Topics in statistics III/IV (MATH3361/4071) Michaelmas term, 2020-2021

Handout 2: Basic tools for asymptotics in statistics

Lecturer & author: Georgios Karagiannis georgios.karagiannis@durham.ac.uk
References: [3]

1 Consistency

Note 1. 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 $X_{1:n}$. 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.

Definition 2. [Weak consistency / consistency in probability] 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) of Handout 1 is defined on the true parameter θ, i.e. $P = P_\theta$.

Definition 3. [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 4. [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. $P = P_\theta$

2 Law of Large Numbers

Note 5. 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 6!!!

Theorem 6. *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* $\bar{X}_n \xrightarrow{qm} E(X)$

Proof. It is given as an Exercise 11 in the Exercise sheet. □

Note 7. Often it is easy to show that the 2nd moment is finite, and hence consistence in quadratic mean is feasible, and presents interest.

Example 8. (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 !!!):

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

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

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

Solution. 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 $\hat{b}_n \xrightarrow{\mathrm{qm}} b$ 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 $\hat{a}_n \xrightarrow{\mathrm{qm}} a$, 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 #16 ; 17

3 Central Limit Theorems

Note 9. 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.

Note 10. We present a basic version of the Central Limit Theorem (CLT) that assumes IID random variables (vectors). [1] In Appendix A, we discuss about the so called Edgeworth-Expansions, which are another version of the CLT considering higher order terms.

3.1 Central Limit Theorem (IID case)[2], [3]

Theorem 11. *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*

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

Proof. It is given as an Exercise 13 in the Exercise sheet. □

Example 12. 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] Generalisations of the above CLT exist, (such as the Lindeberg-Feller version of the CLT which does not require "identically distributed" variables) see [1]. –not examinable

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

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

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

$$\sqrt{n}(\boldsymbol{p} - \boldsymbol{\pi}) \xrightarrow{D} \mathrm{N}(0, \mathrm{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^T) \tag{3.1}$$

2. From (3.1) it is

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

Exercise sheet

Exercise #15 ; 7

4 Slutsky Theorems

Note 13. 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 14. *(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*

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

Proof. Out of the scope. □

Example 15. 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 14(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)

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

Example 16. 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 14(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 17. *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 14(1) and 14(3). □

Example 18. 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 17, and given a function $f : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ with $f(v, u) = v^T u$.

Example 19. 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 17. 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
$$\frac{\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma}}{\sqrt{\frac{S_n^2}{\sigma^2}}} \xrightarrow{D} \mathrm{N}(0,1) \qquad \Longrightarrow \qquad \sqrt{n}\frac{\bar{X}_n - \mu}{S_n} \xrightarrow{D} \mathrm{N}(0,1)$$

Remark 20. There are analogous Slutsky theorems for convergence in Probability (Theorem 21), and for convergence almost surely (Theorem 22).

Theorem 21. *(Slutsky theorems for convergence in Probability)*

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

Proof. Out of the scope. □

Theorem 22. *(Slutsky theorems for convergence almost surely)*

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

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

Proof. Out of the scope. □

Definition 23. Random variables X_n and Y_n are asymptotically equivalent if and only if $(X_n - Y_n) \xrightarrow{P} 0$. We will denote it as $X_n \sim Y_n$.

Remark 24. Essentially Theorem 14(2) says that asymptotically equivalent random variables have the same asymptotic distributions.

Exercise sheet

Exercise #18

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

Note 25. The following notation [2] is useful in order to denote the order of magnitude of specified quantities.

Definition 26. 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 27. *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 $f(X_n) = o_P(|X_n|^p)$
2. if $X_n \xrightarrow{P} 0$ and $f(x) = O(|x|^p)$ as $x \to 0$, then $f(x) = O_P(|X_n|^p)$

Proof. Omitted; see [5] □

Proposition 28. *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 [5] □

Fact 29. *Some rules*

$$o_P(1) + o_P(1) = o_P(1)\,; \qquad o_P(1) + O_P(1) = O_P(1) \tag{5.1}$$

$$o_P(1)O_P(1) = o_P(1)\,; \qquad \frac{1}{1 + o_P(1)} = O_P(1) \tag{5.2}$$

$$o_P(O_P(1)) = o_P(1) \tag{5.3}$$

more rules

$$O_P(R_n) = R_n O_P(1) \qquad o_P(R_n) = R_n o_P(1) \tag{5.4}$$

$$O_P(c_n)O_P(d_n) = O_P(c_n d_n); \qquad O_P(c_n)o_P(d_n) = o_P(c_n d_n) \tag{5.5}$$

$$O_P(c_n) + O_P(d_n) = O_P(\max(c_n, d_n)) \tag{5.6}$$

even more rules

$$\textit{If} \quad h_n \to 0, \quad \textit{and} \quad X_n = O_P(h_n) \quad \textit{then} \quad X_n = o_P(1). \tag{5.7}$$

$$O_P(\frac{1}{\sqrt{n}}) = o_P(1) \tag{5.8}$$

Proof. The proofs are omitted, but they are consequences of Slutsky theorems, Taylor expansion, Markov inequality, Theorem 27, Proposition 28, and probability calculus. □

Example 30. Show that the statistics G^2 and X^2 from the Goodness of fit test of log-linear models for contingency tables are asymptotically equivalent.

Hint-1: Denote $G^2 = 2\sum_{\forall i,j,k} n_{ijk} \log(\frac{n_{ijk}}{\hat{\mu}_{ijk}})$, $X^2 = \sum_{\forall i,j,k} \frac{(n_{ijk} - \hat{\mu}_{ijk})^2}{\hat{\mu}_{ijk}}$, and $\hat{\mu}_{ijk}$ are MLE under model M described in the null hypothesis oft he GoF test.

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

Hint-3: Assume that you know that: $\sqrt{n}(p_i - \pi_i) \xrightarrow{D}$ Normal and $\hat{\pi}_i - \pi_i \xrightarrow{D}$ Normal

Solution. Consider just one index for simplicity; aka $G^2 = 2\sum_{\forall i} n_i \log(\frac{n_i}{\hat{\mu}_i})$, and $X^2 = \sum_{\forall i} \frac{(n_i - \hat{\mu}_i)^2}{\hat{\mu}_{ijk}}$I need to show

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

It is

$$
\begin{aligned}
G^2 =& 2\sum_{\forall i} n_i \log\frac{n_i}{\hat{\mu}_i} = 2n\sum_{\forall i} p_i \log(1+\frac{p_i-\hat{\pi}_i}{\hat{\pi}_i}) \\
\overset{\text{Taylor}}{=}& 2n\sum_{\forall i}(\hat{\pi}_i+(p_i-\hat{\pi}_i))\left(\frac{p_i-\hat{\pi}_i}{\hat{\pi}_i}-\frac{1}{2}\frac{(p_i-\hat{\pi}_i)^2}{\hat{\pi}_i^2}+O((p_i-\hat{\pi}_i)^3)\right) \\
=& 2n\sum_{\forall i}\left((p_i-\hat{\pi}_i)-\frac{1}{2}\frac{(p_i-\hat{\pi}_i)^2}{\hat{\pi}_i}+\frac{(p_i-\hat{\pi}_i)^2}{\hat{\pi}_i}-\overbrace{\frac{1}{2}\frac{(p_i-\hat{\pi}_i)^3}{\hat{\pi}_i}+O((p_i-\hat{\pi}_i)^3)}^{=O((p_i-\hat{\pi}_i)^3)}\right) \\
\overset{\text{re-arrange}}{=}& 2n\sum_{\forall i}\left((p_i-\hat{\pi}_i)-\frac{1}{2}\frac{(p_i-\hat{\pi}_i)^2}{\hat{\pi}_i}+\frac{(p_i-\hat{\pi}_i)^2}{\hat{\pi}_i}+O(((p_i-\pi_i)-(\hat{\pi}_i-\pi_i))^3)\right) \\
=& n\sum_{\forall i}\frac{(p_i-\hat{\pi}_i)^2}{\hat{\pi}_i}+2nO_P(n^{-3/2}) \qquad (5.9) \\
=& X^2+O_P(n^{-1/2}) = X^2+o_P(1)
\end{aligned}
$$

Regarding (5.9) 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 27 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.

References

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

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

[3] Robert J Serfling. *Approximation theorems of mathematical statistics*, volume 162. John Wiley & Sons, 2009.

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

[5] A. W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Appendix

A 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 11, 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 (2) 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 :

$$F_n(x) = \Phi(x) - \phi(x)\frac{\kappa_3 H_2(x)}{6\sqrt{n}} - \phi(x)(\frac{\kappa_4 H_3(x)}{24n} + \frac{\kappa_3^2 H_5(x)}{72n}) + O(n^{-3/2}) \tag{A.1}$$

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

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

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 (A.1) 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

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

[4] For more details see [1, 4].

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

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

- 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.

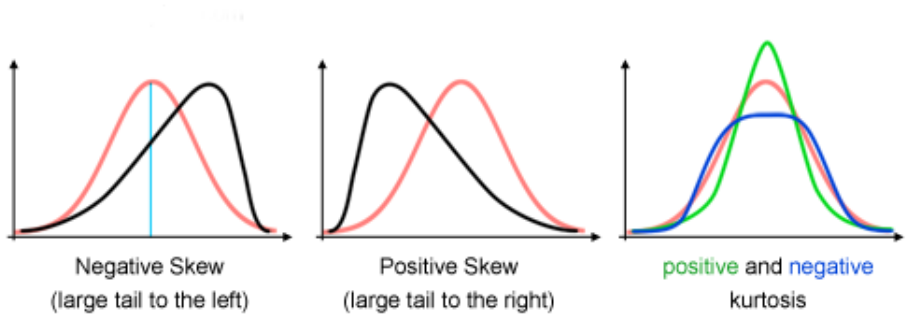


Figure A.1: Skewness and Kurtosis

- 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 (A.1) 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, (A.1) is not bounded between $[0, 1]$. Other expansions such as the Saddlepoint expansions [4] provide better approximations (which are out of the scope).
- The proof of (A.1) 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 (A.1). A proof is available in [1]. The asymptotic quantiles based on (A.1) can be found in a similar manner via Cornish–Fisher expansion. However, these technicalities are out of the module scope.

Example 31. 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

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

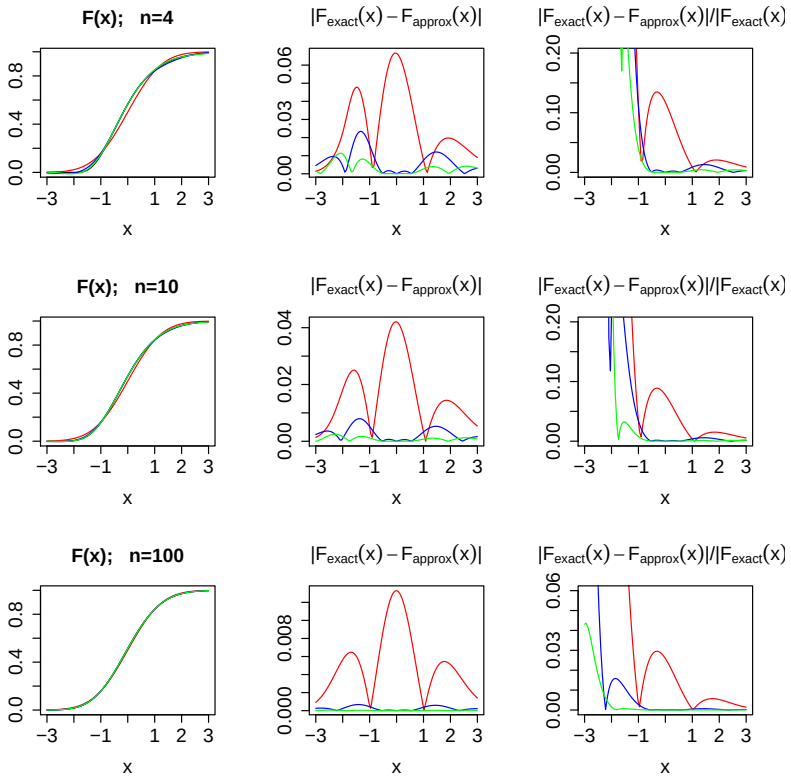
- The CDF of Z_n according to the Edgeworth exp. up to the $1/\sqrt{n}$ term

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

- 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 #18