Special Mathematics Lecture
Statistics

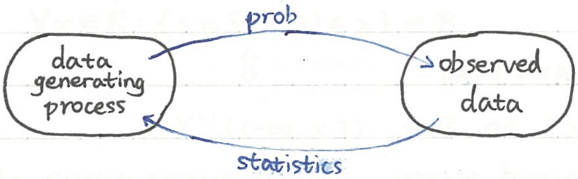
Table of content

Handwritten notes taken by L. Zhang

Statistics

Course is based on [CB] Statistical inference.

Probability and Statistics



I Probability

I.1 Probability space

Ingredients

1) $S = \Omega$ is a set called the SAMPLE SPACE set of all outcomes of an experiment

2) Any subset $A \subset S$ is called an EVENT

3) An EVENT SPACE $\mathcal{B} = \mathcal{F}$ is a collection of events satisfying:

$$\left.\begin{array}{l} 1)\ \phi \text{ and } S \in \mathcal{B} \quad \phi \text{ is the empty set} \\ 2)\ \text{If } A \in \mathcal{B} \text{ then } A^c = S \setminus A \ \text{S minus A or the complement of A in S} \in \mathcal{B} \\ 3)\ \text{If } A_j \in \mathcal{B} \text{ for } j = 1, 2, 3, \cdots \ \text{infinite but countable family then } \bigcup_{j=1}^{\infty} A_j \in \mathcal{B} \end{array}\right\} \overset{\text{sigma}}{\sigma}\text{ - ALGEBRA}$$

Recall: $\left.\begin{array}{l} (A \cup B)^c = A^c \cap B^c \\ (A \cap B)^c = A^c \cup B^c \end{array}\right\}$ De Morgan's Law

Examples

1) If $S = \{a_1, a_2, \cdots, a_n\}$ then $\mathcal{B} = \{\text{all subset of } S\}$ power set of S

2) If $S = \mathbb{R}$, then we choose $\mathcal{B}$ as the family of subsets generated by intervals $(a, b) \forall a < b$

$\mathcal{B}$ is called the BOREL ALGEBRA of $\mathbb{R}$ denoted by σ_B.

(We can do the same for $[a, b]$)

⇒ ∃ some subsets of S to which we cannot give a weight

4) a PROBABILITY FUNCTION "WEIGHT" $P = \mathbb{P}$ is a map $P: \mathcal{B} \mapsto \mathbb{R}$ such that (s.t.)

$$\left.\begin{array}{l} 1)\ P(A) \geq 0 \quad \forall A \in \mathcal{B} \\ 2)\ P(S) = 1 \text{ and } P(\phi) = 0 \\ 3)\ \text{If } A_j \cap A_k = \phi \ \forall j, k \text{ then } P\left(\bigcup_{j=1}^{\infty} A_j\right) = \sum_{j=1}^{\infty} P(A_j) \end{array}\right\} \text{axioms of probability OR Holmogorov's axioms}$$

Propositions

1) $P(A) \leq 1$ and $P(A^c) = 1 - P(A) \quad \forall A \in \mathcal{B}$

2) $P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad \Leftrightarrow$

$P(A \cap B) = P(A) + P(B) - P(A \cup B) \geq P(A) + P(B) - 1$ Bonfevioni's inequality

3) If $B \subset A$ then $P(B) \leq P(A)$

4) If $C_j \in \mathcal{B}$ s.t. $\bigcup_{j=1}^{\infty} C_j = S$ and $C_j \cap C_k = \phi \;\; \forall j, k$ partition of S

then $P(A) = \sum_{j=1}^{\infty} P(A \cap C_j) \;\; \forall A \in \mathcal{B}$

Def. $(S$ $^{\text{sample}}_{\text{space}}$, $\mathcal{B}$ $^{\text{event}}_{\text{space}}$, P $^{\text{probability}}_{\text{function}}$ $)$ is called a PROBABILITY SPACE.

Exercise : Find the number of arrangement of n objects from S. p. 14~15 [CB]

	without replacement	with replacement
ordered		
unordered		

Ex. 1.2.20

Def. If $A, B \in \mathcal{B}$ and if $P(B) > 0$

we define the CONDITIONAL PROBABILITY of A given B by

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Observe that $P(A \cap B) = P(A|B)\,P(B)$

$\qquad\qquad\quad \shortparallel$

$\qquad\qquad P(B \cap A) = P(B|A)\,P(A)$

$\Rightarrow \qquad P(A|B) = P(B|A)\dfrac{P(A)}{P(B)}$

Example of 2 dice

$A = \{$black dice is $1\}$; $B = \{$red dice is $1\}$

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1/36}{1/6} = \frac{1}{6}$$

More generally : (Bayes Rule) Exercise : check for its application

If $\{C_j\}_{j=1}^{\infty}$ is a partition of S

$$P(C_j | B) = \frac{P(B|C_j)\,P(C_j)}{\underbrace{\sum_k P(B|C_k)\,P(C_k)}_{\overset{\shortparallel}{P(B \cap C_k)}}} \Bigg\} = P(B) \quad \text{if } P(B) \neq 0$$

Def. Let $A, B \in \mathcal{B}$, they are INDEPENDENT if $P(A \cap B) = P(A)\,P(B)$

Def. A collection of events $A_1, A_2, \cdots \in \mathcal{B}$ are MUTUALLY INDEPENDENT if

$$\forall A_{i_1}, A_{i_2}, \cdots, A_{i_n} : P\left(\bigcap_{j=1}^{n} A_{i_j}\right) = \prod_{j=1}^{n} P(A_{i_j})$$

I.2 Random variable

Def. a RANDOM VARIABLE X on a probability space $(S, \mathcal{B}, P)$ is a function $X : S \mapsto \mathbb{R}$ satisfying

$$\forall x \in \mathbb{R} : \{ s \in S \mid X(s) \leq x \} \in \mathcal{B}$$

$$\Updownarrow$$

$$X^{-1}((-\infty, x]) \in \mathcal{B} \Leftrightarrow \{X \leq x\} \in \mathcal{B}$$

Def. The CUMULATIVE DISTRIBUTION FUNCTION (CDF) associated to X is the function $F_X : \mathbb{R} \mapsto [0,1]$ defined for any $x \in \mathbb{R}$ by

$$F_X(x) := P(\{X \leq x\})$$

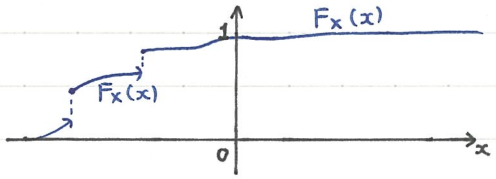
Properties

1) $F_X(x) \leq F_X(y)$ if $x \leq y$ (increasing / monotonic non-decreasing)

2) $\lim_{x \to -\infty} F_X(x) = 0$; $\lim_{x \to \infty} F_X(x) = 1$

3) $\lim_{\varepsilon \searrow 0} F_X(x+\varepsilon) = F_X(x)$ (right continuous)

One has



Thm. Whenever $F : \mathbb{R} \mapsto [0,1]$ satisfies properties 1) ~ 3), there exists $(S, \mathcal{B}, P)$ and a random variable X such that

$$F_X = F.$$

⚠ non-unique

Exercise: Summary ~~t~~ one distribution in the handout for this lecture, writing the properties and an example for this distribution.
(do in pairs)

Remark: From F_X one gets

- $P(a < X \leq b) = P(\{\omega \in S \mid X(\omega) \in (a,b]\}) = F_X(b) - F_X(a)$
- Since $P(X<a) = \lim_{\varepsilon \searrow 0} F_X(a-\varepsilon)$

$$\Rightarrow P(a \leq X \leq b) = F_X(b) - \lim_{\varepsilon \searrow 0} F_X(a-\varepsilon)$$

$$\Rightarrow \underbrace{P(X=a)}_{} = F_X(a) - \lim_{\varepsilon \searrow 0} F_X(a-\varepsilon)$$

see the figure in p. 3

↳ non-zero whenever F_X has a jump at a

Def.

1) X is CONTINUOUS if "F_X has no jump", or
$\lim_{\varepsilon \searrow 0} F_X(a-\varepsilon) = F_X(a) \ \forall a \in \mathbb{R}$

2) X is (ABSOLUTELY) CONTINUOUS if
$\exists f_X : \mathbb{R} \mapsto \mathbb{R}$ such that $F_X(x) = \int_{-\infty}^{x} f_X(y)\,dy$; $f_X \in L^1(\mathbb{R})$
↳ PROBABILITY DENSITY FUNCTION (PDF)

3) X is DISCRETE if $X(S) \subset \mathbb{R}$ is finite or countable.
⇕
(in bijection with $\mathbb{N}$)

4) X can be SINGULAR CONTINUOUS,
or a mixture of abs. continuous, sing. continuous and discrete.

Remark

When X is discrete, the function

$f_X : \mathbb{R} \mapsto \mathbb{R}$, $f_X(x) := F_X(x) - \lim_{\varepsilon \searrow 0} F_X(x-\varepsilon)$ is called the PROBABILITY MASS FUNCTION (pmf).

Observe that $\sum_{x \in \mathbb{R}} f_X(x) = 1$.

And for the pdf $\int_{-\infty}^{\infty} f_X(y)\,dy = 1$.

Use of pdf and pmf:

If $I \subset \mathbb{R}$ (Borel subset of $\mathbb{R}$), then

$$P(x \in I) = P(\{x \in S \mid X(x) \in I\}) = \begin{cases} \int_I f_X(y)\,dy & \text{in continuous case} \\ \sum_{x \in I} f_X(x) & \text{in discrete case} \end{cases}$$

Remark:

The set $\{X(s) \mid s \in S\}$ is called the IMAGE of the random variable X.

We denote it by $\mathrm{Im}(X)$ or $\mathrm{Ran}(X)$

Remark:

If a pdf or a pmf depends on some parameters $\theta_1, \theta_2, \cdots, \theta_n$, then we write

$f(\cdot \mid \theta_1, \cdots, \theta_n) : \mathbb{R} \mapsto \mathbb{R}$

Remark:

The range of a discrete r.v. is countable, while the range of a countinuous one is not countable.

I.3 Transformations and expectations

We always use $(S, \mathcal{B}, P)$ for a prob. space and X for a random variable.

Let $g : \mathbb{R} \mapsto \mathbb{R}$ and consider $g(X) := g \circ X$

$$\begin{tikzcd} S \arrow[r, "X"] \arrow[rr, bend right, "g \circ X"'] & \mathbb{R} \arrow[r, "g"] & \mathbb{R} \end{tikzcd}$$

Lemma: $g(X)$ is a random variable if $g : \mathbb{R} \mapsto \mathbb{R}$ is BOREL MEASURABLE.

It means that $\forall A \in \sigma_B : g^{-1}(A) \in \sigma_B$

← Borel algebra on $\mathbb{R}$

Remark: This is a condition satisfied by almost all functions (for example continuous, or piecewise continuous functions).

Def. If X is discrete or (absolutely) continuous with pmf or pdf f, then the EXPECTED VALUE or MEAN of X is given by

$E(X) := \int_{-\infty}^{\infty} x f_X(x)\,dx$ or $\sum_x x f(x)$

if it converges absolutely

$(:\Leftrightarrow \int_{-\infty}^{\infty} |x| f_X(x)\,dx < \infty$ or $\sum_x |x| f(x) < \infty)$

← Exercise: find an example when $E(X)$ does not exist.

Thm. If X is continuous or discrete and g Borel measurable, then

$E(g(X)) = \int_{\mathbb{R}} g(x) f_X(x)\,dx$

whenever it converges absolutely.

Def. $\mathrm{Var}(X) = E((X - E(X))^2)$

is the VARIANCE whenever it exists.

$\sqrt{\mathrm{Var}(X)}$ is called the STANDARD DERIVATION.

Remark: If $E(X)=\mu$ and $Var(X)=\sigma^2$, then by setting

$$Z := \frac{X-\mu}{\sigma} \quad \text{(the STANDARD FORM)}$$

one has $E(Z)=0$ and $Var(Z)=1$.

Def. The k^{th}-moment of X are defined by $E(X^k)$, and the central k^{th}-moment of X by $E((X-E(X))^k)$ whenever they exist.

Also $M_X(t) := E(e^{tX})$ for $t \in \mathbb{R}$ is called the moment generating function (mgf).

I.4 Multiple random variables

We are going to consider n random variables $X_1, X_2, \cdots, X_N : S \mapsto \mathbb{R}$

$$\Leftrightarrow \qquad \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix} =: \underline{X} : S \mapsto \mathbb{R}^N$$

Def. The joint cumulative distribution function associated with $\underline{X}$ is a function $F_{\underline{X}} : \mathbb{R}^N \longmapsto [0,1]$ defined for $\underline{x} = (x_1, \cdots, x_n)$ by

$$\begin{aligned} F_{\underline{X}}(x) &:= F_{X_1,\cdots,X_N}(x_1, \cdots, x_n) \\ &:= P(\{s \in S \mid X_1(s) \leq x_1, X_2(s) \leq x_2, \cdots, X_n(s) \leq x_n\}) \\ &= P(X_j \leq x_j \ \forall j \in \{1, \cdots, N\}) \end{aligned}$$

If $\underline{X}$ takes only a countable number of values in $\mathbb{R}^N$, the joint probability mass function is defined by

$$f_{\underline{X}}(x_1, \cdots, x_N) = P(X_1 = x_1, \cdots, X_N = x_N)$$

$\underline{X}$ is an absolutely continuous random vector if

$$F_{\underline{X}}(x_1, \cdots, x_N) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \cdots \int_{-\infty}^{x_N} \underset{\uparrow}{f_{\underline{X}}}(y_1, y_2, \cdots, y_N)\, dy_1\, dy_2 \cdots dy_N$$

joint probability density function

If $A \subset \mathbb{R}^N$ (Borel subset) then

$$P(X \in A) = \iint_A \cdots \int f_X(x_1, \cdots, x_n)\, dx_1 \cdots dx_n$$

Def. $X_1, \cdots, X_N$ are independent r.v. if

$$P(X_1 \leq x_1, X_2 \leq x_2, \cdots, X_N \leq x_n) = P(X_1 \leq x_1) P(X_2 \leq x_2) \cdots P(X_N \leq x_n)$$

$$\Updownarrow$$

$$F_{\underline{X}}(x_1, \cdots, x_n) = \prod_{j=1}^{N} F_{X_j}(x_j)$$

$\Updownarrow$ ← only in the case of discrete or abs. continuous

$$f_{\underline{X}}(x_1, \cdots, x_n) = \prod_{j=1}^{N} f_{X_j}(x_j)$$

Remark: From $F_{\underline{X}}$ we can obtain F_{X_1} by the formula

$$\underset{\text{marginal cdf}}{F_{X_1}(x_1)} = \lim_{y\to\infty} P(X_1 \leq x_1, X_2 \leq y, X_3 \leq y, \cdots, X_N \leq y)$$

And similarly in the continuous case

$$\underset{\text{marginal pdf}}{f_{X_1}(x_1)} = \underbrace{\int_{\mathbb{R}} \cdots \int_{\mathbb{R}}}_{N-1 \text{ integrals}} f_{\underline{X}}(x_1, x_2, \cdots, x_n)\, dx_2 \cdots dx_N$$

As before $g: \mathbb{R}^N \longmapsto \mathbb{R}$ continuous

$$E(g(\underline{X})) = \iiint_{\mathbb{R}^N} g(\underline{x}) f_{\underline{X}}(\underline{x})\, \underbrace{d\underline{x}}_{dx_1 dx_2 \cdots dx_N}$$

Def. Conditional pdf or pmf for $\overbrace{X_{k+1}, \cdots, X_N}^{N-k \text{ random v.}}$ knowing $X_1, \cdots, X_k$ is given by

$$f(x_{k+1}, \cdots, x_N | x_1, \cdots, x_k) = \frac{f_{\underline{X}}(x_1, \cdots, x_N)}{f_{X_1, \cdots, X_k}(x_1, \cdots, x_k)}$$

if the denominator is not 0.

(Example 4.2.2 + 4.2.4 for motivation)

I.5 Covarience and correlation

Consider X and Y random variables on S,

with $E(X) =: \mu_X$, $E(Y) =: \mu_Y$, $Var(X) := \sigma_X^2$, $Var(Y) := \sigma_Y^2$ We assume they exist.

Def. The covarience of X and Y is defined by

$$Cov(X,Y) = E\big((X-\mu_X)(Y-\mu_Y)\big) = E(XY) - \sigma_X \sigma_Y$$

The carelation of X and Y is defined by

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \in [-1, 1]$$

Lemma: If X and Y are independent then $Cov(X,Y) = \rho_{XY} = 0$ easy exercise

⚠ The converse is not true.

The covarience measures a linear relationship:

$$|\rho_{XY}| = 1 \text{ iff } P(Y = aX + b) = 1 \ \exists a, b \in \mathbb{R}$$

The pages 3,4 in Appendix 3 are helpful and easy.

II. Random sample

II.1 Basic definations

Def. A random sample of size N is a family of random variables

1) all having the same pdf $\substack{\text{continuous}\\\text{case}}$ or pmf $\substack{\text{discrete}\\\text{case}}$ and

2) all muturally independent.

We speak about $\underset{\downarrow}{\overset{\uparrow}{\text{iid}}}$ random variables.

1) identically distributed

⚠ The condition of independence is very strong ; We take it as a first approximation.

$\Rightarrow f_{\underline{X}}(x_1, \cdots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n)$ because of independence

Def. If $Y : \mathbb{R}^N \longrightarrow \mathbb{R}^d$ is Boral measurable ($\Leftarrow$ continuous)

then $Y(X_1, \cdots, X_N)$ is a real $\overset{\text{if}}{d=1}$ or vector $\overset{\text{if}}{d>1}$ valued random variable

↳ no dependence explicitly on other variables (μ, etc)

called a statistic.

Examples

1) Sample mean: $\overline{X} = \frac{1}{N} \sum_{j=1}^{N} X_j$

2) Sample variance: $S^2 = \boxed{\frac{1}{N-1}} \sum_{j=1}^{N} (X_j - \overline{X})^2$

3) Sample standard divation: $S = \sqrt{S^2}$

Thm. Let $\underline{X} = (X_1, \cdots, X_N)$ be a random sample, with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$.

↖ arbitrary ↗

Then $E(\overline{X}) = \mu$; $Var(\overline{X}) = \frac{\sigma^2}{N}$; $E(S^2) = \sigma^2$

(Thm 5.2.6)

A useful relation: $M_{\overline{X}}(t) = \left(M_{X_i}(t/N)\right)^N$ if M_{X_i} exist

II.2 Sample from a normal distribution

Let $X_1, \cdots, X_N$ be a random sample with $X_j \underset{\uparrow}{\sim} n(\mu, \sigma^2)$ Normal distribution of mean μ and variance σ^2

having or following

with the pdf of $n(\mu, \sigma^2)$ given by

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

One has $E(X_j) = \mu$ and $Var(X_j) = \sigma^2$

Propositions

1) $\overline{X}$ and S^2 are independent random variables (Thm 5.3.1)

2) $\overline{X} \sim n(\mu, \sigma^2/N)$

3) $(N-1)S^2/\sigma^2$ has a chi squared distribution with parameter $N-1$. (χ^2_{N-1})

Remark: If $Y \sim n(\mu, \sigma^2)$, then $\frac{Y-\mu}{\sigma} \sim n(0,1)$ re-scaling. Thus

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{N}} = \sqrt{N}\,\frac{\bar{X}-\mu}{\sigma} \sim n(0,1)$$

If we compute $\sqrt{N}\,\frac{\bar{X}-\mu}{S} = \frac{(\bar{X}-\mu)/(\sigma/\sqrt{N})}{\sqrt{S^2/\sigma^2}}$ $\begin{matrix} \to \sim n(0,1) \\ \to \sim \chi^2_{N-1}/N-1 \end{matrix}$

The ratio is the student's distribution (≡ t-distribution) and does not depend on σ.

⇝ We can deduce μ.

II.3 Order statistics

Def. The order statistics of a random sample $X_1, \cdots, X_N$ are the sample values placed in ascending order:

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(N)}.$$

Examples: $X_{(1)} = \min\limits_{j \in \{1,\cdots,N\}} X_j$; $\cdots$; $X_{(N)} = \max\limits_{j \in \{1,\cdots,N\}} X_j$

Based on them one has

- Sample range: $R := X_{(N)} - X_{(1)}$
- Sample median: $M := \begin{cases} X_{(\frac{N+1}{2})} & \text{if } N \text{ odd} \\ \frac{1}{2}(X_{(N/2)} + X_{(N/2+1)}) & \text{if } N \text{ even} \end{cases}$

⚠ Sample median ≠ sample mean

- For $p \in (0,1)$ we define the 100p^th sample percentile to be the observation s.t. Np observations are smaller and $N(1-p)$ are larger

$$= \begin{cases} X_{(\{Np\})} & \text{if } \frac{1}{2N} < p < \frac{1}{2} \\ X_{(N+1-\{N(1-p)\})} & \text{if } \frac{1}{2} < p < 1 - \frac{1}{2N} \end{cases} \text{ with}$$

$$i - \tfrac{1}{2} \leq b < i + \tfrac{1}{2} \text{ with } i \in \mathbb{N} \Rightarrow \{b\} := i \quad \text{nearest integer}$$

One often uses

lower quartile := 25th sample percentile

upper quartile := 75th sample percentile

No. 4

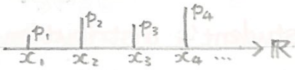
Date 2019 5 15

Computation related to $X_{(j)}$

Prop. (discrete case)

Suppose $X_1, \cdots, X_N$ are iid with X_j discrete with

$\mathrm{Ran}(X_j) = \{x_i\}_{i \in \mathbb{N}}$



and s.t. $x_i \leq x_{i+1}$ and $f_{X_j}(x_i) =: p_i$ with $\sum_{i \in \mathbb{N}} p_i = 1$. Then

$$P(X_{(j)} \leq x_i) = \sum_{k=j}^{N} \binom{N}{k} P_i^k (1-P_i)^{N-k} \quad \text{with}$$

$$P_i = F_{X_j}(x_i) = P(X_j \leq x_i) = \sum_{k=1}^{i} p_k$$

Proof

Set $Y := \overset{\text{number}}{\#}\{X_j \leq x_i \text{ for } j = 1, \cdots, N\}$ number of X_j taking a value smaller than x_i

But $\{X_j\}$ are i.i.d and $P(X_j \leq x_i) = P_i$

Then $Y \sim$ binomial (N, P_i).

Then $\{X_{(j)} \leq x_i\} = \{Y \geq j\}$ key point (both are subsets of $S \times \cdots \times S$)

$$\Rightarrow \underbrace{P(X_{(j)} \leq x_i)}_{\text{want to compute}} = P(Y \geq j) \quad \text{because of binomial}$$

$$= \sum_{k=j}^{N} P(Y=k) \overset{\downarrow}{=} \sum_{k=j}^{N} \binom{N}{k} P_i^k (1-P_i)^{N-k} \quad (*)$$ □

Prop. (continuous case) (Thm. 5.4.4)

If X_j has cdf F and pdf f, then

$$\overset{}{f_{X_{(j)}}}(x) = \frac{N!}{(j-1)!(N-j)!} f(x) F(x)^{j-1} (1-F(x))^{N-j}$$

↑ pdf for having j r.v. X_k taking a value smaller than x.

II.4 Computing with random samples

Aim: compute $(*)$ numerically with random samples.

Idea: use the weak law of large number (from Appendix 3).

Consider $\underline{X}$ of i.i.d. r.v. with $E(X_j) = \mu$ and $\mathrm{Var}(X_j) = \sigma^2$

Then $\overline{X_N} = \frac{1}{N} \sum_{j=1}^{N} X_j$ (sample mean);

$\overline{X_N} \xrightarrow{N \to \infty} \mu$ in probability

$\Leftrightarrow \forall \varepsilon > 0,\ P(|\overline{X_N} - \mu| \geq \varepsilon) \xrightarrow{N \to \infty} 0$

1) Consider $X_1, \cdots, X_N$ discrete r.v. satisfying the prescribed distribution of the previous propersition.

2) Set $Y_k := \begin{cases} 1 & \text{if at least } j \ X_\ell \text{ take a value } \leq x_i \\ 0 & \text{otherwise} \end{cases}$

3) Observe that $Y_k \sim \text{Bernoulli}(q)$ with

$q = P(X_{(j)} \leq x_i)$, and $E(Y_k) = q$

4) Then $\{Y_k\}$ by repeating the experiment is an i.i.d family with $E(Y_k) = q$.

By weak law of large number, $Y \xrightarrow[\text{experiments}]{\text{for many}} q$

What is missing : how to generate random numbers following a prescribed distributio

There are solutions. See section 5.6 (exercise : report on this)

Ⅲ Data reduction

Idea: Suppose $\underline{X}$ is a random sample with $X_j \sim \underset{\leftarrow \text{certain distribution}}{f}(\cdot\,|\overset{\swarrow\text{unknown parameter}}{\theta})$. e.g. $n(\underset{\uparrow \atop \theta}{\mu}, \underset{\uparrow \atop \theta}{\sigma^2})$ only interested in one parameter not in the others

Can we find a statistic $T(\underline{X})$ which keeps all information on θ?

(with the aim of reducing the necessary information)

Ⅲ.1 Sufficient statistics

Sufficient principle: $T(X)$ is a sufficient statistic for θ if two sample points $\underline{x}$ and $\underline{y}$ with $T(\underline{x}) = T(\underline{y})$ One gets same influence on θ.

Def. $T(\underline{X})$ is a sufficient statistic for θ if the conditional pmf or pdf of the sample $\underline{X}$ given $T(\underline{X})$ does not depend on θ.

Example: Consider $X_i \sim \text{Bernoulli}(\theta)$ with $\theta \in (0,1)$ and $T(\underline{X}) := \sum_{j=1}^{N} X_j$ with N fixed.

e.g. $\underline{x} = (0,1,0,0,0,1,1,0,\cdots,1)$ e.g. $T(\underline{x}) = 25$ number of 1 in $\underline{x}$

Observe that $T(\underline{X}) \sim \text{binormial}(N,\theta)$.

Set $t := \sum_{i=1}^{N} x_i$. Then the conditional pmf of $\underline{X}$ given $T(\underline{X})$ is

$$\frac{\prod_{j=1}^{N} \text{Bern}(x_j|\theta)}{\text{binomial}(t|N,\theta)} = \frac{\prod_{j=1}^{N} \theta^{x_j}(1-\theta)^{1-x_j}}{\binom{N}{t}\theta^t(1-\theta)^{N-t}} = \frac{\theta^{\sum_{j=1}^{N} x_j}(1-\theta)^{\sum_{j=1}^{N}(1-x_j)}}{\binom{N}{t}\theta^t(1-\theta)^{N-t}} = \frac{\theta^t(1-\theta)^{N-t}}{\binom{N}{t}\theta^t(1-\theta)^{N-t}}$$

$$= \frac{1}{\binom{N}{t}} \text{ indep of } \theta.$$

$\Rightarrow$ T is a sufficient statistics for θ $\Rightarrow$ any $\underline{x}, \underline{y}$ with $\sum x_j, \sum y_j$ provides the same information on θ.

Thm. (Factorization Thm)

A statistic $T(\underline{X})$ is sufficient for θ if and only if

$$f_{\underline{X}}(\underline{x}|\theta) = g(T(\underline{x})|\theta)\,\underset{\nwarrow \text{indep of } \theta}{h(\underline{x})} \quad \forall \underline{x} \in \mathbb{R}^N$$

Example: $X_j \sim (\mu, \sigma^2)$ and consider $\theta = \mu$. σ^2 is known.

$$f_{\underline{x}}(\underline{x}|\mu,\sigma^2) = \prod_{j=1}^{N}(2\pi\sigma^2)^{-\frac{1}{2}}\exp\left(-\frac{(x_j-\bar{x}+\bar{x}-\mu)^2}{2\sigma^2}\right) \quad \bar{x} := \frac{1}{N}\sum_{j=1}^{N} x_j$$

$$= (2\pi\sigma^2)^{-\frac{N}{2}}\exp\left(-\frac{\sum_{j=1}^{N}(x_j-\bar{x})^2 + N(\bar{x}-\mu)^2}{2\sigma^2}\right)$$

$$= \underbrace{e^{-N(\bar{x}-\mu)^2/2\sigma^2}}_{g(\bar{x}|\mu)}\underbrace{(2\pi\sigma^2)^{-\frac{N}{2}}e^{-\sum_j(x_j-\bar{x})^2/2\sigma^2}}_{\text{indep of } \mu}$$

$\Rightarrow$ $T(\underline{X}) := \bar{X}$ is a sufficient statistic for μ.

Remarks:

1) Sufficient statistics are θ-dependent and model dependent(X_j).

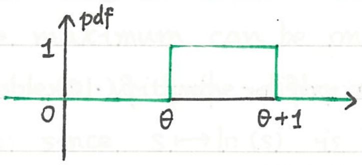
2) Vector valued $T(\underline{X})$ and vector parameters $\boldsymbol{\theta}$ are possible.

3) We can look at minimal sufficient statistics. (maybe not unique)

(any other sufficient statistics should be a function of a minimal one)

Def. A statistic $T(\underline{X})$ whose pdf or pmf does not depend on θ is called an ancillary statistic.

Example: Let $\underline{X} = (X_1, \cdots, X_N)$ a random sample with $X_j \sim \text{unif}(\theta, \theta+1)$ with



$$f_{X_j}(x) = \begin{cases} 1 & \text{if } x \in (\theta, \theta+1) \\ 0 & \text{otherwise} \end{cases}$$

Then we set $\underset{\text{range statistic}}{R(\underline{X})} = \overset{\text{maximal stat.}}{X_{(N)}} - \underset{\text{minimal stat.}}{X_{(1)}}$

Then the pdf of $R(\underline{X})$ is the function

$x \longmapsto N(N-1)x^{N-2}(1-x)$ for $x \in (0,1)$ exercise: someone could show this.

which is θ-independent.

$\Rightarrow R(\underline{X})$ is an ancillary statistic.

⚠ A sufficient statistic can be related to an ancillary statistic.

Example: in the previous example,

$(R(\underline{X}), M(\underline{X}) := \frac{1}{2}(X_{(1)} + X_{(N)}))$ is a sufficient statistic for θ.

It means that both information are necessary,

and $R(\underline{X})$ alone does not ~~d~~say anything on θ.

III.2 Likelihood principle

Def. Let $\underline{X} := (X_1, \cdots, X_N)$ be a r. sample with joint pdf or pmf $f_{\underline{X}}(\cdot|\theta)$.

For a given observation $\underline{x}$, the function

$$\theta \longmapsto L(\theta|\underline{x}) := f_{\underline{X}}(\underline{x}|\theta)$$

is called the likelihood function at $\underline{x}$.

Likelihood principle

If $\underline{x}$ and $\underline{y}$ satisfy $L(\theta|\underline{x}) = cL(\theta|\underline{y})$ $\forall\theta$ and a fixed c,

then the inference on θ from $\underline{x}$ and $\underline{y}$ should be the same.

Example: For $\underline{X} = (X_1, \cdots, X_N)$ with $X_j \sim n(\mu, \sigma^2)$, one has

$$f_{\underline{X}}(\underline{x}|\mu,\sigma^2) = e^{-N(\bar{x}-\mu)^2/2\sigma^2}(2\pi\sigma^2)^{-N/2}e^{-\sum_j (x_j-\bar{x})^2/2\sigma^2}$$

For $\theta = \mu$, one has

$$L(\theta|\underline{x}) = cL(\theta|\underline{y}) \text{ if } \bar{x} = \bar{y} \text{ and } \overset{\text{indep of }\theta}{\overset{\downarrow}{c}} = \exp\left(\sum_j \frac{-(x_j-\bar{x})^2 + (y_j-\bar{y})^2}{2\sigma^2}\right)$$

Thus from this principle, if $\bar{x} = \bar{y}$ the inference on θ is same from $\underline{x}$ or $\underline{y}$.

IV Point estimation

Aim: infer expressions for some parameters θ from a random sample $\underline{X}$ or from a realized measurement $\underline{x}$.

2 parts: 1) finding estimation for θ (several methods)

2) evaluation of this estimations

Framework: $\underline{X} = (X_1, \cdots, X_N)$ a r. s. with $X_j = X$ a pdf or pmf $f(\cdot|\theta)$.

$f_{\underline{X}}(\cdot|\theta)$ is the joint pdf or pmf.

1. 1 Method of moments ($k \in \mathbb{N}$)

$$\mu_k := E(X^k) = \int_{\mathbb{R}} x^k f(x|\theta)\, dx$$

Set $m_k := \frac{1}{N}\sum_{j=1}^{N} X_j^k$ and solve the system of equation

$$\begin{cases} m_1 = \mu_1 \\ m_2 = \mu_2 \\ \vdots \\ m_K = \mu_K \leftarrow \text{depending on the size of } \theta \end{cases}$$

Example (used for estimating the crime rate)

$X_j = X \sim \text{binomial}(k, p)$ with $k \in \mathbb{N}$, $p \in (0,1)$

↑number of crimes reported to the police every day

↑ ↑reporting rate of crime

"real" number of crime every day

$\mu_1 = kp$; $\mu_2 = kp(1-p) + kp^2$ (from table)

$$\begin{cases} m_1 = \overline{X} = kp \\ m_2 = \frac{1}{N}\sum_{j=1}^{N} X_j^2 = kp(1-p) + kp^2 \end{cases}$$

The solution is

$$k = \frac{\overline{X}^2}{\overline{X} - \frac{1}{N}\sum_j (X_j - \overline{X})^2} \; ; \; p = \frac{\overline{X}}{k}$$

1.2 Maximum likelihood estimator (MLE)

Recall $L(\theta|\underline{x}) = f_{\underline{X}}(\underline{x}|\theta)$

Def: For any fixed sample point $\underline{x}$ let $\hat{\theta}(\underline{x})$ be the point where $\theta \longmapsto L(\theta|\underline{x})$ takes its maximal value. There might be more than one. Then the maximum likelihood estimator for θ on a sample $\underline{X}$ is $\hat{\theta}(\underline{X})$.

Then given $\underline{x}$ we estimate θ by $\hat{\theta}(\underline{x})$.

Justification: The maximum likelihood estimator is the parameter point which is observed most likely by the sample.

Drawbacks:

- requires heavy computation (compute derivatives and study the Hessian matrix)
- not always unique
- the maximum can be on the boundary of the parameter space
- problems with the discrete parameters.

Remark: since $s \mapsto \ln(s)$ is strictly increasing, one can also study $\theta \longmapsto \ln L(\theta|\underline{x})$ and get the same maximum.

One speaks about log likelihood functions.

Example

$$X_j \sim n(0,1)$$

$$L(\theta|\underline{x}) = \frac{1}{(2\pi)^{N/2}} e^{-\frac{1}{2}\sum_{j=1}^{N}(x_j-\theta)^2}$$

One has

$$\frac{\partial L}{\partial \theta}(\theta|\underline{x}) = 0 \iff \sum_{j=1}^{N}(x_j\ \theta) = 0 \iff \theta = \frac{1}{N}\sum_{j=1}^{N} x_j$$

and it turns out to be a maximum. Thus

$$\hat{\theta}(\underline{x}) = \frac{1}{N}\sum_{j=1}^{N} x_j$$

1.3 Bayes estimator

Idea: we suppose that θ follows a certain prob. distribution (called prior distribution) which is going to be updated with the sample to a posterior distribution. The update is based on Bayes formula.

Recall that Bayes formula reads:

If $\{C_j\}$ is a partition of the sample space S $(:\iff \bigsqcup_j C_j = S)$ then

$$P(C_j|B) = \frac{P(B|C_j)P(C_j)}{\sum_k P(B|C_k)P(C_k)}$$

Example

Let $\underline{X} = (X_1, \cdots, X_N)$ a r.s. with $X_j \sim \text{Bernoulli}(\overset{=\theta\in(0,1)}{p})$ and suppose $p \sim \overset{\text{prior distribution}}{\text{beta}}(\underbrace{\alpha,\beta}_{\text{fixed}})$

Set $Y = \sum_{j=1}^{N} X_j \sim$ bino~~x~~mial (N,p)

$$P(C_j, B) \rightsquigarrow f(\theta|Y=y) \quad \text{posterior distribution}$$

$$\frac{P(B|C_j)P(C_j)}{\sum_k P(B|C_k)P(C_k)} \rightsquigarrow \frac{\text{binomial}(y|N,p)\,\text{beta}(p|\alpha,\beta)}{\int \text{binomial}(y|N,s)\,\text{beta}(s|\alpha,\beta)\,ds}$$

This computation will be finished next week.

$P(C_j) \rightsquigarrow \theta \longmapsto \pi(\theta)$ prior distribution

$P(C_j|B) \rightsquigarrow \theta \longmapsto \pi_{post}(\theta|Y=y)$ posterior distribution

$P(B|C_j) \rightsquigarrow y \longmapsto f_Y(y|\theta)$

$\sum_k P(B|C_k)P(C_k) \rightsquigarrow y \longmapsto \int f_Y(y|\theta)\pi(\theta)d\theta$

Suppose $Y = \sum_{j=1}^{N} X_j \sim \text{binomial}(N, \underset{\theta\in[0,1]}{\underset{\downarrow}{p}})$

$\pi(p) \sim \text{beta}(\underset{\text{fixed}}{\underbrace{\alpha,\beta}})$

$$\Rightarrow \pi_{post}(p|y) = \frac{\text{binomial}(y|N,p)\,\text{beta}(p|\alpha,\beta)}{\int_0^1 \text{binomial}(y|N,s)\,\text{beta}(s|\alpha,\beta)\,ds} \begin{matrix}\}①\\ \}②\end{matrix}$$

$$① = \binom{N}{y} p^y (1-p)^{N-y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

$$= \binom{N}{y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{y+\alpha-1}(1-p)^{N-y+\beta-1}$$

↙ Gamma function:

$$② = \int ① ds = \binom{N}{y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(y+\alpha)\Gamma(N-y+\beta)}{\Gamma(N+\alpha+\beta)}$$

$$\Rightarrow \pi_{post}(p|y) = \frac{①}{②} = \frac{\Gamma(N+\alpha+\beta)}{\Gamma(y+\alpha)\Gamma(N-y+\beta)} p^{y+\alpha-1}(1-p)^{N-y+\beta-1}$$

$$= \text{beta}(y+\alpha, N-y+\beta)(p)$$

By taking expectation of the prior and posterior distributions

$$E_{prior}(p) = \frac{\alpha}{\alpha+\beta}$$

$$E_{post}(p) = \frac{y+\alpha}{N+\alpha+\beta} = \frac{N}{\alpha+\beta+N}\frac{y}{N} + \frac{\alpha+\beta}{\alpha+\beta+N}\boxed{\frac{\alpha}{\alpha+\beta}}$$ ↪ estimation on p without any prior idea

The posterior expectation is a linear combination of prior expectation and experimental expectation.

Remark: As N becoming large, the prior expectation is becoming less and less importa

Remark: The linear combination is not an accident;

$\rightsquigarrow$ notion of conjugate family of a distribution.

IV.2 Evaluating estimators

Different estimators for θ can give you different values for θ.

Which one should we choose?

2.1 Mean square error

Framework $\underline{X} = (X_1, \cdots, X_N)$ with $X_j = X \sim f(\cdot|\theta)$.

Let $W = W(\underline{X})$ be an estimator ($\equiv$ statistic) for θ.

Def. The mean square error (MSE) of W is defined by

$$E((W-\theta)^2\underset{\uparrow}{)} \quad \text{still a function of } \theta$$

⚠ We made the choice of $s \mapsto s^2$. Later there will be a generation.

Observe that

$$\begin{aligned} E((W-\theta)^2) &= E((W-E(W)+E(W)-\theta)^2) \\ &= E((W-E(W))^2) + \underbrace{E(2(W-E(W))\overbrace{(E(W)-\theta)}^{\text{constant}})}_{\overset{\shortparallel}{0}} + E(\underbrace{(E(W)-\theta)^2}_{\text{constant}}) \\ &= Var_\theta(W) - (E_\theta(W)-\theta)^2 \end{aligned}$$

Def. If W is an estimator for θ, we set

$$Bias_\theta(W) := E_\theta(W) - \theta$$

If $Bias_\theta(W) = 0$ we say that W is unbiased.

$\Rightarrow E_\theta((W-\theta)^2) = Var_\theta(W) + (Bias_\theta(W))^2$

Remark: In Section II.1, we saw that if $E(X) = \mu$ and $Var(X) = \sigma^2$ then

$$E(\underset{\substack{\uparrow \\ \text{sample mean}}}{\bar{X}}) = \mu \quad \text{and} \quad E(\underset{\substack{\uparrow \\ \text{sample variance}}}{S^2}) = \sigma^2$$

$\Rightarrow E(\bar{X})$ and $E(S^2)$ are unbounded estimators for μ and σ^2 resp.

Now if $X \sim n(\mu, \sigma^2)$ then $E((\bar{X}-\mu)^2) = Var(\bar{X}) = \frac{\sigma^2}{N}$

and $E((s^2-\sigma^2)^2) = \frac{e}{N-1}\sigma^4$

Question: Can we find better with a smaller MSE estimator for μ and σ^2?

Remark: for $X \sim n(\mu, \sigma^2)$ consider

$$\tilde{S}^2 := \frac{1}{N}\sum_{j=1}^{N}(X_j - \bar{X})^2 = \frac{N-1}{N}S^2$$

$\Rightarrow E(\tilde{S}^2) = \frac{N-1}{N}\sigma^2 \qquad \Rightarrow \tilde{S}^2$ is biased

But $E((\tilde{S}^2 - \sigma^2)^2) = \frac{2N-1}{N^2}\sigma^4 < \frac{2}{N-1}\sigma^4$

$\Rightarrow \tilde{S}^2$ has a lower MSE, despite being biased.

⇝ Finding a better estimator is not a simple question.

We shall consider only unbiased estimators.

Def. An estimator W^* for θ parameter with value unknown is the best unbiased estimator for θ if $\mathrm{Var}_\theta(W^*) \leq \mathrm{Var}_\theta(W)$ for any estimator W for θ for any θ value of parameter θ with W^* and W unbiased.

Thm. (7.3.19) The best unbiased estimator is unique if it exists.

But it does not tell you how to find it.

Question: Can we get $\mathrm{Var}_\theta(W) = 0$?

Thm: (7.3.9 + 7.3.10) (Gramer-Rao inequality) (based on Cauchy-Schwartz inequality)

If $\frac{d}{d\theta}E_\theta(W) \equiv \frac{d}{d\theta}\int_{\mathbb{R}^N} W(\underline{x}) f_{\underline{X}}(\underline{x}|\theta)\,d\underline{x} = \int_{\mathbb{R}^N}\frac{\partial}{\partial\theta}[W(\underline{x}) f_{\underline{X}}(\underline{x}|\theta)]\cdot d\underline{x}$

and $\mathrm{Var}_\theta(W) < \infty$ then

$$\mathrm{Var}_\theta(W) \geqslant \frac{\left(\frac{d}{d\theta}\overbrace{E_\theta(W)}^{=\theta}\right)^2 \Big\}=1}{\underbrace{NE_\theta[(\frac{\partial}{\partial\theta}\ln f_{\underline{X}}(\cdot|\theta))^2]}_{\substack{\text{information number}\\ \text{or Fisher information}}}}$$

If $\mathrm{Var}_\theta(W)$ saturates satisfies equality in this inequality,

then W is the best unbiased estimater for any value of θ.

⚠ Otherwise for two W the $\mathrm{Var}_\theta(W)$ may cross.

Remark: The following relation holds:

If $\frac{d}{d\theta}E\left(\frac{\partial}{\partial\theta}\ln f(\cdot|\theta)\right) = \int\frac{\partial}{\partial\theta}\left[\frac{\partial}{\partial\theta}\ln f(\cdot|\theta)\right] f(x|\theta)\,dx$ then

$$E_\theta\left[\left(\frac{\partial}{\partial\theta}\ln f(\cdot|\theta)\right)^2\right] = -E_\theta\left(\frac{\partial^2}{\partial\theta^2}\ln f(\cdot|\theta)\right)$$

2.2 Loss function optimality

- So far, $\mathrm{MSE} = E((W-\theta)^2) = E(\mathcal{L}(W,\theta))$ with $\mathcal{L}(s,\theta) = (s-\theta)^2$

but other function $\mathcal{L}$ = loss function could be chosen. For example:

- $\mathcal{L}(s,\theta) = |s-\theta|$
- $\mathcal{L}(s,\theta) = \frac{(s-\theta)^2}{|\theta|+1}$
- $\mathcal{L}(s,\theta) = \begin{cases}(s-\theta)^2 & \text{for } s<\theta\\ 10(s-\theta)^2 & \text{for } s\geqslant\theta\end{cases}$
- $\mathcal{L}(s,\theta) = \frac{s}{\theta} - 1 - \ln\left(\frac{s}{\theta}\right)$ ...

Then the risk function is defined by $R(\theta, W) := E(\mathcal{L}(W,\theta))$

and we want to have a value of R close to 0.

Remark: the nice feature $E((W-\theta)^2) = \mathrm{Var}(W) + \mathrm{Bias}(W)^2$

will not be possible in general.

⇒ minimizing $R(\theta, W)$ can be complicated for other function $\mathcal{L}$.

⇒ It is a complicated question.

V Hypothesis test

Def. A hypothesis is a statement about a population of parameter !
Two complementary hypotheses are called the null hypothesis (denoted by H_0) and the alternative hypothesis (denoted by H_1)

Example: Let $\Theta_0 \subset \overset{\text{parameter space}}{\Theta}$

$$H_0 : \theta \in \Theta_0 \quad \text{and} \quad H_1 : \theta \in \underbrace{\Theta \setminus \Theta_0}_{\Theta_0^c}$$

Def. A hypothesis test is a rule which specifies for which values of $\underline{x}$ the hypothesis H_0 is accepted or rejected.

Def. The subset of the sample space for which H_0 is rejected is denoted by R and called the rejection region. More precisely:

$$R = \{\underline{x} \mid H_0 \text{ is rejected based on } \underline{x}\}$$

Example: A hypothesis test consisting in checking if the static $W(\underline{x}) \in [1,3]$ (for example $\bar{x} \in [1,3]$) in which case we accept H_0
(for example $H_0 : \underset{\text{"}\theta\text{"}}{W(\underline{X})} = 1.5$)

V.1 Finding tests

1) Likelyhood ratio test:

If $\underline{X} = (X_1, \cdots, X_N)$ with $X_j \sim f(\cdot|\theta)$, and recall that

$$L(\theta|\underline{x}) = f_{\underline{x}}(\underline{x}|\theta) = \prod_{j=1}^{N} f(x_j|\theta)$$

The Likelihood ratio test (LRT) for testing $\theta \in \Theta_0 \subset \Theta$ consists in defining

$$\lambda(\underline{x}) = \frac{\sup_{\theta \in \Theta_0} L(\theta|\underline{x})}{\sup_{\theta \in \Theta} L(\theta|\underline{x})}$$

$\cdot \in [0,1]$
is big if $\theta \in \Theta_0$ is likely

The rejection region is $R = \{\underline{x} \mid \lambda(\underline{x}) \leq c\}$ for a fixed $c \in (0,1)$.
(recall $H_0 : \theta \in \Theta_0$)

Example 1: $X_j \sim n(\theta, 1)$, $H_0 : \theta = \overset{\text{given}}{\theta_0}$. Then

$$\lambda(\underline{x}) = \frac{(2\pi)^{-\frac{N}{2}} \exp\left(-\sum_j (x_j \overset{-\bar{x}+\bar{x}}{-} \theta_0)^2/2\right)}{(2\pi)^{-\frac{N}{2}} \exp\left(-\sum_j (x_j - \bar{x})^2/2\right)} = \exp\left(-N(\bar{x}-\theta_0)^2/2\right)$$

$$\lambda(\underline{x}) \leq c \Leftrightarrow -N(\bar{x}-\theta_0)^2/2 \leq \ln(c) \Leftrightarrow |\bar{x}-\theta_0| \geq \sqrt{-\tfrac{2}{N}\ln(c)}$$

Example 2: Suppose $X_j \sim n(\mu, \overset{\text{unknown}}{\sigma^2})$; $H_0 : \mu = \mu_0$ (given)

$\lambda(\underline{x}) = \cdots \le c \iff \boxed{\frac{\bar{x}-\mu_0}{S/\sqrt{N}}} \ge c$ (computation done in Appendix 7)

$\sqrt{\text{sample variance}}$ ↗ S ; ↳ follows a student t distribution

2) Bayesian test

Prior distribution π for θ, and we want to compute π_{post} for θ.

$\rightsquigarrow$ prob. distribution for θ

Then we are going to accept $H_0 : \theta \in \Theta_0$ if

$$P_{post}(\theta \in \Theta_0 \mid \underline{x}) = \int_{\Theta_0} \pi_{post}(\theta \mid \underline{x})\, d\theta \ge c$$

Example: $X_j \sim n(\theta, \sigma^2)$; $\pi \sim n(\mu, \tau^2)$ with σ^2, μ, τ^2 known.

↖ not an accident

$H_0 : \theta \le \theta_0$ fixed

$$\pi_{post}(\cdot \mid \underline{x}) = \pi_{post}(\cdot \mid \bar{x}) \overset{\text{Ex 7.22}}{=} n\left(\frac{N\tau^2\bar{x}+\sigma^2\mu}{N\tau^2+\sigma^2}, \frac{\sigma^2\tau^2}{N\tau^2+\sigma^2}\right)$$

If we choose $c = \frac{1}{2}$, we get by symmetry

$$P_{post}(\theta \le \theta_0 \mid \underline{x}) \ge \frac{1}{2} \iff \frac{N\tau^2\bar{x}+\sigma^2\mu}{N\tau^2+\sigma^2} \le \theta_0$$

$$\iff \bar{x} \le \theta_0 + \frac{\sigma^2(\theta_0-\mu)}{N\tau^2}$$

3) Union-intersection and intersection-union tests

Case 1: $H_0 : \theta \in \bigcap_{\gamma \in \Gamma} \Theta_{0,\gamma} \Rightarrow R = \bigcup_{\gamma \in \Gamma} R_\gamma$ (U.-I. test)

$\Gamma \rightarrow$ family of condition

Case 2: $H_0 : \theta \in \bigcup_{\gamma \in \Gamma} \Theta_{0,\gamma} \Rightarrow R = \bigcap_{\gamma \in \Gamma} R_\gamma$ (I.-U. test)

V.2 Evaluating the tests

Idea: evaluate the possible mistakes of a hypothesis test:

		Decision (on $\underline{x}$)	
		accept H_0	reject H_0
"Truth"	H_0	♡	Type-I error
	H_1	Type-II error	♡

If $\theta \in \Theta_0$ and R denotes the rejection region (for H_0), then an error of type I takes place if $\underline{x} \in R$. Its probability is $P_\theta(\underline{x} \in R)$

Conversely if $\theta \in \Theta_0^c$, the prob. of a type II error is $1 - P_\theta(\underline{x} \in R)$

Def. The power function of a hypothesis test $_{(c)}$ with the rejection region R is

$$\beta : \theta \longmapsto P_\theta(\underset{\sim}{x} \in R)$$

Ideal situation : if $\theta \in \Theta_0$, $\beta(\theta) \sim 0$

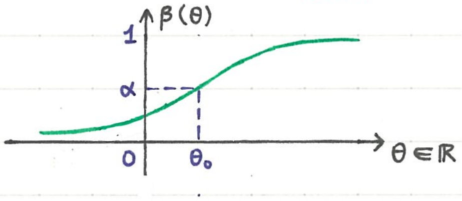
if $\theta \in \Theta_0^c$, $\beta(\theta) \sim 1$

Example : $X_j \sim n(\theta, \underset{\nwarrow \text{known}}{\sigma^2})$; $H_0 : \theta \leqslant \theta_0 \Leftrightarrow \Theta_0 = (-\infty, \theta_0]$

The hypothesis H_0 is rejected if $\frac{\bar{X} - \theta_0}{\sigma/\sqrt{N}} \geqslant \underset{\nwarrow \text{fixed later on}}{c}$. Thus

$$\beta(\theta) = P_\theta\left(\frac{\bar{X}-\theta_0}{\sigma/\sqrt{N}} \geqslant c\right) = P_\theta\left(\frac{\bar{X}-\theta}{\sigma/\sqrt{N}} \geqslant c + \frac{\theta_0 - \theta}{\sigma/\sqrt{N}}\right)$$

$$= P_\theta\left(Z \geqslant c + \frac{\theta_0-\theta}{\sigma/\sqrt{N}}\right) \text{ with } Z \sim n(0,1)$$



with $\alpha = P(Z \geqslant c)$

Experimentor's wish 1 :

Error type I with prob $\leqslant 0.1$

$\Leftrightarrow \beta(\theta) \leqslant 0.1 \ \forall \theta \leqslant \theta_0$ $\quad \Leftrightarrow \beta(\theta_0) = 0.1$ since β is increasing

$\Leftrightarrow P_\theta(Z \geqslant c) = 0.1$ $\quad \overset{\text{table}}{\Leftrightarrow} \underline{c = 1.28}$

Experimentor's wish 2 :

For $\theta \geqslant \theta_0 + \sigma$, Error of type II should have a prob. $\leqslant 0.2$

$\Leftrightarrow \beta(\theta_0 + \sigma) = 0.8$

$\Leftrightarrow P_\theta\left(Z \geqslant c + \frac{-\sigma}{\sigma/\sqrt{N}}\right) = 0.8 \overset{\text{tables}}{\Leftrightarrow} -0.84 \geqslant 1.28 - \underset{\cdots}{\sqrt{N}}$

$\Leftrightarrow \underline{N \geqslant 5}$

⚠ It is not always possible to adjust both errors.

We usually concentrate on error of type I.

Def. For any $\alpha \in [0,1]$, a test of power function β

is a size α-test if $\sup\limits_{\theta \in \Theta_0} \beta(\theta) = \alpha$, and

is a level α-test if $\sup\limits_{\theta \in \Theta_0} \beta(\theta) \leqslant \alpha$.

} way to measure the prob. of type-I error.

One extra make-up class on June 25, same time

Usually we fix $\alpha = 0.05,\ 0.01$ or 0.1. This provides an upper bound for type-I error

In example 1 of Section V.1 with

$\underline{X} = (X_1, \cdots, X_N),\ X_j \sim n(\theta, 1)$

we obtained from the LRT that $\lambda(x) \leq c \in (0,1) \Leftrightarrow$

$\sqrt{-2\ln(c)} \leq \left| \frac{\bar{x} - \theta_0}{1/\sqrt{N}} \right| \sim n(0,1)$ for

$H_0 : \theta = \theta_0$ for a fixed θ_0

Thus, if we set $Z \sim n(0,1)$ and $P(Z \geq z_{\alpha/2}) \overset{\text{def}}{=} \frac{\alpha}{2}$

Thus, we can look for $z_{\alpha/2}$ s.t. $P(Z \geq \sqrt{-2\ln(c)}) = \frac{\alpha}{2}$

given by tables $(\Leftrightarrow P(|Z| \geq \sqrt{-2\ln(c)}) = \alpha)$

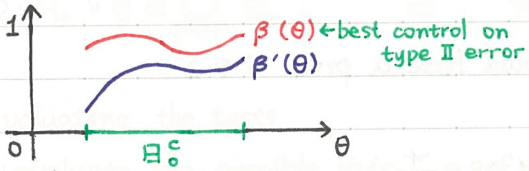
$\Leftrightarrow \sqrt{-2\ln(c)} = z_{\alpha/2}$ $\Leftrightarrow c = e^{-z_{\alpha/2}^2/2}$

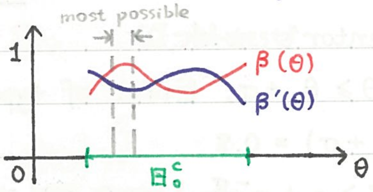
$\rightsquigarrow$ a level-α test

Def. Let C_α be the set of all level-α sets for an level-α test H_0.
A test in C_α is the uniformly most powerful (UMP) C_α-test if the corresponding power function β satisfies

$\beta(\theta) \geq \beta'(\theta)$ for any $\theta \in \Theta_0^c$ and

for the power function β' of any other test in C_α





Remark. Since different power function can cross, it is rarely possible to define the UMP C_α-test. 2 solutions:

1) Further divide Θ_0^c and take the UMP C_α-test on each part;

2) The plausible part of Θ_0^c is neglected.

V.3 p-values ($\equiv$ statistical significance)

$\leadsto$ give less arbitrariness to the value c

Aim: A p-value reports the result of a test on a more continuous scale rather than "accept H_0" or "reject H_0"

Def: Consider $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_0^c$ and let $W(\underline{X})$ be a (test) statistic such that larges values of $W(\underline{X})$ give evidence that H_1 is true. (for example $W(\underline{X}) = |X - \theta|$) For any $\underline{x}$ we set

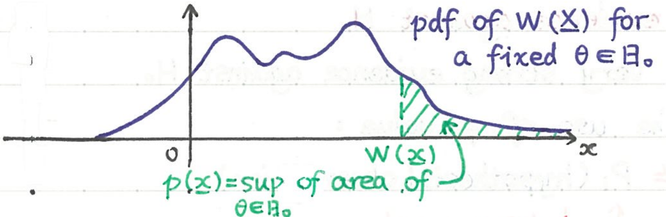
$$p(\underline{x}) := \sup_{\theta \in \Theta_0} P_\theta\left(W(\underline{X}) \geq W(\underline{x})\right) \in [0,1]$$

Then we consider $p(\underline{X})$ and call it a p-value (for $W(\underline{X})$).

Remarks:

1) $p(\underline{X})$ is a random variable.

2) Given $\underline{x}$, $p(\underline{x})$ gives the probability of equal or more extreme values of $W(\underline{X})$, knowing that H_0 is true. $\Leftrightarrow \theta \in \Theta_0$


pdf of $W(\underline{X})$ for
a fixed $\theta \in \Theta_0$
0
$W(\underline{x})$
x
$p(\underline{x}) = \sup\limits_{\theta\in\Theta_0}$ of area of

$$\overset{\text{unknown}}{} \qquad \overset{\text{fixed}}{}$$

Example 1: $X_j \sim n(\overset{\downarrow}{\mu}, \overset{\downarrow}{\sigma}^2)$, $H_0 : \mu = \overset{\downarrow}{\mu_0}$ and σ^2 arbitrary.

Set $W(\underline{X}) := \dfrac{|\overline{X} - \mu_0|}{s/\sqrt{N}}$ for which large values give evidence of H_1. . recall $E(\overline{X}) = \mu$

and which follows a |student t-distribution| with parameter $N-1$ indep. of σ^2

$$p(\underline{x}) = \sup_{\sigma^2 \in \mathbb{R}} \left(P\ W(\underline{X}) \geq \frac{|\overline{x} - \mu_0|}{s/\sqrt{N}} \right) = \underset{\substack{\uparrow \\ \text{because of abs. and symmetry}}}{2} P\Big(\underset{\substack{\uparrow \\ \text{student T dist.}}}{T_{N-1}} \geq \underset{\nwarrow \text{ found in tables}}{\frac{|\overline{x} - \mu_0|}{s/\sqrt{N}}}\Big)$$

Example 2: $X_j \sim n(\mu, \sigma^2)$, $H_0 : \mu \leq \mu_0 \Leftrightarrow \mu \in \Theta_0 := (-\infty, \mu_0)$

Again $W(\underline{X}) := \dfrac{\overline{X} - \mu_0}{s/\sqrt{N}}$ with the property on $W(\underline{X})$ satisfied. Then

$$\Rightarrow p(\underline{X}) = \sup_{\substack{\mu \leq \mu_0 \\ \sigma^2 \geq 0}} P(W(\underline{X}) \geq W(\underline{x})) = \sup_{\cdots} P\left(\frac{\overline{X} - \mu_0 + \mu - \mu}{s/\sqrt{N}} \geq W(\underline{x})\right)$$

$$= \sup P\left(\frac{\overline{X} - \mu}{s/\sqrt{N}} \geq W(\underline{x}) + \overset{\curvearrowright \geq 0}{\boxed{\frac{\mu_0 - \mu}{S/\sqrt{N}}}}\right) = \sup P\left(\underbrace{\frac{\overline{X} - \mu}{s/\sqrt{N}}}_{\sim T_{N-1}} \geq \overbrace{\frac{\overline{x} - \mu_0}{s/\sqrt{N}}}^{W(\underline{x})}\right)$$

Remark

Here, all computations were explicit, but it is not always the case.
In general, it is not so explicit but a computer can compute $p(\underline{x})$ easily.

Interpretation $(p(\underline{x}) \in [0,1])$

The p-value $p(\underline{x})$ should be interpreted in terms of repetition of the same experiment. $p(\underline{x})$ gives the prob. that the new value $W(\underline{x}')$ will be further away in the prob. distribution of $W(\underline{X})$, assuming that H_0 is correct.

p	interpretation
$p > 0.1$	No evidence against H_0 $\Rightarrow$ the data appear to be consistent with H_0
$\boxed{0.05} < p \leq 0.1$	weak evidence against H_0
$0.01 < p \leq 0.05$	moderate evidence against H_0
$p < 0.01$	strong or very strong evidence against H_0

⚠ many controversies about the use of p-value :

P_i (observation | hypothesis) $\neq$ P_i (hypothesis | observation)

↑ p-value ↑ what we want

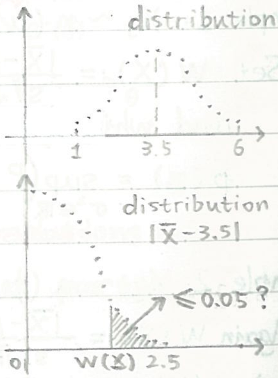
Example: Roll a dice

H_0: the dice is fair (same prob of $\frac{1}{6}$ for each face)

$\underline{X} = (X_1, \cdots, X_N)$, X_j = discrete unif. dist. $\{1, 2, \cdots, 6\}$

$W(\underline{X}) = \left| \frac{1}{N} \sum_{j=1}^{N} X_j - 3.5 \right| = |\overline{X} - 3.5|$ large values of $W(\underline{X})$ gives evidence of H_1

If the surface of the shade is smaller than 0.05 then we reject H_0, which means that we consider that the dice is not fair.



Remark: it is a rather weak approach.

And if H_0 is rejected, it does not say anything on the dice.

Next week: Tue regular lecture (same room & time)
Wed tree-based method

No. 9

Date 2019 6 19

$\underline{\text{VI}}$ Interval estimation

(linked to hypothesis test)

Interval estimation covers several related notions:

→ · confidence intervals

· credible intervals (Bayes approach)

· likelihood intervals

· tolerance intervals

$\underline{\text{VI}}$.1 Confidence intervals

Def: A confidence interval for a parameter θ is a pair of statistics $L(\underset{\sim}{X}), U(\underset{\sim}{X})$ with $L(\underset{\sim}{X}) < U(\underset{\sim}{X})$ (Lower & Upper)

together with a confidence coefficient γ given by

$$\gamma := \inf_{\theta} P\,(\theta \in [L(\underset{\sim}{X}), U(\underset{\sim}{X})])$$ we always take the worst scenario.

↖ if there exists additional parameters, we also take the infimum on them

Remarks:

- One can accept that $L(\underset{\sim}{X}) = -\infty$ and $U(\underset{\sim}{X}) = +\infty$ but not both.
- Usually γ is denoted by $1-\alpha$ and we speak about a $(1-\alpha)$ confidence.

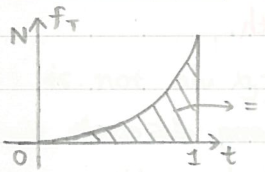
Example:

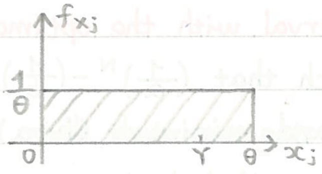
Let $\underset{\sim}{X} = (X_1, \cdots, X_N)$ with $X_j \sim \text{uniform}\,(0, \theta)$

Let $Y := \max\{X_1, \cdots, X_N\} = X_{(N)}$ last ordered statistics and set $T := \frac{Y}{\theta}$

What is the distribution of T? T has a distribution f_T with

$\underline{f_T(t) = N t^{N-1}}$ for $t \in [0, 1]$ (exercise based on § II.3)





Let us consider 2 possible confidence coefficients:

1) $[L(\underset{\sim}{X}), U(\underset{\sim}{X})] = [aY, bY]$ for $1 \leq a < b$

2) $[L(\underset{\sim}{X}), U(\underset{\sim}{X})] = [Y+c, Y+d]$ for $0 \leq c < d$

For 1)

$$P(\theta\in[aY,bY]) = P(aY\le\theta\le bY) = P(a\le\tfrac{\theta}{Y}\le b)$$

$$=P(\tfrac{1}{b}\le\tfrac{Y}{\theta}\le\tfrac{1}{a}) = \int_{1/b}^{1/a} N t^{N-1}\,dt = (\tfrac{1}{a})^N-(\tfrac{1}{b})^N$$

$$\inf_\theta P(\theta\in[aY,bY]) = (\tfrac{1}{a})^N-(\tfrac{1}{b})^N$$ no dependence on θ!

For 2)

$$P(\theta\in[Y+c,Y+d]) = P(Y+c\le\theta\le Y+d)$$

$$=P(1-\tfrac{d}{\theta}\le\tfrac{Y}{\theta}\le 1-\tfrac{c}{\theta}) = \int_{1-d/\theta}^{1-c/\theta} N t^{N-1}\,dt = (1-\tfrac{c}{\theta})^N-(1-\tfrac{d}{\theta})^N$$

$$\inf_\theta P(\theta\in[Y+c,Y+d]) = \inf_\theta\left((1-\tfrac{c}{\theta})^N-(1-\tfrac{d}{\theta})^N\right)=0$$

The 2nd choice is not really good

since for any c and d, the confident coefficience is 0.

For 1), if we impose that

$$(\tfrac{1}{a})^N-(\tfrac{1}{b})^N = 1-\alpha \text{ for a given } \alpha,$$

we can find some a and b.

no uniqueness: different possible choices for (a,b)

Def. A random variable $Q(\underline{X},\theta)$ this is not a statistic (explicit dependence on θ) is called a pivotal quantity or a pivot if its distribution function does not depend on θ.

Exemple: $\frac{Y}{\theta}$

Def. Among all statistics $L(\underline{X})$, $U(\underline{X})$ satisfying

$$L(\underline{X})<U(\underline{X}) \text{ and } \inf_\theta P(\theta\in[L(\underline{X}),U(\underline{X})]) = 1-\overset{\text{(fixed)}}{\alpha}$$

the ones having the shortest length $U(\underline{X})-L(\underline{X})$ define the $(1-\alpha)$-confidence interval with the optimal length.

Exercise: find a and b such that

$$(\tfrac{1}{a})^N-(\tfrac{1}{b})^N = 1-\alpha \text{ and minimum } (b-a)\text{? among } 1\le a<b$$

VI.2 Relation with hypothesis test

Example: Suppose $X_i\sim n(\mu,\sigma^2)$ and $H_0: \mu=\overset{\text{given}}{\mu_0}$

Test: rejecting H_0 if $\left|\underbrace{\frac{\bar{X}-\mu}{\sigma/\sqrt{N}}}_{=:\,Z\,\sim\, n(0,1)}\right|^{\text{known}} = |Z|\ge Z_{\alpha/2}$ with $P(|Z|\ge Z_{\alpha/2})=\frac{\alpha}{2}$

$\Rightarrow$ This test is an α-level test.

$\Leftrightarrow$ Accept H_0 if $\left|\frac{\overline{X}-\mu_0}{\sigma/\sqrt{N}}\right| < Z_{\alpha/2}$

$\Leftrightarrow \quad -Z_{\alpha/2} < \frac{\overline{X}-\mu_0}{\sigma/\sqrt{N}} < Z_{\alpha/2}$

$\Leftrightarrow \overline{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{N}} < \mu_0 < \overline{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{N}}$ with

$P(\overline{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{N}} < \mu_0 < \overline{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{N}} \mid \mu = \mu_0) \overset{\text{true for any } \mu_0}{=} 1-\alpha$

$\Leftrightarrow P(\overline{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{N}} < \mu < \overline{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{N}}) = 1-\alpha$

$\Leftrightarrow P(\mu \in [\overline{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{N}}, \overline{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{N}}]) = 1-\alpha$

Thus, we have obtained a $(1-\alpha)$ confidence interval from a hypothesis test.

More generally:

Thm. For any $\theta \in \underset{\text{parameter space}}{\overset{\subset \mathbb{R}}{\Theta}}$ set $H_0 : \theta = \theta_0$ and consider a level α-test $\mathcal{T}_{\theta_0}$ for H_0.

Let $\mathcal{A}(\theta_0) = \{\underline{x} \mid \mathcal{T}_{\theta_0}(\underline{x}) = H_0 \text{ accepted}\}$ be the acceptance region of this test (the complement of the rejection region).

For any $\underline{x}$, set

$C(\underline{x}) := \{\theta \in \Theta \mid \underline{x} \in \mathcal{A}(\theta)\}$

Then $C(\underline{x})$ is a $(1-\alpha)$ confidence set.

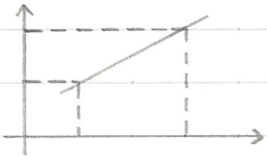
[Thm. 9.2.2] proof not long

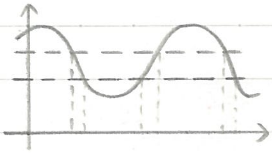
Remark:

1) $C(\underline{x})$ is not always an interval

(but there are conditions such that one gets an interval)

2) It is possible to look for an optimal balance between the confident coefficient and the length of $C(\underline{X})$.





VII Asymptotic evaluation (let $N \to \infty$)

VII.1: Consistency, sufficiency, and robustness

We shall consider a family of estimators (= statistics) $\{W_n\}_{n\in\mathbb{N}}$

with $W_j(X_1, X_2, \cdots, X_j)$

$\sim f(\cdot|\theta)$

Example: $W_1 = X_1$, $W_2 = \frac{1}{2}(X_1 + X_2)$, $W_3 = \frac{1}{3}(X_1 + X_2 + X_3) \cdots$

$\overline{X}_n := W_n = \frac{1}{n}\sum_{j=1}^{n} X_j$

Def. A sequence of estimators $\{W_n\}_{n\in\mathbb{N}}$, with $W_j = W_j(X_1, X_2, \cdots, X_j)$,

is a consistent sequence for the parameter θ if

$$\forall \varepsilon > 0\ \forall \theta \in \Theta: \lim_{n\to\infty} P(|W_n - \theta| < \varepsilon) = 1$$

$$\Leftrightarrow \lim_{n\to\infty} P(|W_n - \theta| \geq \varepsilon) = 0$$

similar to "convergence in probability"

Example: $X_j \sim n(\theta, 1)$ and $W_n = \overline{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j$

Recall that $\frac{\overline{X}_n - \theta}{1/\sqrt{n}} \sim n(0,1)$ $\Rightarrow \overline{X}_n \sim n(\theta, \frac{1}{n})$

$$P(|\overline{X}_n - \theta| < \varepsilon) = \int_{\theta-\varepsilon}^{\theta+\varepsilon} \left(\frac{n}{2\pi}\right)^{\frac{1}{2}} e^{-\frac{n}{2}(s-\theta)^2} ds$$

$$\overset{s-\theta=t}{=} \int_{-\varepsilon}^{\varepsilon} \left(\frac{n}{2\pi}\right)^{\frac{1}{2}} e^{-\frac{n}{2}t^2} dt$$

$$\overset{\sqrt{n}t = x}{=} \int_{-\varepsilon\sqrt{n}}^{\varepsilon\sqrt{n}} \left(\frac{1}{2\pi}\right)^{\frac{1}{2}} e^{-\frac{x^2}{2}} dx$$

$$\xrightarrow[\varepsilon \text{ fixed}]{n\to\infty} 1$$

Remark: We cannot do these computations $\overset{\text{always}}{}$ so explicitly, but we're lucky! ◡̈

$$P(|W_n - \theta| < \varepsilon) = P((W_n - \theta)^2 < \varepsilon^2) \qquad \text{(Appendix 3)}$$

$$\underset{\left(\substack{\text{Markov}\\ \text{inequality}}\right)}{\leq} \frac{E((W_n-\theta)^2)}{\varepsilon^2} \underset{\substack{\uparrow \\ \text{see IV.2}}}{=} \frac{1}{\varepsilon^2}\left(\text{Bias}_\theta(W_n)^2 + \text{Var}_\theta(W_n)\right)$$

Thm. If for $\theta \in \Theta$ and if

$$\lim_{n\to\infty} \text{Bias}_\theta(W_n) = 0 \text{ and } \lim_{n\to\infty} \text{Var}_\theta(W_n) = 0$$

Then $\{W_n\}$ is a consistent family for θ.

Corollory: If $\underline{X} = (X_1, X_2, \cdots)$ with $E(X_j) = \theta < \infty$ and $\text{Var}(X_j) < \infty$,

then $\overline{X}_n$ is a consistent sequence of estimator for θ.

(Based on §II.1)

What about efficiency? It is measured with variance.

Def. Let $\{W_n\}$ be a sequence of estimators for θ, and let $\{k_n\}$ be a natural family of scaling parameters. Natural = coming from other reasons e.g. $k_n = \sqrt{n}$

1) If $\lim_{n\to\infty} k_n \operatorname{Var}(W_n) = J^2$ then J^2 is called the limiting variance.

2) If $k_n(W_n - \theta) \xrightarrow[\text{in distribution}]{n\to\infty} n(0,\sigma^2)$, then σ^2 is called the asymptotic variance.

$\Updownarrow$ Recall in App. 3

$$\forall x \in \mathbb{R} : F_n(x) \xrightarrow{n\to\infty} \frac{1}{\sqrt{2\pi}\,\sigma} \int_{-\infty}^{x} e^{\frac{s^2}{2\sigma^2}}\, ds$$

cdf of $k_n(W_n - \theta)$

Remark: $J^2 = \sigma^2$ in general but not always. (See example 10.1.10)

The asymptotic parameter σ will be a measurement of the efficiency.

Question: Is there a "best" σ^2?

Def. $\{W_n\}_{n\in\mathbb{N}}$ is asymptotically efficient to θ if

$$\sqrt{n}(W_n - \theta) \xrightarrow[\text{in distribution}]{n\to\infty} n(0, v(\theta)) \quad \text{with}$$

$$v(\theta) := \frac{1}{E_\theta\left(\left(\frac{\partial}{\partial\theta}\ln(f(\cdot|\theta))\right)^2\right)}$$

related to Gramer-Rao bound (see 10.6.2)

Thm. If $X_j \sim f(\cdot|\theta)$ with $f(\cdot|\theta)$ satisfying some weak technical assumption and if $\hat{\theta}_n(X_1, \cdots, X_n)$ is the maximum likelihood estimator for θ, then $\{\hat{\theta}_n\}$ is a consistent and asymptotically efficient family of estimators.

Maximum likelihood estimator: see § IV.2

Likelihood function $\theta \mapsto L(\theta|\underline{x}) := f(\underline{x}|\theta)$ and look for its global maximum

Def. If $\{W_n\}$ and $\{V_n\}$ are 2 sequences of estimators for θ satisfying

$$\sqrt{n}(W_n - \theta) \xrightarrow[\text{in distr.}]{n\to\infty} n(0, \sigma_W^2) \quad \text{and}$$

$$\sqrt{n}(V_n - \theta) \xrightarrow[\text{in distr.}]{n\to\infty} n(0, \sigma_V^2)$$

then the asymptotic relative efficiency (ARE) of $\{V_n\}$ with respect to $\{W_n\}$ is

$$\operatorname{ARE}(\{V_n\}, \{W_n\}) := \frac{\sigma_W^2}{\sigma_V^2}$$

Remark: $\frac{v(\theta)}{\sigma_V^2} \leq 1$

We should look for a sequence of estimators such that this ratio is close to 1.

Question: Is there always a sequence of estimators with the best σ^2?

Unfortunately no.

What about robustness? (See the Appendix 10)

Idea: What happens if $X_j \not\sim f(\cdot|\theta)$ for some j (rare events)?

Based on this idea, there should be a trade-off between efficiency & robostness.

⇝ many books.

Example: Consider the sample $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 100\}$

Outlier: a point that differs a lot from the others ↑

Then sample mean $\bar{x} \approx 14$ not robust

sample median $= 6$ more robust

If we compute the ARE

(If the pdf f is symetric ⇒ median = mean)

$$\underset{\overset{\uparrow}{\circledast}}{\text{ARE}(\text{median}}, \underset{\overset{\uparrow}{\text{based on } \bar{X}_n}}{\text{mean})} = \begin{cases} 0.64 & \text{normal} \\ 0.82 & \text{logistic} \\ 2 & \text{double exponential} \end{cases}$$

⊛ define $M_n :=$ median for a sample $(X_1, \cdots, X_n)$

and then $\sqrt{n}\,(M_n - \theta) \xrightarrow[\text{in distr.}]{n \to \infty} n(0, \sigma^2_{\text{median}})$

Then for normal and logistic distributions, mean has a more efficient behavior ;

for double exponential distribution, median is more efficient.

↳ because of heavy tail

One way to take the best of both sequences of estimation is

to consider a mixture :

Consider $\sum_{j=1}^{n} \rho(X_j - a)$ and $\rho(x) = \begin{cases} \frac{1}{2}x^2 & \text{if } |x| \le k \\ k|x| - \frac{1}{2}k^2 & \text{if } |x| \ge k \end{cases}$ related to mean / related to median

and set $\hat{\theta}_n(X_1, \cdots, X_n)$ for the

minimizer (as a function of a) of this expression.

k is a parameter that can be fixed freely.

With this new sequence of estimators we can compute (for $k = 1.5$)

	normal	logistic	d.exp
ARE (new, mean)	0.96	1.08	1.37 ← new is an improvement
ARE (new, median)	1.51	1.31	0.68

↖ stable ↗

Conclusion: $\{\hat{\theta}_n\}$ is a balance between mean and median, but is more robust.

$(n \to \infty)$

Last course on 17 H.

"report" until end of July

VIII Analysis of variance and regression

VIII.1 "One way" ANOVA

Idea: Consider data like

$$\begin{array}{c|cccc} & \multicolumn{4}{c}{\xrightarrow{\text{Treatments}}} \\ & 1 & 2 & \cdots & k \\ \hline \text{Observation} \downarrow & y_{11} & y_{21} & & y_{k1} \\ & y_{12} & y_{22} & & y_{k2} \\ & \vdots & \vdots & & \vdots \\ & y_{1n_1} & y_{2n_2} & & y_{kn_k} \end{array}$$

with $\{n_i\}$ independent.

We shall assume that the corresponding r.v. follow

$$Y_{ij} = \overset{\text{unknown}}{\theta_i} + \overset{\text{noise or error}}{\varepsilon_{ij}}$$

$i = 1, \cdots, k$

$j = 1, \cdots, n_i$

Remark: we could consider

$$Y_{ij} = \underset{\text{additional parameter}}{\mu} + \tau_i + \varepsilon_{ij}$$

but usually impose that $\sum_{i=1}^{k} \tau_i = 0$ which fixes one parameter.

Def. The model $Y_{ij} = \theta_i + \varepsilon_{ij}$ is called a oneway ANOVA if

1) ε_{ij} is a r.v. following $n(0, \sigma^2)$ σ is independent of i, j for any i, j.

2) ε_{ij} and $\varepsilon_{i'j'}$ are independent for any $(i, j) \neq (i', j')$.

Remark: These assumptions can also be weakened if necessary.

Def. ANOVA null assumption is $H_0 : \theta_1 = \theta_2 = \cdots = \theta_k$. very strong assumption

$\Rightarrow H_1 : \theta_j \neq \theta_k$ for ≥ 1 pair (θ_j, θ_k) with $j \neq k$

Let us set

$$\mathcal{A} = \mathcal{A}_k = \{\underline{a} = (a_1, a_2, \cdots, a_k) \in \mathbb{R}^k \text{ with } \sum_{j=1}^{k} a_j = 0\}$$

Examples: $a = (1, -1, 0, \cdots, 0)$ or $a = (1, -\frac{1}{2}, -\frac{1}{2}, 0, \cdots, 0)$, etc

Def. Consider $\underline{t} := (t_1, \cdots, t_k)$ a set of k parameters or of k r.v.

If $\underline{a} \in \mathcal{A}_k$, then $\sum_{j=1}^{k} a_j t_j \equiv \underline{a} \cdot \underline{t}$ is called a contrast.

Lemma: $H_0 \Leftrightarrow \forall \underline{a} \in \mathcal{A} : \underline{a} \cdot \underline{\theta} = 0$ $\Rightarrow$ easy $\Leftarrow$ as an exercise

Corollary: $H_1 \Leftrightarrow \exists\, \underline{a} \in \mathcal{A} : \underline{a} \cdot \underline{\theta} \neq 0$

Under the ANOVA assumption one has

$$Y_{ij} \sim n(\theta_i, \sigma^2) \quad \forall i = 1, \cdots, k, \ \forall j = 1, \cdots, n_i$$

$$\Rightarrow \overline{Y_i} := \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} \sim n(\theta_i, \frac{\sigma^2}{n_i})$$

Then for any $a \in \mathbb{R}^k$ consider $\sum_{i=1}^{k} a_i \overline{Y}_i =: \underline{a} \cdot \underline{\overline{Y}}$ one has

$$\underline{a} \cdot \underline{\overline{Y}} \sim n\left(\sum_{i=1}^{n} a_i \theta_i, \sigma^2 \sum_{i=1}^{k} \frac{a_i^2}{n_i}\right) \quad (\text{Ex. } 11.8)$$

If σ^2, is not known, then we can define the sample variance

$$s_i^2 = \frac{1}{n_i-1}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y}_i)^2 \sim \chi^2_{n_i-1} \quad (\S\,\text{II}.2)$$

Since σ^2 is the same for all experiments, one can set

$$s^2 = \frac{1}{N-k}\sum_{i=1}^{k}(n_i-1)\,s_i^2 = \frac{1}{N-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y}_i)^2$$ and

$(N-k)\,s^2/\sigma^2 \sim \chi^2_{N-k}$ sum of $\chi^2_{n_i-1}$ (Lemma 5.3.2) with

$$N := \sum_{i=1}^{k} n_i$$

Then
$$\frac{\sum_{i=1}^{k} a_i\overline{Y}_i - \sum_{i=1}^{k} a_i\theta_i}{\sqrt{s^2\sum_{i=1}^{k}\frac{a_i^2}{n_i}}} \sim t_{N-k}$$

Now for any $\underline{a} \in \mathbb{R}^k$ fixed, a hypothesis test could be

Reject $H_0 : \sum_{i=1}^{k} a_i\theta_i = 0$ if $\left|\frac{\underline{a}\cdot\overline{\underline{Y}} - \underline{a}\cdot\underline{\theta}}{\sqrt{s^2\sum_{i=1}^{k}\frac{a_i^2}{n_i}}}\right| > t_{N-k,\frac{\alpha}{2}}$ for a given α.

Equivalentally,

∴ a confident interval with confidence coefficient $1-\alpha$ is given by

$$\sum_{i=1}^{k} a_i\overline{y}_i - t_{N-k,\frac{\alpha}{2}}\sqrt{s^2\sum_{i=1}^{k}\frac{a_i^2}{n_i}} \le \sum_{i=1}^{k} a_i\theta_i \le \sum_{i=1}^{k} a_i\overline{y}_i + t_{N-k,\frac{\alpha}{2}}\sqrt{s^2\sum_{i=1}^{k}\frac{a_i^2}{n_i}}$$

Key fact: If we choose $\underline{a} = (1,-1,0,\cdots,0)$ then we are testing $H_0 : \theta_1 - \theta_2 = 0$ or for $\underline{a} = (1,-\frac{1}{2},-\frac{1}{2},0,\cdots,0)$ then we are testing $H_0 : \theta_1 = \frac{1}{2}(\theta_2+\theta_3)$.

Observe that all data are used for the computation of s^2.

⚠ If we choose 2 different $\underline{a}, \underline{a}' \in \mathbb{R}^k$, the 2 computations for $H_0 : \underline{a}\cdot\underline{\theta} = 0$ and $H_0' : \underline{a}'\cdot\underline{\theta} = 0$ are not independent, which implies that the 2 confidence coefficients are not both $1-\alpha$.

Remark: The test for the ANOVA null assumption $H_0 : \theta_1 = \theta_2 = \cdots = \theta_k$ can be obtained by the hypothesis union-intersection

$$H_0 \Leftrightarrow \underline{\theta} \in \bigcap_{\underline{a}\in\mathcal{A}} \{\underline{\xi} \in \mathbb{R}^k \mid \underline{a}\cdot\underline{\xi} = 0\}$$

To obtain a criterion for this intersection corresponds to a maximization problem. See 11.2.4

An α-level test for rejecting H_0 is given by

$$\frac{\sum_{i=1}^{k} n_i(\overline{Y}_i - \overline{\overline{Y}})^2}{s^2} > (k-1)\,\underset{\uparrow\ \text{F distribution}}{F_{k-1,n-k;\alpha}} \quad \text{with } \overline{\overline{Y}} := \frac{1}{N}\sum_{i,j} Y_{ij}$$

from which one can obtain a $(1-\alpha)$ confidence interval. Called ANOVA F-test

VIII. 2 Simple Linear Regression

Idea: Consider relation of the form: $f(x_i)$

$$\underset{\text{response}}{Y_i} = \underbrace{\alpha + \beta}_{\text{2 unknown coeff.}} \underset{\text{predictor}}{x_i} + \underset{\text{noise or error}}{\varepsilon_i}$$

Remark: It is called linear regression because it is linear in the coefficients α, β (and not because of x_i)

Remark: Once some data are collected, just evaluating α and β corresponds to "data fitting" but there is no statistical inference.

i	x	Y
1	x_1	y_1
2	x_2	y_2
$\vdots$	$\vdots$	$\vdots$
n	x_n	y_n

For any random variables $\{(X_i, Y_i)\}_{i=1}^{n}$, let us set

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad S_{XX} = \sum_{i=1}^{n} (X_i - \bar{X})^2$$

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \qquad S_{YY} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \qquad S_{XY} = \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$$

About data fitting

Lemma: (least squares) One has

$$\min_{\alpha,\beta} \sum_{i=1}^{n} \left(y_i - (\alpha + \beta x_i)\right)^2 = \sum_{i=1}^{n} \left(y_i - (a + b x_i)\right)^2 \text{ for}$$

$$b = S_{xy} / S_{xx} \text{ and } a = \bar{y} - b\bar{x}$$

Remark: The previous computation is quite natural if we think that y_i is a function of x_i with relation $y = \alpha + \beta x$, but if we just collect (x_i, y_i), it is not clear that this is the best choice.

In order to do some statistical inference let us assume that

$$Y_i = \alpha + \beta x_i + \varepsilon_i$$

with $\{\varepsilon_i\}$ indep. r. v., with $E(\varepsilon_i) = 0$ and $var(\varepsilon_i) = \sigma^2$ independent of i

Then $E(Y_i) = \alpha + \beta x_i$ and $var(Y_i) = \sigma^2$.

Consider a linear estimator for β of the form $\sum_{i=1}^{n} \overset{\swarrow \text{constants}}{d_i} Y_i$

It is unbiased if

$$E\left(\sum_{i=1}^{n} d_i Y_i\right) = \beta \quad \Rightarrow \beta = \sum_{i=1}^{n} d_i(\alpha + \beta x_i) = \alpha \sum_{i=1}^{n} d_i + \beta \sum_{i=1}^{n} d_i x_i$$

is true for any β. Thus

$$\sum_{i=1}^{n} d_i = 0 \text{ and } \sum_{i=1}^{n} d_i x_i = 1$$

What about the best linear unbiased estimator (BLUE) ?

The BLUE is the one with the smallest variance :

$$\mathrm{Var}\left(\sum_{i=1}^{n} d_i Y_i\right) = \sum_{i=1}^{n} d_i^2 \,\mathrm{Var}(Y_i) = \sigma^2 \sum_{i=1}^{n} d_i^2$$

$\Rightarrow$ We have to minimize $\sum_{i=1}^{n} d_i^2$ "distance" from origin intersection of 2 "planes" under the condition

$$\sum_{i=1}^{n} d_i = 0 \text{ and } \sum_{i=1}^{n} d_i x_i = 1$$

The solution : $d_i = \dfrac{(x_i - \bar{x})}{S_{xx}}$

$$\Rightarrow \mathrm{Var}\left(\sum_{i=1}^{n} d_i Y_i\right) = \frac{1}{S_{xx}^2} \underbrace{\sum_{i=1}^{n} (x_i - \bar{x})^2}_{S_{xx}} \sigma^2 = \frac{\sigma^2}{S_{xx}}$$

Then we get (after an experiment)

$$b := \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{S_{xx}} y_i \underset{\Uparrow}{=} \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{S_{xx}} = \frac{S_{xy}}{S_{xx}}$$

$$\sum (x_i - \bar{x}) \bar{y} = \bar{y} \sum (x_i - \bar{x}) = 0$$

This coefficient is the one obtained by data fitting.

We can do the same for α and one gets the BLUE for α gives $\alpha = \bar{y} - b\bar{x}$.

If we do further analysis, we need to impose more on the distribution of ε_i.

We consider the normal model : $\varepsilon_i \sim n(\underset{\uparrow}{0}, \underset{\uparrow}{\sigma^2})$

imposed

In this context, $Y_i \sim n(\alpha + \beta x_i, \sigma^2)$

Lemma : Assume $Y_i \sim n(\alpha + \beta x_i, \sigma^2)$ then $(\beta = \underset{\uparrow}{\sum} d_i Y_i$ for $\beta)$

the sample distribution for $\hat{\alpha}$ and $\hat{\beta}$ given by the BLUE are given by

$$\hat{\beta} \sim n\left(\beta, \frac{\sigma^2}{S_{xx}}\right) \text{ and } \hat{\alpha} \sim n\left(\alpha, \frac{\sigma^2}{nS_{xx}} \sum_{i=1}^{n} x_i^2\right)$$

The sample variance S^2 given by

$$S^2 = \frac{1}{n-2} \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta} x_i)^2$$

The factor $1/(n-2)$ is imposed by the requirement that S^2 is unbiased.

satisfies $\dfrac{n-2}{\sigma^2} S^2 \sim \chi^2_{n-2}$

In addition, S^2 is independent of $\hat{\alpha}$ and $\hat{\beta}$,

but $\hat{\alpha}$ and $\hat{\beta}$ are not independent. [Thm. 11.3.3]

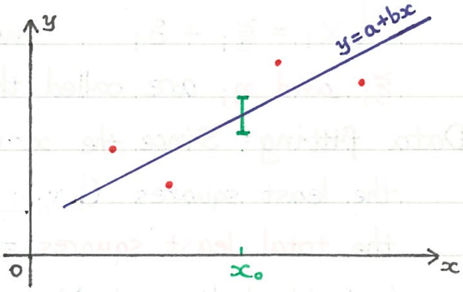
Corollory : $\dfrac{(\hat{\beta}-\beta)/\sqrt{\sigma^2/S_{xx}}}{\sqrt{S^2/\sigma^2}} = \dfrac{\hat{\beta}-\beta}{S/\sqrt{S_{xx}}} \sim t_{n-2}$ student t dist. ⊛

$$\frac{(\hat{\alpha}-\alpha)\sqrt{\frac{\sigma^2}{nS_{xx}}\sum x_i^2}}{\sqrt{S^2/\sigma^2}} = \frac{\hat{\alpha}-\alpha}{S\sqrt{\sum x_i^2/nS_{xx}}} \sim t_{n-2}$$

From ⊛ with the hypothesis $H_0 : \hat{\beta}=\beta$

we get the confidence interval with confident coef. $(1-\alpha)$ as

$$\left[\hat{\beta}-t_{n-2,\alpha/2}\frac{S}{\sqrt{S_{xx}}},\ \hat{\beta}+t_{n-2,\alpha/2}\frac{S}{\sqrt{S_{xx}}}\right]$$



What about estimating Y for a given x_0?

If we fix x_0, the point estimator for the corresponding Y is $Y=\hat{\alpha}+\hat{\beta}x_0$. One has

$$E(Y) = E(\hat{\alpha})+x_0 E(\hat{\beta}) = \alpha+\beta x_0$$

$$\mathrm{Var}(Y) = \mathrm{Var}(\hat{\alpha}+\hat{\beta}x_0) = \mathrm{Var}(\hat{\alpha})+x_0^2\,\mathrm{Var}(\hat{\beta})+2x_0\,\mathrm{Cov}(\hat{\alpha},\hat{\beta})$$

$$= \cdots = \sigma^2\left(\frac{1}{n}+\frac{(x_0-\bar{x})}{S_{xx}}\right)$$

As a consequence

$$Y=\hat{\alpha}+\hat{\beta}x_0 \sim n\left(\alpha+\beta x_0,\ \sigma^2\left(\frac{1}{n}+\frac{(x_0-\bar{x})}{S_{xx}}\right)\right)$$

In order to estimate the parameter σ^2 (which is usually unknown) we use the sample variance S^2, which is independent of $\hat{\alpha}$ and $\hat{\beta}$, and get

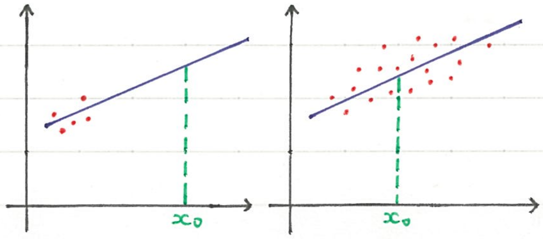
$$\frac{\hat{\alpha}+\hat{\beta}x_0-(\alpha+\beta x_0)}{S\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})}{S_{xx}}}} \sim t_{n-2}$$

The related confidence interval with confident coeff. $1-\alpha$ is given by

$$\left[\hat{\alpha}+\hat{\beta}x_0-t_{n-2,\alpha/2}\,S\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})}{S_{xx}}},\ \hat{\alpha}+\hat{\beta}x_0+t_{n-2,\alpha/2}\,S\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})}{S_{xx}}}\right]$$

Remark: we have a smaller interval either by taking n large, or by fixing $x_1,\cdots,x_n$ s.t. $x_0 \cong \bar{x}$.

We can also maximize S_{xx}.



$\overline{\underline{\text{IX}}}$ Regression models

$\overline{\underline{\text{IX}}}$.1 Regression with errors in the variables

Idea: Before $Y_i = \alpha + \beta x_i + \varepsilon_i$ with fixed x_i

Now X_i is going to be a random variable.

Def. EIV model

$$\begin{cases} Y_i = \underbrace{\alpha + \beta \xi_i}_{=: \eta_i} + \varepsilon_i & \text{with } \varepsilon_i \sim n(0, \sigma_\varepsilon^2) \\ X_i = \xi_i + \delta_i & \text{with } \delta_i \sim n(0, \sigma_\delta^2) \end{cases}$$

ξ_i and η_i are called the latent variables.

Data fitting: Since the x-variable has some errors

the least squares (based on vertical distance) is replaced by

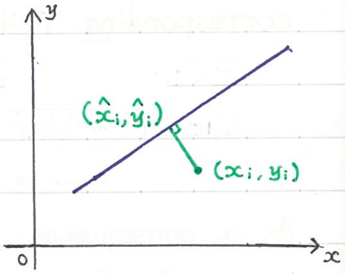
the total least squares $= \sum_{i=1}^{n} \left((x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 \right)$

(with the line given $a+bx$) $= \frac{1}{1+b^2} \sum_{i=1}^{n} \left((y_i - (a + b x_i) \right)^2$

By minimization over a and b, one gets

$$a = \bar{y} - b\bar{x}$$

$$b = \frac{-(S_{xx} - S_{yy}) + \sqrt{(S_{xx} - S_{yy})^2 + 4 S_{xy}^2}}{2 S_{xy}}$$



Recall: $Y_i \sim n(\alpha + \beta \xi_i, \sigma_\varepsilon^2)$;

$X_i \sim n(\xi_i , \sigma_\delta^2)$

This leads to the likelihood function for the sample $\{(X_i, Y_i)\}_{i=1}^{n}$

$$L(\alpha, \beta, \xi_1, \cdots, \xi_n, \sigma_\varepsilon^2, \sigma_\delta^2 | \underline{x}, \underline{y}) =$$

$$= \frac{1}{(2\pi)^n} \frac{1}{(\sigma_\varepsilon^2 \sigma_\delta^2)^{n/2}} \exp\left(-\sum \frac{(x_i - \xi_i)^2}{2\sigma_\delta^2}\right) \exp\left(-\sum_{i=1}^{n} \frac{(y_i - (\alpha + \beta \xi_i))^2}{2\sigma_\varepsilon^2}\right)$$

Reports until end of July
⇝ #237 of this building

Fall 2019: functional analysis
Spring 2020: graph theory
Fall 2020: mathematical methods in machine learning

No. 13

Date 2019 · 7 · 17

Compute the MLE for the different parameters

In this general setting no local maximum.

One assumption: $\sigma_\delta^2 = \lambda \sigma_\varepsilon^2$ for a fixed $\lambda \in \mathbb{R}$

With this assumption

$$L(\cdots | \underline{x}, \underline{y}) = \frac{1}{(2\pi)^n} \frac{\lambda^{n/2}}{\sigma_\delta^{2n}} \left(\exp - \sum_i \frac{(x_i - \xi_i)^2 + \lambda (y_i - \alpha - \beta \xi_i)^2}{2\sigma_\delta^2} \right)$$

We are going to look at local maxima for this function, as a function of the parameters (see Chapter IV on point estimators with MLE)

1) By taking derivatives with respect to ξ_i, and put the derivative equal to 0, one gets that a local maximum is observed for

$$\hat{\xi}_i = \frac{x_i + \lambda \beta (y_i - \alpha)}{1 + \lambda \beta^2}$$

By substituting these expressions in L, one finds

$$L(\alpha, \beta, \hat{\xi}_1, \cdots, \hat{\xi}_n, \sigma_\delta^2 | \underline{x}, \underline{y}) = \frac{1}{(2\pi)^n} \frac{\lambda^{n/2}}{\sigma_\delta^{2n}} \exp\left(- \frac{\lambda}{2\sigma_\delta^2 (1 + \lambda\beta^2)} \sum_i (y_i - \alpha - \beta x_i)^2 \right)$$

For computing the MLE for α and β we set

$$y_i^* := \sqrt{\lambda}\, y_i \; ; \quad \alpha^* := \sqrt{\lambda}\, \alpha \; ; \quad \beta^* := \sqrt{\lambda}\, \beta \qquad \text{(rescaling)}$$

One gets

$$L(\alpha^*, \beta^*, \hat{\xi}_1, \cdots, \hat{\xi}_n, \sigma_\delta^2 | \underline{x}, \underline{y}) = \frac{1}{(2\pi)^n} \frac{\lambda^{n/2}}{\sigma_\delta^{2n}} \exp\left(- \frac{1}{2\sigma_\delta^2 (1 + \lambda\beta^2)} \sum_i (y_i^* - \alpha^* - \beta^* x_i)^2 \right)$$

(similar expression as at the beginning of the section for data fitting)

By the result for data fitting one gets

$$\hat{\alpha} = \frac{\overline{y^*} - \beta^* \overline{x}}{\sqrt{\lambda}} = \overline{y} - \beta \overline{x}$$

$$\hat{\beta} = \frac{\beta^*}{\sqrt{\lambda}} = \frac{-(S_{xx} - S_{y^*y^*}) + \sqrt{(S_{xx} - S_{y^*y^*})^2 + 4 S_{xy^*}^2}}{2 S_{xy^*}}$$

$$= \frac{-(S_{xx} - \lambda S_{yy}) + \sqrt{(S_{xx} - \lambda S_{yy})^2 + 4\lambda S_{xy}^2}}{2\lambda S_{xy}}$$

Remark:

For $\lambda = 1$, we get the result of the data fitting (obtained with the total least square)

This can be considered as a justification of the total least square.

Remark: From

$$L(\hat\alpha,\hat\beta,\hat\xi_1,\cdots,\hat\xi_n,\sigma_\delta^2\,|\,\underline{x},\underline{y})=\frac{1}{(2\pi)^{n/2}}\frac{\lambda^{n/2}}{\sigma_\delta^{2n}}\exp\left(-\frac{\lambda}{2\sigma_\delta^2(1+\lambda\hat\beta^2)}\sum_i(y_i-\hat\alpha-\hat\beta x_i)^2\right)$$

We can differenciate it with respect to σ_δ^2, and find the critical point.

One gets

$$\hat\sigma_\delta^2 \underset{\text{for } \sigma_\delta^2}{\overset{\text{MLE}}{}} =\frac{1}{n}\frac{\lambda}{2(1+\lambda\hat\beta^2)}\sum_i(y_i-\hat\alpha-\hat\beta x_i)^2$$

What about confidence interval for β? $\rightsquigarrow$ very complicated

One way to get an approximate solution: Consider

$$\hat\sigma_\beta^2 := \frac{(1+\lambda\hat\beta^2)(S_{xx}S_{yy}-S_{xy}^2)}{(S_{xx}-\lambda S_{yy})^2+4\lambda S_{xy}^2}$$

is a $\underbrace{\text{consistent}}_{\text{Chap VII}}$ estimator for σ_β^2 $\rightarrow$ when sample size $\to\infty$, $\hat\sigma_\beta^2 \xrightarrow{\text{in precise sense see Chap VII}} \sigma_\beta^2$

Then by the central limit thm, one has

we don't know it

$$\frac{\hat\beta-\beta}{\sigma_\beta/\sqrt{n}}\xrightarrow{n\to\infty} n(0,1)$$

from which one obtains the approximate $(1-\alpha)$ confidence interval

$$\left[\hat\beta-z_{\alpha/2}\frac{\hat\sigma_\beta}{\sqrt{n}},\ \hat\beta+z_{\alpha/2}\frac{\hat\sigma_\beta}{\sqrt{n}}\right]$$

⚠ It is not a $(1-\alpha)$ confidence interval,

but for n large, it converges to a $(1-\alpha)$ confidence interval.

IX.2 Logistic regression (0,1 model)

Model: $\{Y_i\}$ independent variables with $Y_i\sim\text{Bernoulli}(\pi_i)$ with $E(Y_i)=\pi_i=P(Y_i=1)$

$$\pi_i=\frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}\in[0,1] \iff \ln\underbrace{\frac{\pi_i}{1-\pi_i}}=\alpha+\beta x_i$$

$$=\frac{\text{prob of success}}{\text{prob of failure}} =: \text{odds}$$

Remark: We cannot draw a graph (x_i,y_i) and use the least squares but we can use the MLE.

One has $L(\alpha,\beta\,|\,\underline{x},\underline{y})=\prod_i\underbrace{\pi(x_i)^{y_i}(1-\pi(x_i))^{1-y_i}}_{\text{pmf for Bernoulli}}$ under the assumption of independence of measurements

with $\pi_i=\pi(x_i)$

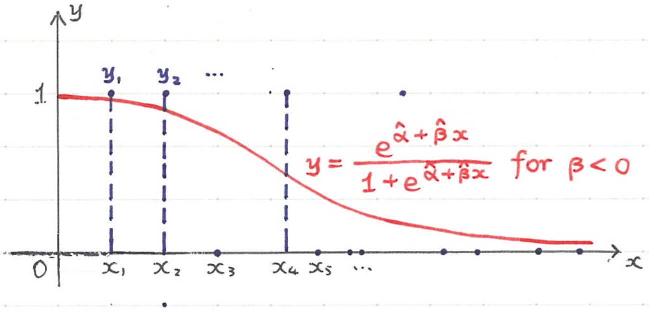
and π the function $x\longmapsto\pi(x):=\dfrac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$

We can then compute

$$\frac{\partial}{\partial\alpha}L(\alpha,\beta\,|\,\underline{x},\underline{y})=0 \text{ and } \frac{\partial}{\partial\beta}L(\alpha,\beta\,|\,\underline{x},\underline{y})=0$$

and solve the system.

We can not solve this system explicitly, but a computer can do it easily.
Suppose we have obtained $\hat{\alpha}$ and $\hat{\beta}$ numerically, then we can plot the result



Remember $E(Y) = \pi(x)$

We can get some confidence intervals.

If we want to consider several measures for a given x,

one has to use a binomial distribution.

More precisely, if n_i independent, Bernoulli observations are measured at x_i,

then Bernoulli (π_i) has to be replaced by binomial (n_i, π_i). $\hookrightarrow = \pi(x_i)$

Then $L(\alpha, \beta | \underline{x}, \underline{y}) = \prod_i \binom{n_i}{y_i} \pi(x_i)^{y_i} (1-\pi(x_i))^{n_i - y_i}$

with y_i the number of success at x_i,

and we can compute the MLE for α and for β (with a computer)

Application: see (3.2) of Appendix 13

Conclusion for the course

We have only touched the surface of statistics, but we have opened many doors, and you can continue in these directions.