MS = PARAMETRIC INFERENCE

In Probability Theory (PT), knowing the general population $(\Omega, \mathcal{F}, \mathbb{P})$, we deduce the distribution of the characteristics of the phenomenon. PT studies randomness [→ Direct problem]

- **Probability Statement:** Previous studies showed that a treatment was 80% effective on patients. We can anticipate that for a study on 50 patients, on average 40 will be cured, and at least 35 will be cured with 94.9% of chances.

In Mathematical Statistics (MS), we deal with the [Inverse problem]: given the observed composition of a random sample, what can be said about the general population?

- **Statistical Statement:** We observe that 45 out of 50 patients were cured. We conclude that we are 95% confident that in other studies the drug will be effective on between 82% and 98% of patients.

In MS, the Random Experiment is run n times, each time obtaining an independent observation of a generic random variable X. After n experiments, you collect

$\mathcal{L}_n = \{X_1, \dots, X_n\}$ = called a RANDOM SAMPLE.

In parametric statistics, we assume an underlying probability model $(\Omega, \mathcal{F}, \mathbb{P}_\theta)$, where $\mathbb{P}_\theta$ is a probability measure known up to a parameter $\theta \in \Theta \subset \mathbb{R}^d$.

↑ parameter space ≡ range of possible values for θ.

↳ We have a family of 'suspects' $\mathcal{P} = \{\mathbb{P}_\theta\}_{\theta \in \Theta}$, and we want to select the one that explains the data $\mathcal{L}_n$ the best.

- Technical point: We observe random variables $X_1, \dots, X_n$ with distribution P_θ induced by $\mathbb{P}_\theta$. We can identify $(\Omega, \mathcal{F}, P_\theta)$ with $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P_\theta)$ in the future.

Def: A function $S = S(X_1, \dots, X_n)$ of the data is called a STATISTICS.

Statistics are random variables

Another technical point, it should be a measurable function.

I - POINT ESTIMATION

I.1. Definition & Properties of an estimator

Estimators of the unknown parameter θ are statistics $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$, i.e. (measurable) functions of the random sample $\mathcal{L}_n$ taking values in the parameter space Θ.

An estimator is also a random variable, and it is meaningful to talk about its distribution, mean, variance... Probability statements will allow us to compare the accuracy of different estimators.

Def : (i) The BIAS of an estimator $\hat{\theta}$ of $\theta \in \Theta$ is

$$b(\hat{\theta}) := \mathbb{E}\hat{\theta} - \theta$$

↑ The expectation is taken under the true distribution p_θ. When needed, we emphasize the dependence on θ by writing $\mathbb{E}_\theta$ for $\mathbb{E}$.

(ii) An estimator $\hat{\theta}$ is UNBIASED if $b(\hat{\theta}) = 0$

(iii) The Mean Square Error (MSE) of $\hat{\theta}$ is

$$MSE(\hat{\theta}) = \mathbb{E}(\hat{\theta}-\theta)^2$$

Bias-Variance decomposition of the MSE:

$$\begin{aligned} MSE(\hat{\theta}) &= \mathbb{E}(\hat{\theta} - \mathbb{E}\hat{\theta} + \mathbb{E}\hat{\theta} - \theta)^2 \\ &= \mathbb{E}\left\{(\hat{\theta}-\mathbb{E}\hat{\theta})^2 + (\mathbb{E}\hat{\theta}-\theta)^2 \right. \\ &\qquad \left. + 2(\hat{\theta}-\mathbb{E}\hat{\theta})(\mathbb{E}\hat{\theta}-\theta)\right\} \\ &= \mathbb{E}\{(\hat{\theta}-\mathbb{E}\hat{\theta})^2\} + (\mathbb{E}\hat{\theta}-\theta)^2 \\ &= \mathrm{Var}\,\hat{\theta} + b(\hat{\theta})^2 \\ &= \text{Variance} + \text{Bias}^2 \end{aligned}$$

Conclusion: From an MSE point of view, a biased estimator may perform better than an unbiased one.

- Efficiency of an estimator. To measure the performance of an estimator $\hat{\theta}$, a standard approach is to compute its MSE, and select the estimator with the smallest MSE. A natural question: is there an estimator $\hat{\theta}^*$ minimizing the MSE for any $\theta \in \Theta$?

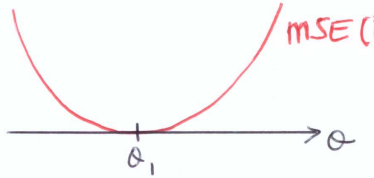
A disappointing answer: in general, no.

To prove this, suppose that there is such an estimator, and call it $\hat{\theta}^*$.

Take an arbitrary $\theta_1 \in \Theta$, and consider $\hat{\theta}_1 := \theta_1$ (estimator $\hat{\theta}_1$ is trivial, and does not depend on the random sample $\mathcal{L}_n$)

Then $MSE(\hat{\theta}_1) = \mathbb{E}_\theta(\hat{\theta}_1 - \theta)^2 = (\theta_1 - \theta)^2$

↑ equal to zero for $\theta = \theta_1$



However, $\hat{\theta}^*$ is the best estimator in the sense that it has the smallest MSE for any $\theta \in \Theta$. Thus

$$\begin{array}{ccc} MSE(\hat{\theta}_1) & \geq & MSE(\hat{\theta}^*) \\ \| & & \\ 0 \text{ for } \theta=\theta_1 & & \end{array}$$

$\Rightarrow$ Necessarily, $\mathbb{E}_{\theta_1}(\hat{\theta}^* - \theta_1)^2 = 0$

↑ But the choice of θ_1 is arbitrary: we must have $\mathbb{E}_{\theta_1}(\hat{\theta}^* - \theta_1)^2 = 0 \quad \forall \theta_1 \in \Theta$

($(\hat{\theta}^* - \theta_1)^2$: a positive random variable, with zero mean)

$\Rightarrow \hat{\theta}^* = \theta_1$ a.s. $\forall \theta_1 \in \Theta$

Conclusion: In general, $\hat{\theta}^* = \theta_1$ a.s. $\forall \theta_1 \in \Theta$ is impossible, except in trivial (and uninteresting) cases where observations $X_1, \dots, X_n$ uniquely determine the value of the parameter θ.

Ex: $X_i \sim \mathcal{U}(\theta, \theta+1)$, with $\theta \in \mathbb{Z}$; take $\hat{\theta}^* = \lfloor X_1 \rfloor$

In practice, we usually compare the performance of estimators within a reasonable class; for example the class of unbiased estimators.

Ex: For a function $b(\theta)$ of $\theta \in \Theta$, let

$$K_b := \{ \text{estimators } \hat{\theta} \mid E_\theta \hat{\theta} = \theta + b(\theta), \theta \in \Theta \}$$

= class of estimators with bias $b(\theta)$.

In particular, K_0 = class of unbiased estimators.

Def: An estimator $\hat{\theta}^* = \hat{\theta}^*(X_1, \dots, X_n)$ from a class K of estimators of θ is called **EFFICIENT in K** if

$$\forall \hat{\theta} \in K, \quad E_\theta(\hat{\theta}^* - \theta)^2 \leq E_\theta(\hat{\theta} - \theta)^2$$
$$MSE(\hat{\theta}^*) \leq MSE(\hat{\theta})$$

↑ Estimators efficient in K_0 are simply called efficient.

Theorem: If it exists, an efficient estimator in K_b is unique. (a.s.)

proof = Suppose that both $\hat{\theta}_1^*$ and $\hat{\theta}_2^*$ are efficient in K_b. Then for $i = 1, 2$,

$$E_\theta(\hat{\theta}_i^* - \theta)^2 = \min_{\hat{\theta} \in K_b} E_\theta(\hat{\theta} - \theta)^2 =: R_\theta, \quad \forall \theta \in \Theta.$$

Consider $\hat{\theta}_0^* := \frac{1}{2}(\hat{\theta}_1^* + \hat{\theta}_2^*)$.

→ Clearly $\hat{\theta}_0^* \in K_b$

→ Making use of the equality $\left(\frac{a_1+a_2}{2}\right)^2 + \left(\frac{a_1-a_2}{2}\right)^2 = \frac{a_1^2+a_2^2}{2}$, with $a_i = \hat{\theta}_i^* - \theta$, $i = 1, 2$, we have

- $\frac{a_1+a_2}{2} = \frac{\hat{\theta}_1^* + \hat{\theta}_2^* - 2\theta}{2} = \hat{\theta}_0^* - \theta$
- $\frac{a_1-a_2}{2} = \frac{\hat{\theta}_1^* - \hat{\theta}_2^*}{2}$
- $\frac{a_1^2+a_2^2}{2} = \frac{1}{2}\{(\hat{\theta}_1^* - \theta)^2 + (\hat{\theta}_2^* - \theta)^2\}$

Taking $E_\theta\{\dots\}$ yields

$$\underbrace{E_\theta(\hat{\theta}_0^* - \theta)^2}_{\geq R_\theta \text{ since not efficient in } K_b} + \frac{1}{4} E_\theta(\hat{\theta}_1^* - \hat{\theta}_2^*)^2 = \frac{1}{2}\{E_\theta(\hat{\theta}_1^* - \theta)^2 + E_\theta(\hat{\theta}_2^* - \theta)^2\} = R_\theta$$

$$\Rightarrow E_\theta \underbrace{(\hat{\theta}_1^* - \hat{\theta}_2^*)^2}_{\text{positive RV}} \leq 0 \Rightarrow \hat{\theta}_1^* = \hat{\theta}_2^* \text{ a.s.} \quad \blacksquare$$

× Asymptotic properties of an estimator.

↘ An estimator $\hat{\theta}_n = \hat{\theta}_n(X_1, \dots, X_n)$ is called (↖ we emphasize the dependence on the sample size)

WEAKLY (resp. STRONGLY) CONSISTENT if $\hat{\theta}_n \to \theta$ in probability (resp. a.s.) as $n \to +\infty$.

↘ An estimator is called asymptotically unbiased if $b(\hat{\theta}_n) \to 0$ as $n \to +\infty$.

(and similarly for asymptotically efficient, and so on)

I.2 Sufficient Statistics

- Goal: Construct an estimator $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$ of θ based on $\underline{X} = (X_1, \dots, X_n)$, where the X_i are iid with distribution P_θ, and density $f_\theta(x)$ (the case of discrete RVs is treated similarly, replacing densities $f_\theta(x)$ by probability masses $\mathbb{P}_\theta(X_i = x)$). The joint density of $(X_1, \dots, X_n)$ is $f_\theta(\underline{x}) = \prod_{i=1}^{n} f_\theta(x_i)$, where $\underline{x} = (x_1, \dots, x_n)$.

- A statistic cannot contain more information about θ than the initial sample itself [there is a reduction of information]. A sufficient statistic will throw away the "useless" information about θ, and keeping all the "usefull" information.
  ↳ How can we express this statement mathematically?

- Let $S = S(X_1, \dots, X_n)$ be a statistic.

- Consider the following experiment:

Statistician 1 (S1)	Statistician 2 (S2)
* Given $(X_1, \dots, X_n)$	* Not given $(X_1, \dots, X_n)$
↓	
* Constructs $S = S(X_1, \dots, X_n)$ and gives it to Statistician 2	
↓	* Observes S
	* knows its distribution
* Constructs an estimator of θ	↓
	Wants to construct an estimator of θ.

To have as much information as S1, S2 should be able to generate a new sample $(X_1', \dots, X_n')$ distributed like $(X_1, \dots, X_n)$, and then use this new sample to construct his estimator (or, even better, construct an estimator directly from the observed value of S).

Assuming discrete $X_1, \dots, X_n$ and S, we have

$$\mathbb{P}_\theta(\underline{X} = \underline{x}) = \mathbb{P}_\theta(\underline{X} = \underline{x},\ S(\underline{X}) = s(\underline{x}))$$
$$= \underbrace{\mathbb{P}_\theta(\underline{X} = \underline{x} \mid S(\underline{X}) = s(\underline{x}))}_{\text{This term usually depends on } \theta. \text{ However, if it does not, statistician 2 can generate a new sample from this conditional distribution!}} \underbrace{\mathbb{P}_\theta(S(\underline{X}) = s(\underline{x}))}_{\text{known to S2}}$$

⇒ Take this as a definition of a sufficient stat!

Def: A statistic $S = S(X_1, \dots, X_n)$ is called a SUFFICIENT STATISTIC (SS) for θ if the conditional distribution of $\underline{X}$ given S does not depend on θ.

↳ Note that if φ is a one-to-one function, the $\varphi(S)$ is also a SS for θ.

* **Example**: Let . $X_1, \dots, X_n$ be iid $P(\lambda)$
  . $S = X_1 + \dots + X_n$

We show that S is a sufficient statistic for λ

Recall that $S \sim P(n\lambda)$

Fix an integer $s \geq 0$, and take $\underline{x} = (x_1, \dots, x_n)$.

Then . $P_\lambda(\underline{X}=\underline{x} \mid S(\underline{X})=s) = 0$ if $\sum_{i=1}^n x_i \neq s$.

. If $s = \sum_{i=1}^n x_i$, we have

$$P_\lambda(\underline{X}=\underline{x} \mid S(\underline{X})=s) = \frac{P_\lambda(\underline{X}=\underline{x})}{P_\lambda(S=s)}$$

$$= \frac{\prod_{i=1}^n \frac{\lambda^{x_i}}{x_i!} e^{-\lambda}}{\frac{(n\lambda)^s}{s!} e^{-n\lambda}}$$

$$= \frac{s!}{(n\lambda)^s} \frac{\lambda^{\sum x_i}}{\prod_{i=1}^n x_i!}$$

Multinomial distribution with $\sum_{i=1}^n x_i$ independent trials and n equally likely outcomes. ⟶

$$= \frac{(x_1+\cdots+x_n)!}{x_1!\cdots x_n!} \left(\frac{1}{n}\right)^{x_1} \cdots \left(\frac{1}{n}\right)^{x_n}$$

↑ independent of λ ∎

We proved this the direct way, i.e. making use of the definition of a SS. But there are simpler ways to find SS [what if the question was not show that S is a SS, but find a SS for λ ?], one such way is using the following theorem.

Theorem [NEYMAN-FISHER FACTORISATION]

Suppose that $X_1, \ldots, X_n$ are iid with distribution P_θ, and density f_θ, so that the joint density of $\underline{X} = (X_1, \ldots, X_n)$ is $f_\theta(\underline{x}) = \prod_{i=1}^n f_\theta(x_i)$, where $\underline{x} = (x_1, \ldots, x_n)$.
A necessary & sufficient condition for a statistic $S = S(\underline{X})$ to be an SS for θ is that for some functions $\psi(s, \theta)$ and $h(x)$, we have $f_\theta(\underline{x}) = h(\underline{x})\,\psi(S(\underline{x}), \theta)$.

The theorem holds as well in the discrete case, replacing $f_\theta(\underline{x})$ by $P_\theta(\underline{X}=\underline{x})$.

x **Example** (continued) : $X_1, \ldots, X_n$ iid $\sim P(\lambda)$.

$$f_\lambda(\underline{x}) = \prod_{i=1}^n \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} = \frac{e^{-n\lambda}\,\lambda^{\sum x_i}}{\prod_i x_i!} = h(\underline{x})\,\psi(S(\underline{x}), \lambda),$$

with $h(\underline{x}) = \frac{1}{\prod_i x_i!}$, $S(\underline{x}) = \sum_{i=1}^n x_i$, and $\psi(s, \lambda) = \lambda^s e^{-n\lambda}$.

We conclude that $S(\underline{X}) = \sum_{i=1}^n X_i$ is a SS for λ (but we knew this already). ∎

x **Example** : $X_1, \ldots, X_n$ iid $\sim \mathcal{N}(\mu, \sigma^2)$. $\theta = (\mu, \sigma^2)$

$$f_\theta(\underline{x}) = \frac{1}{(2\pi\sigma^2)^{-\frac{n}{2}}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2\right\}$$

$$= \frac{1}{(2\pi\sigma^2)^{-\frac{n}{2}}} \exp\left\{-\frac{1}{2\sigma^2}\left[\underbrace{\sum x_i^2}_{=S_2} - 2\mu\underbrace{\sum x_i}_{=S_1} + n\mu^2\right]\right\}$$

$$= \frac{1}{(2\pi\sigma^2)^{-\frac{n}{2}}} \exp\left\{-\frac{1}{2\sigma^2}\left[S_2 - 2\mu S_1 + n\mu^2\right]\right\}$$

$$= h(\underline{x})\,\psi(S(\underline{x}), \theta),$$

with $S = (S_1, S_2)$, $h(\underline{x}) \equiv 1$, and $\psi(s, \theta) = \ldots$.

$\Rightarrow$ S is a SS for (μ, σ^2).
In other words, out of all the information contained in $X_1, \ldots, X_n$, we only need $\sum X_i$ and $\sum X_i^2$ for the estimation of μ and σ^2. ∎

× **Example**: $X_1, \dots, X_n$ iid $X_i \sim \mathcal{U}(0, \theta)$.

The X_is have density $f_\theta(x) = \begin{cases} 1/\theta & \text{if } x \in [0,\theta] \\ 0 & \text{otherwise} \end{cases}$

The joint density of $X_1, \dots, X_n$ is

$$f_\theta(\underline{x}) = \prod_{i=1}^n f_\theta(x_i) = \begin{cases} 1/\theta^n & \text{if all } x_i \in [0,\theta],\ i \le n \\ 0 & \text{otherwise} \end{cases}$$

$$= \theta^{-n}\, \mathbb{1}\left(0 \le x_{(1)},\ x_{(n)} \le \theta\right)$$

Making use of the notation $x_{(1)} = \min_{1 \le i \le n} x_i$

$x_{(n)} = \max_{1 \le i \le n} x_i$

$$\Rightarrow f_\theta(\underline{x}) = \underbrace{\theta^{-n}\, \mathbb{1}(x_{(n)} \le \theta)}_{\psi(S(\underline{x}),\, \theta)} \underbrace{\mathbb{1}(0 \le x_{(1)})}_{h(\underline{x})}$$

with $S(\underline{x}) = x_{(n)}$.

Therefore, $S(\underline{X}) = \max_{1 \le i \le n} X_i$ is a SS for θ. ∎

× **Example**: $X_1, \dots, X_n$ iid $X_i \sim \gamma(r, \lambda)$, λ known, r unknown

Then

$$f_r(\underline{x}) = \frac{\lambda^{nr}}{(\Gamma(r))^n} \left(\prod_{i=1}^n x_i^{r-1}\right) \exp\left\{-\lambda \sum_{i=1}^n x_i\right\}.$$

Writing $\prod_{i=1}^n x_i^{r-1} = \exp\left\{(r-1) \sum_{i=1}^n \ln x_i\right\}$, we see that $S(\underline{X}) = \sum_{i=1}^n \ln X_i$ is an SS; $\exp S(\underline{X}) = \prod_{i=1}^n X_i$ as well, but not the sample mean. ∎

Corollary: If T is a statistic
ψ is a function
$S := \psi(T)$ is a SS for θ
Then T is also a SS for θ.

↑ This follows immediately from the Neyman-Fisher factorization theorem.

↳ A MINIMAL SUFFICIENT STATISTICS (MSS) is a SS that is a function of all the others.

Example (continued): $X_1, \dots, X_n \sim \mathcal{P}(\lambda)$.

We saw on page 10 that $S_1(X_1, \dots, X_n) = \sum_{i=1}^n X_i$ is an SS for λ.

Obviously, $S_2(X_1, \dots, X_n) = \alpha \sum_{i=1}^n X_i$, $\alpha \neq 0$

$S_3(X_1, \dots, X_n) = \left(\sum_{i=1}^n X_i,\ X_n\right)$

$S_4(X_1, \dots, X_n) = (X_1, \dots, X_n)$

are also SS for λ.

It turns out that you can construct/express S_1 as a function of S_2, S_3, S_4. However, you cannot obtain S_3 or S_4 from S_1. It turns out that $\sum X_i$ is a minimal SS for λ. Note that S_2 also is a MSS for λ. The MSS is not unique, but the partition it creates is. The MSS are the ones that generate the coarsest partition: the ones that achieve the greatest data reduction, without loss of information.

- There exists a criterion for finding MSS: if $\frac{f_\theta(\underline{x})}{f_\theta(\underline{y})}$ is independent of $\theta \iff T(\underline{x}) = T(\underline{y})$; then T is a MSS for θ.

The following result shows that SS can improve the efficiency of an estimator.

↳ cf page 5 for the definition of an efficient estimator.

Theorem (RAO-BLACKWELL)

Let. $\hat\theta \in K_b$ (class of estimators with bias $b(\theta)$)
. S a sufficient statistic for θ

Then, the conditional expectation $\hat\theta_S := \mathbb{E}_\theta(\hat\theta \mid S)$ satisfies:

(i) $\hat\theta_S$ is a function of S only

(ii) $\hat\theta_S \in K_b$

(iii) $\forall \theta \in \Theta$, $\mathbb{E}_\theta(\hat\theta_S - \theta)^2 \leq \mathbb{E}_\theta(\hat\theta - \theta)^2$, with equality iff $\mathbb{P}_\theta(\hat\theta_S = \hat\theta) = 1$.

↑ Applying $\mathbb{E}(.\mid S)$ decreases the MSE of the original estimator $\hat\theta$.

proof: (i) $\hat\theta_S = \mathbb{E}(\hat\theta \mid S) = \int \hat\theta(x)\, \mathbb{P}_\theta(\underline{X} \in d\underline{x} \mid S) = \int \hat\theta(x)\, \mathbb{P}(\underline{X} \in d\underline{x} \mid S)$ ← since S is an SS

$\Rightarrow \hat\theta_S$ is itself a statistic.

(ii) $\mathbb{E}_\theta \hat\theta_S = \mathbb{E}_\theta\, \mathbb{E}_\theta(\hat\theta \mid S) = \mathbb{E}_\theta \hat\theta = \theta + b(\theta)$ since $\hat\theta \in K_b$.
Hence $\hat\theta_S \in K_b$ as well.

(iii) $\mathbb{E}_\theta(\hat\theta - \theta)^2 = \mathbb{E}_\theta(\hat\theta - \hat\theta_S + \hat\theta_S - \theta)^2$
$= \mathbb{E}_\theta(\hat\theta - \hat\theta_S)^2 + \mathbb{E}(\hat\theta_S - \theta)^2$
$+ 2\,\underbrace{\mathbb{E}(\hat\theta - \hat\theta_S)(\hat\theta_S - \theta)^2}_{=0}$
$\geq \mathbb{E}(\hat\theta_S - \theta)^2$

with equality iff $\mathbb{E}_\theta(\hat\theta - \hat\theta_S)^2 = 0$; i.e. iff $\hat\theta = \hat\theta_S$ a.s. ∎

* **Remark:** Recall that if $S = \varphi(T)$ is a SS for θ, for T = statistic and φ = some function, then T is also a SS for θ.

Consider $\hat\theta_T := \mathbb{E}_\theta(\hat\theta \mid T)$. Then

$\mathbb{E}(\hat\theta_T - \theta)^2 = \mathbb{E}(\hat\theta_T - \hat\theta_S + \hat\theta_S - \theta)^2$
$= \underbrace{\mathbb{E}(\hat\theta_T - \hat\theta_S)^2}_{\geq 0} + \mathbb{E}(\hat\theta_S - \theta)^2$
$+ 2\,\underbrace{\mathbb{E}(\hat\theta_T - \hat\theta_S)(\hat\theta_S - \theta)}$

$\mathbb{E}\{\hat\theta_S - \theta\,[\underbrace{\mathbb{E}(\hat\theta_T'' - \hat\theta_S \mid S)}]\}$

$\mathbb{E}(\hat\theta_T'' \mid S) - \hat\theta_S = 0$

since $\mathbb{E}(\hat\theta_T'' \mid S) = \mathbb{E}\big(\mathbb{E}[\hat\theta \mid T] \mid S\big)$
↑ fine partition; ↑ $S = \varphi(T)$ coarser partition

$= \mathbb{E}(\hat\theta \mid S) = \hat\theta_S$.

Conclusion: $\mathbb{E}(\hat\theta_S - \theta)^2 \leq \mathbb{E}(\hat\theta_T - \theta)^2$

⇒ The smaller the conditioning, the better ⇒ MSS are preferred!

Example: Let. $X_1, \dots, X_n \sim \mathcal{P}(\lambda)$ i.i.d
. $\hat\lambda = X_1 \in K_0$ (since $\mathbb{E}\hat\lambda = \mathbb{E}X_1 = \lambda$)

$\mathrm{MSE}(\hat\lambda) = \mathrm{Var}\,\hat\lambda = \mathrm{Var}\,X_1 = \lambda$.

Since $S = \sum_{i=1}^n X_i$ is an SS for λ, we can consider the estimator

$\hat\lambda_S := \mathbb{E}_\lambda(\hat\lambda \mid S) = \mathbb{E}_\lambda(X_1 \mid S) = \frac{S}{n} = \overline{X}$

↑ since $(X_1 \mid S = m) \sim \mathrm{Bi}(m, \frac{1}{n})$

$\mathbb{E}\hat{\lambda}_S = \lambda$ so that $\hat{\lambda}_S \in K_0$, and

$MSE(\hat{\lambda}_S) = Var\ \hat{\lambda}_S = \frac{\lambda}{n} \ll Var\ \hat{\lambda}$ ■

× Let • S be a sufficient statistic for Θ
• T be a MSS for Θ, so that $T = \varphi(S)$ for some function φ.

Consider $U := S - \underbrace{\mathbb{E}_\theta(S|T)}_{\text{function of } T = \varphi(S) \Rightarrow \text{function of } S}$,

and we conclude that we can write $U = g(S)$, for some function g.

Next, note that $\mathbb{E}_\theta U = \mathbb{E}_\theta S - \mathbb{E}_\theta \mathbb{E}_\theta(S|T) = 0$.

Denoting G_θ the distribution of U under P_θ, we have that

$$\int g(s)\, G_\theta(ds) = 0 \quad , \quad \theta \in \Theta \qquad (*)$$

If this implies that $g(s) = 0$,

we would have that $U \equiv 0 = S - \mathbb{E}_\theta(S|T)$.

$\Rightarrow S = \mathbb{E}_\theta(S|T)$

$\Rightarrow S$ is a function of T

$\Rightarrow S$ is also a MSS

Conclusion: if you find a sufficient statistic S whose distribution G_θ satisfies (*), then S is a MSS.

Distributions satisfying (*) have a name: they are called COMPLETE.

(under this terminology, any complete SS is minimal)

(but the converse is not always true)

Def: A family $\{G_\theta\}_{\theta\in\Theta}$ of distributions on $\mathbb{R}^m$ is said to be COMPLETE if, given a function $g: \mathbb{R}^m \to \mathbb{R}$

$$\int g(\underline{s})\, G_\theta(d\underline{s}) = 0 \Rightarrow G_\theta(\{\underline{s} \mid g(\underline{s}) = 0\}) = 1, \ \forall \theta$$

A statistic $S = S(X_1, \dots, X_n) \in \mathbb{R}^m$ is said to be COMPLETE if the family of its distributions $\{G_\theta\}_{\theta\in\Theta}$ induced by $\{P_\theta\}_{\theta\in\Theta}$ is complete.

$\Rightarrow$ complete SS are minimal + conditioning on MSS achieve greater reduction of the MSE $\Rightarrow$ you should be searching for complete SS.

× Example = $G_\theta = Bi(n, \theta)$, for $\theta \in \Theta = (0,1)$. Then

$$0 = \int g(s)\, G_\theta(ds) = \sum_{k=0}^{n} \binom{n}{k} g(k)\, \theta^k (1-\theta)^{n-k}$$

$$= (1-\theta)^n \sum_{k=0}^{n} g(k) \binom{n}{k} \left(\frac{\theta}{1-\theta}\right)^k$$

$$= \text{polynomial of order } n \text{ in } u = \frac{\theta}{1-\theta} \in [0, \infty)$$

So $\forall u \in [0,\infty)$, we have that $\sum_{k=0}^{n} \alpha_k u^k = 0$. This implies that all coefficients α_k must vanish; i.e. $g(k) = 0$ for $k = 0, \dots, n$. We obtain $G_\theta(\{0 \le k \le n \mid g(k) = 0\}) = 1$ ■

× Example = $X_1, \dots, X_n \sim \mathcal{P}(\lambda)$, with $S(X_1, \dots, X_n) = \sum_{i=1}^{n} X_i$.

Then $S \sim \mathcal{P}(n\lambda) = G_\theta$. For a function g satisfying:

$$0 = \int g(s)\, G_\theta(ds) = \sum_{k\ge 0} g(k)\, G_\theta(S = k) = \sum_{k\ge 0} g(k) \frac{(n\lambda)^k}{k!} e^{-n\lambda}$$

$$\Leftrightarrow \sum_{k\ge 0} \frac{g(k)}{k!} z^k = 0 \ , \ \forall z \ge 0 \Rightarrow g(k) = 0 \ , \ k \ge 0$$

■

Theorem: If • S is a complete SS for θ
• $\hat{\theta} \in K_b$ (class of estimators with bias K_b)
Then $\hat{\theta}_S := \mathbb{E}_\theta(\hat{\theta} \mid S)$ is the unique efficient estimator in K_b.

In the case $\hat{\theta} \in K_0$, this result is known as the LEHMANN–SHEFFÉ THEOREM: $\hat{\theta}_S = \mathbb{E}(\hat{\theta} \mid S)$ is the unique minimum variance unbiased estimator.

x **Example** (continued): $X_1, \dots, X_n \sim \mathcal{P}(\lambda)$, with $\hat{\lambda} = X_1$.
Take $S = \sum_{i=1}^n X_i$ = complete SS.
page 16 page 10

Then $\hat{\lambda}_S = \mathbb{E}_\lambda(\hat{\lambda} \mid S) = \frac{1}{n}\sum_{i=1}^n X_i = \bar{X} \in K_0$ is the unique minimum variance unbiased estimator. It is efficient in the class K_0. ∎

x **Example:** $X_1, \dots, X_n \sim B(p)$ iid
Consider $\hat{p} = X_1 \in K_0$, and $S = \sum_{i=1}^n X_i$ = SS for p.
check this using the NF factorization.

We can form $\hat{p}_S := \mathbb{E}(\hat{p} \mid S) = \mathbb{E}(X_1 \mid S)$.
Note that $S \sim Bi(n, p)$, and we saw on page 16 that this family of distributions is complete. We deduce that S is a complete SS, and that $\hat{p}_S = \mathbb{E}(X_1 \mid S)$ is the unique minimum variance unbiased estimator of p. We compute:

$$\hat{p}_S = \mathbb{E}(X_1 \mid S) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}(X_i \mid S) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n X_i \mid S\right) = \frac{1}{n}\sum_{i=1}^n X_i$$

since $\mathbb{E}(X_1 \mid S) \overset{a.s.}{=} \mathbb{E}(X_i \mid S)$ $\forall i$ ∎

I.2. Cramer-Rao Inequality.

- Given an estimator $\hat{\theta}$, can we say how efficient it is? What is the smallest MSE one can hope for?
- Let $X_1, \dots, X_n$ be a random sample (so that $X_1, \dots, X_n$ are iid RVs), with $X_i \sim P_\theta$, for $\theta \in \Theta$.
- Suppose that the X_i are absolutely continuous, with density f_θ. The density of $(X_1, \dots, X_n) = \underline{X}$ is then
$$f_\theta(\underline{x}) = \prod_{i=1}^n f_\theta(x_i).$$
We put $$\ell(\underline{x}, \theta) := \log f_\theta(\underline{x}) = \sum_{i=1}^n \underbrace{\log f_\theta(x_i)}_{=: \ell(x_i, \theta)}.$$
- Note that if X is discrete, the density f_θ is replaced by probability masses. Results will be presented/proved in the absolutely continuous case, but hold in the discrete case as well.

Ex: (i) $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ iid $\Rightarrow \theta = (\mu, \sigma^2)$.
Then $$\ell(x_i; \theta) = \log\left\{\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu)^2\right)\right\}$$
$$\ell(\underline{x}; \theta) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2,$$
since $$f_\theta(\underline{x}) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x_i - \mu)^2\right\}$$

(ii) $X_1, \dots, X_n \sim \mathcal{P}(\lambda)$ iid $\Rightarrow \theta = \lambda$, and
$$\ell(x_i; \theta) = \log\left\{\frac{\lambda^{x_i}}{x_i!} e^{-\lambda}\right\}, \quad x_i \in \mathbb{N}$$
$$\ell(\underline{x}; \theta) = -n\lambda - \sum_{i=1}^n \log(x_i!) + \log\lambda \sum_{i=1}^n x_i.$$

* **Score.** Provided $f_\theta(x)$ is differentiable with respect to θ, the score of the random sample $\mathcal{L}_n$ is given by

$$S_n(\theta) := \frac{\partial}{\partial\theta}\, \ell(\underline{X};\theta) = \frac{1}{f_\theta(x)}\left\{\frac{\partial}{\partial\theta} f_\theta(x)\right\}$$

Random quantity (it inherits the randomness of $X_1, \dots, X_n$)

The derivative with respect to θ will be denoted $\ell'(\underline{X};\theta)$ when there is no confusion.

Note that the score has zero mean:

$$\mathbb{E}_\theta\{S_n(\theta)\} = \mathbb{E}_\theta\{\ell'(\underline{X};\theta)\}$$

$$= \int \ell'(\underline{x};\theta) f_\theta(x)\, dx \qquad f'_\theta(x) = \frac{d f_\theta(x)}{dx}$$

$$= \int \frac{f'_\theta(x)}{f_\theta(x)} f_\theta(x)\, dx$$

$$= \frac{d}{d\theta}\int f_\theta(x)\, dx \qquad \text{Under the assumption that we can exchange } \int \text{ and derivative}$$

$$= 0.$$

Def: The variance of the score function $s_n(\theta) = \ell'(\underline{X};\theta)$ is called **FISHER INFORMATION**, and is denoted $I_n(\theta)$:

$$I_n(\theta) = \mathrm{Var}_\theta\{S_n(\theta)\} = \mathrm{Var}_\theta\{\ell'(\underline{X};\theta)\} = \mathbb{E}_\theta\{(\ell'(\underline{X};\theta))^2\}.$$

This quantity turns out to be the amount of "information" carried by the random sample $(X_1, \dots, X_n)$. A small value of $I_n(\theta)$ [small variance] indicates that the sample carries little information about θ. This will be better understood when we discuss max likelihood.

↳ **Alternative expression of Fisher information.**

We saw previously that $\mathbb{E}_\theta\{\ell'(\underline{X};\theta)\} = 0$. Thus

$$0 = \frac{\partial}{\partial\theta}\left\{\mathbb{E}_\theta\{\ell'(\underline{X};\theta)\}\right\} = \int \frac{d}{d\theta}\left\{f_\theta(\underline{x})\,\ell'(\underline{x},\theta)\right\} d\underline{x} \qquad \text{Technical!}$$

$$= \int \left\{ f_\theta(\underline{x})\,\ell''(\underline{x};\theta) + \ell'(\underline{x},\theta)\frac{d f_\theta(\underline{x})}{d\theta}\right\} d\underline{x}$$

Note that $\dfrac{d \log f_\theta(\underline{x})}{d\theta} = \dfrac{f'_\theta(\underline{x})}{f_\theta(\underline{x})}$,

so that $f'_\theta(\underline{x}) = f_\theta(\underline{x})\,\ell'(\underline{x};\theta)$

$$= \int \left\{ f_\theta(\underline{x})\,\ell''(\underline{x};\theta) + f_\theta(\underline{x})\left[\ell'(\underline{x};\theta)\right]^2\right\} d\underline{x}$$

$$= \int \left\{\ell''(\underline{x};\theta) + \left[\ell'(\underline{x};\theta)\right]^2\right\} f_\theta(\underline{x})\, d\underline{x}$$

$$= \mathbb{E}_\theta\{\ell''(\underline{X};\theta)\} + \underbrace{\mathbb{E}_\theta\{(\ell'(\underline{X};\theta))^2\}}_{= I_n(\theta)}$$

We conclude that:

$$I_n(\theta) = \mathbb{E}_\theta\{(\ell'(\underline{X};\theta))^2\} = -\,\mathbb{E}_\theta\{\ell''(\underline{X};\theta)\}$$

(Fisher information for n observations)

Consequence: Since $X_1, \dots, X_n$ are iid, $\ell''(\underline{X};\theta) = \sum_{i=1}^{n} \ell''(x_i;\theta)$, and

$$I_n(\theta) = n\, I_1(\theta)$$

For iid observations only!

Theorem (CRAMER-RAO INEQUALITY)

Suppose that $\ell(\underline{x}, \theta)$ is continuously differentiable in $\theta \in \Theta$, and that the Fisher information $I_n(\theta)$ is a continuous function of θ. If $\hat{\theta} \in K_b$, then

$$\mathrm{Var}_\theta \hat{\theta} \geq \frac{(1 + b'(\theta))^2}{I_n(\theta)}.$$

in the case $\hat{\theta} \in K_0$, one obtains $\mathrm{Var}_\theta \hat{\theta} \geq \frac{1}{I_n(\theta)}$:

the larger the Fisher information, the smaller the variance one can hope for.

Equivalently, one obtains

$$\mathbb{E}_\theta(\hat{\theta} - \theta)^2 \geq \frac{(1 + b'(\theta))^2}{I_n(\theta)} + b^2(\theta).$$

Remark: in the case of iid observations, $I_n(\theta) = n\, I_1(\theta)$, CR indicates that in "regular" cases (ie in cases where the conditions of the theorem are met), one cannot hope to construct estimators whose variance decreases faster than n^{-1}. Note that the CR bound is just a bound. It does not need to be attained.

× Example : $X_1, \dots, X_n \sim \mathcal{P}(\lambda)$

- Then $f_\lambda(x) = \mathbb{P}_\lambda(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}$, so that

$$f'_\lambda(x) = \frac{x\,\lambda^{x-1}}{x!} e^{-\lambda} - \frac{\lambda^x}{x!} e^{-\lambda} = \frac{\lambda^x}{x!}\left(\frac{x}{\lambda} - 1\right) e^{-\lambda}$$

$$\Rightarrow \ell'(x;\lambda) = \frac{f'_\lambda(x)}{f_\lambda(x)} = \frac{x}{\lambda} - 1 = \text{continuous function of } \lambda.$$

- Moreover, $I_1(\lambda) = \mathbb{E}_\lambda\{(\ell'(X;\lambda))^2\} = \mathbb{E}_\lambda\left(\frac{X}{\lambda} - 1\right)^2 = \frac{\mathrm{Var}\, X}{\lambda^2}$

$$= \frac{1}{\lambda},$$

so that $I_1(\lambda)$ is a continuous function of λ.

- Now, for $\hat{\lambda} = \frac{1}{n}\sum_{i=1}^n X_i = \overline{X} \in K_0$, we see that

$$\mathrm{Var}\,\hat{\lambda} = \frac{\lambda}{n} = \frac{1}{I_n(\lambda)}$$

(recall: $I_n(\lambda) = n\, I_1(\lambda)$)

$\Rightarrow \hat{\lambda} = \overline{X}$ has the smallest possible variance amongst the class of unbiased estimator; it is efficient. But we knew this already from page 17. ∎

× Example : $X_1, \dots, X_n \sim \mathcal{U}(0, \theta)$ iid. (not differentiable in θ.)

Consider the estimator $\hat{\theta} = \frac{n+1}{n} X_{(n)}$, where $X_{(n)} = \max_{1 \leq i \leq n} X_i$.

$$\mathbb{E}\, X_{(n)} = \int_0^\infty (1 - F_{X_{(n)}}(u))\, du = \int_0^\theta \left(1 - \left(\frac{u}{\theta}\right)^n\right) du = \frac{n\theta}{n+1}$$

(Since $\mathbb{P}(X_{(n)} \leq u) = \mathbb{P}(X \leq u)^n = \left(\frac{u}{\theta}\right)^n$, for $u \in [0, \theta]$.)

$$\Rightarrow \mathbb{E}\,\hat{\theta} = \theta \Rightarrow \hat{\theta} \in K_0.$$

The distribution of $X_{(n)}^2$ is

$$F_{X_{(n)}^2}(x) = \mathbb{P}(X_{(n)}^2 \leq x) = \mathbb{P}(X_{(n)} \leq \sqrt{x}) = F_{X_{(n)}}(\sqrt{x}).$$

Thus,

$$\mathbb{E}(X_{(n)}^2) = \int_0^\infty (1 - F_{X_{(n)}}(\sqrt{u}))\, du = 2\int_0^\infty u(1 - F_{X_{(n)}}(u))\, du$$

$$= n\theta^2/(n+2).$$

$$\mathrm{Var}_\theta \hat{\theta} = \left(\frac{n+1}{n}\right)^2 \frac{n\theta^2}{n+2} - \theta^2 = \frac{\theta^2}{n(n+2)} \leftarrow \text{order } n^{-2}!$$

In regular cases, we only have order n^{-1}. ∎

proof = • We saw on page 19 that $\mathbb{E}_\theta\{\ell'(\underline{X};\theta)\}=0$.

Thus $\mathbb{E}\{(\theta+b(\theta))\,\ell'(\underline{X};\theta)\}=0$ ——— (*)

• Next, consider the function $a(\theta) := \mathbb{E}_\theta\hat{\theta} = \theta + b(\theta)$.

$$= \int \hat{\theta}(\underline{x})\, f_\theta(\underline{x})\, dx.$$

Under the conditions of the theorem, it is possible to show that $a(\theta)$ is a differentiable function of θ, and that

$$a'(\theta) = \int \hat{\theta}(x)\, f_\theta'(\underline{x})\, dx = 1 + b'(\theta).$$

$$= \int \hat{\theta}(x)\, \ell'(\underline{x};\theta)\, f_\theta(\underline{x})\, dx \quad \text{since } \ell'(\underline{x};\theta) = \frac{f_\theta'(\underline{x})}{f_\theta(\underline{x})}$$

$$= \mathbb{E}_\theta\{\hat{\theta}\,\ell'(\underline{X};\theta)\}$$

We obtain $\mathbb{E}_\theta\{\hat{\theta}\,\ell'(\underline{X};\theta)\} = 1 + b'(\theta)$ ——— (**)

• Substracting (*) from (**) gives

$$\mathbb{E}_\theta\{(\hat{\theta} - [\theta + b(\theta)])\,\ell'(\underline{X};\theta)\} = 1 + b'(\theta).$$

$$\Rightarrow (1+b'(\theta))^2 \leq \underbrace{\mathbb{E}_\theta(\hat{\theta} - [\theta+b(\theta)])^2}_{=\mathrm{Var}_\theta\hat{\theta}}\ \underbrace{\mathbb{E}\{(\ell'(\underline{X};\theta))^2\}}_{=I_n(\theta)}$$

Cauchy-Bunyakovsky inequality: $E|XY| \leq \sqrt{EX^2\,EY^2}$

And we obtain indeed that $\mathrm{Var}_\theta\hat{\theta} \geq \dfrac{(1+b'(\theta))^2}{I_n(\theta)}$. ∎

II - INTERVAL ESTIMATION

II.1. The general principle.

Point estimation returns a single value as an estimate of a parameter. It is often desirable to express the degree of confidence we have in this estimation. We use the same notation as in section I.

Definition. Let $\alpha \in (0,1)$. A CONFIDENCE INTERVAL for θ of level $(1-\alpha)$ is a statistic I taking values in the intervals of $\mathbb{R}$, such that for all $\theta \in \Theta$,

$$\mathbb{P}_\theta(\theta \in I) = 1-\alpha.$$

(θ = parameter of interest; statistic = i.e. does not depend on θ, and function of $X_1, \dots, X_n$; I is a random interval.)

× **Example:** $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ iid, σ^2 known, μ unknown.

We want to construct a confidence interval for μ.

Consider the point estimate $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n X_i = \bar{X}$. Then

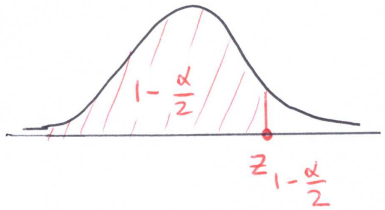
$\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$, and $\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$.

(A PIVOTAL STATISTICS; i.e. a statistics whose distribution does not depend on any parameters.)

For $0<\alpha<1$, we can find $z_{1-\alpha/2}$ such that

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

($z_{1-\alpha/2}$: quantile of the $\mathcal{N}(0,1)$ distribution)



Ex: for • $\alpha = 0.05$, $z_{1-\frac{\alpha}{2}} \simeq 1.96$
• $\alpha = 0.1$, $z_{1-\frac{\alpha}{2}} \simeq 1.645$

Re-arranging the terms gives

$$\mathbb{P}\left(\bar{X} - z_{1-\frac{\alpha}{2}} \sigma n^{-1/2} \leq \mu \leq \bar{X} + z_{1-\alpha/2} \sigma n^{-1/2} \right) = 1-\alpha$$

$\Rightarrow$ The interval $\boxed{I = \bar{X} \pm z_{1-\frac{\alpha}{2}} \sigma n^{-1/2}}$ is a $(1-\alpha)$ C.I. for μ.

The interval is centered around the point estimate.

Note that the length of I decreases to 0 as $n \to \infty$: at a given confidence level, the more observations you collect, the smaller the confidence interval.

→ What if σ is unknown?

Recall: $T := \dfrac{Z}{\sqrt{U/k}}$ with $\left.\begin{matrix} Z \sim \mathcal{N}(0,1) \\ U \sim \chi^2(k) \end{matrix}\right\}$ independent

has a Student distribution with k degrees of freedom $T \sim t(k)$.

Result: Let $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ iid. Then

(i) $\bar{X}$ and $S^2 = \dfrac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$ are independent

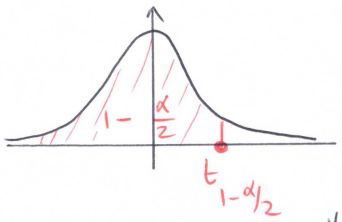
(ii) $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$

We obtain

$$T = \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2} \Big/ (n-1)}} = \frac{n^{1/2}(\bar{X}-\mu)}{S} \sim t(n-1).$$

↑ A pivotal statistics.

Let $t^{n-1}_{1-\alpha/2}$ be the $(1-\frac{\alpha}{2})$-quantile of the $t(n-1)$ distribution



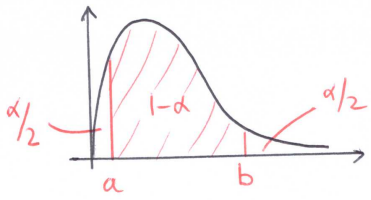
$$\mathbb{P}(T \leq t_{1-\alpha/2}) = 1-\alpha/2, \quad T \sim t(n-1)$$

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

The interval $\boxed{J = \bar{X} \pm t_{1-\alpha/2} S n^{-1/2}}$ is an (exact) $(1-\alpha)$ confidence interval for μ.

→ Note that we can use the distribution of $\frac{(n-1)S^2}{\sigma^2}$ to construct a confidence interval for σ^2: select a and b such that $\mathbb{P}\left(a \leq \frac{(n-1)S^2}{\sigma^2} \leq b \right) = 1-\alpha$. A possible choice is:

$\frac{S}{\sigma^2} \sim \chi^2(n-1)$



Re-arranging terms, we get that $\left[\frac{(n-1)S^2}{b}, \frac{(n-1)S^2}{a} \right]$ is a $(1-\alpha)$ C.I. for σ^2. ∎

Remark: The construction of a confidence interval relies on a pivotal statistics. If little is know about its distribution, or if its distribution depends on the parameter of interest we relax the original definition, requiring only a lower bound on the confidence level, and we construct I such that

$$\mathbb{P}_\theta(\theta \in I) \geqslant 1-\alpha.$$

For example, suppose that $X_1, \dots, X_n$ are iid, taking values in $[a, b]$ (known). Consider $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$ as a point estimate of the mean μ of X_1. We construct a CI for μ, with coverage at least $1-\alpha$, for some $\alpha \in (0,1)$. Chebyshev inequality implies that $\mathbb{P}(|\overline{X} - \mu| \geqslant \varepsilon) \leqslant \frac{\mathrm{Var}\,\overline{X}}{\varepsilon^2}$.

In other words, $\mu \in [\overline{X} - \varepsilon, \overline{X} + \varepsilon]$ with probability $\geqslant 1 - \frac{\mathrm{Var}\,\overline{X}}{\varepsilon^2}$.

Putting $\alpha = \frac{\mathrm{Var}\,\overline{X}}{\varepsilon^2}$, we obtain that

$$\mu \in \left[\overline{X} - \sqrt{\frac{\mathrm{Var}\,\overline{X}}{\alpha}},\ \overline{X} + \sqrt{\frac{\mathrm{Var}\,\overline{X}}{\alpha}}\right] \text{ with proba } \geqslant 1-\alpha$$

↖ To be a confidence interval, the bounds of the interval should not depend on the parameter(s) of the distribution. Recall that for bounded RVs taking values in $[a, b]$, we have that $\mathrm{Var}\,X_1 \leqslant \frac{(b-a)^2}{4}$ (Popoviciu ineq.)

Thus $\mathrm{Var}\,\overline{X} \leqslant \frac{(b-a)^2}{4n}$.

And we obtain (a, b are assumed to be known.)

$$\mu \in \left[\overline{X} - \frac{b-a}{\sqrt{2n\alpha}},\ \overline{X} + \frac{b-a}{\sqrt{2n\alpha}}\right] \text{ w.p. } \geqslant 1-\alpha \quad \text{—— (1)}$$

The interval does not depend on the shape of the distribution of the X_is, and relies on the crude Chebyshev bound. We can obtain a better interval if using Hoeffding inequality:

For independent RVs $X_1, \dots, X_n \in [a, b]$, we have

$$\mathbb{P}\left(|S_n - \mathbb{E}S_n| \geqslant \varepsilon\right) \leqslant 2\exp\left\{-\frac{2\varepsilon^2}{n(b-a)^2}\right\},$$

(For a proof of this result, see chapter SL: VAPNIK CHERVONENKIS THEORY)

where $S_n = \sum_{i=1}^n X_i$.

$$\mathbb{P}(|\overline{X} - \mu| \geqslant \varepsilon) = \mathbb{P}(|S_n - \mathbb{E}S_n| \geqslant n\varepsilon) \leqslant 2\exp\left\{-\frac{2n\varepsilon^2}{(b-a)^2}\right\} =: \alpha,$$

so that

$$\mu \in \left[\overline{X} - (b-a)\sqrt{\frac{1}{2n}\log\frac{2}{\alpha}},\ \overline{X} + (b-a)\sqrt{\frac{1}{2n}\log\frac{2}{\alpha}}\right]$$

with proba $\geqslant 1-\alpha$ —— (2)

Compare intervals (1) and (2): for $\alpha \in (0,1)$, $\frac{1}{\sqrt{\alpha}} > \sqrt{\log\frac{2}{\alpha}}$, so the length of interval (2) is smaller.

⇒ The length of a confidence interval at a given confidence level is a key property: one should aim at constructing the shortest possible CIs.

II.2. Asymptotic Confidence Intervals

If little is known about the distribution of the pivotal statistic (or if its law is not tractable), a second option to construct confidence intervals is to rely on asymptotic properties.

Definition. Let $\alpha \in (0,1)$. An ASYMPTOTIC CONFIDENCE INTERVAL for θ of level $(1-\alpha)$ is a statistic I_n taking values in the intervals of $\mathbb{R}$, and such that for any $\theta \in \Theta$,

$$\mathbb{P}_\theta(\theta \in I_n) \xrightarrow[n\to+\infty]{} 1-\alpha .$$

Useful in particular when a CLT-type of result applies to some estimator $\hat{\theta}$ of θ. Indeed, assuming that one has $n^{1/2}(\hat{\theta}-\theta) \xrightarrow{d} \mathcal{N}(0,\sigma^2(\theta))$, then $Z_n := \frac{n^{1/2}(\hat{\theta}-\theta)}{\sigma(\theta)} \xrightarrow{d} \mathcal{N}(0,1)$ is asymptotically pivotal, and one can use the asymptotic normal distribution to construct a CI for θ (provided one has a consistent estimator for $\sigma(\theta)$).

Example: Asymptotic CIs for a Binomial proportion.

Let $X_1, \dots, X_n \sim B(p)$ iid, and consider the following empirical estimator of p, given by $\hat{p} = \frac{1}{n}\sum_{i=1}^n X_i = \bar{X}$.

A consistent estimator of p since the SLLN ensures that $\hat{p} \xrightarrow{a.s.} \mathbb{E}X_1 = p$, as $n \to +\infty$.

$\hat{p} \in K_0$ since $\mathbb{E}\hat{p} = p$, and $\mathrm{Var}\,\hat{p} = \frac{p(1-p)}{n}$.

$\hat{p}$ = sample mean $\Rightarrow$ CLT applies and

$$\frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \xrightarrow{d} \mathcal{N}(0,1).$$

The variance of $\hat{p}$ depends on the unknown parameter p. We replace p by its empirical estimate $\hat{p}$. An application of the CLT + Slutsky theorem ensures that

$$\frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}} \xrightarrow{d} \mathcal{N}(0,1).$$

Recall that if $X_n \xrightarrow{d} X$, $Y_n \xrightarrow{a} c$, then $X_n Y_n \xrightarrow{d} cX$.

Select $z_{1-\alpha/2}$ such that for $Z \sim \mathcal{N}(0,1)$, $\mathbb{P}(|Z| \le z_{1-\frac{\alpha}{2}}) = 1-\alpha$, and consider

$$\mathbb{P}\left(-z_{1-\frac{\alpha}{2}} \le Z \le z_{1-\alpha/2}\right) \approx \mathbb{P}\left(-z_{1-\alpha/2} \le \frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}} \le z_{1-\frac{\alpha}{2}}\right).$$

The random interval

$$\left[\hat{p} - z_{1-\frac{\alpha}{2}} n^{-1/2}\sqrt{\hat{p}(1-\hat{p})}\ ,\ \hat{p} + z_{1-\frac{\alpha}{2}} n^{-1/2}\sqrt{\hat{p}(1-\hat{p})}\right] =: I_W$$

is an asymptotic $(1-\alpha)$ CI for p.

This interval is known as the WALD INTERVAL, and is of nearly universal use. However, as we now explain, its properties are rather poor: even for relatively large values of n, the Wald interval can be quite far from the nominal $(1-\alpha)$ level, and its properties vary greatly from one value of p to another, at fixed n. Its poor coverage properties are a direct consequence of the normal approximation we used.

To capture this precisely, we need to make use of EDGEWORTH EXPANSIONS, which are refinements of the Central limit theorem. Edgeworth expansions quantify the error made when approximating the distribution of some quantity with its asymptotic normal distribution. We have:

$$\mathbb{P}\left(\frac{n^{1/2}(\hat{p}-p)}{\sqrt{p(1-p)}} \leqslant x\right) = \Phi(x) + \frac{1}{6}(1-2p)(1-x^2)\phi(x)(npq)^{-1/2} + \left(\frac{1}{2} - g(p,x)\right)\phi(x)(npq)^{-1/2} + O(n^{-1})$$

- $\Phi(x)$: The asymptotic normal distribution
- $\frac{1}{6}(1-2p)(1-x^2)\phi(x)(npq)^{-1/2}$: → skewness error: $n^{-1/2}$ term that corrects for the lack of skewness
- $\left(\frac{1}{2} - g(p,x)\right)\phi(x)(npq)^{-1/2}$: → rounding error: the price to pay for approximating a discrete distribution with a continuous one

This rounding error is responsible for the poor coverage properties of the Wald interval: if the original distribution had a density, the rounding error would disappear from the Edgeworth expansion.

[Ref] Theorem 23.1 in Bhattacharya & Rao (1976). Normal Approximation and Asymptotic Expansions. Wiley, New York.

In the expansion above, $g(p,x) = h(np + z(npx)^{-1/2})$,
where $h(u) = u - \lfloor u \rfloor$ = fractional part of u.
= highly oscillating function.

Put $z_n := \dfrac{n^{1/2}(\hat{p}-p)}{\sqrt{\hat{p}(1-\hat{p})}}$ and $w_n := \dfrac{n^{1/2}(\hat{p}-p)}{\sqrt{p(1-p)}}$

By definition,

$$I_W = \left\{ p \in [0,1] \;\middle|\; -z_{1-\frac{\alpha}{2}} \leqslant z_n \leqslant z_{1-\frac{\alpha}{2}} \right\}.$$

Note that

$$z_n = \frac{w_n}{\sqrt{1 + (1-2p)\frac{w_n}{\sqrt{npq}} - \frac{w_n^2}{n}}},$$

($q := 1-p$)

so that

$$I_W = \left\{ p \in [0,1] \;\middle|\; -z_{1-\frac{\alpha}{2}} \leqslant \frac{w_n}{\sqrt{1 + (1-2p)\frac{w_n}{\sqrt{npq}} - \frac{w_n^2}{n}}} \leqslant z_{1-\frac{\alpha}{2}} \right\},$$

and re-arranging terms, we get

$$I_W = \left\{ p \in [0,1] \;\middle|\; \ell_W \leqslant w_n \leqslant u_W \right\},$$

(ℓ_W: "lower", u_W: "upper")

with

$$(\ell_W, u_W) = \frac{z\left(\frac{1}{2}-p\right)n^{1/2} \pm z\sqrt{n\left(\frac{z^2}{4n} + pq\right)}}{(z^2+n)\sqrt{pq}}$$

(writing z in place of $z_{1-\frac{\alpha}{2}}$ for simplicity.)

This expression of the Wald interval will allow us to use the EE on the previous page to deduce its coverage properties. To do so, we are required to obtain expansions for (ℓ_W, u_W), and for $\Phi(\ell_W)$, $\phi(\ell_W)$, $\Phi(u_W)$, $\phi(u_W)$. After calculations, we obtain:

$$(\ell_W, u_W) = \pm\lambda_1 + \lambda_2 n^{-1/2} \pm \lambda_3 n^{-1} + O(n^{-3/2}),$$

with $\lambda_1 = z$, $\lambda_2 = \dfrac{z^2(\frac{1}{2}-p)}{\sqrt{pq}}$, $\lambda_3 = z^3\left(\dfrac{1}{8pq}-1\right)$

Moreover,

$$\Phi(u_W) = \Phi(z) + \lambda_2 \phi(z) n^{-1/2} + O(n^{-1})$$
$$\Phi(\ell_W) = \Phi(-z) + \lambda_2 \phi(z) n^{-1/2} + O(n^{-1}),$$

and

$$\phi(u_W) = \phi(z) - z\,\lambda_2 \phi(z) n^{-1/2} + O(n^{-1})$$
$$\phi(\ell_W) = \phi(z) + z\,\lambda_2 \phi(z) n^{-1/2} + O(n^{-1}).$$

We finally obtain:

$$\mathbb{P}(p \in I_W) = \mathbb{P}\left(z_n \in [-z_{1-\frac{\alpha}{2}}, z_{1-\frac{\alpha}{2}}]\right)$$
$$= \mathbb{P}\left(W_n \in [\ell_W, u_W]\right)$$
$$= \Phi(u_W) + \left\{\left(\tfrac{1}{2} - g(p,u_W)\right) + \tfrac{1}{6}(1-2p)(1-u_W^2)\right\} \times \phi(u_W)(npq)^{-1/2}$$
$$- \Phi(\ell_W) - \left\{\left(\tfrac{1}{2} - g(p,\ell_W)\right) + \tfrac{1}{6}(1-2p)(1-\ell_W^2)\right\} \times \phi(\ell_W)(npq)^{-1/2}$$

$$= \Phi(z) + \lambda_2 \phi(z) n^{-1/2} - \Phi(-z) - \lambda_2 \phi(z) n^{-1/2} + O(n^{-1})$$
$$= (1-\alpha) + O(n^{-1})$$

The $O(n^{-1/2})$ skewness terms cancel each other.

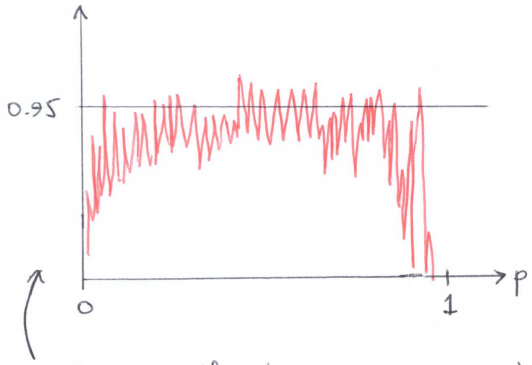
⇒ It remains the $O(n^{-1/2})$ terms corresponding to the rounding error:

$$\mathbb{P}(p \in I_W) = (1-\alpha) + \left\{g(p,\ell_W) - g(p,u_W)\right\}\phi(z)(npq)^{-1/2} + O(n^{-1})$$

nominal asymptotic level — persistent oscillations

The persistent $O(n^{-1/2})$ oscillations appearing in the expansion of the coverage probability of the Wald interval are responsible for its poor properties. Note that if the original distribution was absolutely continuous instead of discrete, this rounding error term would disappear, and the coverage probability would be of order n^{-1}, improving greatly on its coverage properties.

For a fixed n, the function $(1-\alpha) + \{g(p,\ell_W) - g(p,u_W)\} \times \phi(z)(npq)^{-1/2}$ plotted as a function of p typically looks like:



The approximation of the true coverage probability can be improved by considering second order EE.

[Ref] L.D. Brown, T.T. Cai & A. DasGupta. (2002). Confidence Intervals for a Binomial Proportion and Asymptotic Expansions. The Annals of Statistics, Vol 30, No1, p. 160-201.

This study highlights one important fact: when constructing asymptotic confidence intervals, there are two properties one should have in mind: the coverage properties of the interval, and its length.

x Many alternatives to the Wald interval exist. The most popular alternative is the Wilson interval:

$$I_{\text{wilson}} := \left\{ p \in [0,1] \;\middle|\; -z_{1-\frac{\alpha}{2}} \leq W_n \leq z_{1-\frac{\alpha}{2}} \right\},$$

which has the explicit form

Compare its definition with the Wald interval, on page 32.

$$I_{\text{wilson}} = \frac{n\hat{p} + z^2/2}{n + z^2} \pm \frac{z n^{1/2}}{n + z^2} \left(\hat{p}(1-\hat{p}) + \frac{z^2}{4n} \right)^{1/2}.$$

This interval is not centered around the point estimate $\hat{p}$, and is known to have better coverage properties than the Wald interval (despite the fact that, obviously, its coverage probability still contains oscillating terms).

x Other intervals include: Agresti-Coull, likelihood ratio, Jeffreys, ... See Brown et al (2002) for their definitions, and for a theoretical comparison of their properties.