Law of large numbers (LLN)

The average converges towards the expectation.

CLT

Xn_bar, the average of Xis tends to a standard normal (as n tends to infinity):

- E[X] for its mean
- Var(X)/n for its variance

Central limit theorem (CLT):

$$\sqrt{n}\,\frac{\bar{X}_n - \mu}{\sigma} \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1).$$

(Equivalently, $\sqrt{n}\,(\bar{X}_n - \mu) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,\sigma^2)$.)

Another useful tool: Hoeffding's inequality

What if n is not large enough to apply CLT?

Theorem (Hoeffding, 1963)
Let n be a positive integer and $X, X_1, \ldots, X_n$ be i.i.d. r.v. such that $\mu = \mathbb{E}[X]$ and

$X \in [a, b]$ almost surely ($a < b$ are given numbers)

Then,

$$\mathbb{P}[|\bar{X}_n - \mu| \geq \varepsilon] \leq 2e^{-\frac{2n\varepsilon^2}{(b-a)^2}}. \quad \forall \varepsilon > 0$$

This holds even for small sample sizes n.

$X_i \overset{iid}{\sim} Ber(p)$ $\qquad$ $\mathbb{P}\left(|\bar{X}_n - \mu| \geq \frac{c}{\sqrt{n}}\right) \leq 2e^{-2c^2}$

Addition, multiplication, division

. . . only for a.s. and $\mathbb{P}$. . .
Assume

$$T_n \xrightarrow[n\to\infty]{\text{a.s./}\mathbb{P}} T \quad \text{and} \quad U_n \xrightarrow[n\to\infty]{\text{a.s./}\mathbb{P}} U$$

Then,

- $T_n + U_n \xrightarrow[n\to\infty]{\text{a.s./}\mathbb{P}} T + U$,
- $T_n U_n \xrightarrow[n\to\infty]{\text{a.s./}\mathbb{P}} TU$,
- If in addition, $U \neq 0$ a.s., then $\frac{T_n}{U_n} \xrightarrow[n\to\infty]{\text{a.s./}\mathbb{P}} \frac{T}{U}$.

In general, these rules **do not** apply to convergence (d).

Slutsky's theorem

Some partial results exist for convergence in distribution on the form of *Slutsky's theorem*.

Let $(X_n), (Y_n)$ be two sequences of r.v., such that:

$$\text{(i) } T_n \xrightarrow[n\to\infty]{(d)} T \qquad \text{and} \qquad \text{(ii) } U_n \xrightarrow[n\to\infty]{\mathbb{P}} u$$

where T is a r.v. and u is a given real number (deterministic limit: $\mathbb{P}(U = u) = 1$). Then,

- $T_n + U_n \xrightarrow[n\to\infty]{(d)} T + u$,
- $T_n U_n \xrightarrow[n\to\infty]{(d)} Tu$,
- If in addition, $u \neq 0$, then $\dfrac{T_n}{U_n} \xrightarrow[n\to\infty]{(d)} \dfrac{T}{u}$.

Continuous mapping theorem

$$X_n \xrightarrow{\mathrm{d}} X \quad \Rightarrow \quad g(X_n) \xrightarrow{\mathrm{d}} g(X);$$

$$X_n \xrightarrow{\mathrm{p}} X \quad \Rightarrow \quad g(X_n) \xrightarrow{\mathrm{p}} g(X);$$

$$X_n \xrightarrow{\mathrm{a.s.}} X \quad \Rightarrow \quad g(X_n) \xrightarrow{\mathrm{a.s.}} g(X).$$

Taking functions

Continuous functions (for all three types) . If f is a continuous function:

$$T_n \xrightarrow[n\to\infty]{\text{a.s./}\mathbb{P}/(d)} T \quad \Rightarrow \quad f(T_n) \xrightarrow[n\to\infty]{\text{a.s/P/d}} f(T).$$

Continuous Mapping Theorem

Example: Recall that by LLN, $\bar{R}_n \xrightarrow[n\to\infty]{\mathbb{P},\text{ a.s.}} p$. Therefore

$$f(\bar{R}_n) \xrightarrow[n\to\infty]{\mathbb{P},\text{ a.s.}} f(p) \quad \text{for any continuous } f$$

(Only need f to be continuous around p: f(x)=1/x works if $p > 0$)

We also have by CLT: $\sqrt{n}\dfrac{\bar{R}_n - p}{\sqrt{p(1-p)}} \xrightarrow[n\to\infty]{(d)} Z$, $Z \sim \mathcal{N}(0,1)$. So

$$f\left(\sqrt{n}\,(\bar{R}_n - p)\right) \xrightarrow[n\to\infty]{(d)} f(Y) \quad Y \sim \mathcal{N}(0, p(1-p))$$

⚠ not the limit of $\sqrt{n}[f(\bar{R}_n) - f(p)]$!!

Delta (Δ) method

Expectation

$$\mathbf{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx$$

Variance

Définition

$$\operatorname{Var}(X) = \operatorname{E}\left[(X - \mu)^2\right].$$

This definition encompasses random variables that are ge
itself:

$$\operatorname{Var}(X) = \operatorname{Cov}(X, X).$$

The variance is also equivalent to the second cumulant of
expression for the variance can be expanded:

$$\begin{aligned}
\operatorname{Var}(X) &= \operatorname{E}\left[(X - \operatorname{E}[X])^2\right] \\
&= \operatorname{E}\left[X^2 - 2X\operatorname{E}[X] + \operatorname{E}[X]^2\right] \\
&= \operatorname{E}\left[X^2\right] - 2\operatorname{E}[X]\operatorname{E}[X] + \operatorname{E}[X]^2 \\
&= \operatorname{E}\left[X^2\right] - \operatorname{E}[X]^2
\end{aligned}$$

Propriétés

$$\operatorname{Var}(aX + bY) = a^2\operatorname{Var}(X) + b^2\operatorname{Var}(Y) + 2ab\operatorname{Cov}(X, Y),$$
$$\operatorname{Var}(aX - bY) = a^2\operatorname{Var}(X) + b^2\operatorname{Var}(Y) - 2ab\operatorname{Cov}(X, Y),$$

Variance of Xn, average of Xi

Var = Var Xi/n

Covariance

$$\operatorname{cov}(X,Y) = \operatorname{E}\left[(X - \operatorname{E}[X])(Y - \operatorname{E}[Y])\right], \quad \textit{(Eq.1)}$$

where $\operatorname{E}[X]$ is the expected value of X, also known as the mean of X. T
can be simplified to the expected value of their product minus the produc

$$\begin{aligned}
\operatorname{cov}(X,Y) &= \operatorname{E}[(X - \operatorname{E}[X])\,(Y - \operatorname{E}[Y])] \\
&= \operatorname{E}[XY - X\operatorname{E}[Y] - \operatorname{E}[X]Y + \operatorname{E}[X]\operatorname{E}[Y]] \\
&= \operatorname{E}[XY] - \operatorname{E}[X]\operatorname{E}[Y] - \operatorname{E}[X]\operatorname{E}[Y] + \operatorname{E}[X]\operatorname{E}[Y] \\
&= \operatorname{E}[XY] - \operatorname{E}[X]\operatorname{E}[Y].
\end{aligned}$$

$$\begin{aligned}
\operatorname{cov}(X,a) &= 0 \\
\operatorname{cov}(X,X) &= \operatorname{var}(X) \\
\operatorname{cov}(X,Y) &= \operatorname{cov}(Y,X) \\
\operatorname{cov}(aX,bY) &= ab\,\operatorname{cov}(X,Y) \\
\operatorname{cov}(X+a,Y+b) &= \operatorname{cov}(X,Y) \\
\operatorname{cov}(aX+bY,cW+dV) &= ac\,\operatorname{cov}(X,W) + ad\,\operatorname{cov}(X,V) + bc\,\operatorname{cov}(Y,W) + bd\,\operatorname{cov}(Y,V)
\end{aligned}$$

For a sequence $X_1, \ldots, X_n$ of random variables in real-valued, and constants $a_1, \ldots, a_n$, we have

$$\sigma^2\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2\sigma^2(X_i) + 2\sum_{i,j\,:\,i<j} a_i a_j \operatorname{cov}(X_i, X_j) = \sum_{i,j} a_i a_j \operatorname{cov}(X_i, X_j)$$

Hoeffding's Covariance Identity [edit]

A useful identity to compute the covariance between two random variables X, Y is the Hoeffding's Covariance Identity:[7]

$$\operatorname{cov}(X,Y) = \int_{\mathbb{R}}\int_{\mathbb{R}} xy(F_{(X,Y)}(x,y) - F_X(x)F_Y(y))\,dx\,dy$$

where $F_{(X,Y)}(x,y)$ is the joint distribution function of the random vector (X,Y) and $F_X(x), F_Y(y)$ are the marginals.

Joint distributions

f(x,y) = f(y|x).f(x)

f(x,y) = f(x|y).f(y)

Conditional distribution

f(y|x) = f(x,y) / f(x)

f(x|y) = f(x,y) / f(y)

Marginal distributions

Marginal distribution f(x) = integrale over y de f(x,y)dy

Marginal distribution f(x) = Ey [f(x|y)]

Bayes' theroem

P(B|A) = P(A|B)* P(B)/ P(A)

Statistical model

- Space of x does not depend of parameter
- Space of normal is R, etc
- Proba integrates to 1

- Usually, we will assume that the statistical model is *well specified*, i.e., defined such that $\exists \theta$ such that $P = P_\theta$

Identifiability

The parameter θ is called *identifiable* iff the map $\theta \in \Theta \mapsto \mathbb{P}_\theta$ is injective, i.e.,

$$\theta \neq \theta' \Rightarrow \mathbb{P}_\theta \neq \mathbb{P}_{\theta'}$$

or equivalently:

$$\mathbb{P}_\theta = \mathbb{P}_{\theta'} \Rightarrow \theta = \theta'$$

ESTIMATION

Estimators

The estimator is expressed in function of KNOWN variables (and not of true parameters…)

- An estimator $\hat{\theta}_n$ of θ is *asymptotically normal* if

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

The quantity σ^2 is then called *asymptotic* variance of $\hat{\theta}_n$.

Jensen's inequality when applying continuous mapping theorem to estimators (-> biaised estimator):

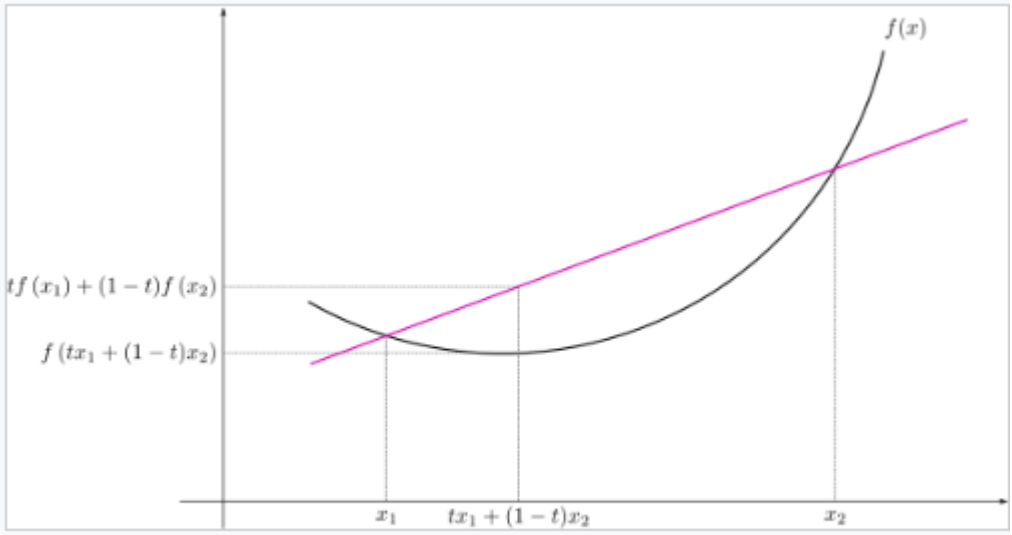
Estimator

- Density of T_1:
$$f(t) = \lambda e^{-\lambda t}, \quad \forall t \geq 0.$$
- $\mathbb{E}[T_1] = \frac{1}{\lambda}$.
- Hence, a natural estimate of $\frac{1}{\lambda}$ is
$$\bar{T}_n := \frac{1}{n}\sum_{i=1}^{n} T_i.$$
- A natural estimator of λ is
$$\hat{\lambda} := \frac{1}{\bar{T}_n} \xrightarrow[n\to\infty]{a.s.,\,\mathbb{P}} \lambda$$

$\mathbb{E}\left[\frac{1}{\bar{T}_n}\right] > \frac{1}{\mathbb{E}[\bar{T}_n]} = \lambda$

$\lambda > 0$

Jensen's inequality



Jensen's inequality generalizes the statement that a secant line of a convex function lies above the graph.

In the context of probability theory, it is generally stated in the following form: if X is a random variable and φ is a convex function, then

$\varphi(\mathbf{E}[X]) \leq \mathbf{E}[\varphi(X)]$.

Quadratic risk

- We want estimators to have low bias and low variance at the same time.
- The *Risk* (or *quadratic risk*) of an estimator $\hat{\theta}_n \in \mathbb{R}$ is

$$R(\hat{\theta}_n) = \mathbb{E}\left[|\hat{\theta}_n - \theta|^2\right]$$

$\mathbb{E}[(\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n] + \mathbb{E}[\hat{\theta}_n] - \theta)^2] = \mathbb{E}[(\hat{\theta}_n - \mathbb{E}(\hat{\theta}_n))^2] + \mathbb{E}[(\mathbb{E}(\hat{\theta}_n) - \theta)^2]$ $+ 2\,\mathbb{E}[(\hat{\theta}_n - \mathbb{E}(\hat{\theta}_n))(\mathbb{E}(\hat{\theta}_n) - \theta)]$

Var $(\hat{\theta}_n)$ — bias $(\hat{\theta}_n)^2$

- Low quadratic risk means that both bias and variance are small:

  quadratic risk= VARIANCE + BIAS^2

Confidence interval

- Define estimator of parameter of interest theta
- Define estimator distribution (using CLT for example)
- Apply constraint P(theta in I) = 1- alpha on the known distribution

- Use constraint on distribution P(theta in I) = 1- alpha to get I

Combining the first two questions, by setting

$$q = \Phi^{-1}\left(0.975\right) = 1.96,$$

we see that

$$\mathbf{P}\left(\sqrt{\frac{n}{\lambda}}\left(\overline{X}_n - \lambda\right) \in [-q, q]\right) \to \mathbf{P}\left(Z \in [-q, q]\right) = 2\Phi\left(q\right) - 1 = 2 \times 0.975 - 1 = 0.95.$$

Hence, we have

$$\mathcal{I}_\lambda := \left[\overline{X}_n - 1.96\sqrt{\frac{\lambda}{n}}, \overline{X}_n + 1.96\sqrt{\frac{\lambda}{n}}\right],$$

$$\sqrt{\frac{n}{\theta}}\left(\overline{X}_n - \theta\right) \sim \mathcal{N}\left(0, 1\right),$$

so together with looking up the quantile value for a symmetric 90% confidence interval for a Gaussian random variable $Z \sim \mathcal{N}\left(0, 1\right)$,

$$\mathbf{P}\left(|Z| \leq 1.6448\right) \approx 0.9,$$

we obtain

$$\mathbf{P}\left(\left|\sqrt{\frac{n}{\theta}}\left(\overline{X}_n - \theta\right)\right| \leq 1.6448\right) = 0.9,$$

and hence can set

$$\mathcal{I}_1 = \left[\overline{X}_n - \frac{1.6448\sqrt{\theta}}{\sqrt{n}}, \overline{X}_n + \frac{1.6448\sqrt{\theta}}{\sqrt{n}}\right].$$

Confidence interval?

- For a fixed $\alpha \in (0,1)$, if $q_{\alpha/2}$ is the $(1-\alpha/2)$-quantile of $\mathcal{N}(0,1)$, then with probability $\simeq 1-\alpha$ (if n is large enough !),

$$\bar{R}_n \in \left[p - \frac{q_{\alpha/2}\sqrt{p(1-p)}}{\sqrt{n}}, p + \frac{q_{\alpha/2}\sqrt{p(1-p)}}{\sqrt{n}} \right].$$

- It yields

$$\lim_{n\to\infty} \mathbb{P}\left(\left[\bar{R}_n - \frac{q_{\alpha/2}\sqrt{p(1-p)}}{\sqrt{n}}, \bar{R}_n + \frac{q_{\alpha/2}\sqrt{p(1-p)}}{\sqrt{n}} \right] \ni p \right) = 1-\alpha$$

- But this is **not** a confidence interval because it depends on p!

- To fix this, there are 3 solutions.

Solution 1: Conservative bound

- Note that no matter the (unknown) value of p,

$$p(1-p) \leq \frac{1}{4}$$

- Hence, roughly with probability at least $1-\alpha$,

$$\bar{R}_n \in \left[p - \frac{q_{\alpha/2}}{2\sqrt{n}}, p + \frac{q_{\alpha/2}}{2\sqrt{n}} \right].$$

- We get the asymptotic confidence interval:

$$\mathcal{I}_{\text{conserv}} = \left[\bar{R}_n - \frac{q_{\alpha/2}}{2\sqrt{n}}, \bar{R}_n + \frac{q_{\alpha/2}}{2\sqrt{n}} \right]$$

- Indeed

$$\lim_{n \to \infty} \mathbb{P}(\mathcal{I}_{\text{conserv}} \ni p) \geq 1 - \alpha$$

Solution 2: Solving the (quadratic) equation for p

- We have the system of two inequalities in p:

$$\bar{R}_n - \frac{q_{\alpha/2}\sqrt{p(1-p)}}{\sqrt{n}} \leq p \leq \bar{R}_n + \frac{q_{\alpha/2}\sqrt{p(1-p)}}{\sqrt{n}}$$

- Each is a quadratic inequality in p of the form

$$(p - \bar{R}_n)^2 \leq \frac{q_{\alpha/2}^2 p(1-p)}{n}$$

We need to find the roots $p_1 < p_2$ of

$$(1 + \frac{q_{\alpha/2}^2}{n})p^2 - (2\bar{R}_n + \frac{q_{\alpha/2}^2}{n})p + \bar{R}_n^2 = 0$$

- This leads to a new confidence interval $\mathcal{I}_{\text{solve}} = [p_1, p_2]$ such that:

$$\lim_{n\to\infty} \mathbb{P}(\mathcal{I}_{\text{solve}} \ni p) = 1-\alpha$$

(it's complicated to write in generic way so let us wait to have values for n, α and $\bar{R}_n$ to plug-in)

Solution 3: plug-in

- Recall that by the LLN $\hat{p} = \bar{R}_n \xrightarrow[n\to\infty]{\mathbb{P},\text{a.s.}} p$
- So by Slutsky, we also have

$$\sqrt{n}\frac{\bar{R}_n - p}{\sqrt{\hat{p}(1-\hat{p})}} \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1)$$

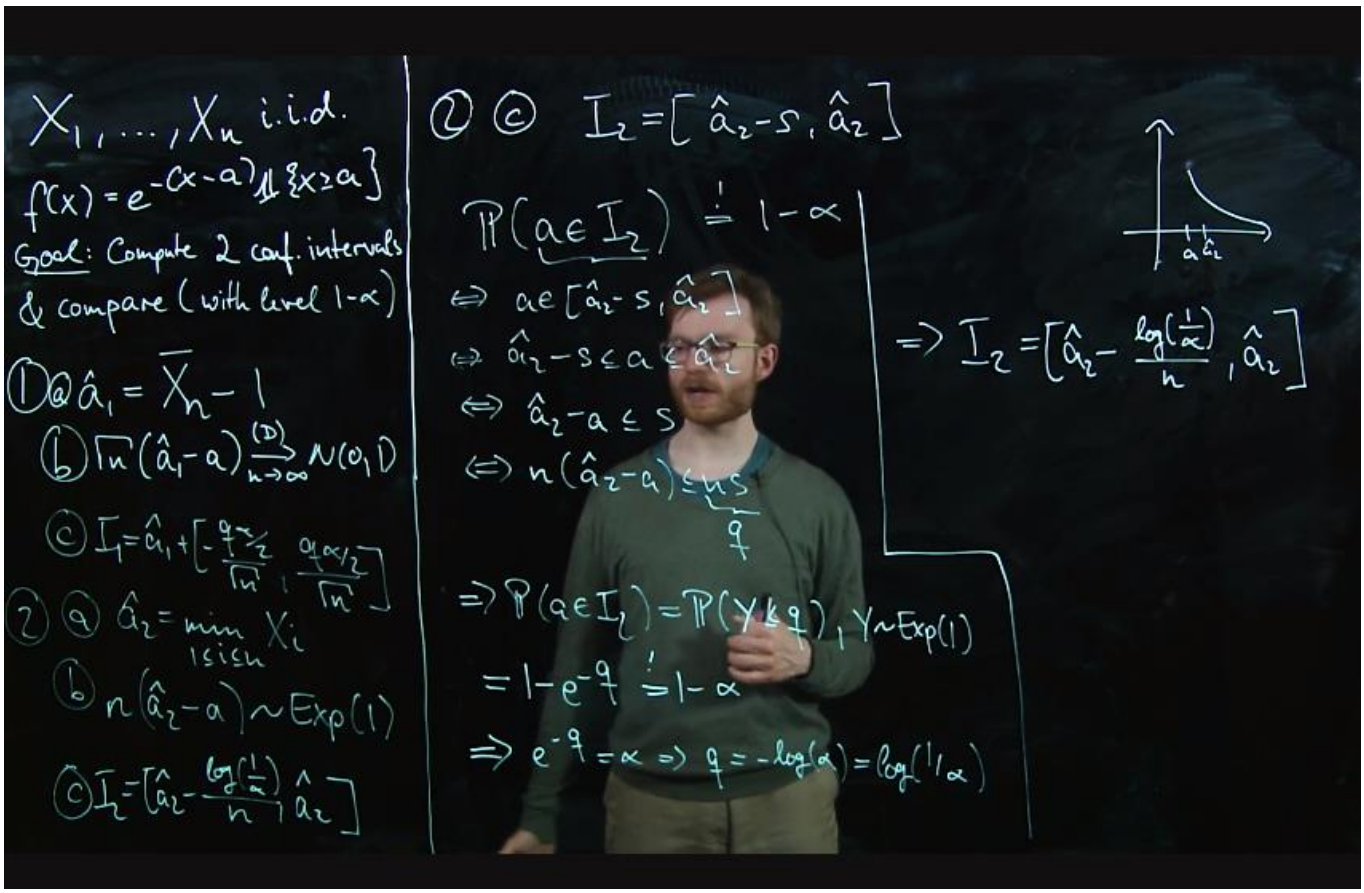
- This leads to a new confidence interval:

$$\mathcal{I}_{\text{plug-in}} = \left[\bar{R}_n - \frac{q_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}, \bar{R}_n + \frac{q_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}\right]$$

such that

$$\lim_{n\to\infty} \mathbb{P}(\mathcal{I}_{\text{plug-in}} \ni p) = 1-\alpha$$

CI


$X_1, \dots, X_n$ i.i.d.
$f(x) = e^{-(x-a)} \mathbb{1}_{\{x \geq a\}}$
Goal: Compute 2 conf. intervals & compare (with level $1-\alpha$)
① @ $\hat{a}_1 = \bar{X}_n - 1$
ⓑ $\sqrt{n}(\hat{a}_1 - a) \xrightarrow[n\to\infty]{(D)} N(0,1)$
ⓒ $I_1 = \hat{a}_1 + \left[-\frac{q_{\alpha/2}}{\sqrt{n}}, \frac{q_{\alpha/2}}{\sqrt{n}}\right]$
② @ $\hat{a}_2 = \min_{1 \leq i \leq n} X_i$
ⓑ $n(\hat{a}_2 - a) \sim Exp(1)$
ⓒ $I_2 = \left[\hat{a}_2 - \frac{\log(\frac{1}{\alpha})}{n}, \hat{a}_2\right]$
② ⓒ $I_2 = [\hat{a}_2 - s, \hat{a}_2]$
$\mathbb{P}(a \in I_2) \overset{!}{=} 1-\alpha$
$\Leftrightarrow a \in [\hat{a}_2 - s, \hat{a}_2]$
$\Leftrightarrow \hat{a}_2 - s \leq a \leq \hat{a}_2$
$\Leftrightarrow \hat{a}_2 - a \leq s$
$\Leftrightarrow n(\hat{a}_2 - a) \leq ns$
q
$\Rightarrow \mathbb{P}(a \in I_2) = \mathbb{P}(Y \leq q), Y \sim Exp(1)$
$= 1 - e^{-q} \overset{!}{=} 1-\alpha$
$\Rightarrow e^{-q} = \alpha \Rightarrow q = -\log(\alpha) = \log(1/\alpha)$
a $\hat{a}_2$
$\Rightarrow I_2 = \left[\hat{a}_2 - \frac{\log(\frac{1}{\alpha})}{n}, \hat{a}_2\right]$

Delta method

Applying g() to our RV whose convergence is known to get the RV of interest

The Delta method ← this is important

Let $(Z_n)_{n\geq 1}$ sequence of r.v. that satisfies

$$\sqrt{n}(Z_n - \theta) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, \sigma^2)$$

for some $\theta \in \mathbb{R}$ and $\sigma^2 > 0$ (the sequence $(Z_n)_{n\geq 1}$ is said to be *asymptotically normal around* θ).

Let $g : \mathbb{R} \to \mathbb{R}$ be continuously differentiable at the point θ. Then,

- $(g(Z_n))_{n\geq 1}$ is also asymptotically normal; around $g(\theta)$
- More precisely,

$$\sqrt{n}\,(g(Z_n) - g(\theta)) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, (g'(\theta))^2 \sigma^2).$$

DELTA METHOD ATTENTION variance of g(Z) = g'(theta)^2 * sigma^2

Multivariate Delta method

Let $(T_n)_{n\geq 1}$ sequence of random vectors in $\mathbb{R}^d$ such that

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

for some $\theta \in \mathbb{R}^d$ and some covariance $\Sigma \in \mathbb{R}^{d\times d}$.

Let $g : \mathbb{R}^d \to \mathbb{R}^k$ $(k \geq 1)$ be continuously differentiable at θ. Then,

$$\sqrt{n}\,(g(T_n) - g(\theta)) \xrightarrow[n\to\infty]{(d)} \mathcal{N}_k\big(0, \nabla g(\theta)^T \Sigma \nabla g(\theta)\big),$$

where $\nabla g(\theta) = \dfrac{\partial g}{\partial \theta}(\theta) = \left(\dfrac{\partial g_j}{\partial \theta_i}\right)_{\substack{1\leq i\leq d \\ 1\leq j\leq k}} \in \mathbb{R}^{d\times k}$.

Hypothesis testing

Errors

- *Rejection region* of a test ψ:

$$R_\psi = \{x \in E^n : \psi(x) = 1\}.$$

- *Type 1 error* of a test ψ (rejecting H_0 when it is actually true):

$$\begin{array}{rccc} \alpha_\psi & : & \Theta_0 & \to & \mathbb{R} \\ & & \theta & \mapsto & \mathbb{P}_\theta[\psi = 1]. \end{array}$$

- *Type 2 error* of a test ψ (not rejecting H_0 although H_1 is actually true):

$$\begin{array}{rccc} \beta_\psi & : & \Theta_1 & \to & \mathbb{R} \\ & & \theta & \mapsto & \end{array}$$

- *Power* of a test ψ:

$$\pi_\psi = \inf_{\theta \in \Theta_1} \left(1 - \beta_\psi(\theta)\right).$$

Level, test statistic and rejection region

- A test ψ has *level* α if

$$\alpha_\psi(\theta) \leq \alpha, \quad \forall \theta \in \Theta_0.$$

- A test ψ has *asymptotic level* α if

$$\lim_{n \to \infty} \alpha_\psi(\theta) \leq \alpha, \quad \forall \theta \in \Theta_0.$$

- In general, a test has the form

$$\psi = \mathbb{1}\{T_n > c\},$$

for some statistic T_n and threshold $c \in \mathbb{R}$.

- T_n is called the *test statistic*. The rejection region is $R_\psi =$

Bernoulli experiment

- Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Ber}(p)$, for some unknown $p \in (0,1)$.
- We want to test:

$$H_0\colon\ p = 1/2 \text{ vs. } H_1\colon\ p \neq 1/2$$

with asymptotic level $\alpha \in (0,1)$.

- Let $T_n = \left| \sqrt{n}\ \dfrac{\hat{p}_n - 0.5}{\sqrt{.5(1-.5)}} \right|$, where $\hat{p}_n$ is the MLE. $\bar{X}_n$
- If H_0 is true, then by CLT,

$$\mathbb{P}[T_n > q_{\alpha/2}] \xrightarrow[n\to\infty]{} 0.05$$

- Let $\psi_\alpha = \mathbb{1}\{T_n > q_{\alpha/2}\}$.

Tn is the test statistic

Define Tn :

- Define parameter estimator
- Define estimator distribution

- Define alpha wrt to its definition (proba under the null to reject the null)

- Tn of the form Tn > c or |Tn| > c (or < c if H1 is of the form < b)
- Tn in terms of the estimator corrected to have a known distribution, and Tn rejects the null hypothesis

- **Define test Tn under H0 or at the boundary, or replace true parameter by estimator (ATTENTION Tn not in function of the true parameter)**

p-value

Definition

The (asymptotic) *p-value* of a test ψ_α is the smallest (asymptotic) level α at which ψ_α rejects H_0. It is random, it depends on the sample.

Golden rule

p-value $\leq \alpha \quad \Leftrightarrow \quad H_0$ is rejected by ψ_α, at the (asymptotic) level α.

The smaller the p-value, the more confidently one can reject H_0.

p-value is a mass, related to a particular sample

p-value is the smallest alpha - alpha is the probability under the null that the null is rejected.

p-value is the mass for the realization of Tn (to the right, and maybe multiplied by 2 if it is a 2 sided test) ! Tn of the form Tn> c, and Tn comparable to Z for example

For one sided tests, p-value is taken with parameter = at the boundary of the parameter space between H0 and H1 (where Type I error is max).

Maximum likelihood estimator

MLE is the theta for which l'(theta)=0

Test first if l'' (resp Hessian) is negative (concave function, which means we are the MAX)

Matrix is concave if it's negative definite, ie:

- check eigenvalues are negative (in R: function EIGEN (mat) with mat square)

or

- check trace (sum of diagonal elements) is negative AND determinant is positive (in R DET(mat) with mat square)

Multivariate concave functions

More generally for a *multivariate* function: $h : \Theta \subset \mathbb{R}^d \to \mathbb{R}$, $d \geq 2$, define the

- *gradient* vector: $\nabla h(\theta) = \begin{pmatrix} \frac{\partial h}{\partial \theta_1}(\theta) \\ \vdots \\ \frac{\partial h}{\partial \theta_d}(\theta) \end{pmatrix} \in \mathbb{R}^d$
- *Hessian* matrix:

$$\mathbf{H}h(\theta) = \begin{pmatrix} \frac{\partial^2 h}{\partial \theta_1 \partial \theta_1}(\theta) & \cdots & \frac{\partial^2 h}{\partial \theta_1 \partial \theta_d}(\theta) \\ & & \\ \frac{\partial^2 h}{\partial \theta_d \partial \theta_1}(\theta) & \cdots & \frac{\partial^2 h}{\partial \theta_d \partial \theta_d}(\theta) \end{pmatrix} \in \mathbb{R}^{d \times d}$$

h is concave $\Leftrightarrow$ $x^\top \mathbf{H}h(\theta)x \leq 0 \quad \forall x \in \mathbb{R}^d,\ \theta \in \Theta.$

h is strictly concave $\Leftrightarrow$ $x^\top \mathbf{H}h(\theta)x < 0 \quad \forall x \in \mathbb{R}^d,\ \theta \in \Theta.$

Examples:

- $\Theta = \mathbb{R}^2$, $h(\theta) = -\theta_1^2 - 2\theta_2^2$ or $h(\theta) = -(\theta_1 - \theta_2)^2$
- $\Theta = (0, \infty)$, $h(\theta) = \log(\theta_1 + \theta_2)$,

Lagrange multipliers

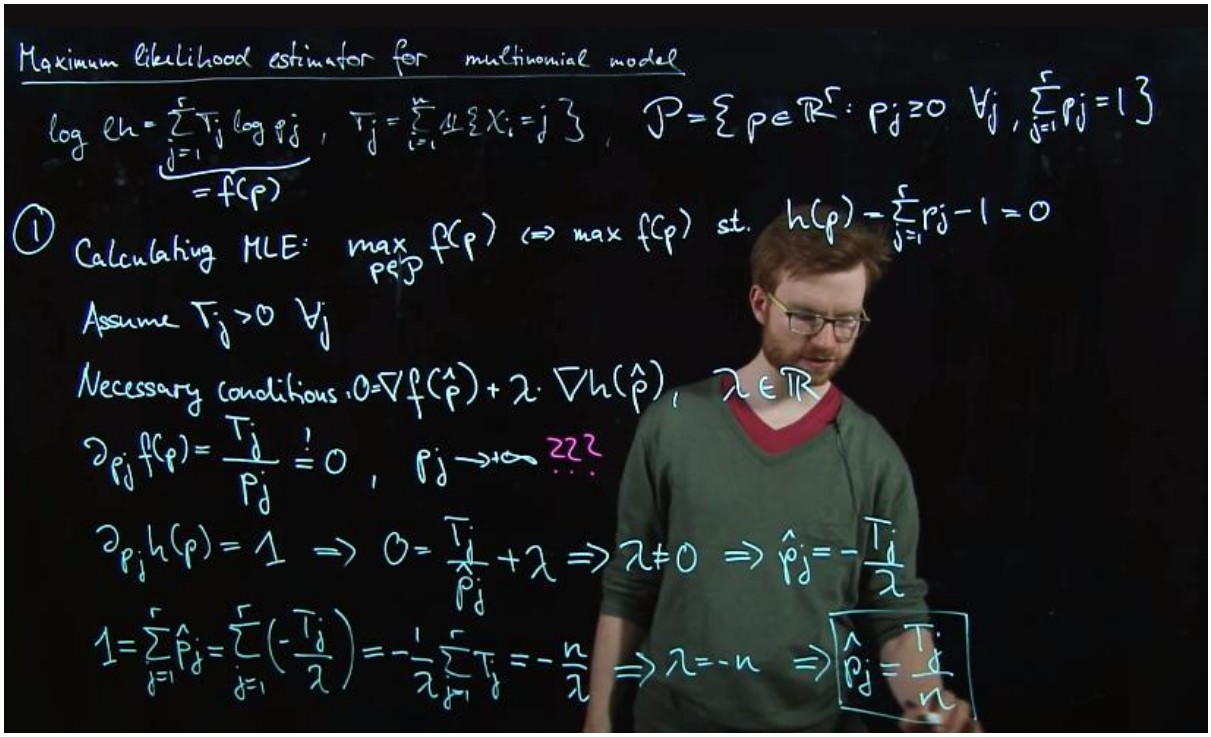
In **mathematical optimization, the method of Lagrange multipliers is a strategy for finding the local maxima and minima of a function subject to equality constraints** (i.e., subject to the condition that one or more equations have to be satisfied exactly by the chosen values of the variables).

maximize $f(x, y)$

subject to $g(x, y) = 0$.

(Sometimes an additive constant is shown separately rather than being included in g, in which case the constraint is written $g(x, y) = c$, as in Figure 1.) We assume that both f and g have continuous first partial derivatives. We introduce a new variable (λ) called a **Lagrange multiplier** (or **Lagrange undetermined multiplier**) and study the **Lagrange function** (or **Lagrangian** or **Lagrangian expression**) defined by

$$\mathcal{L}(x, y, \lambda) = f(x, y) - \lambda \cdot g(x, y),$$



Maximum likelihood estimator asymptotic normality

4 Conditions

- Model identified / parameter identifiable: if I give you 2 parameters theta, they give me 2 different distributions and inversely

Asymptotic normality of the MLE

Theorem

Let $\theta^* \in \Theta$ (the *true* parameter). Assume the following:

1. The parameter is identifiable. ✓
2. For all $\theta \in \Theta$, the support of $\mathbb{P}_\theta$ does not depend on θ; ✓
3. θ^* is not on the boundary of Θ; ✓
4. $I(\theta)$ is invertible in a neighborhood of θ^*; ✓
5. A few more technical conditions. ✓

Then, $\hat{\theta}_n^{MLE}$ satisfies:

- $\hat{\theta}_n^{MLE} \xrightarrow[n\to\infty]{\mathbb{P}} \theta^*$ w.r.t. $\mathbb{P}_{\theta^*}$;
- $\sqrt{n}\left(\hat{\theta}_n^{MLE} - \theta^*\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}_d\left(0, I(\theta^*)^{-1}\right)$ w.r.t. $\mathbb{P}_{\theta^*}$.

Fisher information

Fisher information is defined using only one Xi (X1) for the log likelihood calculation.

Fisher Information

Definition: Fisher information

Define the log-likelihood for one observation as:

$$\ell(\theta) = \log L_1(X, \theta), \qquad \theta \in \Theta \subset \mathbb{R}^d$$

Assume that ℓ is a.s. twice differentiable. Under some regularity conditions, the *Fisher information* of the statistical model is defined as:

$$I(\theta) = \mathbb{E}\left[\nabla \ell(\theta) \nabla \ell(\theta)^\top\right] - \mathbb{E}\left[\nabla \ell(\theta)\right] \mathbb{E}\left[\nabla \ell(\theta)\right]^\top = -\mathbb{E}\left[\mathbf{H}\ell(\theta)\right].$$

If $\Theta \subset \mathbb{R}$, we get:

$$I(\theta) = \mathsf{var}\left[\ell'(\theta)\right] = -\mathbb{E}\left[\ell''(\theta)\right]$$

$$\mathcal{I}(\theta) = \mathsf{Cov}\left(\nabla \ell(\theta)\right) = -\mathbb{E}\left[\mathbf{H}\ell(\theta)\right].$$

$$\mathcal{I}(\theta) = \int_{-\infty}^{\infty} \frac{\left(\frac{\partial}{\partial \theta} f_\theta(x)\right)^2}{f_\theta(x)}\, dx.$$

Method of moments

Moments estimator

Let

$$\begin{array}{rcccl} M & : & \Theta & \to & \mathbb{R}^d \\ & & \theta & \mapsto & M(\theta) = (m_1(\theta), \ldots, m_d(\theta)) \, . \end{array}$$

Assume M is one to one:

$$\theta = M^{-1}(m_1(\theta), \ldots, m_d(\theta)).$$

Definition

Moments estimator of θ:

$$\hat{\theta}_n^{MM} = M^{-1}(\hat{m}_1, \ldots, \hat{m}_d),$$

provided it exists.

Asymptotic variance of the method of moments estimator:

Generalized method of moments

Applying the multivariate CLT and Delta method yields:

Theorem

$$\sqrt{n}\left(\hat{\theta}_n^{MM} - \theta\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}\left(0, \Gamma(\theta)\right) \qquad \text{(w.r.t. } \mathbb{P}_\theta\text{)},$$

where $\Gamma(\theta) = \left[\frac{\partial M^{-1}}{\partial \theta}\big(M(\theta)\big)\right]^\top \Sigma(\theta) \left[\frac{\partial M^{-1}}{\partial \theta}\big(M(\theta)\big)\right]$.

No statistical model is needed

Definition

replace $\mathbb{E}$ with $\frac{1}{n}\sum_{i=1}^{n}$

- Define $\hat{\mu}_n$ as a minimizer of:

$$\mathcal{Q}_n(\mu) := \frac{1}{n}\sum_{i=1}^{n} \rho(X_i, \mu).$$

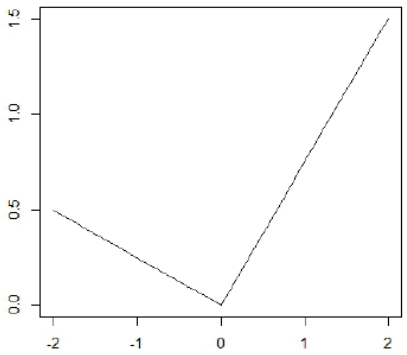
- Examples: Empirical mean, empirical median, empirical quantiles, MLE, etc.

Examples (2)

If $E = \mathcal{M} = \mathbb{R}$, $\alpha \in (0,1)$ is fixed and $\rho(x,\mu) = C_\alpha(x-\mu)$, for all $x \in \mathbb{R}, \mu \in \mathbb{R}$: μ^* is a α-quantile of $\mathbb{P}$.

Check function

$$C_\alpha(x) = \begin{cases} -(1-\alpha)x \text{ if } x < 0 \\ \alpha x \text{ if } x \geq 0. \end{cases}$$



Statistical analysis

- Let $J(\mu) = +\dfrac{\partial^2 Q}{\partial \mu \partial \mu^\top}(\mu) \quad \left(= +\mathbb{E}\left[\dfrac{\partial^2 \rho}{\partial \mu \partial \mu^\top}(X_1, \mu)\right] \text{ under some regularity conditions}\right)$.

- Let $K(\mu) = \text{Cov}\left[\dfrac{\partial \rho}{\partial \mu}(X_1, \mu)\right]$.

- **Remark:** In the log-likelihood case (write $\mu = \theta$),

$$J(\theta) = K(\theta) = I(\theta) \quad \text{(Fisher information)}$$

Asymptotic normality

Let $\mu^* \in \mathcal{M}$ (the *true* parameter). Assume the following:

1. μ^* is the only minimizer of the function $\mathcal{Q}$;
2. $J(\mu)$ is invertible for all $\mu \in \mathcal{M}$;
3. A few more technical conditions.

Then, $\hat{\mu}_n$ satisfies:

- $\hat{\mu}_n \xrightarrow[n\to\infty]{\mathbb{P}} \mu^*$;
- $\sqrt{n}\,(\hat{\mu}_n - \mu^*) \xrightarrow[n\to\infty]{(d)} \mathcal{N}\left(0,\ J(\mu^*)^{-1} K(\mu^*) J(\mu^*)^{-1}\right)$.

HYPOTHESIS TESTING

Our tests were based on CLT (and sometimes Slutsky)...

- What if data is Gaussian, σ^2 is unknown and Slutsky does not apply? T-test
- Can we use asymptotic normality of MLE? Wald's test
- Tests about multivariate parameters $\theta = (\theta_1, \ldots, \theta_d)$ (e.g.: $\theta_1 = \theta_2$)? Implicit hypotheses
- More complex tests: "Does my data follow a Gaussian distribution"? Goodness of fit.

PARAMETRIC HYPOTHESIS TESTING

Small sample size

Distribution of Sample Variance of Gaussian: The Chi-Squared Distribution

It is also the length of a Gaussian vector, distance from the center (norm of a Gaussian vector)

The χ^2 distribution

Definition

For a positive integer d, the χ^2 *(pronounced "Kai-squared") distribution with d degrees of freedom* is the law of the random variable $Z_1^2 + Z_2^2 + \ldots + Z_d^2$, where $Z_1, \ldots, Z_d \overset{iid}{\sim} \mathcal{N}(0,1)$.

Examples:

- If $Z \sim \mathcal{N}_d(\mathbf{0}, I_d)$, then $\|Z\|_2^2 \sim \chi^2_d$
- $\chi^2_2 = \text{Exp}(1/2)$.

pdf of $X \sim \chi^2_k = \Gamma(\frac{k}{2}, \frac{1}{2})$

$\frac{1}{2^{k/2}\Gamma(\frac{k}{2})} x^{k/2-1} e^{-x/2}, \quad x > 0$

- Recall that the sample variance is given by

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

- Cochran's theorem states that for $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, if S_n is the sample variance, then
  - $\bar{X}_n \perp\!\!\!\perp S_n$; for all n.
  - $\frac{nS_n}{\sigma^2} \sim \chi^2_{n-1}$.

$\mathbb{E}[S_n] = \frac{n-1}{n}\sigma^2$

  - We often prefer the unbiased estimator of σ^2:

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

$$\mathbb{E}[\tilde{S}_n] = \frac{n}{n-1}\,\mathbb{E}\left[\frac{\sigma^2}{n}\chi^2_{n-1}\right] = \frac{n\sigma^2}{n-1}\cdot\frac{n-1}{n} = \sigma^2$$

Student's t-distribution : Estimating the mean of a Gaussian

- where sample size n is small or big
  and
- variance is unknown.

Definition

For a positive integer d, the *Student's T distribution with d degrees of freedom* (denoted by t_d) is the law of the random variable $\frac{Z}{\sqrt{V/d}}$, where $Z \sim \mathcal{N}(0,1)$, $V \sim \chi^2_d$ and $Z \perp\!\!\!\perp V$ (Z is independent of V).

Student's T test (one sample, two-sided)

- Let $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ where both μ and σ^2 are unknown
- We want to test:

$$H_0 : \mu = 0, \quad \text{vs} \quad H_1 : \mu \neq 0$$

- Test statistic:

$$T_n = \frac{\bar{X}_n}{\sqrt{\tilde{S}_n/n}} = \frac{\sqrt{n}\bar{X}_n/\sigma}{\sqrt{\tilde{S}_n/\sigma^2}}$$

under H_0 $\sim \mathcal{N}(0,1)$

- Since $\sqrt{n}\bar{X}_n/\sigma \sim \mathcal{N}(0,1)$ (under H_0) and $\tilde{S}_n/\sigma^2 \sim \frac{\chi^2_{n-1}}{n-1}$ are independent by Cochran's theorem, we have:

$$T_n \sim t_{n-1}$$

- Student's test with (non asymptotic) level $\alpha \in (0,1)$:

$$\psi_\alpha = \mathbb{1}\{|T_n| > q_{\alpha/2}\},$$

where $q_{\alpha/2}$ is the $(1-\alpha/2)$-quantile of t_{n-1}.

Student's T test (one sample, one-sided)

- We want to test:

$$H_0 : \mu \leq \mu_0, \quad \text{vs} \quad H_1 : \mu > \mu_0$$

- Test statistic:

$$T_n = \frac{\bar{X}_n - \mu_0}{\sqrt{\tilde{S}_n}} \sim t_{n-1} \quad \text{under } H_0$$

under H_0.

- Student's test with (non asymptotic) level $\alpha \in (0, 1)$:

$$\psi_\alpha = \mathbb{1}\Big\{ T_n > q_\alpha \Big\},$$

where q_α is the $(1-\alpha)$-quantile of t_{n-1}

Two-sample T-test

- Back to our cholesterol example. What happens for small sample sizes?
- We want to know the distribution of

$$\frac{\bar{X}_n - \bar{Y}_m - (\Delta_{\mathrm{d}} - \Delta_{\mathrm{c}})}{\sqrt{\frac{\hat{\sigma}_{\mathrm{d}}^2}{n} + \frac{\hat{\sigma}_{\mathrm{c}}^2}{m}}}$$

- We have approximately

$$\frac{\bar{X}_n - \bar{Y}_m - (\Delta_{\mathrm{d}} - \Delta_{\mathrm{c}})}{\sqrt{\frac{\hat{\sigma}_{\mathrm{d}}^2}{n} + \frac{\hat{\sigma}_{\mathrm{c}}^2}{m}}} \sim t_N$$

where

$$N = \frac{\left(\hat{\sigma}_{\mathrm{d}}^2/n + \hat{\sigma}_{\mathrm{c}}^2/m\right)^2}{\frac{\hat{\sigma}_{\mathrm{d}}^4}{n^2(n-1)} + \frac{\hat{\sigma}_{\mathrm{c}}^4}{m^2(m-1)}} \geq \min(n, m)$$

(Welch-Satterthwaite formula)

W-S formula or shorthand formula N = min(n,m)

T-test is non-asymptotic & works for small sizes but requires Gaussian

Non-asymptotic Two-Sample Test using t-statistic

Assume

- $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu_X, \sigma_X^2)$,
- $Y_1, \ldots, Y_m \overset{iid}{\sim} \mathcal{N}(\mu_Y, \sigma_Y^2)$,
- $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ are independent.

Then, for any n and m, the distribution of the test statistic below is approximated by a t-distribution:

$$\frac{\overline{X}_n - \overline{Y}_m - (\mu_X - \mu_Y)}{\sqrt{\hat{\sigma}_X^2/n + \hat{\sigma}_Y^2/m}} \overset{\text{approx.}}{\sim} t_N$$

where the degrees of freedom N is given by the **Welch-Satterthwaite formula** :

$$\min(n,m) \leq N = \frac{\left(\hat{\sigma}_X^2/n + \hat{\sigma}_Y^2/m\right)^2}{\frac{\hat{\sigma}_X^4}{n^2(n-1)} + \frac{\hat{\sigma}_Y^4}{m^2(m-1)}} \leq n+m$$

A test based on the MLE

- Consider an i.i.d. sample $X_1, \ldots, X_n$ with statistical model $(E, (\mathbb{P}_\theta)_{\theta \in \Theta})$, where $\Theta \subseteq \mathbb{R}^d$ $(d \geq 1)$ and let $\theta_0 \in \Theta$ be fixed and given. θ^* is the true parameter
- Consider the following hypotheses:

$$\begin{cases} H_0 : & \theta^* = \theta_0 \\ H_1 : & \theta^* \neq \theta_0. \end{cases}$$

- Let $\hat{\theta}^{MLE}$ be the MLE. Assume the MLE technical conditions are satisfied.

- If H_0 is true, then

$$\sqrt{n}\, I(\theta_0)^{1/2} \times \left(\hat{\theta}_n^{MLE} - \theta_0\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}_d\left(0, I_d\right)$$

$\sqrt{n}\, I(\theta^*)^{1/2}$

$\sqrt{n}\, I(\hat{\theta}^{MLE})^{1/2}$

Wald's test

Wald's test

- Hence,

$$\underbrace{n\left(\hat{\theta}_n^{MLE}-\theta_0\right)^\top I(\hat{\theta}^{MLE})\left(\hat{\theta}_n^{MLE}-\theta_0\right)}_{T_n} \xrightarrow[n\to\infty]{(d)} \chi^2_d$$

- Wald's test with asymptotic level $\alpha \in (0,1)$:

$$\psi = \mathbb{1}\{T_n > q_\alpha\},$$

where q_α is the $(1-\alpha)$-quantile of χ^2_d (see tables).

- Remark: Wald's test is also valid if H_1 has the form "$\theta > \theta_0$" or "$\theta < \theta_0$" or "$\theta = \theta_1$" ...

But less powerful

Likelihood ratio test

A test based on the log-likelihood

- Consider an i.i.d. sample $X_1, \ldots, X_n$ with statistical model $(E, (\mathbb{P}_\theta)_{\theta \in \Theta})$, where $\Theta \subseteq \mathbb{R}^d$ $(d \geq 1)$.

- Suppose the null hypothesis has the form

$$H_0 : (\theta_{r+1}, \ldots, \theta_d) = (\theta_{r+1}^{(0)}, \ldots, \theta_d^{(0)}),$$

for some fixed and given numbers $\theta_{r+1}^{(0)}, \ldots, \theta_d^{(0)}$.

- Let

$$\hat{\theta}_n = \underset{\theta \in \Theta}{\operatorname{argmax}} \ \ell_n(\theta) \quad \text{(MLE)}$$

and

$$\hat{\theta}_n^c = \underset{\theta \in \Theta_0}{\operatorname{argmax}} \ \ell_n(\theta) \quad \text{("constrained MLE")}$$

where $\Theta_0 = \{\theta \in \Theta : (\theta_{r+1}, \ldots, \theta_d) = (\theta_{r+1}^{(0)}, \ldots, \theta_d^{(0)})\}$

Likelihood ratio test

Test statistic:

$$T_n = 2\left(\ell_n(\hat{\theta}_n) - \ell_n(\hat{\theta}_n^c)\right).$$

Wilks' Theorem

Assume H_0 is true and the MLE technical conditions are satisfied. Then,

$$T_n \xrightarrow[n\to\infty]{(d)} \chi^2_{d-r}$$

Likelihood ratio test with asymptotic level $\alpha \in (0,1)$:

$$\psi = \mathbb{1}\{T_n > q_\alpha\},$$

where q_α is the $(1-\alpha)$-quantile of χ^2_{d-r} (see tables).

NB: (d-r) is the number of fixed parameters in the constraint

Implicit hypothesis: hypothesis on function of parameters

Implicit hypotheses

- Let $X_1, \ldots, X_n$ be i.i.d. random variables and let $\theta \in \mathbb{R}^d$ be a parameter associated with the distribution of X_1 (e.g. a moment, the parameter of a statistical model, etc...)
- Let $g : \mathbb{R}^d \to \mathbb{R}^k$ be continuously differentiable (with $k < d$).
- Consider the following hypotheses:

$$\begin{cases} H_0 : & g(\theta) = 0 \\ H_1 : & g(\theta) \neq 0. \end{cases}$$

- E.g. $g(\theta) = (\theta_1, \theta_2)$ $(k = 2)$, or $g(\theta) = \theta_1 - \theta_2$ $(k = 1)$, or...

Delta method

- Suppose an asymptotically normal estimator $\hat{\theta}_n$ is available:

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}_d(0, \Sigma(\theta)).$$

- Delta method:

$$\sqrt{n}\left(g(\hat{\theta}_n) - g(\theta)\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}_k\left(0, \Gamma(\theta)\right),$$

where $\Gamma(\theta) = \nabla g(\theta)^\top \Sigma(\theta) \nabla g(\theta) \in \mathbb{R}^{k\times k}$.

- Assume $\Sigma(\theta)$ is invertible and $\nabla g(\theta)$ has rank k. So, $\Gamma(\theta)$ is invertible and

$$\sqrt{n}\,\Gamma(\theta)^{-1/2}\left(g(\hat{\theta}_n) - g(\theta)\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}_k(0, I_k).$$

Wald's test for implicit hypotheses

- Then, by Slutsky's theorem, if $\Gamma(\theta)$ is continuous in θ,

$$\sqrt{n}\,\Gamma(\hat{\theta}_n)^{-1/2}\left(g(\hat{\theta}_n)-g(\theta)\right)\xrightarrow[n\to\infty]{(d)}\mathcal{N}_k\left(0,I_k\right).$$

- Hence, if H_0 is true, i.e., $g(\theta)=0$,

$$\underbrace{n g(\hat{\theta}_n)^\top \Gamma^{-1}(\hat{\theta}_n) g(\hat{\theta}_n)}_{T_n}\xrightarrow[n\to\infty]{(d)}\chi^2_k.$$

- Test with asymptotic level α:

$$\psi=\mathbb{1}\{T_n>q_\alpha\},$$

where q_α is the $(1-\alpha)$-quantile of χ^2_k (see tables).

GOODNESS OF FIT (NON PARAMETRIC HYPOTHESIS TESTING)

Chi squared test

Multinomial distribution

χ^2 test

- If H_0 is true, then $\sqrt{n}(\hat{\mathbf{p}} - \mathbf{p}^0)$ is asymptotically normal, and the following holds.

$\sqrt{n}\, I(p^\circ)^{1/2}(\hat{p} - p^\circ) \xrightarrow[n\to\infty]{} \mathcal{N}_K(0, I_{K-1})$

Theorem Under H_0 :

$$\underbrace{n \sum_{j=1}^{K} \frac{\left(\hat{\mathbf{p}}_j - \mathbf{p}_j^0\right)^2}{\mathbf{p}_j^0}}_{T_n} \xrightarrow[n\to\infty]{(d)} \chi^2_{K-1}.$$

- χ^2 test with asymptotic level α: $\psi_\alpha = \mathbb{1}\{T_n > q_\alpha\}$, where q_α is the $(1-\alpha)$-quantile of χ^2_{K-1}.
- Asymptotic p-value of this test: $p - \text{value} = \mathbb{P}\left[Z > T_n | T_n\right]$, where $Z \sim \chi^2_{K-1}$ and $Z \perp\!\!\!\perp T_n$.

More general setting

χ^2-Test for a Family of Distributions :

Now, we return to the following more general statistical set-up.

Let $X_1, \ldots, X_n \overset{iid}{\sim} \mathbf{P}$ denote iid discrete random variables supported on $\{0, \ldots, K\}$. We will decide between the following null and alternative hypotheses.

$$H_0: \quad \mathbf{P} \in \{\mathrm{Bin}\,(K, \theta)\}_{\theta \in (0,1)}.$$

$$H_1: \quad \mathbf{P} \notin \{\mathrm{Bin}\,(K, \theta)\}_{\theta \in (0,1)}.$$

Let f_θ denote the pmf of the distribution $\mathrm{Bin}\,(K, \theta)$, and let $\hat{\theta}$ denote the MLE of the parameter θ from the previous problem.

Further, let N_j denote the number of times that j ($j \in \{0, 1, \ldots, K\}$) appears in the data set $X_1, \ldots, X_n$ (so that $\sum_{j=0}^{K} N_j = n.$) The χ^2 **test statistic for this hypothesis test** is defined to be

$$T_n := n \sum_{j=0}^{K} \frac{\left(\frac{N_j}{n} - f_{\hat{\theta}}\,(j)\right)^2}{f_{\hat{\theta}}\,(j)}.$$

This statistic is different from before. Previously, under the null hypothesis, $\mathbf{P}\,(X = j) = p_j$ for some fixed p_j. Here, instead, we use $f_{\hat{\theta}}\,(j)$ to estimate $\mathbf{P}\,(X = j)$. This statistic still converges in distribution to a χ^2 distribution, but the number of degrees of freedom is smaller.

Degrees of Freedom for χ^2 Test for a Family of Distribution

More generally, to test if a distribution $\mathbf{P}$ is described by some member of a family of discrete distributions $\{\mathbf{P}_\theta\}_{\theta \in \Theta \subset \mathbb{R}^d}$ where $\Theta \subset \mathbb{R}^d$ is d-dimensional, with support $\{0, 1, 2, \ldots, K\}$ and pmf f_θ, i.e. to test the hypotheses:

$$H_0: \quad \mathbf{P} \in \{\mathbf{P}_\theta\}_{\theta \in \Theta}$$

$$H_1: \quad \mathbf{P} \notin \{\mathbf{P}_\theta\}_{\theta \in \Theta},$$

then if indeed $\mathbf{P} \in \{\mathbf{P}_\theta\}_{\theta \in \Theta \subset \mathbb{R}^d}$ (*i.e.*, the null hypothesis H_0 holds), and if in addition some technical assumptions hold, then we have that

$$T_n := n \sum_{j=0}^{K} \frac{\left(\frac{N_j}{n} - f_{\hat{\theta}}\,(j)\right)^2}{f_{\hat{\theta}}\,(j)} \xrightarrow[n \to \infty]{(d)} \chi^2_{(K+1)-d-1}.$$

Note that $K + 1$ is the support size of $\mathbf{P}_\theta$ (for all θ.)

In our example testing for a binomial distribution, the parameter θ is one-dimensional, i.e. $d = 1$. Therefore, under the null hypothesis H_0, it holds that

$$T_n \xrightarrow[n \to \infty]{(d)} \chi^2_{(K+1)-1-1} = \chi^2_{K-1}.$$

In R

```
M <- as.table(rbind(c(762, 327, 468), c(484, 239, 477)))
dimnames(M) <- list(gender = c("F", "M"),
                    party = c("Democrat","Independent", "Republican"))
(Xsq <- chisq.test(M))  # Prints test summary
Xsq$observed   # observed counts (same as M)
Xsq$expected   # expected counts under the null
Xsq$residuals  # Pearson residuals
```

```
Xsq$stdres      # standardized residuals
```

CDF and empirical CDF

Let $X_1, \ldots, X_n$ be i.i.d. real random variables. Recall the cdf of X_1 is defined as:

$$F(t) = \mathbb{P}[X_1 \leq t], \quad \forall t \in \mathbb{R}.$$

It completely characterizes the distribution of X_1.

Definition

The *empirical cdf* (a.k.a sample cdf) of the sample $X_1, \ldots, X_n$ is defined as:

$$\begin{aligned} F_n(t) &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_i \leq t\} \\ &= \frac{\#\{i = 1, \ldots, n : X_i \leq t\}}{n}, \quad \forall t \in \mathbb{R}. \end{aligned}$$

Consistency

By the LLN, for all $t \in \mathbb{R}$,

$$F_n(t) \xrightarrow[n\to\infty]{a.s.} F(t).$$

Glivenko-Cantelli Theorem (*Fundamental theorem of statistics*)

$$\sup_{t\in\mathbb{R}} |F_n(t) - F(t)| \xrightarrow[n\to\infty]{a.s.} 0.$$

Asymptotic normality

By the CLT, for all $t \in \mathbb{R}$,

$$\sqrt{n}\ (F_n(t) - F(t)) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, F(t)(1-F(t))).$$

Donsker's Theorem

If F is continuous, then

$$\sqrt{n} \sup_{t\in\mathbb{R}} |F_n(t) - F(t)| \xrightarrow[n\to\infty]{(d)} \sup_{0\le t\le 1} |\mathbb{B}(t)|,$$

where $\mathbb{B}$ is a Brownian bridge on $[0, 1]$.

KS Test: testing against a specific distribution Po with known parameters

Donker's theorem is valid if known parameters of Po distribution

Kolmogorov-Smirnov test

- Let $T_n = \sup_{t \in \mathbb{R}} \sqrt{n} \left| F_n(t) - F^0(t) \right|$.
- By Donsker's theorem, if H_0 is true, then $T_n \xrightarrow[n \to \infty]{(d)} Z$, where Z has a known distribution (supremum of a Brownian bridge).
- **KS test with asymptotic level α:**

$$\delta_\alpha^{KS} = \mathbb{1}\{T_n > q_\alpha\},$$

where q_α is the $(1-\alpha)$-quantile of Z (obtained in tables).

- p-value of KS test: $\mathbb{P}[Z > T_n | T_n]$.

Let $Y_1, \ldots, Y_n$ be iid random variables with continuous cdf F. Consider the distribution of the statistic

$$T_n = \sqrt{n} \sup_{t \in \mathbb{R}} |F_n(t) - F(t)|.$$

This statistic is **pivotal**, *i.e.*, for any fixed n, the distribution of T_n does **not** depend on the distribution of Y_i. Let P_n^{KS} denote the distribution of T_n.

The number x in the n-th row of the column labeled by the level α table in the slide "K-S table" is such that

$$P_n^{KS}\left(\frac{T_n}{\sqrt{n}} > x\right) = \alpha.$$

Computatinal issues

- In practice, how to compute T_n ?
- F^0 is non decreasing, F_n is piecewise constant, with jumps at $t_i = X_i, i = 1, \ldots, n$.
- Let $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ be the reordered sample.
- The expression for T_n reduces to the following practical formula:

$$T_n = \sqrt{n} \max_{i=1,\ldots,n} \left\{ \max \left(\left| \frac{i-1}{n} - F^0(X_{(i)}) \right|, \left| \frac{i}{n} - F^0(X_{(i)}) \right| \right) \right\}.$$

Other goodness of fit tests

We want to measure the distance between two functions: $F_n(t)$ and $F(t)$. There are other ways, leading to other tests:

- Kolmogorov-Smirnov:

$$d(F_n, F) = \sup_{t \in \mathbb{R}} |F_n(t) - F(t)| \qquad L_\infty$$

- Cramér-Von Mises:

$$d^2(F_n, F) = \int_{\mathbb{R}} [F_n(t) - F(t)]^2 \, dF(t) \qquad L_2$$

$$:= \mathbb{E}_{X \sim F} [F_n(X) - F(X)]^2$$

- Anderson-Darling:

$$d^2(F_n, F) = \int_{\mathbb{R}} \frac{[F_n(t) - F(t)]^2}{F(t)(1 - F(t))} dF(t)$$

KL Test: Test for Gaussian with unknown parameters

Kolmogorov-Lilliefors test (1)

Instead, we compute the quantiles for the test statistic:

$$\sup_{t\in\mathbb{R}} \left|F_n(t) - \Phi_{\hat{\mu},\hat{\sigma}^2}(t)\right|$$

They do not depend on unknown parameters!

This is the Kolmogorov-Lilliefors test.

Recall that the KL test statistic is given by

$$T_n^{\mathrm{KL}} = \max_{i=1,\ldots,n}\{\max\left(\left|\frac{i-1}{n} - \Phi_{\hat{\mu},\widehat{\sigma^2}}(x_i)\right|, \left|\frac{i}{n} - \Phi_{\hat{\mu},\widehat{\sigma^2}}(x_i)\right|\right)\}.$$

To find $\Phi_{\hat{\mu},\widehat{\sigma^2}}(x_i)$, we make change of variables:

$$\Phi_{\hat{\mu},\widehat{\sigma^2}}(x_i) = \frac{x_i - \hat{\mu}}{\sqrt{\widehat{\sigma^2}}}.$$

Then we use the following formula to find $T_5^{\mathrm{KL}}/\sqrt{5}$:

$$\max_{i=1,\ldots 5}\left\{\max\left(\left|\frac{i-1}{5} - \Phi_{\hat{\mu},\widehat{\sigma^2}^{\mathrm{unbiased}}}(x_i)\right|, \left|\frac{i}{5} - \Phi_{\hat{\mu},\widehat{\sigma^2}^{\mathrm{unbiased}}}(x_i)\right|\right)\right\}$$

We take Phi of the normalized xi, where xi is the ith observation of the sample, ordered.

Comparison of tests

- K-L test is more likely to reject than the K-S test

- KS and KL tests are non-asymptotic (valid for small n, tables available, pivotal test statistic that does not depend on the sample data distribution/ of the true parameters)

- Chi squared test is only asymptotic

For example, the Kolmogorov-Smirnov test is a normality test for $\mathcal{F} = \{\mathcal{N}(0,1)\}$– that is, when $\mathcal{F}$ consists of a single Gaussian distribution. The Kolmogorov-Lilliefors test is a normality test with $\mathcal{F} = \{\mathcal{N}(\mu,\sigma^2)\}_{\mu\in\mathbb{R},\sigma^2>0}$– that is, when $\mathcal{F}$ consists of all Gaussian distributions. The χ^2 test studied on this page is also a normality test with $\mathcal{F} = \{\mathcal{N}(\mu,\sigma^2)\}_{\mu\in\mathbb{R},\sigma^2>0}$.

Quantiles quantiles plots (QQ plots) : visual informal goodness of fit

Quantile-Quantile (QQ) plots (1)

- Provide a visual way to perform GoF tests
- Not formal test but quick and easy check to see if a distribution is plausible.
- Main idea: we want to check visually if the plot of F_n is close to that of F or equivalently if the plot of F_n^{-1} is close to that of F^{-1}.
- More convenient to check if the points

$$\left(F^{-1}(\frac{1}{n}), F_n^{-1}(\frac{1}{n})\right), \left(F^{-1}(\frac{2}{n}), F_n^{-1}(\frac{2}{n})\right), \ldots, \left(F^{-1}(\frac{n-1}{n}), F_n^{-1}(\frac{n-1}{n})\right)$$

  are near the line $y = x$.
- F_n is not technically invertible but we define

$$F_n^{-1}(i/n) = X_{(i)},$$

  the ith largest observation.

Quantile-Quantile (QQ) plots (3)

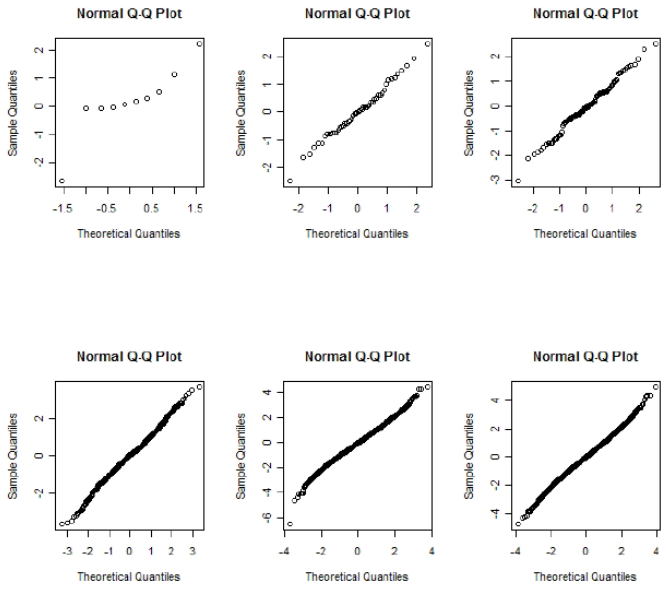


Figure 2: QQ-plots for samples of sizes $10, 50, 100, 1000, 5000, 10000$ from a t_{15} distribution. The upper-left figure is for sample size 10, the lower-right is for sample 10000.

Bayes' formula

- Bayes' formula states that:

proportional up to constant that does not depend on θ

$$\pi(\theta|X_1,\ldots,X_n) \propto \pi(\theta)L_n(X_1,\ldots,X_n|\theta), \quad \forall\theta\in\Theta.$$

- The constant does not depend on θ:

$$\pi(\theta|X_1,\ldots,X_n) = \frac{\pi(\theta)L_n(X_1,\ldots,X_n|\theta)}{\int_\Theta \pi(\theta)L_n(X_1,\ldots,X_n|\theta)d\theta}, \quad \forall\theta\in\Theta.$$

Non informative priors

- We can still use a Bayesian approach if we have no prior information about the parameter. How to pick prior π?
- Good candidate: $\pi(\theta)\propto 1$, i.e., constant pdf on Θ.
- If Θ is bounded, this is the uniform prior on Θ.
- If Θ is unbounded, this does not define a proper pdf on Θ !
- An *improper prior* on Θ is a measurable, nonnegative function $\pi(\cdot)$ defined on Θ that is not integrable. Improper iff $\int\pi(\theta)d\theta=\infty$
- In general, one can still define a posterior distribution using an improper prior, using Bayes' formula.

Jeffreys' prior

- **Jeffreys prior:**
$$\pi_J(\theta) \propto \sqrt{\det I(\theta)}$$
where $I(\theta)$ is the Fisher information matrix of the statistical model associated with $X_1, \ldots, X_n$ in the frequentist approach (provided it exists).

- In the previous examples:
  - Bernoulli experiment: $\pi_J(p) \propto \frac{1}{\sqrt{p(1-p)}}$, $p \in (0,1)$: the prior is Beta($\frac{1}{2}$, $\frac{1}{2}$).
  - Gaussian experiment: $\pi_J(\theta) \propto 1$, $\theta \in \mathbb{R}$ is an improper prior.

- Jeffreys prior satisfies a reparametrization invariance principle: If η is a reparametrization of θ (i.e., $\eta = \phi(\theta)$ for some one-to-one map ϕ), then the pdf $\tilde{\pi}(\cdot)$ of η satisfies:
$$\tilde{\pi}(\eta) \propto \sqrt{\det \tilde{I}(\eta)},$$
where $\tilde{I}(\eta)$ is the Fisher information of the statistical model parametrized by η instead of θ.

Bayesian estimation

- *Bayes estimator:*

$$\hat{\theta}^{(\pi)} = \int_{\Theta} \theta \, \pi(\theta \mid X_1, \ldots, X_n) \, d\theta$$

  This is the *posterior mean*.

- The Bayesian estimator depends on the choice of the prior distribution π (hence the superscript π).

- Another popular choice is the point that maximizes the posterior distribution, provided it is unique. It is called the MAP (maximum a posteriori):

$$\hat{\theta}^{\text{MAP}} = \underset{\theta \in \Theta}{\operatorname{argmax}} \; \pi(\theta \mid X_1, \ldots, X_n)$$

$$L_n(X_1, \ldots, X_n \mid \theta) \, \pi(\theta)$$

LINEAR REGRESSION

$$\mathbf{X} = \begin{pmatrix} 1 & x_1^{(1)} & x_2^{(2)} \\ & \vdots & \\ 1 & x_n^{(1)} & x_n^{(2)} \end{pmatrix}.$$

There exists a unique least-squares estimator $\hat{\beta}$ if $rank\,(\mathbf{X}) = 3$, since β has three components. The first and fourth choices each yield square 3×3 matrices $\mathbf{X}$ with zero determinant, so their respective ranks are strictly smaller than 3. In the second choice, the resulting matrix has size 2×3, so its rank can be at most 2. In contrast, the third choice yields a 4×3 matrix with full rank (rank 3).