum_{i=1}^n \mathrm{Var}(Y_i)(X_i - \bar{X}_n)}{(\sum_{i=1}^n (X_i - \bar{X}_n)^2)^2} = \frac{\sum_{i=1}^n \sigma^2 (X_i - \bar{X}_n)^2}{(\sum_{i=1}^n (X_i - \bar{X}_n)^2)^2} = \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X}_n)^2}$$

   Therefore we have $\hat{b}_n \xrightarrow{\text{qm}} b$ if $\sum_{i=1}^n (X_i - \bar{X}_n)^2 \to \infty$.

2. It is

$$\mathrm{E}(\hat{a}_n) = \mathrm{E}\bar{Y}_n - \mathrm{E}\hat{b}_n \bar{X}_n = (a + b\bar{X}_n) - b\bar{X}_n = a$$

   I rearrange $\hat{a}_n$ in a more convenient form

$$\hat{a}_n = \sum_i Y_i \left(\frac{1}{n} - \frac{(X_i - \bar{X}_n)\bar{X}_n}{\sum_{i=1}^n (X_i - \bar{X}_n)^2}\right)$$

   so

$$\mathrm{Var}(\hat{a}_n) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}_n^2}{\sum_{i=1}^n (X_i - \bar{X}_n)^2}\right)$$

   in order to have $\hat{a}_n \xrightarrow{\text{qm}} a$, I need $\mathrm{Var}\hat{a}_n \to 0$ which happens if $\frac{\bar{X}_n^2}{\sum_{i=1}^n (X_i - \bar{X}_n)^2} \to 0$ as $n \to \infty$.

Exercise sheet

Exercise #8
Exercise #10

4 Central Limit Theorems

Motivation Let a sequence of random samples $\{X_1, ..., X_n\}$ from a sampling distribution $\mathrm{d}f(\cdot|\theta)$ labeled by an unknown parameter $\theta \in \Theta$ which you winch to learn. To construct confidence intervals, we would try to specify a statistic $U_n = U_n(\theta, X_1, ..., X_n)$ connecting the samples with unknown parameters θ, and following a tractable sampling distribution. This will allow us to calculate the confidence interval at a specified confidence level. Central Limit Theorems are tools allowing us to find asymptotic distributions in certain cases.

We present a basic version of the Central Limit Theorem (CLT) that assumes IID random variables (vectors). [1] We discuss about the so called Edgeworth-Expansions, which are another version of the CLT considering higher order terms.

4.1 Central Limit Theorem (IID case)

Theorem 24. *Let $X_1, X_2, ...$ IID random vectors $X_i \in \mathbb{R}^d$ with mean $E(X_i) = \mu$ and finite covariance matrix $Var(X_i) < \infty$ for all $i = 1, ...,$ Then*

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{D} N(0, \Sigma)$$

Note: *A web applet about Normal distribution is available*[2]. [3]

Proof. We 'll gonna use again the characteristic function, and its property with the IID variables. It is

$$\sqrt{n}(\bar{X}_n - \mu) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} (X_j - \mu)$$

Hence, for any $t \in \mathbb{R}^d$

$$\begin{aligned} \varphi_{\sqrt{n}(\bar{X}_n - \mu)}(t) &= \varphi_{\frac{1}{\sqrt{n}} \sum_{j=1}^{n} (X_j - \mu)}(t) = \varphi_{\sum_{j=1}^{n} (X_j - \mu)}(\frac{t}{\sqrt{n}}) \\ &= \prod_{j=1}^{n} \varphi_{(X_j - \mu)}(\frac{t}{\sqrt{n}}) \\ &= \left(\varphi_{(X_j - \mu)}(\frac{t}{\sqrt{n}}) \right)^n = \left(\varphi_{(X - \mu)}(\frac{t}{\sqrt{n}}) \right)^n \end{aligned}$$

[1] Generalisations of the above CLT exist, (such as the Lindeberg-Feller version of the CLT which does not require "identically distributed" variables) see [5]. –not examinable

[2] `https://georgios-stats-3.shinyapps.io/demo_multivariatenormaldistribution/`

[3] `https://github.com/georgios-stats/Topics_in_Statistics/tree/master/demo_MultivariateNormalDistribution`

Here, let $\varphi(t) := \varphi_{(X_j-\mu)}(t)$ for notation convenience, as $X_1, X_2, ...$ are IID and hence have the same moments. We use Taylor expansion around 0 as

$$\varphi_{(X-\mu)}(\frac{t}{\sqrt{n}}) = \cancelto{1}{\varphi_{(X-\mu)}(0)} + \cancelto{0}{\dot{\varphi}_{(X-\mu)}(0)}\frac{t}{\sqrt{n}} + t^T \left(\int_0^1 \int_0^1 v\ddot{\varphi}_{(X-\mu)}(0 + vu\frac{t}{n})\mathrm{d}u\mathrm{d}v \right) \frac{t}{n}$$

because $\ddot{\varphi}_X(t)$ is obviously continuous. So

$$\begin{aligned}
\varphi_{\sqrt{n}(\bar{X}_n-\mu)}(t) &= \left(\varphi_{(X-\mu)}(\frac{t}{\sqrt{n}}) \right)^n \\
&= \left(1 + t^T \left(\int_0^1 \int_0^1 v\ddot{\varphi}_{(X-\mu)}(vu\frac{t}{n})\mathrm{d}u\mathrm{d}v \right) \frac{t}{n} \right)^n
\end{aligned}$$

Because $\lim_{n\to\infty}(1+a_n)^n = \exp(\lim_{n\to\infty} na_n)$ if $\lim_{n\to\infty} na_n$ exists (Exercise #27), it is

$$\begin{aligned}
\lim_{n\to\infty} \varphi_{\sqrt{n}(\bar{X}_n-\mu)}(t) &= \lim_{n\to\infty} \left(1 + t^T \left(\int_0^1 \int_0^1 v\ddot{\varphi}_{(X-\mu)}(vu\frac{t}{n})\mathrm{d}u\mathrm{d}v \right) \frac{t}{n} \right)^n \\
&= \exp\left(\lim_{n\to\infty} t^T \left(\int_0^1 \int_0^1 v\ddot{\varphi}_{(X-\mu)}(vu\frac{t}{n})\mathrm{d}u\mathrm{d}v \right) t \right) \\
&= \exp\left(t^T \left(\int_0^1 \int_0^1 v(-\Sigma)\mathrm{d}u\mathrm{d}v \right) t \right) \\
&= \exp(-\frac{1}{2}t^T\Sigma\, t)
\end{aligned} \tag{4.1}$$

This is because $\ddot{\varphi}_{(X-\mu)}(\cdot)$ is continuous so $\lim_{n\to\infty} \ddot{\varphi}_{(X-\mu)}(u\frac{t}{n}) = \ddot{\varphi}_{(X-\mu)}(0) = -\mathrm{E}((X-\mu)^T(X-\mu)) = -\Sigma$.

Since $\lim_{n\to\infty} \varphi_{\sqrt{n}(\bar{X}_n-\mu)}(t) = \exp(-\frac{1}{2}t^T\Sigma\, t)$, aka $\varphi_{\sqrt{n}(\bar{X}_n-\mu)}(t) \to \varphi_Z(t)$ where $Z \sim \mathrm{N}(0,\Sigma)$, it is $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{D} \mathrm{N}(0,\Sigma)$. □

Example 25. Consider an M-way contingency table. Let $\boldsymbol{n} = (n_1, ..., n_N)^T$ be the cell observed counts in a contingency table with probabilities $\pi = (\pi_1, ..., \pi_N)^T$ in a vectorized form. Let $\boldsymbol{p} = (p_1, ..., p_N)^T$ be the sample proportional, where $p_j = n_j/n_+$ with $n_+ = \sum_{j=1}^N n_j$.

1. Show that the asymptotic distribution of the sample proportion is such that
$$\sqrt{n}(\boldsymbol{p} - \boldsymbol{\pi}) \xrightarrow{D} \mathrm{N}(0, \mathrm{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^T)$$
2. What is the asymptotic marginal distribution of the sample proportion in the j-th cell is such that
$$\sqrt{n}(p_j - \pi_j) \xrightarrow{D} \mathrm{N}(0, \pi_j(1-\pi_j))$$

Solution.

1. Denote the i-th observation (aka, sample) by $\xi_i = (\xi_{i,1}, ..., \xi_{i,N})^T$, where

$$\xi_{i,j} = \begin{cases} 1 & \text{, if observation } i \text{ falls in cell } j \\ 0 & \text{, if observation } i \text{ does not fall in cell } j \end{cases}$$

Since its observation falls in only one cell, $\sum_j \xi_{i,j} = 1$ and $\xi_{i,j}\xi_{i,k} = 0$ when $j \neq k$. Therefore p can be considered as the arithmetic mean of $\{\xi_{i,j}\}_{i=1}^n$ IID variables as

$$\boldsymbol{p} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\xi_i}$$

The moments of $\{\xi_i\}$, are equal to

$$\begin{aligned} \mathrm{E}(\boldsymbol{\xi_i}) &= \boldsymbol{\pi} \\ \mathrm{Var}(\boldsymbol{\xi_i}) &= \boldsymbol{\Sigma} \end{aligned}$$

where

$$\begin{aligned} [\boldsymbol{\Sigma}]_{j,j} =& \mathrm{var}(\xi_{i,j}) = \mathrm{E}(\xi_{i,j}^2) - (\mathrm{E}(\xi_{i,j}))^2 = \pi_j(1-\pi_j) \\ [\boldsymbol{\Sigma}]_{j,k} =& \mathrm{cov}(\xi_{i,j}, \xi_{i,k}) = \mathrm{E}(\xi_{i,j}\xi_{i,k}) - \mathrm{E}(\xi_{i,j})\mathrm{E}(\xi_{i,k}) = -\pi_j\pi_k \end{aligned}$$

because

$$\begin{aligned} \mathrm{E}(\xi_{i,j}) =& \mathrm{P}(\xi_{i,j} = 1) = \pi_j \\ \mathrm{E}(\xi_{i,j}^2) =& \mathrm{P}(\xi_{i,j} = 1) = \pi_j \\ \mathrm{E}(\xi_{i,j}\xi_{i,k}) =& 0, \text{ if } j \neq k \end{aligned}$$

Hence

$$\boldsymbol{\Sigma} = \mathrm{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^T$$

Therefore, according tot he CPT

$$\sqrt{n}(\boldsymbol{p} - \boldsymbol{\pi}) \xrightarrow{D} \mathrm{N}(0, \mathrm{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^T) \tag{4.2}$$

2. From (4.2) it is

$$\sqrt{n}(p_j - \pi_j) \xrightarrow{D} \mathrm{N}(0, \pi_j(1-\pi_j))$$

Exercise sheet

Exercise #7
Exercise #27

4.2 Higher order approximations (Edgeworth expansions)

The technical details of this topic require more advanced tools[4], we focus on the concepts rather than the technicalities. Also, we consider the $1D$ case, as the multivariate one involves complex notation.

Let, the standardized $\bar{X}_n$ be

$$Z_n = \sqrt{n}\frac{\bar{X}_n - \mu}{\sigma}$$

with distribution function $F_n(x)$, density function $f_n(x)$, and a-quantile x_a (i.e. $F_n(x_a) = a$).

Theorem 24, gives the good looking result $\lim_{n\to\infty} F_n(x) = \Phi(x)$ where $\Phi(\cdot)$ is the standard Normal distribution function. It is based on the fact that we truncated the Taylor expansion of the characteristic function in the 2nd term...

It would be reasonable to expect that the approximation can be improved if we truncate that Taylor expansion in (4.1) a bit further down and hence take into account higher order terms able to represent the characteristic function more accurately.

Based on this rational, and by truncating the Taylor expansion at the $4th$ term, one can produce the Edgeworth expansions :

$$F_n(x) = \Phi(x) - \phi(x)\frac{\kappa_3 H_2(x)}{6\sqrt{n}} - \phi(x)\left(\frac{\kappa_4 H_3(x)}{24n} + \frac{\kappa_3^2 H_5(x)}{72n}\right) + O(n^{-3/2}) \tag{4.3}$$

$$f_n(x) = \phi(x)\left(1 + \frac{\kappa_3 H_3(x)}{6\sqrt{n}} + \frac{1}{n}\left(\frac{\kappa_4 H_4(x)}{24} + \frac{\kappa_3^2 H_6(x)}{72}\right)\right) + O(n^{-3/2}) \tag{4.4}$$

$$x_a = z_a + \frac{\kappa_3(z_a^2 - 1)}{6\sqrt{n}} + \frac{\kappa_4(z_a^3 - 3z_a)}{24n} - \frac{\kappa_3^2(z_a^5 - 5z_a)}{36n} + O(n^{-3/2}) \tag{4.5}$$

where $\phi(x)$ is the standard Normal PDF, $H_r(x) = (-1)^r \phi^{(r)}(x)/\phi(x)$ are Hermitian polynomials[5], z_a is the a quantile of the standard Normal distribution $\Phi(z_a) = a$, and κ_3, κ_4 are important moments presented below.

Approximation (4.3) takes into account the coefficient of skewness κ_3 (3rd moment) and the coefficient of kurtosis κ_4 (4th moment) which give extra information about the shape or the underling distribution.

Precisely:

- κ_3 is the coefficient of skewness –a measure of the asymmetry of the probability distribution, where

$$\kappa_3 = \frac{\mathrm{E}(X_i - \mu)^3}{\sigma^3} :: \begin{cases} < 0 & \text{large tail to the left} \\ = 0 & \text{symetric} \\ >= & \text{large tail to the right} \end{cases}$$

  hence the term of order $1/\sqrt{n}$ represents the correction for skewness.

[4] For more details see [5, 10].

[5] Hermitian polynomials: $H_2(x) = x^2 - 1$, $H_3(x) = x^3 - 3x$, $H_4(x) = x^4 - 6x^2 + 3$, $H_5(x) = x^5 - 10x^3 + 15x$, $H_6(x) = x^6 - 15x^4 + 45x^2 - 15$.

- κ_4 is the coefficient of kurtosis –a measure of the tailedness of the probability distribution, where

$$\kappa_4 = \frac{\mathrm{E}(X_i - \mu)^4}{\sigma^4} - 3 :: \begin{cases} < 0 & \text{Platykuttic} \\ = 0 & \text{Mesokurtic: (like in the Normal distr)} \\ > 0 & \text{Leptokurtic} \end{cases}$$

  hence the term of order $1/n$ represents the correction for kurtosis.

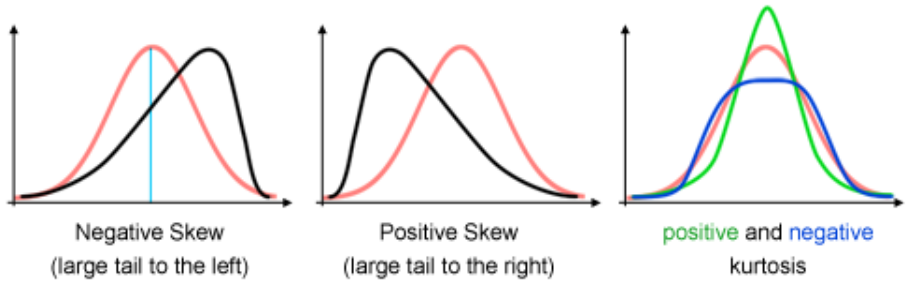


Figure 4.1: Skewness and Kurtosis

- If we truncate before the term of order $(\frac{1}{\sqrt{n}})$ then $F_n(x) = \Phi(x) + O(\frac{1}{\sqrt{n}})$, and we have the approximation of the CLT... I.e., We would approximate the distribution function of Z_n with the symmetric, and mesokurtic Normal distribution.
- However, if the actual distribution is symmetric $\kappa_3 = 0$ then $F_n(x) = \Phi(x) + O(\frac{1}{n})$
- Note that the error in (4.3) is absolute and not relative, and hence for tiny values F_n the approximation might not be reliable. The approximation in the tails, as $|x|$ increases may be poor or even negative. In fact Edgeworth approx. is not a CDF, (4.3) is not bounded between $[0, 1]$. Other expansions such as the Saddlepoint expansions [12] provide better approximations (which are out of the scope).
- The proof of (4.3) involves expanding the characteristic function by considering higher order terms, using inverse Fourier transformation to recover the density function, and integrating to recover the CDF in (4.3). A proof is available in [5]. The asymptotic quantiles based on (4.3) can be found in a similar manner via Cornish–Fisher expansion. However, these technicalities are out of the module scope.

Example 26. Consider r.v. $X_i \sim \mathrm{Ex}(\lambda)$, for $i = 1, ..., n = 3$.

1. For the distribution function of the standardized arithmetic mean $\bar{X}_n$, calculate the CLT approximation, Edgeworth expansion up to the $1/\sqrt{n}$ term, and Edgeworth expansion up to the $1/n$ term, and the exact distribution. Consider $\lambda = 1$ and $n = 3, 10, 100$, and Plot the expansions together for each (λ, n) case.

Hint-1; The standardized variable of X is $(X - \mathrm{E}(X))/\mathrm{var}(X)$

Hint-2: if $X_i \sim \mathrm{Ex}(\lambda)$ for $i = 1, ..., n$ then $S_n = \sum_{i=1}^{n} X_i \sim \mathrm{Ga}(n, \lambda)$ with mean $\mathrm{E}(S_n) = n/\lambda$

Hint-3: $\Gamma(r) = \int_0^\infty x^{r-1} \exp(-x)\mathrm{d}x$, and if r is integer then $\Gamma(r) = (r-1)!$

2. See the plot and discuss ....

Solution. The standardized standardized arithmetic mean $\bar{X}_n$ is $Z_n = \sqrt{n}\frac{\bar{X}_n - \mu}{\sigma}$.

1.

- The exact CDF of Z_n is

$$\mathrm{P}(Z_n \leq \xi) = \mathrm{P}_{\mathrm{Ga}(n,\lambda)}(S_n \leq \sqrt{n}\xi\sigma + n\mu) = F_{\mathrm{Ga}(n,\lambda)}(\sqrt{n}\xi\sigma + n\mu)$$

- The CDF of Z_n according to the CLT is

$$\mathrm{P}(Z_n \leq \xi) \approx \Phi(x)$$

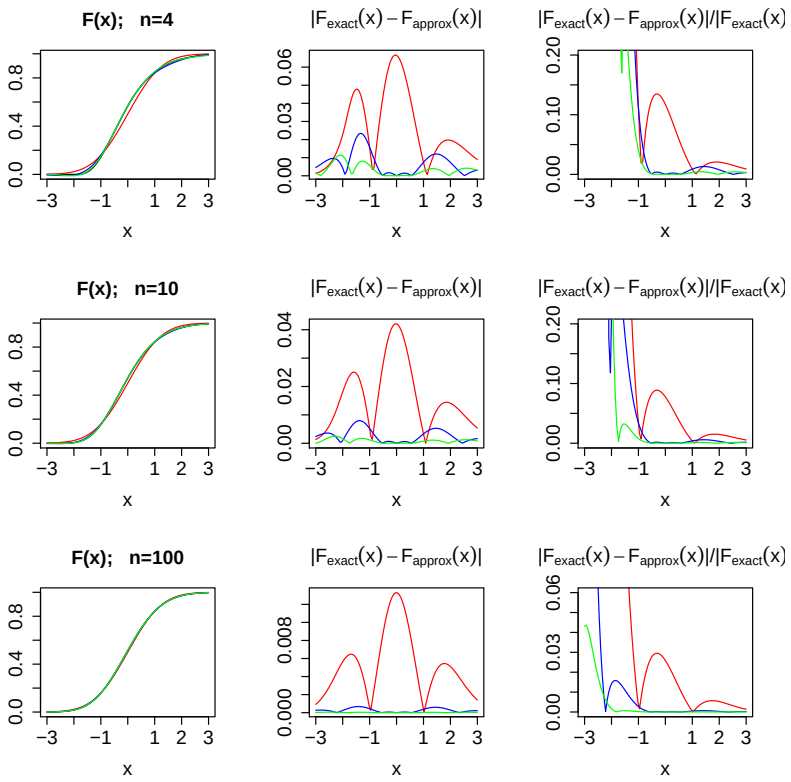
- The CDF of Z_n according to the Edgeworth exp. up to the $1/\sqrt{n}$ term

$$\begin{aligned}\mathrm{P}(Z_n \leq \xi) \approx & \Phi(x) - \phi(x)\frac{\kappa_3 H_2(x)}{6\sqrt{n}} \\ & = \Phi(x) - \phi(x)\frac{\kappa_3(x^2-1)}{6\sqrt{n}}\end{aligned}$$

- The CDF of Z_n according to the Edgeworth exp. up to the $1/n$ term

$$\begin{aligned}\mathrm{P}(Z_n \leq \xi) &\approx \Phi(x) - \phi(x)\frac{\kappa_3 H_2(x)}{6\sqrt{n}} - \phi(x)(\frac{\kappa_4 H_3(x)}{24n} + \frac{\kappa_3^2 H_5(x)}{72n}) \\ &= \Phi(x) - \phi(x)\frac{\kappa_3(x^2-1)}{6\sqrt{n}} \\ &\quad - \phi(x)(\frac{\kappa_4(x^3-3x)}{24n} + \frac{\kappa_3^2(x^5-10x^3+15x)}{72n})\end{aligned}$$

- Here it is $\mathrm{E}(X_i^r) = \int_0^\infty x^r \lambda \exp(-x)\mathrm{d}x = \frac{1}{\lambda^r}\Gamma(2) = \frac{1}{\lambda^r}$, so $\mu = \mathrm{E}(X_i) = 1/\lambda$, $\sigma^2 = \mathrm{var}(X_i) = 1/\lambda^2$, where $\kappa_3 = \frac{\mathrm{E}(X_i-\mu)^3}{\sigma^3} = 2$ and $\kappa_4 = \frac{\mathrm{E}(X_i-\mu)^4}{\sigma^4} - 3 = 6$ ...
- So, now I need to plot them all, in R . The code is available from my GitHub in `https://github.com/georgios-stats/Topics_in_Statistics/tree/master/edworth_ex`



Exact Distr. in black line ; CLT approx. in red line ; Edgeworth exp. up to the $1/\sqrt{n}$ term in blue line ; Edgeworth exp. up to the $1/n$ term green line

2. Well,

   - overall, approximations improve as we consider higher order terms.
   - However.. Higher order approximations are not better uniformly for any x. For a given and fixed n, it does not mean that by adding more terms you will get better approximation throughout the whole domain of x. This is because the expansion is in the limit of n...
   - All the approximation methods get more accurate as n increases.
   - The shapes of the lines (eg red line has always the same bumps) remain the same (of course) because they are controlled by the Hermitian polynomials each of them represents certain behavior of functions[a].

[a] https://en.wikipedia.org/wiki/Hermite_polynomials

Exercise sheet

Exercise #12

5 Slutsky Theorems

In several cases, the statistician knows the asymptotic distribution of some random variables (like arithmetic average $\bar{X}_n$, and standard deviation S_n) but he is actually interested in finding the asymptotic distribution of a function of them (like the t-statistic $T_n = \sqrt{n}(\bar{X}_n - \mu)/S_n$).

- Slutsky theorems can provide the asymptotic distribution of such functions.

Theorem 27. *(Slutsky theorems for convergence in Distribution)*

1. *If $X_n \in \mathbb{R}^d$ a random vector such as $X_n \xrightarrow{D} X$, and $f : \mathbb{R}^d \to \mathbb{R}^k$ is such that $P(X \in C(f)) = 1$, where $C(f)$ is the continuity set of f, then*

$$f(X_n) \xrightarrow{D} f(X)$$

2. *If $X_n \xrightarrow{D} X$ and $(X_n - Y_n) \xrightarrow{P} 0$ then*

$$Y_n \xrightarrow{D} X$$

3. *If $X_n \in \mathbb{R}^d$ where $X_n \xrightarrow{D} X$, and $Y_n \in \mathbb{R}^k$ where $Y_n \xrightarrow{D} c$ and c is a constant then*

$$\begin{bmatrix} X_n \\ Y_n \end{bmatrix} \xrightarrow{D} \begin{bmatrix} X \\ c \end{bmatrix}$$

Example 28. If $X_n \xrightarrow{D} Z$ where $Z \sim \mathrm{N}(0,1)$, what is the asymptotic distribution of $1/X_n$?? In particular, state the probability density.

Solution. Well, function $f(x) = 1/x$ is continuous for $x \in \mathbb{R} - \{0\}$ and discontinuous only for $x \in \{0\}$.

- Because the probability $P_{\mathrm{N}(0,1)}(X \in \mathbb{R} - \{0\}) = 1 - \cancelto{0}{P_{\mathrm{N}(0,1)}(Z \in \{0\})} = 1$ then Theorem 27(1) can be applied. So $1/X_n \xrightarrow{D} 1/Z$.
- So I need to find the distribution of $\xi = 1/Z$ where $Z \sim \mathrm{N}(0,1)$.
- By using random variable transformation (Stat. Concepts 2)

$$\begin{aligned} \pi_\xi(\xi) &= \pi_Z(1/\xi)\left|\frac{\mathrm{d}}{\mathrm{d}\xi} f^{-1}(\xi)\right| = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{1}{\xi} - 0\right)^2\right)\left| - \frac{1}{\xi^2}\right| \\ &= \frac{1}{\sqrt{2\pi}} \frac{1}{\xi^2} \exp\left(-\frac{1}{2}\frac{1}{\xi^2}\right), \ \forall \xi \in \mathbb{R} - \{0\} \end{aligned}$$

Example 29. If $X_n = \frac{1}{n}$, and $f(x) = \begin{cases} 1, & \text{if } x > 0 \\ 0, & \text{if } x \leq 0 \end{cases}$, can we say $X_n \xrightarrow{P} 0 \implies f(X_n) \xrightarrow{D} f(0)$?

Solution. No, we CANNOT. Theorem 27(1) is not applied. This is because of the following

- $X_n \xrightarrow{D} X$ where X is a degenerate random variable in zero (i.e. $P(X = 0)$)
- because the $f(\cdot)$ is not continuous a.s., in fact $\mathrm{P}(X \in C(f)) = 1 - \mathrm{P}(X \in C(f)) = 1 - \mathrm{P}(X = 0) = 0 < 1$

so the assumption is violated.

Corollary 30. *If*

- $X_n \in \mathbb{R}^d$ *such that* $X_n \xrightarrow{D} X$, *and*
- $Y_n \in \mathbb{R}^k$ *such that* $X_n \xrightarrow{D} c$, *where* c *is a constant and*
- $f : \mathbb{R}^d \times \mathbb{R}^k \to \mathbb{R}^r$, *is such that* $P((X, c)^T \in C(f)) = 1$,

then

$$f(X_n, Y_n) \xrightarrow{D} f(X, c)$$

Proof. From Theorems 27(1) and 27(3). □

Example 31. Show that: if $X_n \in \mathbb{R}^d$ such that $X_n \xrightarrow{D} X$, and $Y_n \in \mathbb{R}^d$ such that $Y_n \xrightarrow{D} c$, where c is a constant, then $X_n^T Y_n \xrightarrow{D} X^T c$

Solution. This is straightforward from Theorem 30, and given a function $f : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ with $f(v, u) = v^T u$.

Example 32. Let $X_1, X_2, ...$ be IID random quantities each of them following (the same) distribution with mean $\mathrm{E}(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2 > 0$. Show that:

$$\sqrt{n}\frac{\bar{X}_n - \mu}{S_n} \xrightarrow{D} \mathrm{N}(0, 1)$$

where $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ and $S_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2$.

Solution.

- From the CLT, I know that

$$\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \xrightarrow{D} \mathrm{N}(0, 1)$$

- It is $\mathrm{Var}(X_j) < \infty$ so $\mathrm{E}(X_j^2) < \infty$ and $\mathrm{E}(|X_j|) < \infty$. Then from the weak law of large numbers we have $\bar{X}_n \xrightarrow{D} \mu$ and $\frac{1}{n}\sum_{i=1}^n X_i^2 \xrightarrow{D} \mathrm{E}X_j^2$.
- It is $S_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \bar{X}_n^2$. Because function $f(x,y) = x - y$ is continuous, it is
$$S_n^2 \xrightarrow{D} \mu_2 - \mu^2 = \sigma^2$$
where $\mu_2 = \mathrm{E}X_j^2$ from Theorem 30. Actually, because σ^2 is constant, it is $S_n^2 \xrightarrow{P} \sigma^2$ or $\frac{S_n^2}{\sigma^2} \xrightarrow{P} 1$ or $\frac{S_n^2}{\sigma^2} \xrightarrow{D} 1$
- Because function $f(x,y) = x/\sqrt{y}$ is continuous apart from 0 where $P_{\mathrm{N}(0,1)}(X \in \mathbb{R} - \{0\}) = 1 - \cancel{P_{\mathrm{N}(0,1)}(Z \in \{0\})}^{0} = 1$, then it is
$$\frac{\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma}}{\sqrt{\frac{S_n^2}{\sigma^2}}} \xrightarrow{D} \mathrm{N}(0,1) \qquad \Longrightarrow \qquad \sqrt{n}\frac{\bar{X}_n - \mu}{S_n} \xrightarrow{D} \mathrm{N}(0,1)$$

There are analogous Slutsky theorems for convergence in Probability (Theorem 33), and for convergence almost surely (Theorem 34).

Theorem 33. *(Slutsky theorems for convergence in Probability)*

As a homework, just put $\xrightarrow{P}$ *in the blanks*

1. *If $X_n \in \mathbb{R}^d$ a random vector such as $X_n \xrightarrow{P} X$, and $f : \mathbb{R}^d \to \mathbb{R}^k$ is such that $P(X \in C(f)) = 1$, where $C(f)$ is the continuity set of f, then , and then*
$$f(X_n) \xrightarrow{P} f(X)$$
2. *If $X_n \xrightarrow{P} X$ and $(X_n - Y_n) \xrightarrow{P} 0$ then*
$$Y_n \xrightarrow{P} X$$
3. *If $X_n \in \mathbb{R}^d$ where $X_n \xrightarrow{P} X$, and $Y_n \in \mathbb{R}^k$ where $Y_n \xrightarrow{P} Y$ then*
$$\begin{bmatrix} X_n \\ Y_n \end{bmatrix} \xrightarrow{P} \begin{bmatrix} X \\ Y \end{bmatrix}$$

Theorem 34. *(Slutsky theorems for convergence almost surely)*

As a homework, just put $\xrightarrow{as}$ *in the blanks*

1. *If $X_n \in \mathbb{R}^d$ a random vector such as $X_n \xrightarrow{as} X$, and $f : \mathbb{R}^d \to \mathbb{R}^k$ is such that $P(X \in C(f)) = 1$, where $C(f)$ is the continuity set of f, then , and then*
$$f(X_n) \xrightarrow{as} f(X)$$

2. *If* $X_n \xrightarrow{as} X$ *and* $(X_n - Y_n) \xrightarrow{as} 0$ *then*

$$Y_n \xrightarrow{as} X$$

3. *If* $X_n \in \mathbb{R}^d$ *where* $X_n \xrightarrow{as} X$, *and* $Y_n \in \mathbb{R}^k$ *where* $Y_n \xrightarrow{as} Y$ *then*

$$\begin{bmatrix} X_n \\ Y_n \end{bmatrix} \xrightarrow{as} \begin{bmatrix} X \\ Y \end{bmatrix}$$

Definition 35. Random variables X_n and Y_n are asymptotically equivalent iff $(X_n - Y_n) \xrightarrow{P} 0$.

Remark 36. Essentially Theorem 27(2) says that asymptotically equivalent random variables have the same asymptotic distributions.

Exercise sheet

Exercise #12

6 Mann-Wald notation (Big and Little Oh pee)

The following notation [8] is useful in order to denote the order of magnitude of specified quantities.

Definition 37. Let two sequences of random vectors (X_n) and (Y_n). Then

Little Oh pee we write $X_n = o_P(Y_n)$ iff

$$\frac{X_n}{|Y_n|} \xrightarrow{p} 0$$

namely for any $\epsilon > 0$, and any $\delta > 0$ and a finite $N_\epsilon > 0$ such that

$$P(\frac{|X_n|}{|R_n|} \leq \delta) \geq 1 - \epsilon, \quad \text{for any, } n \geq N_\epsilon$$

Hence it means that X_n converges in probability to zero at a rate Y_n. Little Oh pee gives a strict statement of an upper bound on the rate of convergence of Y_n as n increases.

Big Oh pee we write $X_n = O_P(Y_n)$ iff for any $\epsilon > 0$, there exists a finite $\delta_\epsilon > 0$ and a finite $N_\epsilon > 0$ such that

$$P(\frac{|X_n|}{|R_n|} \leq \delta_\epsilon) \geq 1 - \epsilon, \quad \text{for any, } n \geq N_\epsilon$$

Hence it means that X_n is bounded in probability to zero at a rate Y_n

Theorem 38. *Consider function* $f : \mathbb{R}^k \to \mathbb{R}$ *such that* $f(0) = 0$.

1. if $X_n \xrightarrow{P} 0$ and $f(x) = o(|x|^p)$ as $x \to 0$, then $f(X_n) = o_P(|x|^p)$
2. if $X_n \xrightarrow{P} 0$ and $f(x) = O(|x|^p)$ as $x \to 0$, then $f(x) = O_P(|x|^p)$

Proof. Omitted; see [11] □

Proposition 39. *If $X_n \xrightarrow{D} X$ then $X_n = O_P(1)$, meaning that X_n is bounced in probability.*

This means that X_n cannot be growing arbitrary large in magnitude, *if X_n* converges in distribution to same variable X.

Proof. Omitted; see [11] □

Fact 40. *Some rules*

$$o_P(1) + o_P(1) = o_P(1)\,; \qquad o_P(1) + O_P(1) = O_P(1) \tag{6.1}$$

$$o_P(1)O_P(1) = o_P(1)\,; \qquad \frac{1}{1 + o_P(1)} = O_P(1) \tag{6.2}$$

$$o_P(O_P(1)) = o_P(1) \tag{6.3}$$

more rules

$$O_P(R_n) = R_n O_P(1) \qquad o_P(R_n) = R_n o_P(1) \tag{6.4}$$

$$O_P(c_n)O_P(d_n) = O_P(c_n d_n); \qquad O_P(c_n)o_P(d_n) = o_P(c_n d_n) \tag{6.5}$$

$$O_P(c_n) + O_P(d_n) = O_P(\max(c_n, d_n)) \tag{6.6}$$

even more rules

$$\textit{If} \quad h_n \to 0, \quad \textit{and} \quad X_n = O_P(h_n) \quad \textit{then} \quad X_n = o_P(1). \tag{6.7}$$

$$O_P(\frac{1}{\sqrt{n}}) = o_P(1) \tag{6.8}$$

Proof. The proofs are omitted, but they are consequences of Slutsky theorems, Taylor expansion, Markov inequality, Theorem 38, Proposition 39, and probability calculus. □

Example 41. Show that

$$G^2 - X^2 \xrightarrow{P} 0$$

where $G^2 = 2\sum_{\forall i,j,k} n_{ijk} \log(\frac{n_{ijk}}{\hat{\mu}_{ijk}})$, and $X^2 = \sum_{\forall i,j,k} \frac{(n_{ijk} - \hat{\mu}_{ijk})^2}{\hat{\mu}_{ijk}}$.

Hint-1: Use Taylor expansion: $\log(1 + x) = x - \frac{1}{2}x^2 + O(x^3)$ as $x \to 0$.

Solution. Right... It is

$$\begin{aligned}
G^2 =& 2\sum_{\forall i} \log \frac{n_i}{\hat{\mu}_i} = 2n \sum_{\forall i} p_i \log(1 + \frac{p_i - \hat{\pi}_i}{\hat{\pi}_i}) \\
\overset{\text{Taylor}}{=}& 2n \sum_{\forall i} (\hat{\pi}_i + (p_i - \hat{\pi}_i)) \left((p_i - \hat{\pi}_i) - \frac{1}{2} \frac{(p_i - \hat{\pi}_i)^2}{\hat{\pi}_i} + O((p_i - \hat{\pi}_i)^3) \right) \\
=& 2n \sum_{\forall i} \left((p_i - \hat{\pi}_i) - \frac{1}{2} \frac{(p_i - \hat{\pi}_i)^2}{\hat{\pi}_i} + \frac{(p_i - \hat{\pi}_i)^2}{\hat{\pi}_i} + O((p_i - \hat{\pi}_i)^3) \right) \\
\overset{\text{re-arrange}}{=}& 2n \sum_{\forall i} \left((p_i - \hat{\pi}_i) - \frac{1}{2} \frac{(p_i - \hat{\pi}_i)^2}{\hat{\pi}_i} + \frac{(p_i - \hat{\pi}_i)^2}{\hat{\pi}_i} + O(((p_i - \pi_i) - (\hat{\pi}_i - \pi_i))^3) \right) \\
=& n \sum_{\forall i} \frac{(p_i - \hat{\pi}_i)^2}{\hat{\pi}_i} + 2n O_P(n^{-3/2}) \qquad (6.9) \\
=& X^2 + O_P(n^{-1/2}) = X^2 + o_P(1)
\end{aligned}$$

Regarding (6.9) notice that: (i.) $\sqrt{n}(p_i - \pi_i) \xrightarrow{D}$ Normal implies $p_i - \pi_i \xrightarrow{P} 0$, namely $p_i - \pi_i = o_P(1)$; (ii.) same story for $\hat{\pi}_i - \pi_i = o_P(1)$; (iii.) then given that $p_i - \hat{\pi}_i = o_P(1)$, we use Theorem 38 and get that $O((p_i - \hat{\pi}_i)^3) = O_P(n^{-3/2})$.
So

$$G^2 - X^2 = o_P(1) \implies G^2 - X^2 \xrightarrow{P} 0$$

aka G^2 and X^2 are asymptotically equivalent.

7 Cramer's theorem & the Delta method

Cramer theorem is another implication of Slutsky Theorems. Briefly, it refers to convergence in distribution, and it says (more or less) that smooth differential functions of asymptotically Normal variables (or statistics) are asymptotically Normal too. So, Normality can be transmitted under specific conditions ... like virus.

Theorem 42. *(Cramer Theorem) Let function $g : \mathbb{R}^d \to \mathbb{R}^k$ such that $\dot{g}(x)$ is continuous in a neighborhood of $\mu \in \mathbb{R}^d$. If $X_n \in \mathbb{R}^d$ is a sequence of random vectors such that $n^a(X_n - \mu) \xrightarrow{D} X$, where $a > 0$, then*

$$n^a(g(X_n) - g(\mu)) \xrightarrow{D} \dot{g}(\mu) X$$

Proof. Because $(\frac{1}{n})^a \to 0$ for $a > 0$, and $n^a(X_n - \mu) \xrightarrow{D} X$, then $(X_n - \mu) = \frac{1}{n^a} n^a(X_n - \mu) \xrightarrow{D} 0$, by Slutky theorem. Therefore $(X_n - \mu) \xrightarrow{D} 0 \implies (X_n - \mu) \xrightarrow{P} 0 \implies X_n \xrightarrow{P} \mu$. Hence, I get

$$X_n \xrightarrow{P} \mu$$

Because $\dot{g}(x)$ is continuous in a neighborhood $\{x : |x-\mu| < \delta\}$, then by Mean Value Theorem for $\{|x-\mu| < \delta\}$

$$g(x) = g(\mu) + \int_0^1 \dot{g}(\mu + v(x-\mu))\mathrm{d}u\,(x-\mu). \tag{7.1}$$

So for $|X_n - \mu| < \delta$,

$$n^a(g(X_n) - g(\mu)) = \int_0^1 \dot{g}(\mu + v(X_n-\mu))\mathrm{d}u\, n^a(X_n-\mu). \tag{7.2}$$

Because $X_n \xrightarrow{P} \mu$, it is

$$\int_0^1 \dot{g}(\mu + v(X_n-\mu))\mathrm{d}u \xrightarrow{P} \dot{g}(\mu) \tag{7.3}$$

Slutsky Theorems. Then (7.2) becomes

$$n^a(g(X_n) - g(\mu)) \xrightarrow{D} \dot{g}(\mu)X$$

by using Slutsky theorem on (7.2) and because of $n^a(X_n - \mu) \xrightarrow{D} X$ and (7.3). □

Theorem 43. *(Delta Theorem) Let function $g : \mathbb{R}^d \to \mathbb{R}^k$ such that $\dot{g}(x)$ is continuous in a neighborhood of $\mu \in \mathbb{R}^d$. If $X_n \in \mathbb{R}^d$ is a sequence of random vectors such that $\sqrt{n}(X_n - \mu) \xrightarrow{D} N(0, \Sigma)$ then*

$$\sqrt{n}(g(X_n) - g(\mu)) \xrightarrow{D} N(0, \dot{g}(\mu)\Sigma\dot{g}(\mu)^T)$$

Proof. Using Cramer's Theorem for $a = 1/2$, we have $\sqrt{n}(X_n - \mu) \xrightarrow{D} Z$, where $Z \sim \mathrm{N}(0, \Sigma)$, then $\sqrt{n}(g(X_n) - g(\mu)) \xrightarrow{D} \dot{g}(\mu)Z$. Because $Z \sim \mathrm{N}(0, \Sigma)$ it is $\dot{g}(\mu)Z \sim \mathrm{N}(0, \dot{g}(\mu)\Sigma\dot{g}(\mu)^T)$. So

$$\sqrt{n}(g(X_n) - g(\mu)) \xrightarrow{D} \mathrm{N}(0, \dot{g}(\mu)\Sigma\dot{g}(\mu)^T)$$

□

Example 44. Consider a 2×2 contingency table where $(n_{i,j})$ is the (i,j)th cell count, and π_{ij} is the (i,j)th cell probability.

1. Show that the marginal distribution of the MLE of the odd ratio $\hat{\theta}$ is such that

$$\sqrt{n}(\log(\hat{\theta}) - \log(\theta)) \xrightarrow{D} \mathrm{N}(0, \frac{1}{\pi_{11}} + \frac{1}{\pi_{12}} + \frac{1}{\pi_{21}} + \frac{1}{\pi_{22}}).$$

2. Show that

$$\frac{\log(\hat{\theta}) - \log(\theta)}{\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}} \xrightarrow{D} \mathrm{N}(0,1).$$

Hint: It is $\hat{\theta} = \frac{n_{11}n_{22}}{n_{21}n_{12}} = \frac{p_{11}p_{22}}{p_{21}p_{12}}$, , where $p_{i,j} = n_{i,j}/n$.

Solution.

1.

- In Example 25, we showed that from the CLT, have

$$\sqrt{n}(\boldsymbol{p}-\boldsymbol{\pi}) \xrightarrow{D} \mathrm{N}(0, \mathrm{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^T)$$

  where

$$\mathrm{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^T = \begin{bmatrix} (1-\pi_{11})\pi_{11} & -\pi_{11}\pi_{12} & -\pi_{11}\pi_{21} & -\pi_{11}\pi_{22} \\ -\pi_{11}\pi_{12} & (1-\pi_{12})\pi_{12} & -\pi_{12}\pi_{21} & -\pi_{12}\pi_{22} \\ -\pi_{11}\pi_{21} & -\pi_{12}\pi_{21} & (1-\pi_{21})\pi_{21} & -\pi_{21}\pi_{22} \\ -\pi_{22}\pi_{11} & -\pi_{22}\pi_{12} & -\pi_{22}\pi_{21} & (1-\pi_{22})\pi_{22} \end{bmatrix}$$

  for the whole vectorized quantities $\boldsymbol{\pi} = (\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})$, and $\boldsymbol{p} = (p_{11}, ..., p_{22})$.
- It is $\hat{\theta} = \frac{p_{11}p_{22}}{p_{21}p_{12}} \implies \log(\hat{\theta}) = \log(p_{11}) + \log(p_{22}) - \log(p_{12}) - \log(p_{21})$
- So I can specify $g(x) = \log(x_{11}) + \log(x_{22}) - \log(x_{12}) - \log(x_{21})$
- It is

$$\dot{g}(x) = \frac{\mathrm{d}}{\mathrm{d}x} g(x) = (\frac{1}{x_{11}}, -\frac{1}{x_{12}}, -\frac{1}{x_{21}}, \frac{1}{x_{22}})$$

  and hence $\dot{g}(x)$ is continuous a.s.
- Because all the assumptions of Delta Method are satisfied, it is

$$\sqrt{n}(\log(\hat{\theta}) - \log(\theta)) \xrightarrow{D} \mathrm{N}(0, \dot{g}(\pi)(\mathrm{diag}(\pi) - \pi\pi^T)\dot{g}(\pi)^T)$$

  with

$$\begin{aligned} \dot{g}(\boldsymbol{\pi})(\mathrm{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^T)\dot{g}(\boldsymbol{\pi})^T &= \dot{g}(\boldsymbol{\pi})(1,-1,-1,1)^T \\ &= \frac{1}{\pi_{11}} + \frac{1}{\pi_{12}} + \frac{1}{\pi_{21}} + \frac{1}{\pi_{22}} \end{aligned}$$

2. Using Slutsky theorem, and law of large numbers, similar to Example 32, we find that

$$\frac{\sqrt{\frac{1}{p_{11}} + \frac{1}{p_{12}} + \frac{1}{p_{21}} + \frac{1}{p_{22}}}}{\sqrt{\frac{1}{\pi_{11}} + \frac{1}{\pi_{12}} + \frac{1}{\pi_{21}} + \frac{1}{\pi_{22}}}} \xrightarrow{P} 1$$

   and by using Slutsky theorem as in Example 32, we find

$$\frac{\log(\hat{\theta}) - \log(\theta)}{\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}} \xrightarrow{D} \mathrm{N}(0,1).$$

Exercise. 13 Consider an M-way contingency table. Consider the quantities obs. cell counts, cell probabilities, cell proportions in their vectorised forms as

$$\boldsymbol{n} = (n_1, ..., n_N)^T \; ; \qquad \boldsymbol{p} = (p_1, ..., p_N)^T \qquad \boldsymbol{\pi} = (\pi_1, ..., \pi_N)^T \; ;$$

where $n_+ = \sum_{j=1}^{N} n_j$, and $p_j = n_j / n_+$.

1. Consider a constant matrix $\boldsymbol{C} \in \mathbb{R}^{k \times N}$, and show that

$$\sqrt{n}(\boldsymbol{C} \log(\boldsymbol{p}) - \boldsymbol{C} \log(\boldsymbol{\pi})) \xrightarrow{D} \mathrm{N}(0, \boldsymbol{C} \mathrm{diag}(\boldsymbol{\pi})^{-1} \boldsymbol{C}^T - \boldsymbol{C} 1 1^T \boldsymbol{C}^T) \tag{7.4}$$

2. Consider a 3×3 contingency table with probabilities in a vectorized form as

$$\boldsymbol{\pi} = (\ \pi_{11}\ ,\ \pi_{21}\ ,\ \pi_{31}\ ,\ \pi_{12}\ ,\ \pi_{22}\ ,\ \pi_{32}\ ,\ \pi_{13}\ ,\ \pi_{23}\ ,\ \pi_{33}\ )^T$$

   Also $\boldsymbol{n}$, $\boldsymbol{p}$ are vectorized likewise. Find the joint asymptotic distribution of the vector of different log odd ratios

$$\log(\boldsymbol{\theta}^{\boldsymbol{C}}) = \begin{bmatrix} \log(\frac{\pi_{11}\pi_{22}}{\pi_{21}\pi_{12}}) \\ \log(\frac{\pi_{22}\pi_{33}}{\pi_{23}\pi_{32}}) \end{bmatrix}$$

Remark 45. Exercise above shows that if we wish to find the joint asymptotic distribution of a number of odds ratio $\boldsymbol{\theta}^{\boldsymbol{C}}$, we can write the vector of odds ratios in a vectorized form as

$$\log(\boldsymbol{\theta}^{\boldsymbol{C}}) = \boldsymbol{C} \log(\boldsymbol{\pi})$$

and them use (7.4). Here each row of the matrix $\boldsymbol{C}$ contains zeros except for two $+1$ elements and two -1 elements in the positions multiplied by elements of $\log(\boldsymbol{p})$ to form the given log odds ratio $\log(\boldsymbol{\pi})$.

Exercise sheet

Exercise #11
Exercise #15
Exercise #13
Exercise #14

7.1 Variance stabilizing transformations

Variance stabilizing transformations are applied to statistics (data functions) with purpose to define new statistics whose variation does not depend on the unknown nuisance parameters of the sampling distribution (such as the expected value or higher moments of the statistic).

Consider we wish to perform inference on the parameter θ based on a statistic $S(\theta, X_{1:n})$; i.e. construct confidence interval for θ. It is often inconvenient when the variance (asymptotic or exact) of the statistic $S(\theta, X_{1:n})$ involves unknown quantities such as moments e.g. $\mu_r = \mathrm{E}(S(\theta, X_{1:n})^r)$.

- One way to address this issue is to plug in the sample analogs of such unknown quantities, e.g. $\overline{X^r}$, and use Slutsky theorems to compute the asymptotic distribution. However, such a treatment may increase the variance of the statistic causing problems, i.e. too wide confidence interval for θ; e.g. recall the Z and T statistics from SC2.
- Another way is to transform the statistic $S(\theta, X_{1:n})$ by creating a new one whose variance (asymptotic or exact) is independent of these unknown (and troublesome) quantities. Variable stabilizing transformations use Theorem 43 as a bases to create transformations whose asymptotic variance is independent of θ. Here, we consider the 1D case.

The idea Let $X_n \in \mathbb{R}$ be a sequence of random vectors such that $\sqrt{n}(X_n - \mu(\theta)) \xrightarrow{D} \mathrm{N}(0, \sigma^2(\theta))$ where $\mu(\theta)$ and $\sigma^2(\theta)$ are functions of the parameter of interest θ. The variable stabilizing transformation $g(\cdot)$ is the solution of the differential equation resulting from the Delta method method such that

$$\dot{g}(\mu(\theta))\sigma(\theta) = \cancel{\text{constant}} = 1 \tag{7.5}$$

Then, by Theorem 43,

$$\sqrt{n}(g(X_n) - g(\mu(\theta))) \xrightarrow{D} \mathrm{N}(0, 1)$$

Exercise 46. Consider IID random sample from a Poisson distr $X_i \sim \mathrm{Poi}(\lambda)$, $i = 1, ..., n$. Find the variance stabilizing transformation that will allow us to find asymptotic confidence intervals for the parameter λ.

Hint: If $X_i \sim \mathrm{Poi}(\lambda)$, then $\mathrm{E}(X_i) = \mathrm{Var}(X_i) = \lambda$

Solution. By CLT, it is $\sqrt{n}(\bar{X}_n - \lambda) \xrightarrow{D} \mathrm{N}(0, \lambda)$. To find the stabilising transofrmation, I need to solve (7.5)

$$\dot{g}(\mu(\theta))\sigma(\theta) = 1 \Longrightarrow \dot{g}(\lambda) = 1/\sqrt{\lambda} \Longrightarrow g(\lambda) = 2\sqrt{\lambda}$$

So, because $\dot{g}(\lambda)^2\lambda = 1$ it is $\sqrt{n}(2\sqrt{\bar{X}_n} - 2\sqrt{\lambda}) \xrightarrow{D} \mathrm{N}(0, 1)$.
The asymptotic $1 - a$ CI for $2\sqrt{\lambda}$ is $\{\, 2\sqrt{\bar{X}_n} \pm z_{1-\frac{a}{2}} \frac{1}{2\sqrt{n}} \,\}$

Example 47. Find the variance stabilizing transformation, for the case that $\sigma^2(\theta) = \theta^b$, $b \neq 0$. Does it ring any bell?

Solution. It is

$$g'(\theta)\sigma(\theta) = 1 \Longleftrightarrow g'(\theta) = \theta^{-b}$$

Then

$$g(\theta) = \int \theta^{-b} \mathrm{d}\theta = \begin{cases} \log(\theta) & , \text{if } b = 1 \\ \frac{2}{2-b}\theta^{\frac{2-b}{2}} & , \text{if } b \neq 1 \end{cases}$$

which is the Box-Cox transformation for $\lambda = \frac{2-b}{2}$ (although a little bit shifted by $\frac{1}{\lambda}$).

Exercise sheet

Exercise #17

7.2 Second order Delta method

Implementation of Theorem 43, when $\dot{g}(\mu) = 0$ may be impractical and provide poor approximations because the approximation gets poor. The following theorem shows that we can improve the approximation of Delta method when $\dot{g}(\mu) = 0$, by considering a 2nd order Taylor expansion in (7.1).

Theorem 48. *Let function $g : \mathbb{R} \to \mathbb{R}$ such that $\dot{g}(\mu) = 0$, and $\ddot{g}(x)$ is continuous in an interval of $\mu \in \mathbb{R}$. If $X_n \in \mathbb{R}$ is a sequence of random vectors such that $\sqrt{n}(X_n - \mu) \xrightarrow{D} N(0, \sigma^2)$ then*

$$n(g(X_n) - g(\mu)) \xrightarrow{D} \frac{\sigma^2 \ddot{g}(\mu)}{2} Y$$

where $Y \sim \chi^2_1$.

Proof. The derivation is given as an exercise (Exercise 16). □

Example 49. Consider $X, X_1, X_2, ...$ IID from a distribution with $\mu = \mathrm{E}(X)$ and $\sigma^2 = \mathrm{Var}(X) < \infty$.

1. Find the asymptotic distribution of $\bar{X}_n$, $\bar{X}_n^2$ by using the Delta method.
2. Assume $\mu = \mathrm{E}(X) = 0$. Find$\bar{X}_n^2$ by using the 2nd order Delta method.
3. When $\mu = 0$, discuss how/why the 2nd order Delta method provides more accurate than Delta method.

Solution 50.

1. By using the CPT, it is
$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{D} \mathrm{N}(0, \sigma^2)$$
Let $g(x) = x^2$, and $\dot{g}(x) = 2x$. Then by using Delta method, it is
$$\sqrt{n}(\bar{X}_n^2 - \mu^2) \xrightarrow{D} \mathrm{N}(0, 4\mu^2\sigma^2) \tag{7.6}$$
2. Let $g(x) = x^2$, and $\dot{g}(x) = 2x$. I also compute $\ddot{g}(x) = 2$, and $\dot{g}(\mu) = 0$. By using 2nd order Delta method I get
$$n\bar{X}_n^2 \xrightarrow{D} \sigma^2 \chi^2_1$$
3. If $\mu = 0$

- The exact result is

$$n\frac{\bar{X}_n^2}{\sigma^2} = (\frac{\bar{X}_n - 0}{\sigma/\sqrt{n}})^2 \sim \chi_1^2 \tag{7.7}$$

  as $\bar{X}_n \sim \mathrm{N}(0, \sigma^2/n)$.
- Delta method gives, $\sqrt{n}(\bar{X}_n - 0) \xrightarrow{D} \mathrm{N}(0, 0)$, aka $\bar{X}_n \xrightarrow{D} 0$, the mass is around a point ... strange
- 2nd order Delta method gives $n\bar{X}_n^2 \xrightarrow{D} \sigma^2\chi_1^2$, which is a more reasonable result.
- Of course, that 2nd order Delta method gives the exact result (7.7) here is a coincidence. However it gives an idea that the 2nd order Delta method gives better results than the 1st order Delta method.

Exercise sheet

Exercise #16

8 Uniformly strong law of large numbers

The Uniformly strong law of large numbers (USLLN) is a tool that we will use in order to produce our statistical tools; we will not go through many technical details[6].

§ **Statement:**

Assume $X, X_1, ..., X_n \overset{\text{IID}}{\sim} \mathrm{d}f(\cdot|\theta)$, where $\mathrm{d}f(\cdot|\theta)$ is a distribution labeled by a parameter $\theta \in \Theta$. Consider a function $U(x, \theta)$ measurable function of x for all θ, and let $\mu(\theta) = \mathrm{E}_f(U(X, \theta))$ be continuous on θ. USLLN states that

$$\sup_{\forall\theta\in\Theta} |\frac{1}{n}\sum_{i=1}^{n} U(X_i, \theta) - \mu(\theta)| \xrightarrow{\text{a.s.}} 0 \tag{8.1}$$

which is a much stronger statement than SLLN because it implies SLLN.

§ **Theory:**

USLLN in (8.1) holds under the following scenarios:

Theorem 51. *If Θ is finite and SLLN implies $\frac{1}{n}\sum_{i=1}^{n} U(X_i, \theta) \xrightarrow{a.s.} \mu(\theta)$ at each $\theta \in \Theta$, then*

$$\sup_{\forall\theta\in\Theta} |\frac{1}{n}\sum_{i=1}^{n} U(X_i, \theta) - \mu(\theta)| \xrightarrow{a.s.} 0$$

[6] For more see [5].

Proof. If $A_n(\theta) = |\frac{1}{n}\sum_{i=1}^n U(X_i,\theta) - \mu(\theta)| < \epsilon$ for $\epsilon > 0$, then

$$\begin{aligned}\mathrm{P}(\sup_{\forall\theta} A_n(\theta), \text{ as } n\to\infty) &= \mathrm{P}(\cap_{\forall\theta} A_n(\theta), \text{ as } n\to\infty)\\ &= \prod_{\forall\theta} \underbrace{\mathrm{P}(A_n(\theta), \text{ as } n\to\infty)}_{=1, \text{ if SLLN holds}} = 1\end{aligned}$$

□

If Θ is compact, additional assumptions are needed for (8.1) to hold; these assumptions are stated in Theorem 52.

Theorem 52. *If*

1. Θ *is compact*

   (a) $U(x,\theta)$ *is continuous in* θ *for all* x

   (b) $|U(x,\theta)| \le K(x)$*, for some function* $K(\cdot)$ *such that* $E_f(K(X)) < \infty$*.*

   then

$$\sup_{\forall\theta\in\Theta} |\frac{1}{n}\sum_{i=1}^n U(x_i,\theta) - \mu(\theta)| \xrightarrow{a.s.} 0$$

   Proof. It is omitted □

§ **Usage:**

Assume that $\bar{U}_n(\theta) = \frac{1}{n}\sum_{i=1}^n U(X_i,\theta)$ is a statistic based on a sample $X_1, ..., X_n \overset{\text{IID}}{\sim} \mathrm{d}f(X|\theta_0)$ where θ_0 is the real value of the parameter θ. USLLN is a sufficient condition to ensure the following:

- We can show that:

$$\frac{1}{n}\sum_{i=1}^n U(X_i,\hat{\theta}_n) \xrightarrow{\text{a.s.}} \mu(\theta_0), \quad \text{as } n\to\infty \tag{8.2}$$

  given that $\hat{\theta}_n \xrightarrow{\text{a.s.}} \theta_0$.

  Remark 53. Elaborating ...

$$|\frac{1}{n}\sum_{i=1}^n U(X_i,\hat{\theta}_n) - \mu(\theta_0)| \le |\frac{1}{n}\sum_{i=1}^n U(X_i,\hat{\theta}_n) - \mu(\hat{\theta}_n)| + |\mu(\hat{\theta}_n) - \mu(\theta_0)| \tag{8.3}$$

$$\le \sup_{\forall\theta\in\Theta} |\frac{1}{n}\sum_{i=1}^n U(X_i,\theta) - \mu(\theta)| + |\mu(\hat{\theta}_n) - \mu(\theta_0)| \tag{8.4}$$

  If $\mu(\theta)$ is continuous then $|\mu(\hat{\theta}_n) - \mu(\theta_0)| \xrightarrow{\text{a.s.}} 0$ by Slutsky theorem. However, in (8.3), we observe that we cannot get (8.2) by just using the SLLN on the first term of the right hand side. So we need a stronger condition such as the one in (8.4). If the USLLN can be used there, then I can get (8.2).

- We can show that the maximizer/minimizer $\hat{\theta}_n$ of $\bar{U}_n(\theta)$ converges to the unique maximizer/minimizer θ_0 of $\mu(\theta)$, if $\bar{U}_n(\theta)$ converges to $\mu(\theta)$, an $n \to \infty$. [7]

  A rigorous statement of the above is presented below.

  Theorem 54. *Assume $\Theta \subset \mathbb{R}^d$ is compact, and $\bar{U}_n(\theta)$ and $\mu(\theta)$ are continuous on θ. Let $\bar{U}_n(\theta) \xrightarrow{a.s.} \mu(\theta)$, let θ_0 be the unique maximizer of $\mu(\theta) < \infty$, i.e. $\theta_0 = \arg\max_{\forall\theta} \mu(\theta)$, and let $\hat{\theta}_n$ be a maximizer of $\bar{U}_n(\theta)$, i.e. $\hat{\theta}_n = \arg\max_{\forall\theta} \bar{U}_n(\theta)$. If the USLLN in (8.1) holds then $\hat{\theta}_n \xrightarrow{a.s.} \theta_0$.*

  Proof. For every $\epsilon > 0$, let $S_\epsilon = \{\theta \in \Theta : |\theta - \theta_0| \geq \epsilon\}$. Because S_ϵ is compact, $\mu(\theta)$ is continuous on S, and θ_0 is the unique maximizer, it is

$$\underbrace{\sup_{\theta \in S_\epsilon} \mu(\theta)}_{=c_\epsilon} < \mu(\theta_0).$$

  Consider the event,

$$A_{n,\epsilon} = \{X_{1:n} : \sup_{\theta\in\Theta} |\bar{U}_n(\theta) - \mu(\theta)| < \delta_\epsilon\}. \tag{8.5}$$

  Let $c_\epsilon = \sup_{\theta\in S} \mu(\theta)$, and pick a $\delta_\epsilon > 0$ such that

$$0 < \delta_\epsilon < \frac{\mu(\theta_0) - c_\epsilon}{2} \quad \Longrightarrow \quad c_\epsilon + \delta_\epsilon < \mu(\theta_0) - \delta_\epsilon. \tag{8.6}$$

  Based on the event $A_{n,\epsilon}$ in 8.5, if I pick a δ_ϵ as in (8.6), it is

$$\sup_{\theta\in S_\epsilon} \bar{U}_n(\theta) < \sup_{\theta\in S_\epsilon} \mu(\theta) + \delta_\epsilon = c_\epsilon + \delta_\epsilon < \mu(\theta_0) - \delta_\epsilon \overset{(8.5)}{\leq} \bar{U}_n(\theta_0).$$

  So if $\hat{\theta}_n \in S$ (aka $\hat{\theta}_n$ away from θ_0), then θ_0 would yield a strictly larger value of $\bar{U}_n(\cdot)$, which is a contradiction to $\hat{\theta}_n$ being the maximizer of $\bar{U}_n$. So it has to be $\hat{\theta}_n \notin S$. Then, because

$$\{x_{1:n} : \ |\hat{\theta}_n - \theta_0| < \epsilon\} \supset A_{n,\epsilon}. \tag{8.7}$$

  holds for any $\epsilon > 0$, it is

$$\begin{aligned} \Pr(\{|\hat{\theta}_n - \theta_0| < \epsilon\},\ n \to \infty) &\leq \Pr(A_{n,\epsilon},\ n \to \infty) \\ &= \Pr(\sup_{\forall\theta\in\Theta} |\bar{U}_n(\theta) - \mu(\theta)|,\ n \to \infty) \overset{\text{USLLN}}{=} 1 \end{aligned} \tag{8.8}$$

  Hence, it is$\hat{\theta}_n \xrightarrow{\text{a.s.}} \theta_0$. □

9 Asymptotic properties of maximum likelihood estimator

Recall that a sequence of random samples $\{X_i\}_{i=1}^n$ is actually a sequence of IID random variables. Let $X, X_1, X_2, ..., X_n$ be a sequence of IID random samples (aka random variables) such that

$$X_i \sim \mathrm{d}f(\cdot|\theta)$$

where $f(\cdot|\theta)$ is the PDF of a distribution $\mathrm{d}f(\cdot|\theta)$ labeled by a parameter $\theta \in \Theta$.

The theory

[7] As we will see later, the following result is is used to prove that MLE estimators are consistent.

§ Quantities

Likelihood function is denoted as

$$L_n(\theta) = L_n(\theta|X_1, ..., X_n) = \prod_{i=1}^{n} f(X_i|\theta)$$

Log Likelihood Function is denoted as

$$\ell_n(\theta) = \log(L_n(\theta)) = \sum_{i=1}^{n} \log(f(X_i|\theta))$$

Likelihood equations are denoted as

$$0 = \dot{\ell}_n(\theta) = \frac{\mathrm{d}}{\mathrm{d}\theta}\ell_n(\theta) \tag{9.1}$$

if derivative exists

Maximum likelihood estimator is denoted

$$\hat{\theta}_n = \hat{\theta}_n(X_1, ..., X_n)$$

and is it such that

$$\hat{\theta}_n \quad = \arg\sup_{\forall\theta\in\Theta} L_n(\theta) \quad = \arg\sup_{\forall\theta\in\Theta} \ell_n(\theta)$$

Fisher information denoted as

$$\mathcal{I}(\theta) \quad = \mathrm{E}_\theta(\Psi(X,\theta)\Psi(X,\theta)^T) \quad = \mathrm{E}_\theta((\nabla_\theta \log f(X|\theta))^T(\nabla_\theta \log f(X|\theta)))$$

where

$$\Psi(X,\theta) = (\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(X|\theta))^T; \qquad \dot{\Psi}(X,\theta) = (\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log f(X|\theta))$$

with

$$\mathrm{E}_\theta\Psi(X,\theta) = 0 \tag{9.2}$$

It can be shown that

$$\mathcal{I}(\theta) \overset{\text{calc.}}{=} \mathrm{Var}_\theta(\Psi(X,\theta)) \quad = \mathrm{Var}_\theta((\nabla_\theta \log f(X|\theta))^T) \tag{9.3}$$

$$\overset{\text{calc.}}{=} -\mathrm{E}_\theta(\dot{\Psi}(X,\theta)) \quad = -\mathrm{E}_\theta(\nabla^2_\theta \log f(X|\theta)) \tag{9.4}$$

Observed information at θ^* is denoted as

$$\begin{aligned}\mathcal{J}_n(\theta^*) &= -\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\ell_n(\theta)|_{\theta=\theta_*} \\ &= -\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\sum_{i=1}^{n}\log(f(X_i|\theta))|_{\theta=\theta_*}\end{aligned}$$

It can be shown that[8]:

[8] It can be extended to cases there $X_1, ..., X_n$ are independent but not IID; see[10]. –Not examinable.

- for n samples $\{X_1, ..., X_n\}$, it is

$$\mathcal{I}(\theta) = \frac{1}{n}\mathrm{E}_\theta(\mathcal{J}_n(\theta))$$

  meaning that the Fisher information $\mathcal{I}(\theta)$ is the expected information in $n = 1$ sample.
- by SLLN, as $n \to \infty$,

$$\frac{1}{n}\mathcal{J}_n(\theta) \xrightarrow{a.s.} \mathcal{I}(\theta) \tag{9.5}$$

Example 55. If $X \sim \mathrm{N}(\mu, \sigma^2)$ and $\theta = (\mu, \sigma)$, compute $\Psi(X, \theta)$, $\dot{\Psi}(X, \theta)$, and $\mathcal{I}(\theta)$.

Solution.

$$\log(f(x|(\mu,\sigma))) = -\log(\sqrt{2\pi\sigma^2}) - \frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}$$

$$\frac{\partial}{\partial\mu}\log(f(x|\theta)) = \frac{(x-\mu)}{\sigma^2}; \qquad \frac{\partial}{\partial\sigma}\log(f(x|\theta)) = -\frac{1}{\sigma} + \frac{(x-\mu)^2}{\sigma^3}$$

$$\frac{\partial^2}{\partial\mu^2}\log(f(x|\theta)) = \frac{\partial}{\partial\mu}\frac{(x-\mu)}{\sigma^2} = -\frac{1}{\sigma^2}; \qquad \frac{\partial^2}{\partial\sigma^2}\log(f(x|\theta)) = \frac{1}{\sigma^2} - 3\frac{(x-\mu)^2}{\sigma^4}$$

$$\frac{\partial^2}{\partial\mu\partial\sigma}\log f(x|\theta) = -2\frac{(x-\mu)}{\sigma^3}$$

So

$$\Psi(X,\theta) = \begin{bmatrix} \frac{(X-\mu)}{\sigma^2} \\ -\frac{1}{\sigma} + \frac{(X-\mu)^2}{\sigma^3} \end{bmatrix} \qquad \dot{\Psi}(X,\theta) = \begin{bmatrix} -\frac{1}{\sigma^2} & -2\frac{(X-\mu)}{\sigma^3} \\ -2\frac{(X-\mu)}{\sigma^3} & \frac{1}{\sigma^2} - 3\frac{(X-\mu)^2}{\sigma^4} \end{bmatrix}$$

$$\mathcal{I}(\theta) = -\mathrm{E}(\dot{\Psi}(X,\theta)) \qquad = -\mathrm{E}\begin{bmatrix} -\frac{1}{\sigma^2} & -2\frac{(X-\mu)}{\sigma^3} \\ -2\frac{(X-\mu)}{\sigma^3} & \frac{1}{\sigma^2} - 3\frac{(X-\mu)^2}{\sigma^4} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & 2\frac{1}{\sigma^2} \end{bmatrix}$$

Exercise 56. Prove (9.2), (9.3), and (9.4). Check the solution at `https://en.wikipedia.org/wiki/Fisher_information`.

Kullback-Leibler (KL) information number is defined as

$$\mathrm{KL}(f_0, f_1) = \mathrm{E}_0(\log\frac{f_0(X)}{f_1(X)}) = \int \log\frac{f_0(X)}{f_1(X)} f_0(X)\mathrm{d}X$$

where

- $\mathrm{KL}(f_0, f_1) = \infty$ if $f_0(x) = \infty$ and $f_1(x) = 1$
- $\mathrm{KL}(f_0, f_1) = 0$ if $f_0(x) = 0$ and $f_1(x) = \infty$

KL can be used as a measure of 'distance' (it is not a distance) between two functions (like density functions). In stats, it measures the ability of the likelihood ratio to distinguish between models f_0 and f_1.

Lemma. *(Shannon-Kolmogorov Information Inequality) Let f_0 and f_1 (like $f_0(\cdot) = f(\cdot|\theta_0)$ and $f_1(\cdot) = f(\cdot|\theta_1)$) be PDFs of corresponding distributions with respect to x. Then*

$$KL(f_0, f_1) = E_0 \log \frac{f_0(X)}{f_1(X)} = \int \log \frac{f_0(X)}{f_1(X)} f_0(X) dX \geq 0$$

with the equality iff $f_0(x) = f_1(x)$ a.s.

Proof. Function $\log(\cdot)$ is convex, then Jensen's inequality[9] implies

$$-K(f_0, f_1) = \mathrm{E}_0 \log \frac{f_1(X)}{f_0(X)} \;::\; \begin{cases} < \log \mathrm{E}_0 \frac{f_1(X)}{f_0(X)} & \text{, if } f_1(x) \neq f_0(x) \\ = \log \mathrm{E}_0 \frac{f_1(X)}{f_0(X)} & \text{, if } f_1(x) = f_0(x) \end{cases}$$

But

$$\mathrm{E}_0 \frac{f_1(x)}{f_0(x)} = \int \frac{f_1(x)}{f_0(x)} f_0(x) \mathrm{d}x = \int_{S_0} f_1(x) \mathrm{d}x \leq 1$$

at $S_0 = \{x : f_0(x) > 0\}$. Hence,

$$\mathrm{KL}(f_0, f_1) : \begin{cases} > 0 & \text{, if } f_1(x) \neq f_0(x) \\ = 0 & \text{, if } f_1(x) = f_0(x) \end{cases}$$

□

Theorem 57. *Let distributions df_0 & df_1 with densities $f(\cdot|\theta_0)$ & $f(\cdot|\theta_0)$. Let $X_1, ...X_n$ is a sequence of IID random samples from df_0. Let $\ell_n(\theta_0)$ & $\ell_n(\theta_1)$ be log-likelihoods based on densities $f(\cdot|\theta_0)$ & $f(\cdot|\theta_0)$ and given a realisation of samples $X_1, ...X_n$. Then*

$$E_{f_0}(\ell_n(\theta_0)) \geq E_{f_0}(\ell_n(\theta_1))$$

where $E_{f_0}(\spadesuit) = \int \spadesuit f(X|\theta_0) dX$ denotes expectation w.r.t $f(X|\theta_0)dX$.

Proof. Direct statement from Shannon-Kolmogorov Information Inequality Lemma. □

[9] Jensen's inequality: Consider a function φ, it is

- $\mathrm{E}(\varphi(x)) \leq \varphi(\mathrm{E}(x))$ if $\varphi(\cdot)$ is convex
- $\mathrm{E}(\varphi(x)) \geq \varphi(\mathrm{E}(x))$ if $\varphi(\cdot)$ is concave
- The equality holds if x is constant a.s.

§ On the consistency & asymptotic distribution of the "MLE"

Let $X, X_1, ...X_n$ be a sequence of IID random samples from sampling distribution $\mathrm{d}f(\cdot|\theta_0)$ with density $f(x|\theta_0)$, and θ_0 is the real value of the unknown parameter $\theta \in \Theta$.

MLE (ML equations roots) maximizes the function

$$\frac{1}{n}(\ell_n(\theta) - \ell_n(\theta_0)) = \frac{1}{n}\sum_{i=1}^{n} \overbrace{\log \frac{f(X_i|\theta)}{f(X_i|\theta_0)}}^{=U(X_i,\theta)}$$

From SLLN, it is

$$\frac{1}{n}\log\frac{L_n(\theta)}{L_n(\theta_0)} = \frac{1}{n}\sum_{i=1}^{n} \overbrace{\log \frac{f(X_i|\theta)}{f(X_i|\theta_0)}}^{U(X_i,\theta)=} \xrightarrow{a.s.} \mathrm{E}_{\theta_0} \overbrace{\log \frac{f(X|\theta)}{f(X|\theta_0)}}^{U(X,\theta)=} = -\mathrm{KL}(f_0, f_1) : \begin{cases} < 0 & \text{, if } f(X|\theta) \neq f(X|\theta_0) \\ & \quad a.s. \\ = 0 & \text{, if } f(X|\theta) = f(X|\theta_0) \\ & \quad a.s. \end{cases} \overbrace{\qquad\qquad}^{\mu(\theta)=} \tag{9.6}$$

where $f_\theta(\cdot) = f(\cdot|\theta)$, $f_{\theta_0}(\cdot) = f(\cdot|\theta_0)$, and $U(X_i, \theta)$ & $\mu(\theta)$ show the correspondence to (8.1).

To study the consistency of the MLE of θ; i.e. that $\hat{\theta}_n \xrightarrow{a.s.} \theta_0$.

§§ Case: Θ is finite

Since Θ is finite then USLLN applies in for $\sup_{\forall\theta} |\frac{1}{n}\sum_{i=1}^{n} U(X_i, \theta) - \mu(\theta)| \xrightarrow{a.s.} 0$. Hence, we can deduce that the maximizer of the left-hand side of (9.6), (namely $\hat{\theta}_n$), will be consistent to that of the right-hand side (namely θ_0); i.e. $\hat{\theta}_n \xrightarrow{a.s.} \theta_0$.

§§ Case: Θ is compact

We present the Cramer theorem about the consistency and asymptotic distribution of the ML equations roots.

Proposition 58. *Let $X, X_1, ..., X_n$ be a sequence of IID random samples from a distribution admitting a density $f(\cdot|\theta)$, $\theta \in \Theta$, then*

$$\frac{1}{\sqrt{n}}\dot{\ell}(\theta) \xrightarrow{D} N(0, \mathcal{I}(\theta)) \tag{9.7}$$

Proof. It is $\dot{\ell}(\theta) = n\frac{1}{n}\sum_{i=1}^{n} \Psi(X_i, \theta)$, where $\Psi(X_i, \theta)$ are IID with $\mathrm{E}(\Psi(X_i, \theta)) = 0$, and $\mathrm{Var}(\Psi(X_i, \theta)) = \mathcal{I}(\theta)$. Then by CLT

$$\underbrace{\sqrt{n}\frac{1}{n}\sum_{i=1}^{n} \Psi(X_i, \theta)}_{=\frac{1}{\sqrt{n}}\dot{\ell}(\theta)} \xrightarrow{D} \mathrm{N}(0, \mathcal{I}(\theta))$$

□

The theorem below provides a tool that allows to find the asymptotic distribution of MLE, under certain conditions.

Theorem 59. *(Cramer Theorem) Let $X, X_1, ..., X_n$ be a sequence of IID random samples from distribution with density $f(\cdot|\theta)$, $\theta \in \Theta$, and let θ_0 denote the true value of the parameter.*

If:

C.1 *Θ is an open subset of $\mathbb{R}^d$*

C.2 *$\ddot{\Psi}(x, \theta)$ exists, and is continuous for all x, and the derivative sign can pass under the integral sign.*

C.3 *Each component of $\dot{\Psi}(x, \theta)$ is bounded in absolute value by a function $K(x)$ where $E_{\theta_0}(K(X)) < \infty$*

C.4 *$\mathcal{I}(\theta) = -E_{\theta_0}(\dot{\Psi}(X, \theta))$ is positive definite*

C.5 *(Identifiability ass.) $f(x|\theta) = f(x|\theta)$ implies $\theta = \theta_0$ a.s.*

Then there exists sequence $\hat{\theta}_n$ of roots of the likelihood equations, where

- *it is strongly consistent*
$$\hat{\theta}_n \xrightarrow{a.s.} \theta_0 \tag{9.8}$$
- *has asymptotic distribution such as*
$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{D} N(0, \mathcal{I}(\theta_0)^{-1}) \tag{9.9}$$

Proof.

- Regarding consistency: It follows as a direct consequence of Theorem 54 and (9.6). [10]
- Regarding the asymptotic distribution: I expand $\dot{\ell}(\theta)$ around θ_0 by Taylor expansion (MVT...) as
$$\dot{\ell}_n(\theta) = \dot{\ell}_n(\theta_0) + \int_0^1 \sum_{i=1}^n \dot{\Psi}(X_i, \theta_0 + u(\theta - \theta_0))\mathrm{d}u \ (\theta - \theta_0) \tag{9.10}$$
Let $\hat{\theta}_n$ is any consistent ML equation root such that $\dot{\ell}_n(\hat{\theta}_n) = 0$. Then by setting $\theta = \hat{\theta}_n$, dividing by $\sqrt{n}$, and rearranging a bit the terms in (9.10), I get
$$C_n\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{1}{\sqrt{n}}\dot{\ell}_n(\theta_0) \xrightarrow{D} Z \sim \mathrm{N}(0, \mathcal{I}(\theta_0)^{-1}) \tag{9.11}$$
where
$$C_n = -\int_0^1 \sum_{i=1}^n \dot{\Psi}(X_i, \theta_0 + u(\hat{\theta}_n - \theta_0))\mathrm{d}u$$

[10] The interested student however may have a look at [11].

To derive (9.9) from (9.11), I need to show that $C_n \xrightarrow{a.s.} \mathcal{I}(\theta_0)$.

It is

$$\begin{aligned}|C_n - \mathcal{I}(\theta_0)| \leq & |C_n + \mathcal{I}(\theta_0)| \leq \int_0^1 |\sum_{i=1}^n \dot{\Psi}(X_i, \theta_0 + u(\hat{\theta}_n - \theta_0)) + \mathcal{I}(\theta_0)| \mathrm{d}u \\ & \leq \int_0^1 \underbrace{\sup_{\theta \in \{\theta : |\hat{\theta}_n - \theta_0| \leq \delta\}} |\sum_{i=1}^n \dot{\Psi}(X_i, \theta_0 + u(\hat{\theta}_n - \theta_0)) - \mathrm{E}_{\theta_0}(\dot{\Psi}(X, \theta))|}_{=(ii)} \\ & + \underbrace{|\mathrm{E}_{\theta_0}(\dot{\Psi}(x, \theta)) + \mathcal{I}(\theta_0)|}_{=(i)} \mathrm{d}u\end{aligned}$$

- About (ii)... $\dot{\Psi}(X, \theta)$ is continuous at θ_0 from [C.3] so by definition, $(\forall \epsilon' > 0)(\exists \delta > 0)$ where

$$|\mathrm{E}_{\theta_0}(\dot{\Psi}(X, \theta)) + \mathcal{I}(\theta_0)| < \epsilon \qquad \text{whenever} \quad |\hat{\theta}_n - \theta_0| < \delta \tag{9.12}$$

- About (i)... I can restrict the neighborhood below the sup to be smaller than that of (9.12). Also the conditions of Theorem 52, are satisfied from C.2 and C.3. So $(\forall \epsilon'' > 0)(\exists N > 0)(\forall n > N)$

$$\sup_{\theta \in \{\theta : |\hat{\theta}_n - \theta_0| \leq \delta\}} |\sum_{i=1}^n \dot{\Psi}(X_i, \theta_0 + u(\hat{\theta}_n - \theta_0)) - \mathrm{E}_{\theta_0}(\dot{\Psi}(X, \theta))| < \epsilon''$$

So I get that $(\forall \epsilon > 0)(\exists N > 0)(\forall n > N)$ $|C_n - \mathcal{I}(\theta_0)| < \epsilon = \max(\epsilon', \epsilon'')$, and hence $C_n \xrightarrow{D} \mathcal{I}(\theta_0)$.

By using Slutsky theorems,

$$\cancel{C_n^{-1}} \cancel{C_n} \sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{D} \mathcal{I}(\theta_0) Z \sim \mathrm{N}(0, \mathcal{I}(\theta_0) \cancel{\mathcal{I}(\theta_0)^{-1}} \cancel{\mathcal{I}(\theta_0)}) \to I$$

□

Remark 60. Theorem 59 says that given a set of regularity conditions C.1-C.5, there exists a sequence of the likelihood equation roots (aka $\hat{\theta}_n$ such that $\dot{\ell}_n(\theta)|_{\theta=\hat{\theta}_n} = 0$) which is strongly consistent and asymptotically Normal with Fisher information as covariance matrix. It does not explicitly refers to MLE. MLE may not be the consistent root, or it may not be a likelihood equation root at all.

Remark 61. If the root of likelihood equation (aka $\hat{\theta}_n$ such that $\dot{\ell}_n(\theta)|_{\theta=\hat{\theta}_n} = 0$) is unique and the conditions of Theorem 59 are satisfied, then $\hat{\theta}_n$ refers to the MLE (aka $\hat{\theta}_n = \arg\sup_{\forall \theta} \ell_n(\theta)$). Otherwise MLE may not be the consistent root, or it may not be a likelihood equation at all.

Remark 62. Theorem 59 combined with Delta methods imply that continuous functions of MLE are asymptotically Normal with mean and covariance that can be computed.

Remark 63. The conclusions of Proposition 58, and Cramer's Theorem 59 remain valid even when the samples are independent but not identically distributed. For instance, Theorem 64 is a restatement of Cramer's Theorem 59.

The Theorem below is a re-statement of Cramer's Theorem 59, for independent but non-identically distributed samples.

Theorem 64. *For $i = 1, ..., k$, let $X_{i,1}, ..., X_{i,m_i}$, be a sequence of m_i IID random samples drawn from sampling distribution $df_i(X_{i,1}, ..., X_{i,m_i}|\theta)$ labeled by a d-dimensional parameter $\theta \in \Theta$; namely*

$$X_{1,1}, ..., X_{1,m_1} \overset{IID}{\sim} df_1(\cdot|\theta) \quad ; \cdots ; \quad X_{i,1}, ..., X_{i,m_i} \overset{IID}{\sim} df_i(\cdot|\theta) \quad ; \cdots ; \quad X_{k,1}, ..., X_{k,m_k} \overset{IID}{\sim} df_k(\cdot|\theta)$$

Let the Cramer theorem conditions [C.1-C.5] hold for each $f_i(\cdot|\theta)$. Let the true value of θ be θ_0. Let $n = \sum_{i=1}^{k} m_i$.
Then there exists sequence $\hat{\theta}_n$ of roots of the likelihood equations, where

- *it is strongly consistent*
$$\hat{\theta}_n \xrightarrow{as} \theta_0$$
- *it has asymptotic distribution such as*
$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{D} N(0, \mathcal{I}(\theta_0)^{-1}) \tag{9.13}$$
where $\mathcal{I}(\theta) = \sum_{i=1}^{k} \varpi_i \mathcal{I}^{(i)}(\theta)$ is the expected information matrix, $\mathcal{I}^{(i)}(\theta)$ is the Fisher information matrix for $f_i(\cdot|\theta)$, $\varpi_i = m_i / \sum_{i=1}^{k} m_i$ for $i = 1, ..., k$.

Proof. [4] The proof is omitted and can be found in [4]; we offer the rational of the proof. The log-likelihood is
$$\ell_n(\theta) = \sum_{i=1}^{k} \sum_{j=1}^{m_i} \log(f_i(x_{i,j}|\theta)) \tag{9.14}$$
and the roots of the likelihood equations are such as $0 = \ell_n(\theta)|_{\theta=\hat{\theta}_n}$. Since $\ell_n(\theta)$, and hence its Taylor expansion, is a sum of k independent terms of the kind considered in the proof of Theorem 59, then (9.13) could be true. □

Fact 65. *Cholesky decomposition: Every symmetric, positive definite matrix Σ can be decomposed into a product of a unique lower triangular matrix L and its transpose, i.e. $\Sigma = LL^T$.*

- *The following mathematical procedure can be used to compute the Cholesky factor L of Σ.*

  for $i = 1, ...d$

  for $j = 1, .., d$

$$L_{i,j} = \begin{cases} \sqrt{\Sigma_{i,i} - \sum_{k=1}^{i-1} L_{i,k}^2} & if i = j \\ \frac{1}{L_{j,j}} (\Sigma_{i,i} - \sum_{k=1}^{i-1} L_{i,k} L_{j,k}) & if i > j \\ 0 & if i < j \end{cases}$$

  – *the computations evolve row-wise, i.e. $L_{1,1} \to L_{2,1} \to L_{2,2} \to L_{3,1} \to L_{3,2} \to L_{3,3}$ etc...*

Often it is difficult to derive a confidence interval or a hypothesis test based on Theorem 59 because it requires $\mathcal{I}(\theta)$ to be computationally tractable. Following, we use Slusky theorem to find alternative statistics that can be used as statistics for inference.

Proposition 66. *(or actually an example...) Given that the assumptions [C.1-C.5] of Theorem 59 are satisfied, and that $\mathcal{I}(\theta)$ and $\mathcal{J}_n(\theta)$ are continuous on θ, then*

$$\sqrt{n}\mathcal{I}(\theta_0)^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{D} N(0, I) \tag{9.15}$$

$$\sqrt{n}\mathcal{I}(\hat{\theta}_n)^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{D} N(0, I) \tag{9.16}$$

$$\mathcal{J}_n(\hat{\theta}_n)^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{D} N(0, I) \tag{9.17}$$

where $A^{1/2}$ denotes the lower triangular matrix of the Cholesky decomposition of A; i.e., $A = A^{1/2}(A^{1/2})^T$.

Proof. The proof, is given as an exercise in your Exercise sheet, in the case of a sequence of IID random samples. However for independent but not identically distributed the proof is out of the scope.

- Eq 9.15 results from Cramer Theorem
- Eq. 9.16 results by using Slutsly theorems with (9.8).
- Eq. 9.16 results by using USLLN in Theorem 52, and Slutsky Theorems.

□

Example 67. Consider an M way contingency table $(n_{i,j})$ generated by a Poisson sampling scheme. Consider it is modeled by a log-Linear model with link function

$$\log(\mu) = X^T\beta \tag{9.18}$$

in the vectorized form, where vector $\beta \in \mathbb{R}^d$ contains the unknown coefficients. Consider that identifiability contains have be considered in (9.18). Show that

1. the asymptotic distribution of the MLEs $\hat{\beta}_n$ is such that

$$(X\text{diag}(\mu_n)X^T)^{1/2}(\hat{\beta}_n - \beta_0) \xrightarrow{D} \text{N}(0, I) \tag{9.19}$$

   where $\hat{\mu}_n = \exp(X^T\hat{\beta}_n)$, and β_0 is the true value of β.

2. An $(1-a)100\%$ asymptotic confidence interval for β_0 is

$$\{(n_{i,j}) \,:\, (\hat{\beta}_n - \beta_0)^T(X\text{diag}(\hat{\mu}_n)X^T)(\hat{\beta}_n - \beta_0) \leq \chi^2_{d,1-a}\}$$

Solution 68.

1. Since $\hat{\beta}_n$ is an MLE then it is asymptotically Normal as

$$\mathcal{J}_n(\hat{\theta}_n)^{1/2}(\hat{\beta}_n - \beta_0) \xrightarrow{D} \mathrm{N}(0, I) \tag{9.20}$$

It is

$$\mathcal{J}_n = -\frac{\mathrm{d}^2}{\mathrm{d}\beta^2}\ell_n(\beta)|_{\beta=\hat{\beta}_n} = -\frac{\mathrm{d}^2}{\mathrm{d}\beta^2}\sum_{i=1}^{n}\log(\mathrm{Poi}(n|\mu(\beta)))|_{\beta=\hat{\beta}_n}$$

$$\frac{\mathrm{d}}{\mathrm{d}\beta_j}\ell_n(\underset{\sim}{\mu}(\beta)) = -\sum_i n_i X_{i,j} + \sum_i \exp(\sum_j X_{i,j}\beta_j)X_{i,j}$$
$$\frac{\mathrm{d}^2}{\mathrm{d}\beta_j \mathrm{d}\beta_k}\ell_n(\underset{\sim}{\mu}(\beta)) = \sum_i \exp(\sum_j X_{i,j}\beta_j)X_{i,j}X_{i,k}$$

and

$$\mathcal{J}_n = X^T \mathrm{diag}(\hat{\mu}_n)X$$

Therefore from (9.20)

$$\underbrace{(X^T \mathrm{diag}(\hat{\mu}_n)X)^{1/2}(\hat{\beta}_n - \beta_0)}_{=Z_n} \xrightarrow{D} \mathrm{N}(0, I)$$

2. I use the statistic

$$T_n = (\hat{\beta}_n - \beta_0)^T(X^T \mathrm{diag}(\hat{\mu}_n)X)(\hat{\beta}_n - \beta_0) = Z_n^T Z_n = \sum_{i=1}^{d} Z_{n,i}$$

where $T_n \sim \chi^2_d$ as summation of d standard normal variables. Then

$$1 - a = P_{\chi^2_d}(T_n < q)$$

where $q = \chi^2_{d,1-a}$.

Exercise sheet

Exercise #20

10 The information inequality & Asymptotic efficiency

§ The information inequality

Let $X_1, X_2, ...$ be IID random variables[11] with a distribution f_θ labeled by a parameter $\theta \in \Theta$. Let $\hat{\theta}_n := \hat{\theta}_n(X_{1:n})$ be a sequence of estimators of θ.

In multivariate case the variance matrix comparison $\Sigma_1 \geq \Sigma_2$ is has the following meaning.

Definition 69. For two covariance matrices Σ_1 and Σ_2 we say $\Sigma_1 \geq \Sigma_2$ if $\Sigma_1 - \Sigma_2$ is positive semi-definite[12].

The theorem below provides a lower bound (Cramer-Rao lower bound) of the variance of an estimator $\hat{\theta}_n$ of a multivariate parameter θ.

Theorem 70. *(Information inequality theorem) Let $X_1, ..., X_n \in \mathbb{R}^d$ be a sequence of IID random samples from a distribution $f_\theta(\cdot)$ labeled by an parameter $\theta \in \Theta \subset \mathbb{R}^r$ and admitting PDF $f(\cdot|\theta)$. Consider an estimator $\hat{\theta}_n := \hat{\theta}_n(X_{1:n}) \in \Theta \subset \mathbb{R}^r$ such that $g_n(\theta) = E_{f_\theta}(\hat{\theta}_n)$ exists on Θ.*

Then

$$var_{f_\theta}(\hat{\theta}_n) \geq \frac{1}{n}\dot{g}_n(\theta)\mathcal{I}(\theta)^{-1}\dot{g}_n(\theta)^T \tag{10.1}$$

where $\mathcal{I}(\theta)$ is the Fisher's information matrix.

- *The quantity $\frac{1}{n}\dot{g}_n(\theta)\mathcal{I}(\theta)^{-1}\dot{g}_n(\theta)^T$ is called Cramer-Rao lower bound (CRLB).*
- *we assumed that, $\frac{d}{d\theta}f(x|\theta)$ exists ; $\frac{d}{d\theta}$ can pass under the integral sign in $\int f(X|\theta)dX$ and $\int \hat{\theta}_n(X)f(X|\theta)dX$.*

Proof. Let $\Psi(X_{1:n}, \theta) = (\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(X_{1:n}|\theta))^T = (\frac{\mathrm{d}}{\mathrm{d}\theta}\log\prod_{i=1}^n f(X|\theta))^T = \sum_{i=1}^n \Psi(X_i, \theta)$, where $\Psi(X_i, \theta) = (\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(X_i|\theta))^T$. It is

$$\begin{aligned}
\mathrm{E}_{f_\theta}\Psi(X_{1:n}, \theta) =& 0 \quad \text{(you have proved it before)} \\
\dot{g}_n(\theta) =& \frac{\mathrm{d}}{\mathrm{d}\theta}\int \hat{\theta}_n f(X_{1:n}|\theta)\mathrm{d}X = \int \hat{\theta}_n(x_{1:n})\frac{\frac{\mathrm{d}}{\mathrm{d}\theta}f(X_{1:n}|\theta)}{f(X_{1:n}|\theta)}f(X|\theta)\mathrm{d}X \\
=& \int \hat{\theta}_n \frac{\mathrm{d}}{\mathrm{d}\theta}\log f(X_{1:n}|\theta)f(X_{1:n}|\theta)\mathrm{d}X = \mathrm{E}_{f_\theta}(\hat{\theta}_n\Psi(X_{1:n}, \theta) - \cancelto{0}{\mathrm{E}_\theta\Psi(X_{1:n}, \theta)}) \\
=& \mathrm{cov}_{f_\theta}(\hat{\theta}_n, \Psi(X_{1:n}, \theta))
\end{aligned} \tag{10.2}$$

For any $a, \gamma \in \mathbb{R}^r$, $\xi = a^T\hat{\theta}_n$ and $\zeta = \gamma^T\Psi(X_{1:n}, \theta)$. It is

$$\begin{aligned}
\mathrm{cov}_{f_\theta}(\xi, \zeta) =& a^T\mathrm{cov}_{f_\theta}(\hat{\theta}_n, \Psi(X_{1:n}, \theta))\gamma = a^T\dot{g}_n(\theta)\gamma \\
\mathrm{var}_{f_\theta}(\zeta) =& \gamma^T\mathrm{var}_{f_\theta}(\Psi(X_{1:n}, \theta))\gamma = \gamma^T\left(n\mathcal{I}(\theta)\right)\gamma \\
\mathrm{var}_{f_\theta}(\xi) =& a^T\mathrm{var}_{f_\theta}(\hat{\theta}_n)a
\end{aligned}$$

[11] ...informally a sequence of IID obvervations

[12] Matrix A is called positive semidefinite iff for any z , $zAz^T \geq 0$. It is symbolized as $A \geq 0$

So

$$1 \geq (\mathrm{corr}_{f_\theta}(\xi,\zeta))^2 = \frac{a^T \dot{g}_n(\theta)\gamma}{a^T \mathrm{var}_{f_\theta}(\hat{\theta}_n) a \, \gamma^T \left(n\mathcal{I}(\theta)\right)\gamma}$$

the right maximizes[13] with respect to γ at $\gamma^* = \frac{\mathcal{I}(\theta)^{-1}\dot{g}_n(\theta)^T a}{(a^T\mathcal{I}(\theta)a)^{1/2}}$ for all a and gives

$$1 \geq \frac{a^T \dot{g}_n(\theta)\left(n\mathcal{I}(\theta)\right)^{-1}\dot{g}_n(\theta)^T a}{a^T \mathrm{var}_{f_\theta}(\hat{\theta}_n) a}$$

namely

$$a^T \left(\mathrm{var}_{f_\theta}(\hat{\theta}_n) - \frac{1}{n}\dot{g}_n(\theta)\mathcal{I}(\theta)^{-1}\dot{g}_n(\theta)^T \right) a \geq 0$$

which completes the proof. □

Remark 71. Notice that the CRLB is attained if and only if $\hat{\theta}_n$ and $\Psi(X_{1:n},\theta)$ are linearly related.

Remark 72. Assume $n = 1$ for simplicity. Following Remark 71, CRLB is attained if and only if there are constants $a(\theta)$, $b(\theta)$ such that

$$\begin{aligned}
\Psi(X,\theta) =& a(\theta) + b(\theta)^T\hat{\theta} \Leftrightarrow \\
(\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(X|\theta))^T =& a(\theta) + b(\theta)^T\hat{\theta} \Leftrightarrow \\
(\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(X|\theta))^T =& \int a(\theta) + b(\theta)^T\hat{\theta}\,\mathrm{d}\theta \Leftrightarrow \\
\log f(X|\theta) =& \tilde{a}(\theta) + \tilde{b}(\theta)^T\hat{\theta} + h(X)
\end{aligned}$$

Namely, $\hat{\theta}$ achieves the CRLB for all θ if and only if $\hat{\theta}$ is a natural sufficient statistic of the exponential family of distributions

$$f(X|\theta) = \exp(\tilde{a}(\theta) + \tilde{b}(\theta)^T\hat{\theta}_n)\tilde{h}(X)$$

Example 73. If the estimator $\hat{\theta}_n$ has bias $b_n(\theta) = \mathrm{E}_{f_\theta}\hat{\theta}_n - \theta$ then

$$\mathrm{var}_{f_\theta}(\hat{\theta}_n) \geq \frac{1}{n}(I + \dot{b}_n)\mathcal{I}(\theta)^{-1}(I + \dot{b}_n)^T \tag{10.3}$$

Solution. It is $b_n(\theta) = \underbrace{\mathrm{E}_{f_\theta}\hat{\theta}_n}_{g} - \theta \implies \dot{b}_n(\theta) = \dot{g}_n(\theta) - I$. So replacing the terms in (10.1), I get (10.3).

Proposition 74. *If $\hat{\theta}_n$ is an unbiased estimator of θ, Theorem 70 implies that*

$$var_{f_\theta}(\hat{\theta}_n) \geq \frac{1}{n}\mathcal{I}(\theta)^{-1} \tag{10.4}$$

Definition 75. Best unbiased estimator (BUE) of θ is called the estimator $\hat{\theta}_n$ which has the lowest variance compared to other unbiased estimators. (I.e., its variance equal to the lower bound in (10.4))

[13] Fact: from Linear algebra: It is $\max_{\forall x}\frac{(a^Tx)^2}{x^TBx} = a^TB^{-1}a$ where the maximum is attained at $x^* = \frac{B^{-1}a}{(x^TB^{-1}x)^{1/2}}$

Example 76. (Cont. of Example 55) If $X_1, ..., X_n \overset{IID}{\sim} \mathrm{N}(\mu, \sigma^2)$ then, the Cramer-Rao lower bound of the variance of an unbiased estimator $\hat{\theta}_n$ of $\theta = (\mu, \sigma)$ is

$$\mathrm{var}_{\mathrm{N}(\mu,\sigma^2)}(\hat{\theta}_n)\frac{1}{n}\mathcal{I}(\theta)^{-1} = \frac{1}{n}\begin{bmatrix}\frac{1}{\sigma^2} & 0\\ 0 & 2\frac{1}{\sigma^2}\end{bmatrix}^{-1} = \frac{1}{n}\begin{bmatrix}\sigma^2 & 0\\ 0 & 2\sigma^2\end{bmatrix}$$

§ Asymptotic efficiency

Let $X, X_1, X_2, ...$ be a sequence of IID random variables with a distribution f_θ labeled by a parameter $\theta \in \Theta$. Let $\hat{\theta}_n := \hat{\theta}_n(X_{1:n})$ be a sequence of estimators of θ.

Definition 77. Estimator $\hat{\theta}_n$ is called asymptotically efficient estimator of θ if

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{D} \mathrm{N}(0, \Sigma(\theta))$$

and $\Sigma(\theta) = \mathcal{I}(\theta)^{-1}$ for all $\theta \in \Theta$, whatever the true value of the parameter θ is. Also, $\hat{\theta}_n$ is called asymptotically sub-efficient estimator of θ if $\Sigma(\theta) \neq \mathcal{I}(\theta)^{-1}$ for some $\theta \in \Theta$.[14]

Remark 78. We observe that MLE estimators $\hat{\theta}_n$ are asymptotically efficient under the assumptions C.1-C.5 of Theorem 59.

- Hence, one way to show that another estimator $\tilde{\theta}_n$ is asymptotically efficient estimator of θ, is by showing that it is asymptotically equivalent to the MLE estimator $\hat{\theta}_n$ (if it exist); i.e. $\tilde{\theta}_n - \hat{\theta}_n \xrightarrow{P} 0$ (see, Definition 35, and Theorem 27(2)).

Exercise sheet

Exercise #18

11 Alternative estimators

Often it is difficult (or just bothersome) to calculate the MLE estimator $\hat{\theta}_n$, because it requires to find the roots of the likelihood equation (9.1). Hence, we can resort to alternative estimators which may not be asymptotically efficient.

To address this issue, one one can follow the procedure:

1. Calculate another estimator $\tilde{\theta}_n := \tilde{\theta}_n(X_{1:n})$ for θ (by using an alternative estimator method). Ideally, $\tilde{\theta}_n$ should be tractable (or easier to compute), asymptotically Normal, consistent (although in practice this is violated), but not necessarily asymptotically efficient (like the MLE $\hat{\theta}_n$).

[14] However, there are examples of asymptotically super-efficient estimators... where $\Sigma(\theta) \geq \mathcal{I}(\theta)$. This is because the Information inequality theorem refers to the exact Variance, and not the asymptotic one.

- In Section 11.1, we introduce the moments estimators.

2. Use a recursive procedure, by using the sub-efficient estimator $\tilde{\theta}_n$ as a first guess, with purpose to derive an improved estimator in terms of asymptotic variance.

   - In Section 11.2, we introduce the Newton, and the Fisher scoring algorithms.

11.1 Method of moments

The method of moments is an alternative pretty simple method producing estimators needed in (1).

Let $X, X_1, X_2, ..., X_n$ be a sequence of IID random samples generated from a distribution f_θ labeled by a d-dimensional parameter $\theta \in \Theta \subset \mathbb{R}^d$, and admitting PDF $f(\cdot|\theta)$.

Definition 79. (Method of moments) The (method of) moments estimator $\tilde{\theta}_n := \tilde{\theta}_n(X_{1:n})$ for the parameter θ is produced by solving the equations

$$\bar{g}_j = \mu_j(\theta), \quad \forall j = 1, ..., d$$

where $\bar{g}_j = \frac{1}{n}\sum_{i=1}^n g_j(X_i)$, and $\mu_j(\theta) = \mathrm{E}_f(g_j(X)|\theta)$, for given functions $g_1(\cdot), g_2(\cdot), ..., g_d(\cdot)$.

A popular choice for the functions $g_j(\cdot)$ is

$$g_j(x) = x^j, \quad \text{for } j = 1, ..., d.$$

To make the notation compact, we define $g(x) = (g_1(x), ..., g_d(x))^T$, $\bar{g} = (\bar{g}_1, \bar{g}_2, ..., \bar{g}_d)^T$, and $\mu(\theta) = \mathrm{E}_f(g(X)|\theta) = (\mathrm{E}_f(g_1(X)|\theta), \mathrm{E}_f(g_2(X)|\theta), ..., \mathrm{E}_f(g_d(X)|\theta))$.

Under mild conditions (given below), moment estimators can be consistent, and asymptotically Normal, however they are not necessarily asymptotically efficient.

- The moment estimator is uniquely defined as

  $$\tilde{\theta}_n = \mu^{-1}(\bar{g})$$

  if $\mu(\cdot)$ is continuous and 1-1 (bijective).

- Let θ_0 be the true value of parameter θ. By the CLT, it is

  $$\sqrt{n}(\bar{g} - \mu(\theta_0)) \xrightarrow{D} \mathrm{N}(0, \Sigma(\theta_0))$$

  where $\mu(\theta_0) = \mathrm{E}_f(g(X)|\theta_0)$, and $\Sigma(\theta_0) = \mathrm{var}_f(g(X)|\theta_0)$.

- By Delta method, it is

  $$\sqrt{n}(\tilde{\theta}_n - \theta_0) = \sqrt{n}(\mu^{-1}(\bar{g}) - \mu^{-1}(\mu(\theta_0))) \xrightarrow{D} \mathrm{N}(0, \dot{\mu}_{\theta_0}^{-1}\Sigma(\theta_0)(\dot{\mu}_{\theta_0}^{-1})^T) \tag{11.1}$$

  where $\Sigma_{g,\theta_0} = \mathrm{var}_f(g(X)|\theta_0)$, and $\dot{\mu}_{\theta_0}^{-1}$ is the inverse of the derivative $\dot{\mu}(\theta_0) = \frac{\mathrm{d}}{\mathrm{d}\theta}\mu(\theta)|_{\theta=\theta_0}$, if $\mu(\cdot)$ is continuously differentiable at $\theta = \theta_0$

In (11.1) it is possible that $\dot{\mu}_{\theta_0}^{-1}\Sigma(\theta_0)(\dot{\mu}_{\theta_0}^{-1})^T \neq \mathcal{I}(\theta_0)^{-1}$.

11.2 Improving sub-efficient estimators

The following algorithms are recursive procedures which use sub-efficient estimators (e.g., the moment estimator) as initial guesses, with purpose to produce an improved one.

Consider $X, X_1, X_2, ..., X_n$ be a sequence of IID random samples from a distribution f_θ labeled by a d-dimensional parameter $\theta \in \Theta \subset \mathbb{R}^d$, and admitting PDF $f(\cdot|\theta)$.

Let, $\tilde{\theta}_n$ be a sub-efficient estimator of parameter θ.

Newton's algorithm Set $\breve{\theta}_n^{(0)} = \tilde{\theta}_n$, as an initial guess. For $t = 1, ...,$ iterate

$$\breve{\theta}_n^{(t)} = \breve{\theta}_n^{(t-1)} - \ddot{\ell}_n(\breve{\theta}_n^{(t-1)})^{-1}\dot{\ell}_n(\breve{\theta}_n^{(t-1)}) \tag{11.2}$$

$$\overset{\text{or}}{=} \breve{\theta}_n^{(t-1)} + \mathcal{J}_n(\breve{\theta}_n^{(t-1)})^{-1}\dot{\ell}_n(\breve{\theta}_n^{(t-1)}) \tag{11.3}$$

because $\mathcal{J}_n(\cdot) = -\ddot{\ell}_n(\cdot)$.

Fisher's scoring algorithm Set $\breve{\theta}_n^{(0)} = \tilde{\theta}_n$, as an initial guess. For $t = 1, ...,$ iterate

$$\breve{\theta}_n^{(t)} = \breve{\theta}_n^{(t-1)} + \frac{1}{n}\mathcal{I}(\breve{\theta}_n^{(t-1)})^{-1}\dot{\ell}_n(\breve{\theta}_n^{(t-1)}) \tag{11.4}$$

The connection between the two algorithms is (9.5). In fact, Fisher proposed to replace the observed $\mathcal{J}_n$ (in Newton's algorithm) with the expected $\mathcal{I}$; because he observed that Newton's algorithm (11.2) was getting trapped into undesirable values of θ when the sample size n was small.

One step estimators

Definition 80. One step estimators are the estimators (11.2) and (11.4) using only one iteration $t = 1$, for a given sub-efficient estimator used as an initial guess.

$$\text{Newton alg.} \qquad \breve{\theta}_n = \tilde{\theta}_n - \ddot{\ell}_n(\tilde{\theta}_n)^{-1}\dot{\ell}_n(\tilde{\theta}_n)$$

$$\text{Fisher scoring alg.} \qquad \breve{\theta}_n = \tilde{\theta}_n + \frac{1}{n}\mathcal{I}(\tilde{\theta}_n)^{-1}\dot{\ell}_n(\tilde{\theta}_n)$$

Under conditions, one iteration/step of (11.2) or (11.4) is enough to derive asymptotically efficient estimators.

The following theorem proves that only one iterative step in (11.2) or (11.4) is required to match the asymptotic efficiency of solutions to the likelihood equations, which, from Cramér Theorem 59, have been shown to have asymptotic variance equal to the Cramér-Rao information bound.

Theorem 81. *Let $\tilde{\theta}_n$ be a strongly consistent sequence such that*

$$\sqrt{n}(\tilde{\theta}_n - \theta_0) \xrightarrow{D} N(0, \Sigma(\theta_0))$$

for some $\Sigma(\theta_0) > 0$, where θ_0 is the true value of the parameter. Assume assumptions of Theorem 59 are satisfied. Then the one-step estimators

$$\textit{Newton alg.} \qquad \breve{\theta}_n = \tilde{\theta}_n - \ddot{\ell}_n(\tilde{\theta}_n)^{-1}\dot{\ell}_n(\tilde{\theta}_n) \tag{11.5}$$

$$\textit{Fisher scoring alg.} \qquad \breve{\theta}_n = \tilde{\theta}_n + \frac{1}{n}\mathcal{I}(\tilde{\theta}_n)^{-1}\dot{\ell}_n(\tilde{\theta}_n) \tag{11.6}$$

are asymptotically equivalent to the MLE, and hence asymptotically efficient. [15]

Proof. I 'll show the prove for the statement for (11.5), and you can do the same for (11.6), for practice.

Let $\hat{\theta}_n$ be the MLE (or precisely the consistent likelihood equation root of Theorem 59). To prove the statement of the Cramer Theorem for (11.5), I need to show that

$$\breve{\theta}_n - \hat{\theta}_n \xrightarrow{\mathrm{P}} 0$$

(see Definition 35), which, according to Slutsky Theorem 27 and Cramer Theorem 59, implies that

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

Then, from (11.5)

$$\sqrt{n}(\breve{\theta}_n - \hat{\theta}_n) = \sqrt{n}(\tilde{\theta}_n - \hat{\theta}_n - \ddot{\ell}_n(\tilde{\theta}_n)^{-1}\dot{\ell}_n(\tilde{\theta}_n))$$

Then by using the Mean value theorem (1st order Taylor) to expand $\dot{\ell}_n(\tilde{\theta}_n)$ around $\hat{\theta}_n$, I get

$$\dot{\ell}_n(\tilde{\theta}_n) = \dot{\ell}_n(\hat{\theta}_n) + \int_0^1 \ddot{\ell}_n(\hat{\theta}_n + u(\tilde{\theta}_n - \hat{\theta}_n))\mathrm{d}u\,(\tilde{\theta}_n - \hat{\theta}_n)$$

[15] This theorem can be seen as an exercise. If it was given as an exercise, it would be given with several sub-questions leading to the final result; e.g.

1. Expand $\dot{\ell}_n(\tilde{\theta}_n)$ around $\hat{\theta}_n$ by Mean value theorem
2. Show that $\sqrt{n}(\breve{\theta}_n - \hat{\theta}_n) = \left(I - \left(\frac{1}{n}\ddot{\ell}_n(\tilde{\theta}_n)\right)^{-1} \int_0^1 \frac{1}{n}\ddot{\ell}_n(\hat{\theta}_n + u(\tilde{\theta}_n - \hat{\theta}_n))\mathrm{d}u\right) \times \left(\sqrt{n}(\tilde{\theta}_n - \theta_0) - \sqrt{n}(\hat{\theta}_n - \theta_0)\right)$
3. Show that $\frac{1}{n}\ddot{\ell}_n(\tilde{\theta}_n) \xrightarrow{\mathrm{as}} -\mathcal{I}(\theta_0)$
4. Show that $\sqrt{n}(\breve{\theta}_n - \hat{\theta}_n) \xrightarrow{\mathrm{as}} 0$
5. Show that $\breve{\theta}_n$ asymptotically equivalent to the MLE $\hat{\theta}_n$
6. Show that $\breve{\theta}_n$ is asymptotically efficient.

So

$$\begin{aligned}
\sqrt{n}(\breve{\theta}_n - \hat{\theta}_n) =& \sqrt{n}\left[(\tilde{\theta}_n - \hat{\theta}_n) - \ddot{\ell}_n(\tilde{\theta}_n)^{-1}\left(\dot{\ell}_n(\hat{\theta}_n) + \int_0^1 \ddot{\ell}_n(\hat{\theta}_n + u(\tilde{\theta}_n - \hat{\theta}_n))\mathrm{d}u\,(\tilde{\theta}_n - \hat{\theta}_n)\right)\right] \\
=& \left(I - \ddot{\ell}_n\tilde{\theta}_n)^{-1}\int_0^1 \ddot{\ell}_n(\hat{\theta}_n + u(\tilde{\theta}_n - \hat{\theta}_n))\mathrm{d}u\right)\,\sqrt{n}(\tilde{\theta}_n - \hat{\theta}_n) \\
=& \left(I - \left(\frac{1}{n}\ddot{\ell}_n(\tilde{\theta}_n)\right)^{-1}\int_0^1 \frac{1}{n}\ddot{\ell}_n(\hat{\theta}_n + u(\tilde{\theta}_n - \hat{\theta}_n))\mathrm{d}u\right) \times \left(\sqrt{n}(\tilde{\theta}_n - \theta_0) - \sqrt{n}(\hat{\theta}_n - \theta_0)\right) \\
=& \underbrace{\underbrace{\left(I - \underbrace{\underbrace{\left(\frac{1}{n}\ddot{\ell}_n(\tilde{\theta}_n)\right)}_{\xrightarrow{as}-\mathcal{I}(\theta_0)}}_{\xrightarrow{as}-\mathcal{I}(\theta_0)^{-1}}{}^{-1}\int_0^1 \underbrace{\frac{1}{n}\ddot{\ell}_n(\underbrace{\hat{\theta}_n + u(\tilde{\theta}_n - \hat{\theta}_n)}_{\xrightarrow{as}\theta_0})}_{\xrightarrow{as}\int_0^1 \mathrm{E}_f(\ddot{\ell}_n(\theta_0))\mathrm{d}u=-\mathcal{I}(\theta_0)}\,\mathrm{d}u\right)}_{\xrightarrow{as}0} \times \underbrace{\left(\underbrace{\sqrt{n}(\tilde{\theta}_n - \theta_0)}_{\xrightarrow{D}\mathrm{N}(0,\mathcal{I}(\theta_0)^{-1})} - \underbrace{\sqrt{n}(\hat{\theta}_n - \theta_0)}_{\xrightarrow{D}\mathrm{N}(0,\mathcal{I}(\theta_0)^{-1})}\right)}_{\xrightarrow{D}\mathrm{N}(0,2\mathcal{I}(\theta_0)^{-1})}}_{\text{does not become infinity } a.s.} \\
& \xrightarrow{as}0 \\
=& \xrightarrow{as} 0
\end{aligned}$$

Hence $\sqrt{n}(\breve{\theta}_n - \hat{\theta}_n) \xrightarrow{as} 0$, implying $\sqrt{n}(\breve{\theta}_n - \hat{\theta}_n) \xrightarrow{P} 0$.

Analytically,

- It is
$$\frac{1}{n}\ddot{\ell}_n(\tilde{\theta}_n) \xrightarrow{as} -\mathcal{I}(\theta_0)$$
by using the the USLLN trick (8.2). Namely
$$\begin{aligned}
|\frac{1}{n}\ddot{\ell}_n(\tilde{\theta}_n) - (-\mathcal{I}(\theta_0))| =& |\frac{1}{n}\ddot{\ell}_n(\tilde{\theta}_n) \pm \mathcal{I}(\tilde{\theta}_n) + \mathcal{I}(\theta_0))| \\
& \leq \underbrace{\sup_{\theta\in\{\theta:|\tilde{\theta}_n-\theta_0|\leq\delta\}} |\frac{1}{n}\ddot{\ell}_n(\tilde{\theta}_n) - (-\mathcal{I}(\tilde{\theta}_n))|}_{=(i)} + \underbrace{|\mathcal{I}(\tilde{\theta}_n) - \mathcal{I}(\theta_0)|}_{=(ii)}
\end{aligned}$$
Here, term (i) converges to zero from the USLLN because the assumptions of Theorem 52 are satisfied by the conditions of Cramér Theorem 59. Also (ii) converges to zero from Stutsly theorem.

- It is
$$\hat{\theta}_n + u(\tilde{\theta}_n - \hat{\theta}_n) \xrightarrow{as} \theta_0$$
because $\tilde{\theta}_n \xrightarrow{as} \theta_0$ and $\hat{\theta}_n \xrightarrow{as} \theta_0$ as strongly consistent, and by using Slutsky theorem. From the SLLN, I get
$$\frac{1}{n}\ddot{\ell}_n(\theta_0) \xrightarrow{as} -\mathcal{I}(\theta_0)$$

because

$$\frac{1}{n}\ddot{\ell}_n(\theta_0) = \frac{1}{n}\sum_{i=1}^{n}\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log f(X_i|\theta_0) \xrightarrow{as} \mathrm{E}\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log f(X|\theta_0) = -\mathcal{I}(\theta_0)$$

Consequently,

$$\int_0^1 \frac{1}{n}\ddot{\ell}_n(\theta_0)\mathrm{d}u \xrightarrow{as} \int_0^1 \mathrm{E}_f(\dot{\Psi}(X,\theta_0))\mathrm{d}u = \int_0^1 1\mathrm{d}u \times \mathrm{E}_f(\dot{\Psi}(X,\theta_0)) = -\mathcal{I}(\theta_0)$$

- It is

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

and $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{D} \mathrm{N}(0, \mathcal{I}(\theta_0)^{-1})$ by assumption and by Cramer Theorem 59. So

$$\sqrt{n}(\tilde{\theta}_n - \theta_0) + \sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{D} \mathrm{N}(0, \Sigma(\theta_0) + \mathcal{I}(\theta_0)^{-1})$$

does not became infinity as it is bounded in probability.

- Since $\sqrt{n}(\breve{\theta}_n - \hat{\theta}_n) \xrightarrow{as} 0$ then $\sqrt{n}(\breve{\theta}_n - \hat{\theta}_n) \xrightarrow{P} 0$ which implies that $\breve{\theta}_n$ and $\hat{\theta}_n$ are asymptotic equivalent.
- Because $\breve{\theta}_n$ and $\hat{\theta}_n$ are asymptotic equivalent, they asymptotically follow the same distribution, so

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

as well, which implies that $\breve{\theta}_n$ is asymptotic efficient.

□

Example 82. Consider random sample $X_1, ..., X_n \overset{IID}{\sim} f(a,b)$, $a > 0$, $b > 0$ with PDF

$$f(x|a,b) = \frac{1}{\Gamma(a)b^a}x^{a-1}e^{-x\frac{1}{b}}1(x > 0)$$

Assume that b is known, and a is unknown.

1. Find the moment estimator $\tilde{a}$ of a by using the first raw moments
2. Is the moment estimator $\tilde{a}$ consistent and asymptotically Normal?
3. Find the one step estimator by Fisher scoring algorithm.

Hint-1 Digamma function $\psi(x) = \frac{\mathrm{d}}{\mathrm{d}x}\log\Gamma(x)$

Hint-2 Trigamma function $\psi_1(x) = \frac{\mathrm{d}^2}{\mathrm{d}x^2}\log\Gamma(x)$

Solution.

1. The first raw moment is

$$\mu_1(a) = \mathrm{E}(X) = \int_0^\infty x\frac{1}{\Gamma(a)b^a}x^{a-1}e^{-x/b}\mathrm{d}x = \int_0^\infty \frac{1}{\frac{1}{a}\Gamma(a+1)\frac{1}{b}b^{a+1}}x^{(a+1)-1}e^{-x/b}\mathrm{d}x = ab$$

and the sample one

$$m_1 = \bar{X}$$

From the method of moments I get

$$m_1 = \mu_1(\tilde{a}) \implies \qquad \tilde{a} = \frac{m_1}{b} = \frac{\bar{X}}{b}$$

2. The moment estimator is consistent based on the SLLN, $\bar{X} \xrightarrow{D} E(X)$, and by Slutsky $m_1 = \frac{1}{b}\bar{X} \xrightarrow{D} \frac{1}{b}E(X) = \mu_1$. The asymptotic distribution is

   - by CLT
   $$\sqrt{n}(\bar{X} - \mu_1) \xrightarrow{D} \mathrm{N}(0, ab^2)$$
   - so by Delta method with $g(\bar{X}) = \frac{\bar{X}}{b}$
   $$\sqrt{n}(\frac{\bar{X}}{b} - \frac{\mu_1}{b}) = \sqrt{n}(\tilde{a} - a) \xrightarrow{D} \mathrm{N}(0, \frac{ab^2}{b^2}) \quad (= a)$$

3. For the Fisher algorithm, I need to find $\mathcal{I}(a)^{-1}$. It is

$$\begin{aligned}
\log f(x|a) &= -\log\Gamma(a) - a\log(b) - \frac{1}{b}x + (a-1)\log(x) \\
\frac{\mathrm{d}}{\mathrm{d}a}\log f(x|a) &= -\psi(a) - \log(b) + \log(x) \\
\frac{\mathrm{d}^2}{\mathrm{d}a^2}\log f(x|a) &= -\psi_1(a) \qquad (11.7) \\
\mathcal{I}(a) &= \psi_1(a) \\
\mathcal{I}(a)^{-1} &= 1/\psi_1(a)
\end{aligned}$$

$$\begin{aligned}
\ell_n(\theta) &= -n\log\Gamma(a) - na\log(b) - \frac{1}{b}\sum_{i=1}^{n} x_i + (a-1)\sum_{i=1}^{n}\log(x_i) \\
\dot{\ell}_n(\theta) &= -n\psi(a) - n\log(b) + \sum_{i=1}^{n}\log(x_i) \\
\ddot{\ell}_n(\theta) &= -n\psi_1(a)
\end{aligned}$$

The Fisher recursion is

$$\begin{aligned}
\breve{\theta}_n &= \tilde{\theta}_n + \frac{1}{n}\mathcal{I}(\tilde{\theta}_n)^{-1}\dot{\ell}_n(\tilde{\theta}_n) \\
\breve{\theta}_n &= \frac{\bar{X}}{b} + \frac{1}{\psi_1(\frac{\bar{X}}{b})}(-\psi(\frac{\bar{X}}{b}) - \log(b) + \frac{1}{n}\sum_{i=1}^{n}\log(X_i))
\end{aligned}$$

Additionally for the Newton recursion I need

$$\ddot{\ell}_n(\theta) = -n\psi_1(a)$$

The Newton recursion is

$$\begin{aligned}\breve{\theta}_n &= \tilde{\theta}_n - \ddot{\ell}_n(\tilde{\theta}_n)^{-1}\dot{\ell}_n(\tilde{\theta}_n) \\ &= \frac{\bar{X}}{b} + \frac{1}{\psi_1(\frac{\bar{X}}{b})}(-\psi(\frac{\bar{X}}{b}) - \log(b) + \frac{1}{n}\sum_{i=1}^{n}\log(X_i))\end{aligned}$$

Here, Fisher and Newton recursion lead to the same results, because the 2nd-derivative (11.7) does not depend on the random sample. However, this does not happen always.

Exercise sheet

Exercise #19
Exercise #22

12 Hypothesis tests, and Confidence intervals with likelihood methods

Let $X, X_1, X_2, ..., X_n$ be a sequence of independent random samples generated from a distribution f_θ labeled by a d-dimensional parameter $\theta \in \Theta \subset \mathbb{R}^d$, and admitting PDF $f(\cdot|\theta)$.

Assume that the conditions from the Cramér Theorem 59 are satisfied.

Consider that $\hat{\theta}_n$ is the MLE of θ.

- Although the methods below use the MLE $\hat{\theta}_n$, in fact, any asymptotic equivalent estimator $\clubsuit_n$ of the MLE can be used; e.g., the one-step-estimators with moment estimator initial guess.

  Recall that: asymptotic equivalent $\clubsuit_n - \hat{\theta}_n \xrightarrow{P} 0 \implies$ asymptotic efficient $\sqrt{n}(\clubsuit_n - \theta_0) \xrightarrow{D} \mathrm{N}(0, \mathcal{I}(\theta_0)^{-1})$

- All the tools below are asymptotic equivalent, as you can imagine. However, for smaller samples it seems that the Likelihood ration is the most powerful.

12.1 No nuisance parameters

We present 3 types of HT and CI. Regarding the hypothesis test the rational is depicted in Figure 12.1.

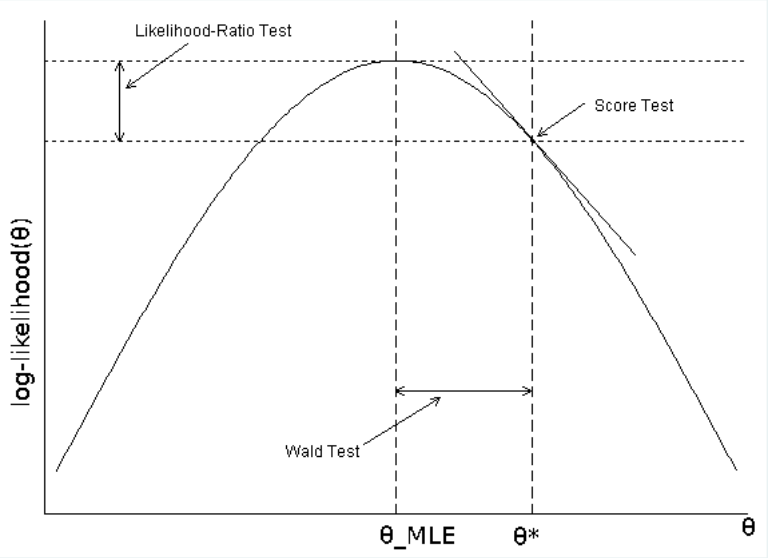


Figure 12.1: Comparison, in 1D , of:
Likelihood ratio : $W_{\mathrm{LR}}(\theta_0) = -2(\ell_n(\theta_0) - \ell_n(\hat{\theta}_n))$
Wald : $W_{\mathrm{Wald}}(\theta_0) = n(\theta_0 - \hat{\theta}_n)^2 \mathcal{I}(\theta_0) = -(\theta_0 - \hat{\theta}_n)^2 \mathrm{E}(\ddot{\ell}_n(\theta_0))$
Score statistic : $W_{\mathrm{Score}}(\theta_0) = n\dot{\ell}_n(\theta_0)/\mathcal{I}(\theta_0) = -\dot{\ell}_n(\theta_0)/\mathrm{E}(\ddot{\ell}_n(\theta_0))$

Likelihood ratio type tools

Pivotal statistic Let the log likelihood ratio statistic be

$$W_{\mathrm{LR}}(\theta) = -2(\ell_n(\theta) - \ell_n(\hat{\theta}_n))$$

If θ_0 is the true value of θ then

$$W_{\mathrm{LR}}(\theta_0) \xrightarrow{D} \chi^2_d$$

and this is used as a Pivotal value.

Theorem 83. *Let $X_1, X_2, ..., X_n$ be independent random samples generated from a distribution f_θ labeled by a d-dimensional parameter $\theta \in \Theta \subset \mathbb{R}^d$, and admitting PDF $f(\cdot|\theta)$. Assume the conditions from the Cramér Theorem 59 are satisfied, and that θ_0 is the true value, then*

$$W_{LR}(\theta_0) = -2(\ell_n(\theta_0) - \ell_n(\hat{\theta}_n)) \xrightarrow{D} \chi^2_d$$

it is where $\hat{\theta}_n$ is the MLE of θ.

Proof. It is given as a homework.[16] □

[16] Layout of the proof:

Hint-1 Expand $\ell_n(\theta_0)$ around $\hat{\theta}_n$ by Taylor expansion

Hint-2 Prove that

$$W_{\mathrm{LR}}(\theta_0) \xrightarrow{a.s} n(\theta_0 - \hat{\theta}_n)^T \mathcal{I}(\theta_0)(\theta_0 - \hat{\theta}_n) \tag{12.1}$$

Hint-3 Prove that $W_{\mathrm{LR}}(\theta_0) \xrightarrow{D} \chi^2_d$

Hypothesis test Consider Hypothesis test

$$\mathrm{H}_0 : \theta = \theta_* \quad \text{vs.} \quad \mathrm{H}_1 : \theta \neq \theta_*$$

Hence the rejection area, at sig. level a, is

$$\begin{aligned}\mathrm{RA}(X_{1:n}) &= \{X_{1:n} : W_{\mathrm{LR}}(\theta_*) \geq \chi^2_{d,1-a}\} \\ &= \{X_{1:n} : -2(\ell_n(\theta_*) - \ell_n(\hat{\theta}_n)) \geq \chi^2_{d,1-a}\}\end{aligned}$$

Confidence intervals The $(1-a)$ confidence interval for θ is

$$\begin{aligned}\mathrm{CI}(\theta) &= \{\theta \in \Theta : W_{\mathrm{LR}}(\theta) \leq \chi^2_{d,1-a}\} \\ &= \{\theta \in \Theta : -2(\ell_n(\theta) - \ell_n(\hat{\theta}_n)) \leq \chi^2_{d,1-a}\}\end{aligned}$$

produced by inverting the $\mathrm{RA}(x_{1:n})$

Comments

- Regarding the HT, the comparison relies on the distance of the log-likelihood ratio $\ell_n(\theta_*)$ and $\ell_n(\hat{\theta}_n)$. The larger the distance is, the biggest doubt about the H_0 based on my data. See Figure 12.1.
- It is more powerful than the other 2 tests, and hence preferable if it can by practically evaluated. The other 2 were derived possibly because ages ago people did not have computers and wanted to use something computationally cheaper.

Wald type tools

Pivotal statistic From the Cramér Theorem 59, as a consequences I get

$$\sqrt{n}\mathcal{I}(\theta_0)^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{D} \mathrm{N}(0, I) \tag{12.2}$$

$$\sqrt{n}\mathcal{I}(\hat{\theta}_n)^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{D} \mathrm{N}(0, I) \tag{12.3}$$

$$\mathcal{J}_n(\hat{\theta}_n)^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{D} \mathrm{N}(0, I) \tag{12.4}$$

whose proof it left as an exercise (See the Exercise sheet).

We can use as a pivotal statistics

$$W_{\mathrm{W}}(\theta_0) = n(\hat{\theta}_n - \theta_0)^T \mathcal{I}(\theta_0)(\hat{\theta}_n - \theta_0) \xrightarrow{D} \chi^2_d \tag{12.5}$$

$$W'_{\mathrm{W}}(\theta_0) = n(\hat{\theta}_n - \theta_0)^T \mathcal{I}(\hat{\theta}_n)(\hat{\theta}_n - \theta_0) \xrightarrow{D} \chi^2_d \tag{12.6}$$

$$W''_{\mathrm{W}}(\theta_0) = (\hat{\theta}_n - \theta_0)^T \mathcal{J}_n(\hat{\theta}_n)(\hat{\theta}_n - \theta_0) \xrightarrow{D} \chi^2_d \tag{12.7}$$

which are derived by Slutsky Theorem, and that $\sum_{i=1}^d z_i^2 \sim \chi^2_d$ if $z_i \sim \mathrm{N}(0,1)$ for $i = 1, ..., d$. They are asymptotically equivalent for large samples (this is obvious by construction). Any of them can be used to construct Hypothesis test, and Confidence intervals. The order of preference is $W_{\mathrm{W}}(\theta_0)$, $W'_{\mathrm{W}}(\theta_0)$, $W''_{\mathrm{W}}(\theta_0)$ when the sample size is not that large, however the proof is out of scope.

Hypothesis test Consider Hypothesis test

$$\mathrm{H}_0 : \theta = \theta_* \quad \text{vs.} \quad \mathrm{H}_1 : \theta \neq \theta_*$$

Hence the rejection area, at sig. level a, is

$$\begin{aligned}\mathrm{RA}(X_{1:n}) &= \{X_{1:n} : W_{\mathrm{Wald}}(\theta_0) \geq \chi^2_{d,1-a}\} \\ &= \{X_{1:n} : n(\hat{\theta}_n - \theta_*)^T \mathcal{I}(\theta_*)(\hat{\theta}_n - \theta_*) \geq \chi^2_{d,1-a}\}\end{aligned}$$

etc...

Confidence intervals The $(1-a)$ confidence interval for θ is

$$\begin{aligned}\mathrm{CI}(\theta) &= \{\theta \in \Theta : W_{\mathrm{Wald}}(\theta) \leq \chi^2_{d,1-a}\} \\ &= \{\theta \in \Theta : n(\hat{\theta}_n - \theta)^T \mathcal{I}(\theta)(\hat{\theta}_n - \theta) \leq \chi^2_{d,1-a}\}\end{aligned}$$

etc...

produced by inverting the $\mathrm{RA}(x_{1:n})$

Comment

- The Wald pivotal statistics are asymptotically equivalent to the LR one.
- Regarding the HT, the comparison relies on the distance of the θ_* and $\hat{\theta}_n$, calibrated by the Information matrix (Fisher or Observed). The larger the distance is, the biggest doubt about the H_0 based on my data. See Figure 12.1.
- Wald type of HT, CI are less expensive than the likelihood ratio ones because they require the computation of the expensive likelihood less number of times.

Score type tools

Pivotal statistic

Definition. The Score statistic is defined as

$$U(\theta) = [\frac{\mathrm{d}}{\mathrm{d}\theta}\ell_n(\theta)]^T = [\sum_{i=1}^{d} \underbrace{\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(X_i|\theta)}_{=\Psi(X_i,\theta)}]^T$$

where I put $\cdot^T$ because $U(\theta)$ is a $d \times 1$ vector.

The asymptotic distribution is

$$\frac{1}{\sqrt{n}}U(\theta) \xrightarrow{D} \mathrm{N}(0, \mathcal{I}(\theta))$$

which results as in Proposition 58. Then similar to above

$$\frac{1}{\sqrt{n}}\mathcal{I}(\theta)^{-1/2}U(\theta) \xrightarrow{D} \mathrm{N}(0,I) \tag{12.8}$$

$$\frac{1}{\sqrt{n}}\mathcal{I}(\hat{\theta}_n)^{-1/2}U(\theta) \xrightarrow{D} \mathrm{N}(0,I) \tag{12.9}$$

$$\mathcal{J}_n(\hat{\theta}_n)^{-1/2}U(\theta) \xrightarrow{D} \mathrm{N}(0,I) \tag{12.10}$$

Therefore the following pivotal statistics can be used

$$W_{\mathrm{Score}}(\theta_0) = \frac{1}{n}U(\theta)^T\mathcal{I}(\theta)^{-1}U(\theta) \xrightarrow{D} \chi^2_d \tag{12.11}$$

$$W'_{\mathrm{Score}}(\theta_0) = \frac{1}{n}U(\theta)^T\mathcal{I}(\hat{\theta}_n)^{-1}U(\theta) \xrightarrow{D} \chi^2_d \tag{12.12}$$

$$W''_{\mathrm{Score}}(\theta_0) = U(\theta)^T\mathcal{J}_n(\hat{\theta}_n)^{-1}U(\theta) \xrightarrow{D} \chi^2_d \tag{12.13}$$

with criterion the above preference order and their tractability.

Hypothesis test Consider Hypothesis test

$$\mathrm{H}_0 : \theta = \theta_* \quad \text{vs.} \quad \mathrm{H}_1 : \theta \neq \theta_*$$

Hence the rejection area, at sig. level a, is

$$\begin{aligned}\mathrm{RA}(X_{1:n}) &= \{X_{1:n} : W_{\mathrm{Score}}(\theta_0) \geq \chi^2_{d,1-a}\} \\ &= \{X_{1:n} : \frac{1}{n}U(\theta_*)^T\mathcal{I}(\theta_*)^{-1}U(\theta_*) \geq \chi^2_{d,1-a}\}\end{aligned} \tag{12.14}$$

etc...

Confidence intervals The $(1-a)$ confidence interval for θ is

$$\begin{aligned}\mathrm{CI}(\theta) &= \{\theta \in \Theta : W_{\mathrm{Score}}(\theta) \leq \chi^2_{d,1-a}\} \\ &= \{\theta \in \Theta : \frac{1}{n}U(\theta)^T\mathcal{I}(\theta)^{-1}U(\theta) \leq \chi^2_{d,1-a}\}\end{aligned} \tag{12.15}$$

etc... produced by inverting the $\mathrm{RA}(X_{1:n})$.

Comments

- The Score pivotal statistics are asymptotically equivalent to the LR one.
- Regarding the HT, the comparison relies on the slop of the log-likelihood at θ_* (aka the $U(\theta_*)$) calibrated by the curvature (Hessian matrix) at θ_*. The larger/steeper the slope, the bigger the distance from the peak (MLE $\hat{\theta}_n$), hence the biggest doubt about the H_0 based on my data. See Figure 12.1.

- Score type of HT, CI are less expensive than the likelihood ratio ones because they require the computation of the expensive likelihood less number of times.
- Score statistic type of HT and CI are computational convenient, in situations when the practitioner wants to calculate the HT or CI for parameter ϕ, which is a function of parameter θ whose Score type HT or CI have already been calculated.

  Let function $g : \mathbb{R}^d \to \mathbb{R}^k$ with $\phi = g(\theta)$. The score statistic

$$U^*(\phi) = [\frac{\mathrm{d}}{\mathrm{d}\phi}\ell_n(\theta)]^T = [\sum_{i=1}^{d}\frac{\mathrm{d}}{\mathrm{d}\phi}f(X_i|\phi)]^T$$

  is such that

$$U^*(\phi) = [\frac{\mathrm{d}}{\mathrm{d}\phi}\ell_n(\phi)]^T = [\frac{\mathrm{d}}{\mathrm{d}\theta}\ell_n(\theta)\frac{\mathrm{d}\theta}{\mathrm{d}\phi}]^T = [\frac{\mathrm{d}\theta}{\mathrm{d}\phi}]^T U(\theta)$$

  has expected value

$$\mathrm{E}(U^*(\phi)) = \mathrm{E}([\frac{\mathrm{d}\theta}{\mathrm{d}\phi}]^T U(\theta)) = 0$$

  has variance

$$\mathrm{var}(U^*(\phi)) = \mathrm{var}([\frac{\mathrm{d}\theta}{\mathrm{d}\phi}]^T U(\theta)) = [\frac{\mathrm{d}\theta}{\mathrm{d}\phi}]^T \mathcal{I}(\theta)[\frac{\mathrm{d}\theta}{\mathrm{d}\phi}]$$

  and hence has asymptotic distribution

$$\frac{1}{\sqrt{n}}U^*(\phi) \xrightarrow{D} \mathrm{N}(0, \underbrace{[\frac{\mathrm{d}\theta}{\mathrm{d}\phi}]^T \mathcal{I}(\theta)[\frac{\mathrm{d}\theta}{\mathrm{d}\phi}]}_{=\mathcal{I}^*(\phi)})$$

  Hence, one can derive HT and CI as in (12.14) and (12.15) by substituting properly [17]. Notice that if the score HT and CI for θ are available then the score HT and CI for the transformation $\phi = g(\theta)$ can by computed by avoiding to recompute the expensive likelihood function

Example 84. [a]Let random sample $x_1, ..., x_n \overset{IID}{\sim} \mathrm{Poi}(\theta)$, $\theta > 0$ with PDF

$$f(x|\theta) = \frac{\theta^x e^{-\theta}}{x!}1(x > 0)$$

For the hypothesis test

$$\mathrm{H}_0 : \theta = \theta_* \quad \text{vs.} \quad \mathrm{H}_1 : \theta \neq \theta_*$$

Calculate

1. the log-likelihood ratio RA at a sig. level
2. the Wald's type RA at a sig. level (the 3 of them)
3. the Score's type RA at a sig. level (the 3 of them)

[17] please write down the derivation

Solution. Ok before that, let's calculate all the quantities required.

$$\begin{aligned}
\log f(x|\theta) \propto & x\log(\theta)-\theta && ;; && 0=\dot\ell(\theta)|_{\theta=\hat\theta} \implies \hat\theta=\bar x\\
\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(x|\theta) = & \frac{x}{\theta}-1 && ;; && \ddot\ell(\theta) = -n\bar x\frac{1}{\theta^2}\\
\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log f(x|\theta) = & -\frac{x}{\theta^2} && ;; && \mathcal{J}_n(\theta) = -\ddot\ell(\theta) = n\bar x\frac{1}{\theta^2}\\
\mathcal{I}(\theta) = & -\mathrm{E}(\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log f(x|\theta)) = \frac{1}{\theta} && ;; && \mathcal{J}_n(\hat\theta) = \frac{n}{\bar x}\\
\ell(\theta) = & n\bar x\log(\theta)-n\theta && ;; && U(\theta)=\dot\ell(\theta)^T = n\bar x\frac{1}{\theta}-n\\
\dot\ell(\theta) = & n\bar x\frac{1}{\theta}-n && ;;
\end{aligned}$$

1. It is

$$\begin{aligned}
\mathrm{CI}(\theta) &= \{\theta\in(0,\infty) : -2(\ell_n(\theta)-\ell_n(\hat\theta_n))\le\chi^2_{1,1-a}\}\\
&= \{\theta\in(0,\infty) : -2((n\bar x\log(\frac{\theta}{\bar x})-n(\theta-\bar x))\le\chi^2_{1,1-a}\}
\end{aligned}$$

   well, here we do not have the condition $n\theta - n\bar x$ like in the contingency tables ...

2. It is

$$\begin{aligned}
\mathrm{RA}(x_{1:n}) &= \{x_{1:n} : n(\hat\theta_n-\theta_*)^T\mathcal{I}(\theta_*)(\hat\theta_n-\theta_*)\ge\chi^2_{1,1-a}\}\\
&= \{x_{1:n} : n(\bar x-\theta_*)^2\frac{1}{\theta_*}\ge\chi^2_{1,1-a}\}
\end{aligned}$$

   and

$$\begin{aligned}
\mathrm{RA}(x_{1:n}) &= \{x_{1:n} : n(\hat\theta_n-\theta_*)^T\mathcal{I}(\hat\theta_n)(\hat\theta_n-\theta_*)\ge\chi^2_{1,1-a}\}\\
&= \{x_{1:n} : n(\bar x-\theta_*)^2\frac{1}{\bar x}\ge\chi^2_{1,1-a}\}
\end{aligned}$$

   and

$$\begin{aligned}
\mathrm{RA}(x_{1:n}) &= \{x_{1:n} : (\hat\theta_n-\theta_*)^T\mathcal{J}_n(\hat\theta_n)(\hat\theta_n-\theta_*)\ge\chi^2_{1,1-a}\}\\
&= \{x_{1:n} : (\bar x-\theta_*)^2\frac{n}{\bar x}\ge\chi^2_{1,1-a}\}
\end{aligned}$$

   where the latter two CI coinside, however this is just a coinsidence...

3. It is

$$\begin{aligned}
\mathrm{RA}(x_{1:n}) &= \{x_{1:n} : \frac{1}{n}U(\theta_*)^T\mathcal{I}(\theta_*)^{-1}U(\theta_*)\ge\chi^2_{1,1-a}\}\\
&= \{x_{1:n} : \frac{1}{n}(n\bar x\frac{1}{\theta_*}-n)^2(\frac{1}{\theta_*})^{-1}\ge\chi^2_{1,1-a}\}\\
&= \{x_{1:n} : n(\bar x\frac{1}{\theta_*}-1)^2\theta_*\ge\chi^2_{1,1-a}\}
\end{aligned}$$

and

$$\text{RA}(x_{1:n}) = \{x_{1:n} : \frac{1}{n} U(\theta_*)^T \mathcal{I}(\hat{\theta}_n)^{-1} U(\theta_*) \geq \chi^2_{1,1-a}\}$$
$$= \{x_{1:n} : n(\bar{x}\frac{1}{\theta_*} - 1)^2 \hat{\theta}_n \geq \chi^2_{1,1-a}\}$$

and

$$\text{RA}(x_{1:n}) = \{x_{1:n} : \frac{1}{n} U(\theta_*)^T \mathcal{J}_n(\hat{\theta}_n)^{-1} U(\theta_*) \geq \chi^2_{1,1-a}\}$$
$$= \{x_{1:n} : n(\bar{x}\frac{1}{\theta_*} - 1)^2 \hat{\theta}_n \geq \chi^2_{1,1-a}\}$$

where the latter two CI coincide, however it is just a coincidence...

Confidence intervals can be computed by inverting the RA, based on the theory learnt in Concepts in Stats 2 (Term 1).

[a] Another example you can find in Example 67.
More examples about these CI and HT, will do in Term 2 when you will apply these in the GLM.

Exercise sheet

Exercise #21
Exercise #23
Exercise #24
Exercise #25
Exercise #26

12.2 With nuisance parameters [18]

Often the sampling distribution f_θ is labeled by an unknown parameter $\theta \in \Theta$, however, the analyst is actually interested in learning only about a part (a sub-parameter, or a lower dimensional function) of it.

Consider the unknown d-dimensional parameter θ is partitioned as $\theta = (\psi, \phi)$, by a d_ψ-dimensional $\psi \in \Psi \subset \mathbb{R}^{d_\psi}$, and d_ϕ-dimensional $\phi \in \Phi \subset \mathbb{R}^{d_\phi}$, obviously $d = d_\psi + d_\phi$. Assume that our interest lies in inference for a sub-parameter (or parameter function) $\psi = \psi(\theta)$, and we do not really care about $\phi = \phi(\theta)$ although we wish to consider uncertainty about it. Then

- $\psi = \psi(\theta)$ is called the parameter of interest
- $\phi = \phi(\theta)$ is called the nuisance parameter

[18] For inference under the presence of nuisance parameters, we present only the profile likelihood methos but not the Wald and Score analogs due to time restrictions. To study about the Wald and Score methods under the presence of nuisance parameters see [10].

Definition 85. Given a joint likelihood $L_n(\theta)$ the profile likelihood $L_{n,p}(\psi)$ of ψ is

$$\begin{aligned} L_{n,p}(\psi) &= \sup_{\forall \phi} L_n(\underbrace{\psi, \phi}_{=\theta}) = \\ &= L_n(\psi, \hat{\phi}_\psi) \end{aligned}$$

where $\hat{\phi}_\psi$ denotes an MLE of ϕ for a given function of ψ and $\hat{\phi}_\psi$ can be a function of ψ .

Analogously we define the profile log-likelihood $\ell_{n,p}(\psi)$ of ψ, as

$$\begin{aligned} \ell_{n,p}(\psi) &= \log(L_{n,p}(\psi)) \\ &= \log(L_n(\psi, \hat{\phi}_\psi)) \\ &= \ell_n(\psi, \hat{\phi}_\psi) \end{aligned}$$

Once the profile log-likelihood $L_{n,p}(\psi)$ of ψ is specified, then we can perform inference (point estimation, CI, HT, etc...) as usual but using $L_{n,p}(\psi)$. Here, we will discuss the point estimation, and the Hypothesis test.

Point estimation The MLE of $\hat{\psi} = \hat{\psi}(x_1, ..., x_n)$ is the

$$\hat{\psi} = \sup_{\forall \psi \in \Psi} \ell_{n,p}(\psi)$$

and it can be found as a root of the equations

$$0 = \frac{\mathrm{d}}{\mathrm{d}\psi} \ell_{n,p}(\psi)|_{\psi=\hat{\psi}}$$

Hypothesis test Consider, we are interested in testing the hypothesis

$$\mathrm{H}_0 : \psi = \psi_* \text{ vs. } \mathrm{H}_1 : \psi \neq \psi_* \tag{12.16}$$

where $\psi_* \in \mathbb{R}^{d_\psi}$, and $\psi := (\psi_1(\theta), ..., \psi_{d_\psi}(\theta))^T$ is a vector of d_ψ smooth functions defined on Θ. We consider the likelihood ratio test, with pivotal statistic

$$\begin{aligned} W_{\mathrm{LR},p}(\psi_*) &= -2 \log\left(\frac{L_{n,p}(\psi_*)}{\sup_{\forall \psi \neq \psi_*} L_{n,p}(\psi)}\right) \\ &= -2 \log\left(\frac{L_n(\psi_*, \hat{\phi}_{\psi_*})}{\sup_{\forall \psi \neq \psi_*} L_n(\psi, \phi)}\right) \\ &= -2(\ell_n(\psi_*, \hat{\phi}_{\psi_*}) - \ell_n(\hat{\theta}_1)) \end{aligned}$$

where and $\hat{\theta}_1 = (\hat{\psi}, \hat{\phi})$ is the MLE of θ under hypothesis H_1 .

Discussion...
To make the connection with Stats Concepts 2; notice that the Hypothesis pair (12.16) can be equivalently be written as

$$\mathrm{H}_0 : \theta \in \Theta_0 \text{ vs. } \mathrm{H}_1 : \theta \in \underbrace{\Theta - \Theta_0}_{=\Theta_1}$$

where Θ_0 results by constraining the parameter space Θ, as $\Theta_0 = \{\theta \in \Theta : \psi(\theta) = \psi_*\}$. Here,

- $d_0 = \dim(\Theta_0) = d_\phi$ is the number of free parameters (to be learned) under H_0, and
- $d_1 = \dim(\Theta_1) = d = d_\psi + d_\phi$ is the number of free parameters (to be learned) under H_1.

Then elaborating a little bit, we get

$$\begin{aligned}
W_{\mathrm{LR},p}(\psi_*) &= -2\log\left(\frac{L_{n,p}(\psi_*)}{\sup_{\forall\psi\neq\psi_*} L_{n,p}(\psi)}\right) \\
&= -2\log\left(\frac{\sup_{\forall\phi} L_n(\psi_*,\phi)}{\sup_{\forall\psi\neq\psi_*}\sup_{\forall\phi} L_n(\psi,\phi)}\right) \\
&= -2\log\left(\frac{\sup_{\{\psi=\psi_*,\forall\phi\}} L_n(\psi_*,\phi)}{\sup_{\{\forall\psi\neq\psi_*,\forall\phi\}} L_n(\psi,\phi)}\right) \\
&= -2\log\left(\frac{\sup_{\theta\in\Theta_0} L_n(\theta)}{\sup_{\theta\in\Theta-\Theta_0} L_n(\theta)}\right) \\
&= -2(\ell_n(\hat{\theta}_0) - \ell_n(\hat{\theta}_1))
\end{aligned}$$

where $\hat{\theta}_0 = (\psi_*, \hat{\phi}_{\psi_*})$ is the MLE under hypothesis H_0, and $\hat{\theta}_1 = (\hat{\psi}, \hat{\phi})$ is the MLE under hypothesis H_1.

Given that H_0 is true, the asymptotic distribution of $W_{\mathrm{LR},p}(\psi_*)$ is

$$W_{LR,p}(\psi_*) \xrightarrow{D} \chi^2_{\underbrace{d_1 - d_0}_{=d_\psi}} \tag{12.17}$$

Then the rejection area is

$$\mathrm{RA} : \{x_{1:n} : \quad W_{LR,p}(\psi_*) \geq \chi^2_{d_\psi, 1-a}\}$$

Discussion...
Precisely, (12.17) is consequence of the Wilk's theorem formally stated below.

Theorem 86. *(Wilk's theorem) Let $X_1, X_2, ..., X_n$ be independent random samples generated from a distribution f_θ labeled by a d-dimensional parameter $\theta \in \Theta$, and admitting PDF $f(\cdot|\theta)$. Assume that the conditions from the Cramér Theorem 59 are satisfied. Consider a hypothesis test*

$$H_0 : \theta \in \Theta_0 \text{ vs. } H_1 : \theta \in \Theta - \Theta_0$$

where $\Theta_0 = \{\theta \in \Theta : \psi(\theta) = \psi_*\}$ *constrains* Θ *with* $\psi := (\psi_1(\theta), ..., \psi_{d_\psi}(\theta))^T$ *is a vector of* d_ψ *smooth functions defined on* Θ*. Then*

$$-2\log(\frac{\sup_{\theta\in\Theta_0} L_n(\theta)}{\sup_{\theta\in\Theta_1} L_n(\theta)}) = -2(\ell_n(\hat{\theta}_0) - \ell_n(\hat{\theta}_1)) \xrightarrow{D} \chi^2_{\underbrace{d_1 - d_0}_{=d_\psi}}$$

where $\hat{\theta}_0$ *is the MLE of* θ *under* H_0*,* d_0 *is the free number of parameters under* H_0*,* $\hat{\theta}_1$ *is the MLE of* θ *under* H_1*, and* d_1 *is the free number of parameters under* H_1*.*

Proof. The proof is a bit technical relying on tricks and does not really offer something for the module purpose; It is omitted. □

Applications of profile likelihood.

Wilk's theorem is often used in model comparison. In such cases, the researcher is interested in checking if a model can be simplified by setting a number of parameters equal to a speficic value.

For instance, consider when you performed model comparison between nested the log-linear models by using likelihood ratio (or devience):

$$\begin{cases} \mathrm{H}_0: & [X, YZ] \\ \mathrm{H}_1: & [XY, YZ] \end{cases} \iff \begin{cases} \mathrm{H}_0: & \log(\mu_{ijk}) = \lambda + \lambda^X_i + \lambda^Y_j + \lambda^Z_k + \lambda^{YZ}_{jk} \\ \mathrm{H}_1: & \log(\mu_{ijk}) = \lambda + \lambda^X_i + \lambda^Y_j + \lambda^Z_k + \lambda^{YZ}_{jk} + \lambda^{XY}_{jk} \end{cases}$$

$$\iff \begin{cases} \mathrm{H}_0: & \lambda^{XY}_{jk} = \lambda^{YZ}_{jk} = \lambda^{XZ}_{ik} = \lambda^{XYZ}_{ijk} = 0 \\ \mathrm{H}_1: & \lambda^{YZ}_{jk} = \lambda^{XZ}_{ik} = \lambda^{XYZ}_{ijk} = 0 \end{cases} \iff \begin{cases} \mathrm{H}_0: & \lambda \in \Theta_0 \\ \mathrm{H}_1: & \lambda \in \Theta_1 \end{cases}$$

Example 87. Let random sample $x_1, ..., x_n \overset{\text{IID}}{\sim} \mathrm{N}(\mu, \sigma^2)$, where μ and σ^2 are unknown. We are interested in inference on μ.

1. Calculate the profile likelihood for μ
2. Find the rejection area for the hypothesis test (at sig. level a)

   $$\mathrm{H}_0 : \mu = \mu_* \text{ vs. } \mathrm{H}_1 : \mu \neq \mu_*$$

   with respect to the t statistic $t = \sqrt{n}\frac{(\bar{x}-\mu_*)}{s}$, $s = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2$

Solution. Ok, I need to perform inference about the parameter of interest μ under the presence of a nuisance parameter σ^2.

1. The profile likelihood is

   $$L_{n,p}(\mu) = \sup_{\forall\sigma^2} L_n(\mu, \sigma^2) = L_n(\mu, \hat{\sigma}^2_\mu)$$

   where $\hat{\sigma}^2_\mu$ is the n MLE of σ^2 for a given μ.

So first I need to find $\hat{\sigma}^2_\mu$. Okay, then, ...

The joint likelihood is

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

The joint log likelihood is

$$\ell_n(\mu, \sigma^2) = -\frac{n}{2}\log(\sigma^2) - \frac{1}{2}\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 + \text{const.}$$

So, to find $\hat{\sigma}^2_\mu$

$$\begin{aligned} 0 =& \frac{\mathrm{d}}{\mathrm{d}\sigma^2}\ell_n(\mu, \sigma^2)|_{\sigma^2=\hat{\sigma}^2_\mu} \\ =& -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2}\frac{1}{(\sigma^2)^2}\sum_{i=1}^{n}(x_i - \mu)^2|_{\sigma^2=\hat{\sigma}^2_\mu} \end{aligned}$$

then

$$\hat{\sigma}^2_\mu = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

Hence the profile likelihood for μ is

$$\begin{aligned} L_{n,p}(\mu) = L_n(\mu, \hat{\sigma}^2_\mu) &= (\frac{1}{2\pi\hat{\sigma}^2_\mu})^{\frac{n}{2}} \exp(-\frac{1}{2}\frac{1}{\hat{\sigma}^2_\mu}\sum_{i=1}^{n}(x_i - \mu)^2) \\ &= (\frac{1}{2\pi}\frac{n}{\sum_{i=1}^{n}(x_i - \mu)^2})^{\frac{n}{2}} \exp(-\frac{1}{2}n) \end{aligned}$$

2. To test

$$\mathrm{H}_0 : \mu = \mu_* \text{ vs. } \mathrm{H}_1 : \mu \neq \mu_*$$

I need to find the log likelihood ratio

$$W_{\mathrm{LR},p}(\mu_*) = -2\log(\frac{\sup_{\mathrm{H}_0} L_{n,p}(\mu)}{\sup_{\mathrm{H}_1} L_{n,p}(\mu)})$$

Under the null hypothesis H_0 it is

$$L_{n,p}(\mu_*) = (\frac{1}{2\pi}\frac{n}{\sum_{i=1}^{n}(x_i - \mu_*)^2})^{\frac{n}{2}} \exp(-\frac{1}{2}n)$$

Under the alternative hypothesis H_1 it is

$$\begin{aligned} \sup_{\mathrm{H}_1} L_{n,p}(\mu) &= \sup_{\forall\mu}((\frac{1}{2\pi}\frac{n}{\sum_{i=1}^{n}(x_i - \hat{\mu})^2})^{\frac{n}{2}} \exp(-\frac{1}{2}n)) \\ &= (\frac{1}{2\pi}\frac{n}{\sum_{i=1}^{n}(x_i - \bar{x})^2})^{\frac{n}{2}} \exp(-\frac{1}{2}n) \end{aligned} \tag{12.18}$$

because the MLE of μ under the H_1 is $\hat{\mu} = \bar{x}$: In fact, under H_1 it is

$$0 = \frac{\mathrm{d}}{\mathrm{d}\sigma^2}\ell_n(\mu, \sigma^2)|_{\sigma^2=\hat{\sigma}^2, \mu=\hat{\mu}} \implies \begin{cases} \hat{\mu} = & \bar{x} \\ \hat{\sigma}^2 = & \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2 \end{cases}$$

or otherwise you can see that $L_{n,p}(\mu)$ maximizes by minimizing the sum-of-squares term. So

$$\begin{aligned} W_{\mathrm{LR},p}(\mu_*) &= -2\log(\frac{\sup_{\mathrm{H}_0} L_{n,p}(\mu)}{\sup_{\mathrm{H}_1} L_{n,p}(\mu)}) = -2\log(\frac{(\frac{1}{2\pi}\frac{n}{\sum_{i=1}^n (x_i-\mu_*)^2})^{\frac{n}{2}}\exp(-\frac{1}{2}n)}{(\frac{1}{2\pi}\frac{n}{\sum_{i=1}^n (x_i-\bar{x})^2})^{\frac{n}{2}}\exp(-\frac{1}{2}n)}) \\ &= n\log(\frac{\sum_{i=1}^n (x_i-\mu_*)^2}{\sum_{i=1}^n (x_i-\bar{x})^2}) \qquad = n\log(\frac{\sum_{i=1}^n (x_i \pm \bar{x} - \mu_*)^2}{\sum_{i=1}^n (x_i-\bar{x})^2}) \\ &= n\log(\frac{\sum_{i=1}^n (x_i-\bar{x})^2 + n(\bar{x}-\mu_*)^2}{\sum_{i=1}^n (x_i-\bar{x})^2}) \\ &= n\log(1 + \frac{1}{n-1}\underbrace{n\frac{(\bar{x}-\mu_*)^2}{\frac{1}{n-1}\sum_{i=1}^n (x_i-\bar{x})^2}}_{=t^2}) \\ &= n\log(1+\frac{1}{n-1}t^2) \xrightarrow{D} \chi^2_{\underbrace{2-1}_{=1}} \end{aligned}$$

where $t = \sqrt{n}\frac{(\bar{x}-\mu_*)}{s}$ with $s = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2$.

Therefore the rejection area at sig. level a is

$$\mathrm{RA}(x_{1:n}) = \{x_{1:n} : n\log(1+\frac{1}{n-1}t^2) \geq \chi^2_{1,1-a}\}$$

References

[1] Apostol, T. M. (1967). *Calculus, Vol. 2: Multi-Variable Calculus and Linear Algebra with Applications to Differential Equations and Probability*. New York: J. Wiley.

[2] Apostol, T. M. (1974). *Mathematical analysis; 2nd ed.* Addison-Wesley Series in Mathematics. Reading, MA: Addison-Wesley.

[3] Bishop, Y. M., S. E. Fienberg, and P. W. Holland (2007). *Discrete multivariate analysis: theory and practice*. Springer Science & Business Media.

[4] Bradley, R. A. and J. J. Gart (1962). The asymptotic properties of ml estimators when sampling from associated populations. *Biometrika 49*(1/2), 205–214.

[5] Feller, W. (1966). *An introduction to probability theory and its applications, Vol. II*. Wiley and Sons, New York.

[6] Lehmann, E. L. (2004). *Elements of large-sample theory*. Springer Science & Business Media.

[7] Lehmann, E. L. and G. Casella (2006). *Theory of point estimation*. Springer Science & Business Media.

[8] Mann, H. B. and A. Wald (1943). On stochastic limit and order relationships. *The Annals of Mathematical Statistics 14*(3), 217–226.

[9] Pawitan, Y. (2001). *In all likelihood: statistical modelling and inference using likelihood*. Oxford University Press.

[10] Severini, T. A. (2000). *Likelihood methods in statistics*. Oxford University Press.

[11] Van der Vaart, A. W. (2000). *Asymptotic statistics*, Volume 3. Cambridge university press.

[12] Young, G. A. and R. L. Smith (2005). *Essentials of statistical inference*, Volume 16. Cambridge University Press.