Inferensteori

course page: bit.ly/2Sw10Ac

rezazesh.sainudiin@math.uu.se : subject 1MS035

Kollin upp kurs:
Theoretical foundation for data scientists
(HT19)

Sets, Maps & Numbers

Sets

A set is a collection of distinct elements.

e.g. $\{\bullet, \circ\}$. We give names to sets, e.g. $A = \{\bullet, \circ\}$.

$A' = \{\bullet, \bullet, \circ\}$ is not a set (multiset)

$\emptyset = \{\}$ is the empty set.

An element belongs to (or doesn't belong to) a set and we write $\in$ or $\notin$ respectively. e.g. $\bullet \in \{\bullet, \circ\}$ and $\circ \notin \{\Delta, \bullet\}$. etc. etc. det här kan du...

Set operations

We can add elements to an existing set by union operation.

e.g. $\{\} \cup \{\times\} = \{\times\}$

$\{\circ, \bullet\} \cup \{\times\} = \{\circ, \bullet, \times\}$

$\{\circ, \Delta\} \cup \{\circ\} = \{\circ, \Delta\}$

Def $A \cup B := \{x \mid x \in A \text{ or } x \in B\}$

Intersection: $A \cap B := \{x \mid x \in A \text{ and } x \in B\}$

set difference: $A \setminus B := \{x \mid x \in A \text{ and } x \notin B\}$

complement

Given a universal set U, $A^c := \{x \mid x \notin A\} = U \setminus A$

Maps

A map or a function associates each element in the set called domain with exactly one element in the set range (codomain)

formally

A function is a specific kind of relation between elements in the domain and range

Inverse Image/map/function

The inverse image of a function $f: X \to Y$ is $f^{[-1]}: Y \to X$ where $f^{[-1]}(y) = \{x \mid x \in X \text{ and } f(x) = y\}$ or more generally, for any $B \subseteq Y$

$f^{[-1]}(B) = \{x \in X \mid f(x) \in B\}$ and $f^{[-1]}: \sigma_Y(Y) \to \sigma_X(X)$

($\sigma_Y(Y)$: collection of subsets of Y)

Note

In this course $\mathbb{N} = \{1, 2, 3, \ldots\}$ and $\mathbb{Z}_+ = \mathbb{Z}_{>0} = \{0, 1, 2, 3, \ldots\}$

Probability

Language

An *experiment* is an activity that produces distinct observable outcomes. The set of such outcomes is called the *sample space* of the experiment, denoted by Ω

An event is a subset of the sample space.

Probability is a function

$P: \{\text{events}\} \to [0,1]$ where P satisfies

1) $\forall$ event a, $0 \le P(a) \le 1$
2) $P(\Omega) = 1$
3) $A \cap B = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B)$
4) $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$ where A_i are pairwise disjoint

Motivation for axioms:

idea of long-term relative frequency of independent experiments (experiment trials)

If we repeat an experiment a large number of times, the fraction of times the event A occurs will be close to $P(A)$.

Formally, let $N(A,n)$ be the number of times A occurs in the first n trials.

$$P(A) = \lim_{n \to \infty} \frac{N(A,n)}{n}$$

axiom i) $0 \le \frac{N(A,n)}{n} \le 1$

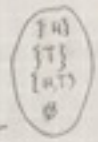
ii) $\frac{N(\Omega,n)}{n} = \frac{n}{n} = 1$ ← "something" happens every time

iii) $A \cap B = \emptyset \Rightarrow N(A \cap B, n) = N(A,n) + N(B,n)$

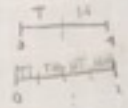
iv) If this is ignored the mathematics you need is much harder

ex1 Tossing a fair coin. Can construct reals by dyadic partitioning.

$\Omega = \{H, T\}$ $\quad 2^\Omega = \mathcal{F}_\Omega$



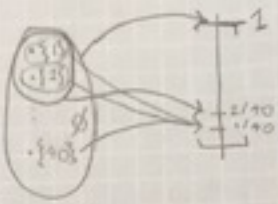
Bernoulli random variable. Bernoulli(θ)



ex2 NZ Lotto (40 balls)

Label of the 1st ball that comes out:

$\Omega = \{1, 2, \dots, 40\}$



De Moivre random variable, $\quad 2^\Omega = \mathcal{F}_\Omega$

$\text{DeMoivre}(\frac{1}{L}, \frac{1}{L}, \dots, \frac{1}{L}) = \text{DeMoivre}(\theta_1, \theta_2, \dots, \theta_L)$

$L = 40$

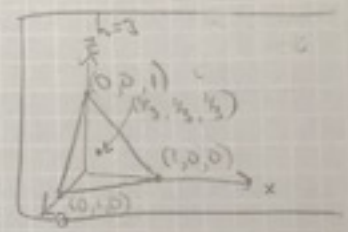
Here we have $\theta_1 = \theta_2 = \dots = \theta_L = \frac{1}{40}$, an Equi-probable DeMoivre variable

So, what is $P(\text{"even number"}) = P(\{2, 4, \dots, 40\}) =$

$= P(\{2\} \cup \{4\} \cup \dots \cup \{40\}) = P(\{2\}) + P(\{4\}) + \dots + P(\{40\}) =$

$= 20 \cdot \frac{1}{40} = \frac{1}{2}$

(action 3 repeatedly)



Properties:

1. $P(A) = 1 - P(A^c)$
2. A, B events. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

—

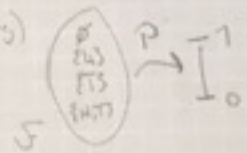
The domain of P is called a sigma field or sigma algebra,

It's denoted by $\mathcal{F}(\Omega)$ or $\mathcal{F}_\Omega$ or just $\mathcal{F}$ if Ω is clear from context.

We see that

i) $\Omega \in \mathcal{F}$, ii) $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$, iii) $A_1, A_2, \dots \in \mathcal{F} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$ by Kolmogorov.

ex Bernoulli experiment (coin flipping)



The triple $(\Omega, \mathcal{F}_\Omega, P)$ is called the probability space.
If $\mathcal{P}$ is a set of probabilities, then $(\Omega, \mathcal{F}_\Omega, \mathcal{P})$ is called a statistical experiment.

Recall

events A, B are independent $\Leftrightarrow P(A \cap B) = P(A)P(B)$

The product experiment

$X_1, X_2 \overset{\text{iid}}{\sim} P_\theta$ (indep., ident. dist.), then $P_\theta(\{X_{i_1} = x_1, X_{i_2} = x_2, \dots, X_{i_n} = x_n\}) = \prod_{i=1}^{n} P_\theta(X_{i_k} = x_k)$

Conditional probability

$P(A|B) := \frac{P(A \cap B)}{P(B)}$ provided $P(B) > 0$

is conditional probability a prob.? Yes, basically B is new Ω

ex Constructing random graph from tossing unfair coins.

Toss a coin with $P(H) = \theta$ iid n^2 times.

Graph:

Let $V = \{v_1, \dots, v_n\}$ be a set of vertices and let $E \subseteq V^2$ and $|E| = k$ (n vertices, k edges). Let

adjacency matrix of graph G_E

$$A = \begin{array}{c} \\ v_1 \\ v_2 \\ \vdots \\ v_n \end{array} \begin{bmatrix} v_1 & v_2 & \cdots & v_n \\ & & & \end{bmatrix}$$

where $A_{ij} = \begin{cases} 1, & \text{if } H \\ 0, & \text{if } T \end{cases}$ in coin toss

so, $E(|E|) = |V|^2 \theta$

Excercises

6.11

$$f(t) = \begin{cases} \lambda e^{-\lambda t}, & t > 0 \\ 0 & \text{otherwise} \end{cases}$$

a) $\lambda = 0.01$

$P(\text{"burns at least } \tau \text{ hours"}) = 1 - F(\tau) = 1 - \int_0^\tau \lambda e^{-\lambda t}\, dt = 1 - \left[-e^{-\lambda t}\right]_{t=0}^{\tau} =$

$= 1 + e^{-\lambda t} - 1 + e^{-0.01 \cdot \tau}$

b) $e^{-0.01\tau} = \frac{1}{2} \iff -0.01\tau = \ln\left(\frac{1}{2}\right) \iff \tau = \frac{\ln\left(\frac{1}{2}\right)}{-0.01} \approx 69.3\,h$

6.12

576 squares, 537 hits

$P(0) = \frac{229}{576}$, $P(1) = \frac{211}{576}$, $P(2) = \frac{93}{576}$, $P(3) = \frac{35}{576}$, $P(4) = \frac{7}{576}$, $P(\geq 5) = \frac{1}{576}$

$\frac{229}{576} = P(0) \approx \lambda e^{-\lambda \cdot 0} = \lambda \Rightarrow \lambda \approx \frac{229}{576}$

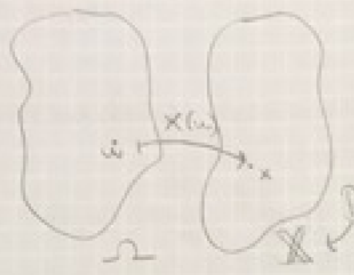
$\lambda = \frac{537}{576}$, "No. of hits" $\sim P_o(\lambda)$

approx $= [576 \cdot D(P_o(\lambda) \cdot i) \text{ for } i \text{ in } (0,1,2,3,4,5)] = [226.7, 211.4, 98.5, 30.6, 7.1, 1.3]$

Which is very close to actual.

Prelude to Decision theory

Estimation in parametric models



Data ~ Random variable (R.V.) → $\mathbb{X}$

prob. mass function (disc) ↘ ↙ prob. dist function (cont.)

distribution

⇕

Problem

Let X be a RV with values in $\mathbb{X}$ and $Law(X) = L(X)$ is law of x.

Assume $L(X)$ is known up to a finite dimensional parameter θ from the parameter space Θ (bold theta):

$$L(X) \in \{P_\theta \mid \theta \in \Theta\}, \text{ here we assume } \Theta \subseteq \mathbb{R}^d \text{ for } d \geq 1.$$

The decision problem is to estimate a function $g(\theta)$ based on a realisation of X.

Typically (WLOG)

$X = (X_1, \ldots, X_n)$, $X_i \overset{iid}{\sim} X_1$

n is called sample size, (initially) $\mathbb{X}$ is countable or $\subseteq \mathbb{R}^n$.

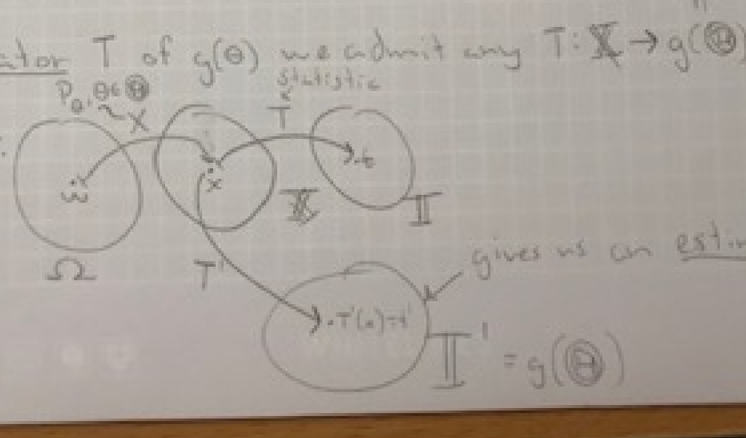
Def

A statistic T is an arbitrary function of the observed R.V. X (data)

Def

As an estimator T of $g(\theta)$ we admit any statistic $T : \mathbb{X} \to g(\Theta)$, where $g(\Theta) = \{g(\theta) \mid \theta \in \Theta\}$.

Concretely:



gives us an estimate of $g(\theta)$

Indicator function

$X(\omega) \mathbb{1}_{(A)}(\omega) := \begin{cases} 1 \text{ if } \omega \in A \\ 0 \text{ if } \omega \notin A \end{cases}$ is Bernoulli

$Law(X_i) \sim Bernoulli(\theta) = Bin(1, \theta)$

Suppose the data vector $x = (x_1, \dots, x_n)$ is a realisation of $X \sim \bigotimes_{i=1}^{n} Bern(\theta)$, i.e. $x \in \mathbb{X} = \{0,1\}^n$, for unknown but fixed $\theta \in \Theta = [0,1]$

Note that $P_\theta(X=x) = \prod_{i=1}^{n} P_\theta(X_i = x_i) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i} \overset{\otimes}{=} \theta^{k}(1-\theta)^{n-k}$.

Let $T(x) = \sum_{i=1}^{n} x_i$

Thus, the prob. of data (i.e. $P_\theta(X=x)$) only depends on the statistic t

Now, consider another statistic: (sample mean)

$T(x) = \frac{1}{n}\sum_{i=1}^{n} x_i =: \overline{x}_n$. $\quad t = n\bar{x}$

Then $\otimes$ becomes $\theta^{k}(1-\theta)^{n-k} \overset{k=n\bar{x}}{=} \theta^{n\bar{x}}(1-\theta)^{n(1-\bar{x})} = \theta^{n\bar{x}}(1-\theta)^{n-n\bar{x}}$

An estimator of θ, $g(\theta)$, (say $g(\theta)=\theta$), should converge towards to the "true" but unknown θ to be estimated, as the sample size $n \to \infty$

Def

A sequence $T_n := T_n(\overbrace{X_1, X_2, \dots, X_n}^{x^{(n)}})$ of estimators (each based on a sample of size n) for a parameter θ is called (asymptotically) consistent if $\forall \varepsilon > 0, \theta \in \Theta : P_{\theta n}(|T_n(x^{(n)}) - \theta| > \varepsilon) \to 0$ as $n \to \infty$ or, in shorter notation:

$T_n = T_n(X^{(n)}) \xrightarrow[n\to\infty]{P} \theta$ if $Law(X^{(n)}) = P_\theta$

ex (Bern. product ex.)

The estimator $T_n(X_1,\dots,X_n)=\overline{X}_n$ is consistent for θ

Proof

Since $X_1,\dots,X_n$ are iid $Be(\theta)$ R.V.s, we have $E(X_1)=\theta$ and the result follows from the law of large numbers

So, consistency can be seen as a minimal requirement on estimators. But this still leaves a lot of consistent estimators to choose from. A quantitative comparison of estimators is made possible by the approach of statistical decision theory.

We choose a loss function $Loss(t,\theta)$ which measures the loss (inaccuracy) when the unknown parameter θ is estimated by t. (this estimate is a realization of T, our estimator $T(X^{(n)})$)

Natural choices for loss when $\Theta \subseteq \mathbb{R}$

Absolute error: $Loss(t,\theta)=|t-\theta|$

Quadratic error: $Loss(t,\theta)=(t-\theta)^2$

or : $Loss(t,\theta)=\mathbb{1}_{[\delta,\infty)}(|t-\theta|)$ for some $\delta>0$ to emphasize the distance being less than δ

Note

Loss is a R.V. : $Loss(T(X),\theta)$ needs to account for randomness.

Def

The risk of an estimator T at parameter θ is

$$R(t,\theta) := E_\theta(Loss(T(X),\theta))$$

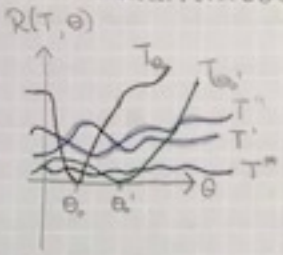
↑ Risk function of T

Note Risk might not exist since the expectation might not exist

Idea

$R(T^*, \theta) = \min_{T \in \mathcal{T}} R(T, \theta)$ for any fixed $\theta \in \Theta$ (class of estimators)

We find an estimator T^* that minimizes the whole risk function simultaneously. If such a T^* can be found, then it is called a uniformly best (UBE) estimator (in $\mathcal{T}$).



In general, such a UBE won't exist

Argument

For each $\theta_0 \in \Theta$, consider $T_{\theta_0}(x) = \theta_0$

so $R(T_{\theta_0}, \theta_0) = 0 \Rightarrow$ best if θ_0 is true

If T^* was UBE, then it would have to compete with T_{θ_0}

Unbiased estimators

Def

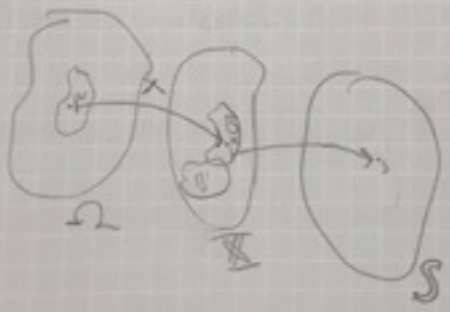
Consider an estimator T s.t. $E_\theta T$ exists.

Then $E_\theta(T) - g(\theta)$ is called bias of the estimator.

If $E_\theta(T) = g(\theta)$ for any $\theta \in \Theta$, then the estimator T is called unbiased for $g(\theta)$

Def

A statistic S is called sufficient for θ if $P_\theta(X \in B \mid S(X) = s)$ is independent of θ, for all values of s and all events B. (for all)



In words, conditional distribution of the data X given that $S(X)$ takes any value does not depend on the parameter θ

For discrete experiments

$$P_\theta(X \in B \mid S(X) = s) = \begin{cases} \dfrac{P_\theta(\{X \in B\} \cap \{S(X) = s\})}{P_\theta(S(X) = s)}, & \text{if } P_\theta(S(X) = s) > 0 \\ 0, & \text{if } P_\theta(S(X) = s) = 0 \end{cases}$$

Let's clarify again:

Originally, the law of X depended on θ ($\text{Law}(X) = P_\theta(\cdot)$).

After the value of the sufficient statistic $S(X) = s$ is known, then the conditional law $P_\theta(\cdot \mid S(X) = t)$ is no longer dependent on θ.

Since we are interested in making inferences about θ ($g(\theta)$), the conditional law is uninteresting for our purpose, so we can disregard it.

After taking $S(X)$ into account, the "remaining randomness" does not depend on θ anymore.

Remark (exercise to prove formally)

The data itself is sufficient, i.e.

if $S(X) = X$, then for $s = x$, $P_\theta(X \in B \mid X = x) = \mathbb{1}_B(x) = \begin{cases} 1, & \text{if } x \in B \\ 0, & \text{if } x \notin B \end{cases}$

Prop

In $\bigotimes_{i=1}^{n} \text{Be}(\theta)$ exp., the sample mean $\overline{X}_n$ is a sufficient statistic.

Proof

$n\overline{X}_n = \sum_{i=1}^{n} X_i \sim \text{Bin}(n, \theta)$

Suppose $\overline{X}_n = t$ where t is one of the possible values. This means $n\overline{X}_n = k$ for some $k \in \{0, 1, \dots, n\}$. Then, for any $x = (x_1, \dots, x_n)$,

$$P_\theta(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n \mid n\overline{X}_n = k) = \frac{P_\theta(X_1 = x_1, \dots, X_n = x_n, n\overline{X}_n = k)}{P_\theta(n\overline{X}_n = k)} \quad (*)$$

If θ fulfills $\sum_{i=1}^{n} x_i = k$, then $(*)$ is $\dfrac{(\theta^k(1-\theta)^{n-k})}{\binom{n}{k}\theta^k(1-\theta)^{n-k}} = \dfrac{1}{\binom{n}{k}}$ which clearly is indep. of θ.

If for this $\bar{x}$, $\sum_{i=1}^{n} x_i = k$, then the numerator is 0 which is indep. of θ, so Θ is indep. of θ. □

Say we want to estimate θ in this example and we limit ourselves to estimators that are functions of the sufficient statistic $\bar{X}_n$:

$$T(x) = h(\bar{X}_n)$$

Additionally, suppose we limit ourselves further to unbiased estimators:

$$E(T(X)) - \theta = 0 \Rightarrow E(T(X)) = \theta$$

Prop.

Under sufficiency and and unbiasedness restrictions on allowed estimators of θ in our $\bigotimes_{i=1}^{n} Be(\theta)$ exp., the only possible estimator is $\bar{X}_n$

Proof $\quad E(\bar{X}_n) = \theta$

$$0 = E_\theta h(\bar{X}_n) - \theta \overset{\downarrow}{=} E_\theta(h(\bar{X}_n) - \bar{X}_n) = \sum_{k=0}^{n} \underbrace{\left(h\left(\tfrac{k}{n}\right) - \tfrac{k}{n}\right)\binom{n}{k}}_{C_k} \theta^k (1-\theta)^{n-k} =$$

$$= (1-\theta)^n \sum_{k=0}^{n} C_k r^k \quad \text{for } r := \frac{\theta}{1-\theta}$$

Now, if $\theta \in [0,1)$, then $r \in [0, \infty)$, hence the above polynomial can only be zero if every C_k is also zero.

Thus,

$$0 = \binom{n}{k}\left(h\left(\tfrac{k}{n}\right) - \tfrac{k}{n}\right) \overset{\binom{n}{k} \neq 0}{\Rightarrow} h\left(\tfrac{k}{n}\right) = \tfrac{k}{n} \text{ for } k \in \{1, \ldots, n\} \text{ and } h(\bar{X}_n) = \bar{X}_n \text{ for}$$

all possible values of $\bar{X}_n$ □

Def

A statistic T is called complete if, for all $h: \mathbb{T} \to \mathbb{R}$:

$[E_\theta(h(T(X))) = 0 \text{ for all } \theta \in \Theta] \Rightarrow$

$\Rightarrow P(\underbrace{h(T(X)) = 0}_{\text{almost sure event}}) = 1 \text{ for all } \theta \in \Theta$

Intuitively, completeness means there is "no superfluous information" or "no redundancy" in the complete statistic T

Consider an event $T(X) \in B$ and suppose $P_\theta(T(X) \in B) = a$ for a indep. of θ.
By taking $h(t) = \mathbb{1}_B(t) - a$, completeness $\Rightarrow P_\theta(\mathbb{1}_B(T(X)) = a) = 1$ for all $\theta \in \Theta$ which means that a is either 0 or 1. Thus, for any event $T(X) \in B$ which has a non-trivial probability $(\neq 0, 1)$, this prob. must depend on θ.

eg/ $\bigotimes_{i=1}^{n} Be(\theta)$

The statistic $T(X) = (X_1, \overline{X_n})$ is sufficient but not complete:

$h(X_1, \overline{X_n}) = X_1 - \overline{X_n}$ has $E(h) = 0$ but $h(X_1, \overline{X_n})$ is not almost surely 0.

Prop

In $\bigotimes_{i=1}^{n} Be(\theta)$, $T(X) = \overline{X_n}$ is sufficient and complete.

Proof

Suppose for some function h,

$E_\theta(h(\overline{X_n})) = 0$

This means that $\sum_{k=0}^{n} h(\frac{k}{n}) \binom{n}{k} \theta^k (1-\theta)^{n-k} = 0 \quad \forall \theta \in [0, 1]$.

Then, $h(\frac{k}{n}) = 0$ for $k \in \{0, 1, \ldots, n\}$ as per the earlier argument used to show that $\overline{X_n}$ is the only unbiased and sufficient estimator.

Limits of R.V.s : Chapter 8 in CSE Book.pdf 151-156

8.1 - conv. of sequence of R.V.

$X_1, \ldots, X_n$ conv. to a R.V. X:

- In distribution

$X_n \rightsquigarrow X$

- In probability

$X_n \xrightarrow{p} X$

- Markov's ineq., Chebyshev's ineq.

③ Suppose T is a sufficient statistic with values in a set $\mathbb{T}$ and $S: \mathbb{T} \to \mathbb{S}$ is a one-to-one mapping with values in $\mathbb{S}$ (ie, there exists an inverse mapping $S^{[-1]}$ such that $S^{[-1]}(S(t)) = t$ for each $t \in \mathbb{T}$). Show that the statistic $S(T(X))$ is sufficient

④ In class it was claimed that the statistic $T(X) = (X_1, \bar{X}_n)$ is sufficient for the $\bigotimes_{i=1}^{n}$ Bernoulli(θ) experiment. Prove this claim.

Problem Set Week 3

1. Show that the estimator $\bar{X}_n$ is consistent for θ in $\bigotimes_{i=1}^{n}$ Bernoulli(θ) experiment
2. Suppose the data X in a statistical experiment can take values in a countable set $\mathbb{X}$ (i.e., Law(X) is discrete). In class it was claimed that the data itself are a sufficient statistic (i.e., $T(X) = X$ is sufficient). Write down the argument that proves this claim (It can be a short paragraph)

5. Let $X_1, \ldots, X_n$ be independent and identically distributed with $\text{Law}(X_i) \sim \text{Poisson}(\lambda)$, $\lambda \in (0, \infty)$ is unknown. Show that the sample mean $\bar{X}_n$ is a sufficient statistic.

Limits of R.V.s : Chapter 8 in CSE Book.pdf 151-156

8.1 - conv. of sequence of R.V.

$X_1, \dots, X_n$ conv. to a R.V. X:

- In distribution

  $X_n \rightsquigarrow X \iff \forall t$ where F_X cont: $\lim_{n\to\infty} F_{X_n}(t) = F_X(t)$ (pointwise convergence of dist. function)

- In probability

  $X_n \xrightarrow{P} X \iff \forall \varepsilon > 0 : \lim_{n\to\infty} P(|X_n - X| > \varepsilon) = 0$ (uniform conv. of dist. function)

- Markov's ineq., Chebyshevs ineq.

$\forall \varepsilon > 0 : P(X \geq \varepsilon) \leq \frac{E(X)}{\varepsilon}$, $P(|X| \geq \varepsilon) \leq \frac{E(|X|)}{\varepsilon}$, $P(|X| > \varepsilon) = P(X^2 \geq \varepsilon^2) \leq \frac{E(X^2)}{\varepsilon^2}$

$P(|X - E(X)| \geq \varepsilon) \leq \frac{V(X)}{\varepsilon^2}$

Problem Set week 3

(1) Show that the estimator $\overline{X_n}$ is consistant for Θ in $\bigotimes_{i=1}^{n} Be(\Theta)$ experiment

(2) Suppose the data X in a stat. experiment can take values in a countable set $\mathbb{X}$ (i.e. $Law(X)$ is discrete).
In class it was claimed that the data itself is a sufficient statistic (i.e. $T(X) = X$ is sufficient). Write down an argument for that claim that proves it.

(3) Suppose T is a sufficient statistic with values in a set $\mathbb{T}$ and $S : \mathbb{T} \to \mathbb{S}$ is a (one to one (bijective)) mapping with values in $\mathbb{S}$ (i.e. $\exists S^{[-1]} : \mathbb{S} \to \mathbb{T}$ s.t. $S^{[-1]}(S(t)) = t \ \forall t \in \mathbb{T}$). Show that the statistic $S(T(X))$ is sufficient.

(4) In class it was claimed that the statistic $T(X) = (X_1, \overline{X_n})$ is sufficient for the $\bigotimes_{i=1}^{n} Be(\Theta)$ experiment. Prove this claim.

(5) Let $X_1, \dots, X_n$ be independent and identically distributed (iid) with $Law(X_i) \sim Po(\lambda)$, $\lambda \in (0, \infty)$ is unknown. Show that the sample mean $\overline{X_n}$ is a sufficient statistic for this data.

Weak law of large numbers

$X_1, X_2, \ldots \overset{IID}{\sim} X_1$, $E(X_1)$ exists. Then, $\overline{X}_n \xrightarrow{P} E(X_1)$

Proof (In the case where $V(X_1) < \infty$)

Let $\varepsilon > 0$. $P(|\overline{X}_n - E(\overline{X}_n)| \geq \varepsilon) \overset{\text{Cheb. ineq.}}{\leq} \frac{V(\overline{X}_n)}{\varepsilon^2} = \frac{1}{n}\frac{V(X_1)}{\varepsilon^2}$ ($X_1, X_2, \ldots \overset{IID}{\sim} X_1$).

We also know that $E(\overline{X}_n) = E(X_1)$ since $X_1, X_2, \ldots \overset{IID}{\sim} X_1$ so $E(X_1) = E(X_2) = \cdots$

so $E(\overline{X}_n) = E\left(\frac{\sum_{i=1}^n X_i}{n}\right) = \frac{\sum_{i=1}^n E(X_1)}{n} = \frac{n E(X_1)}{n} = E(X_1)$.

So $P(|\overline{X}_n - E(X_1)| \geq \varepsilon) = P(|\overline{X}_n - E(\overline{X}_n)| \geq \varepsilon) = \frac{1}{n}\frac{V(X_1)}{\varepsilon^2} \to 0$ as $n \to \infty$.

Thus $\overline{X}_n \to E(X_1)$ □

Central limit theorem (CLT)

If $X_1, X_2, \ldots \overset{IID}{\sim} X_1$ and $E(X_1), V(X_1)$ exist, then

$$\overline{X}_n = \frac{\sum_{i=1}^n X_i}{n} \rightsquigarrow X \sim \text{Normal}\left(E(X_1), \frac{V(X_1)}{n}\right).$$

Cor.

i) $\overline{X}_n - E(X_1) \rightsquigarrow X - E(X_1) \sim \text{Normal}\left(0, \frac{V(X_1)}{n}\right)$

ii) $\sqrt{n}\left(\overline{X}_n - E(X_1)\right) \rightsquigarrow \sqrt{n}\left(X - E(X_1)\right) \sim \text{Normal}(0, V(X_1))$

iii) $Z_n := \frac{\overline{X}_n - E(\overline{X}_n)}{\sqrt{V(\overline{X}_n)}} = \frac{\sqrt{n}\left(\overline{X}_n - E(X_1)\right)}{\sqrt{V(X_1)}} \rightsquigarrow Z \sim \text{Normal}(0,1)$

iv) $\lim_{n\to\infty} P\left(\frac{\overline{X}_n - E(\overline{X}_n)}{\sqrt{V(\overline{X}_n)}} \leq z\right) = \lim_{n\to\infty} P(Z_n \leq z) = P(Z \leq z) = \phi(z) := \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-\frac{x^2}{2}}\,dx$

Approximation

For "large" n ($\approx n > 30$), we can use

$$P\left(\frac{\overline{X}_n - E(\overline{X}_n)}{V(\overline{X}_n)} \leq z\right) \approx P(Z \leq z) = \phi(z)$$

Exercises

1. Prove that the statistic $T(X_1,\dots,X_n)=\frac{\sum_{i=1}^{n}X_i}{n}=\overline{X_n}$ is consistent for θ in $\bigotimes_{i=1}^{n}$ Bernoulli(θ) exp.

Sol.

We have to show that $\overline{X_n}$ is consistent, i.e. $\overline{X_n}\xrightarrow{P}\theta$.

Let $\varepsilon>0$. $S_n=\sum_{i=1}^{n}X_i$ so $\overline{X_n}=\frac{S_n}{n}$. $P(|\overline{X_n}-\theta|>\varepsilon)=$

$=P(|S_n-\theta n|>\varepsilon n)$. Since $V(S_n)=n\theta(1-\theta)$, by Chebyshev:

$P(|\overline{X_n}-\theta|>\varepsilon)=P(|S_n-\theta n|>\varepsilon n)\le\frac{V(S_n)}{\varepsilon^2n^2}=\frac{\theta(1-\theta)}{\varepsilon^2n}\le\frac{\frac{1}{4}}{\varepsilon^2n}=\frac{1}{4\varepsilon^2n}\xrightarrow{n\to\infty}0$ □

Proof of Chebyshev:

$$P(|X-E(X)|>a)=P((X-E(X))^2>a^2)\overset{\text{Markov}}{\le}\frac{1}{a^2}E((X-E(X))^2)=\frac{V(X)}{a^2}$$

2. Suppose X takes values in $\mathbb{X}$ where $|\mathbb{X}|\le|\mathbb{N}|$.

Let $S(X)=s$ if $X=s$. We want to show that

$P_\theta(X\in B\mid S(X)=s)$ is independent of θ. $B\subseteq\mathbb{X}$

$$P_\theta(X\in B\mid S(X)=s)=P_\theta(X\in B\mid X=s)=\begin{cases}\frac{P_\theta(\{X\in B\}\cap\{X=s\})}{P_\theta(X=s)}, & P_\theta(X=s)>0\\ 0, & P_\theta(X=s)=0\end{cases}$$

$$P_\theta(\{X\in B\}\cap\{X=s\})=\begin{cases}P_\theta(X=s), & s\in B\\ 0, & s\notin B\end{cases},\ \text{so}$$

$$P_\theta(X\in B\mid S(X)=s)=\begin{cases}\frac{P_\theta(X=s)}{P_\theta(X=s)}, & P_\theta(X=s)>0,\ s\in B\\ 0, & \text{otherwise}\end{cases}=\begin{cases}1 & \text{if } s\in B\\ 0 & \text{if } s\notin B\end{cases}$$

Alt:

Assume $T(X)=X$, let $t=x$. For an event $B\in\mathcal{F}_{\mathbb{X}}$

$P_\theta(X\in B\mid T(X)=t)=P_\theta(X\in B\mid X=x)=\mathbb{1}_B(x)=\begin{cases}1, & x\in B\\ 0, & x\notin B\end{cases}$.

3. Suppose T sufficient $T: \mathcal{X} \to \mathcal{T}$ and $S: \mathcal{T} \to \mathcal{S}$ inj. Show that S is sufficient.

sol.

$\forall s \in \mathcal{S} := \{s : S(t) = s\}$ and $\forall B \in \mathcal{F}_{\mathcal{X}}$ the cond. prob.

$$\frac{P_\theta(\{X \in B\} \cap \{S(T(X)) = s\})}{P_\theta(S(T(X)) = s)} = \frac{P_\theta(\{X \in B\} \cap \{T(X) = t\})}{P_\theta(T(X) = t)} \quad \text{where } t = S^{-1}(s)$$

$= P_\theta(X \in B \mid T(X) = t)$ which is indep. of θ because T suff. □

Prop.

In the $\bigotimes_{i=1}^{n} Be(\theta)$ experiment, the estimator $\overline{X_n}$ is uniformly best among unbiased estimators for quadratic loss (i.e., for any unbiased estimator T : $R(\overline{X_n}, \theta) = E_\theta\left((\overline{X_n} - \theta)^2\right) \leq E_\theta\left((T(X) - \theta)^2\right) = R(T, \theta)$ for all $\theta \in [0,1]$).

Before we prove this, let's revisit the notion of cond. expectation:

Let Y be a R.V. with finite support in $\mathcal{Y}$, i.e. $|\mathcal{Y}| < \infty$.

Let U be a statistic with values in $\mathcal{U}$ and $u \in \mathcal{U}$, h a real-valued fct. of Y.

Def

The cond. expectation of $h(Y)$ given $U = u$, written $E(h(Y) \mid U(Y) = u)$ is defined as the expectation of $h(Y)$ under the cond. distribution of Y given $U(Y) = u$: $P(Y = y \mid U(Y) = u) = \frac{P(Y = y, U(Y) = u)}{P(U(Y) = u)}$ if $P(U(Y) = u) \neq 0$ (0 otherwise).

$$E(h(Y) \mid U(Y) = u) = \sum_{y \in \mathcal{Y}} h(y) P(Y = y \mid U(Y) = u).$$

In special case $h(y) = 1_B(y)$ for some $B \subseteq \mathcal{Y}$, then

$$E(1_B(y) \mid U(Y) = u) = \sum_{y \in B} P(Y = y \mid U(Y) = u) = P(y \in B \mid U(Y) = u).$$

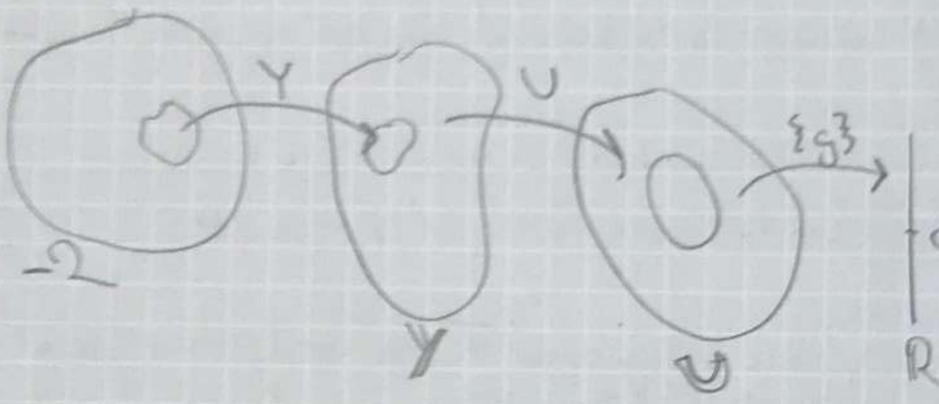
Def

The Conditional expectation can be a R.V.

$E(h(Y)|U)$

Note the following properties of cond. exp. as a R.V.:

Let $\mathcal{M}_U$ be the set of all real-valued R.V.s Z that are functions of U, i.e. $Z = g(U)$ for some function $g: \mathcal{U} \to \mathbb{R}$.



Recall
we assume $|\mathcal{Y}| < \infty$

i) $E(E(h(Y)|U)) = E(h(Y))$

proof

exercise 0

ii) For any $Z \in \mathcal{M}_U$, $E(Zh(Y)|U) = ZE(h(Y)|U)$ "almost surely" / / "with prob. 1" / w.p.1

Intuition:

R.V.s X, Y equal w.p.1 means $1 = P(X = Y)$

iii) For any function $h: \mathcal{Y} \to \mathbb{R}$ s.t. $h = h_1 + h_2$, then $E(h(Y)|U) =$

$= E(h_1(Y)|U) + E(h_2(Y)|U)$ w.p.1

iv) For any $Z \in \mathcal{M}_U$, $E(Z \cdot h(Y)) = E(Z \cdot E(h(Y)|U))$

proof

exercise 1

v) For any $Z \in \mathcal{M}_U$, $E((h(Y) - Z)^2) = E((E(h(Y)|U) - Z)^2) + E(h(Y) - E(h(Y)|U))^2$

$\leq E(h(Y) - E(h(Y)|U))^2$ ⊛

proof

exercise 2 (assuming iii) and iv))

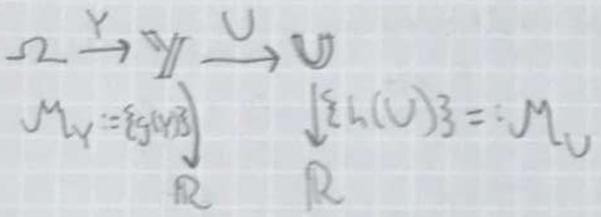
hint: add and subtract $E(h(Y)|U)$ in ⊛

and

$E(E(h(Y)|U)|U) = E(h(Y)|U)E(1|U) = E(h(Y)|U)$

Remark

The conditional expectation can be seen as an operator. Let $\mathcal{M}_Y$ be all $\mathbb{R}$-valued R.V.s which are functions of Y. Then, since U is $U(Y)$ we have $\mathcal{M}_U \subseteq \mathcal{M}_Y$



and both $\mathcal{M}_U$ and $\mathcal{M}_Y$ are vector spaces.

Let $H \in \mathcal{M}_Y$, then the cond. expect. $E(H|U) \in \mathcal{M}_U$ and we define the operator $\Pi(H) := E(H|U)$, i.e. $\Pi: \mathcal{M}_Y \to \mathcal{M}_U$

The "$\leq$" in prop. v) tells us that $\underbrace{E\left(H-\Pi(H)\right)^2}_{E\left[\left(h(Y)-E(h(Y)|U)\right)^2\right]} = \min_{z \in \mathcal{M}_U} \underbrace{E\left(H-z\right)^2}_{E\left(\left(h(Y)-z\right)^2\right)}$.

In English: the cond. expect. of H is the element of $\mathcal{M}_U$ which is closest to H. You can see this as the "projection" Π of H onto the space $\mathcal{M}_U$.

The "$=$" in prop. v) is the orthogonal decomposition

$$E\left((H-Z)^2\right) = E\left(\left(\Pi(H)-Z\right)^2\right) + E\left(\left(H-\Pi(H)\right)^2\right)$$

We are now ready to prove that $\overline{X}_n$ is uniformly best among all unbiased estimators of θ in $\bigotimes_{i=1}^{n} Be(\theta)$ exp. :

Proof

N.T.S. (need to show) that the risk of $\overline{X}_n$ is the lowest.

i.e. $R(\overline{X}_n,\theta) := E_\theta\left((\overline{X}_n-\theta)^2\right) \le E_\theta\left((T(X_1,\dots,X_n)-\theta)^2\right) =: R(T,\theta)$

for all $\theta\in[0,1]$ and all unbiased estimators T.

Define a R.V.: $g(\overline{X}_n) = E_\theta(T\mid\overline{X}_n) = E(T\mid\overline{X}_n)$ and regard it as an estimator of θ

($\overline{X}_n$ is suff. for θ)

by prop. (i), $E(g(\overline{X}_n)) = E\left(E(T\mid\overline{X}_n)\right) = E(T) = \theta$

(T unbiased)

so, g is unbiased.

Before, we showed that $g(\overline{X}_n)=\overline{X}_n$ is the only unbiased estimator of θ ($\overline{X}_n$ is complete)

Hence, $g(\overline{X}_n)=\overline{X}_n$

In prop. (v), set $Z=\theta$, (point-mass R.V $(P(Z=z)=\begin{cases}1, z=0\\ 0, z\neq 0\end{cases})$, so $Z\in M_{U,}$

so $Y=X$, $h(Y)=T$, $U=\overline{X}_n$, $E(h(Y)\mid U) = g(\overline{X}_n)$

to get $E_\theta(T-\theta)^2 \ge E_\theta(g(\overline{x}_n)-\theta)^2$

so, the estimator $g(\overline{X}_n)=\overline{X}_n$ is at least as good as any T. □

For unbiased estimators, the quadratic risk is also the estimators variance:

$$E_\theta(T-\theta)^2 = V(T) := E_\theta\left(T-E_\theta(T)\right)^2$$

$\therefore$ $\overline{X}_n$ for $\bigotimes_{i=1}^{n} Be(\theta)$ exp is also called a uniform minimum variance unbiased estimator (UMVUE)

Bayes estimators

In the Bayesian approach to estimation of the param. θ underpinning the Law of data $X=(X_1,\dots,X_n)$ in an exp, one assumes that prior distribution is given over the parameter space $\Theta \ni \theta$

ex/ In the $\bigotimes_{i=1}^{n} Be(\theta)$ exp., assume that the prior distribution is given in the form of density on $\Theta=[0,1]$.

Def (Integrated risk)

For an estimator T of θ the prior $g(\theta)$ can be used to reduce the risk function $R(T,\theta)$ for each $\theta \in [0,1]$ to a single number, by integration:

$$B(R(T)) = B_r(T) := \int_0^1 R(T,\theta)\, g(\theta)\, d\theta = \int_0^1 E_\theta (T-\theta)^2\, g(\theta)\, d\theta$$

Suppose we have quadratic loss of T under risk R.

$B_r(T)$ is called integrated risk or mixed risk

Def

A Bayes estimator T_B of θ is the estimator that minimizes the integrated risk, i.e. $T_B := \underset{T}{\operatorname{argmin}}\, B_R(T)$

$\underset{T}{\operatorname{argmin}}\, B_R(T) := T$ s.t. $B_R(T) = \min_T B_R(T)$

and $B_R(T_B)$, the minimal integrated risk, is called Bayes Risk

The name "Bayesian" comes from Bayes formula:

$\{D_1,\dots,D_k\}$ partition Ω, then $P(A) = \sum_{i=1}^{k} P(A|D_i)P(D_i)$

Motivation: Consider the case when P is the joint distribution of (X,U) where X is data and U is a R.V. that takes k possible values in $\Theta := \{\theta_1,\dots,\theta_k\}$. Then $A = "X \in A"$ and $D_i = "U=\theta_i"$ (True param.)

$$P(U=\theta_j | X \in A) = \frac{P(X \in A | U=\theta_j) P(U=\theta_j)}{\sum_{i=1}^{k} P(X \in A | U=\theta_i) P(U=\theta_i)}$$

Posterior prob. of θ_j given data ← observed data; prior prob of param $\theta \in \Theta$

The Bayesian approach views $\{P_\theta : \theta \in \Theta\}$ as a family of conditional distributions given θ, awaits a prior dist.

ex In $\bigotimes_{i=1}^{n} Be(\theta)$ exp., consider the family of prior densities for $\theta \in [0,1]$:

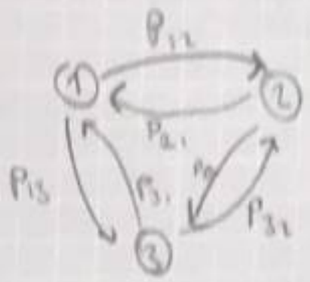
for each $(\alpha,\beta) \in (0,\infty)^2 =: \mathbb{R}^2_{>0}$ $(\alpha>0, \beta>0)$

$$g_{\alpha,\beta}(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}, \quad \theta \in [0,1] \quad \text{where } B \text{ is beta function:}$$

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}, \text{ and } \Gamma(\alpha) := \int_0^\infty x^{\alpha-1} e^{-x}\, dx.$$

Consider the whole family $g_{\alpha,\beta}$ for specification of prior density.

Markov Chains



$$P = \begin{bmatrix} 0 & p_{12} & p_{13} \\ p_{21} & 0 & p_{23} \\ p_{31} & p_{32} & 0 \end{bmatrix}$$

If they "fire", state implies

Let $q_{11}, q_{22}, q_{33} \sim Exp(\lambda_{1,2,3})$

$$Q = \begin{bmatrix} -q_{11} & q_{12} & q_{13} \\ q_{21} & -q_{22} & q_{23} \\ q_{31} & q_{32} & -q_{33} \end{bmatrix}$$

$q_{ij} = q_{ii} p_{ij}$ if $i \neq j$

$$P(t) = e^{Qt}$$

Bayes cont.

Note

We consider the whole family $g_{\alpha\beta}$, induced by $(\alpha,\beta) \in \mathbb{R}^2_{>0}$ to allow a wide range of prior beliefs about Θ.

It will become clear that the Bayesian approaches are very useful even to prove non-Bayesian prop. of estimators.

Prop. In $\bigotimes_{i=1}^{n} Be(\theta)$ exp, let g be an arbitrary prior density for $\theta \in [0,1] =: \Theta$, then the Bayes estimator is

$$T_B(x) : \mathcal{X} \to \Theta = \frac{\int_0^1 \theta^{s(x)+1} (1-\theta)^{n-s(x)} g(\theta)\, d\theta}{\int_0^1 \theta^{s(x)} (1-\theta)^{n-s(x)} g(\theta)\, d\theta} \quad \text{where } s(x) = \sum_{i=1}^{n} x_i$$

Note

T_B is a function of the sufficient statistic $\bar{x}_n = \frac{s(x)}{n}$

Proof

exercise

Def

An estimator T of a parameter θ is called admissible if, for every estimator S of θ, the relation given by

Θ: $R(S,\theta) \leq R(T,\theta)$ $\forall \theta \in \Theta$ implies

Θ: $R(S,\theta) = R(T,\theta)$ — " —

Thus, admissibility of T means that there can be no estimator S which is uniformly at least as good [Θ holds] and strictly better for one $\theta_0 \in \Theta$, because then $R(S,\theta_0) < R(T,\theta_0)$ and thus $(**)$ is contradicted.

So, non-admissibility of T means that T can be improved by another estimator S.

Prop

Suppose that in $\bigotimes_{i=1}^{n} Be(\theta)$ exp, the prior density g is such that $g(\theta) > 0$ for each $\theta \in [0,1]$ with the exception of a finite number of points. Then the Bayes estimator T_B for this prior density is admissable for quadratic loss.

proof

exercise

When trying these 2 exercises, show them to Ruaz.

Def

Minimax Estimators

Let the maximal risk of an estimator T be $M(T) := \max_{\theta \in \Theta} R(T,\theta)$.

An estimator T_M is called Minimax if

$$M(T_M) = \min_{T} M(T) = \min_{T \in \mathcal{T}} \max_{\theta \in \Theta} R(T,\theta)$$

$\mathcal{T}$: some collection of estimators

Prop

In $\bigotimes_{i=1}^{n} Be(\theta)$ exp, the Bayes Estimator $T_{\alpha,\beta}$ for $\alpha = \beta = \frac{\sqrt{n}}{2}$ is a minimax estimator

Proof

exercise

Outline of course:

① Point Estimators
- Moments Estimators { solve { sample moments = theoretical moments
- Maximum likelihood estimators { likelyhood $L(\theta) \propto P(x_1, \dots, x_n | \theta)$; $\hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta)$

↳ CLT ↓

② Set Estimators

point est. $\hat{\theta}_n$

θ_0 $\hat{\theta}_n$ ← set, VERY high prob. to contain θ_0 (true law)

↓

③ Hypothesis testing

(1+2)' Regression : $f_{(y|x)}$

Parametric Inference

Recall the D.T. setting:

Given a param. exp: $(\Omega, \mathcal{F}, \mathcal{P})$, $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, $|\Theta| < \infty$

we are interested in some function $g(\theta)$ based on data x.

Typically, $x = (x_1, \dots, x_n)$, $x_1, \dots, x_n \overset{IID}{\sim} X_1$ has Law(θ) is from a product experiment, $x_1, \dots, x_n \overset{ind}{\sim} F_\theta = F_\theta(x_i; \theta)$ (or $f_\theta = f_\theta(x_i; \theta)$)

e.g. $x_1, \dots, x_n \sim \bigotimes_{i=1}^{n} Be(\theta)$

ex1

$x_1, \dots, x_n \overset{iid}{\sim} N(\mu, \sigma^2)$. Then $\theta = (\theta_1, \theta_2) = (\mu, \sigma^2) \in \Theta = \{(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma \in \mathbb{R}_{>0}\}$

Suppose X_i is the outcome of a blood-alcohol breath-test and we are interested in the fraction of testers with score >1.

What is $g(\theta)$? : $g(\theta) = P(X_1 > 1) = 1 - P(X_1 \le 1) = 1 - P\left(\underbrace{\frac{X_1 - \mu}{\sigma}}_{:= Z \sim N(0,1)} < \frac{1-\mu}{\sigma}\right) =$

$$= 1 - \Phi\left(\frac{1-\mu}{\sigma}\right)$$

ex2

$X \sim \Gamma(\alpha, \beta)$, $\theta = (\theta_1, \theta_2) = (\alpha, \beta) \in \mathbb{R}_{>0}^2$, recall $f(x; \alpha, \beta) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}$.

Often X is used to model lifetime of items.

Then we are interested in an estimate of the mean lifetime (in product exp):

$g(\theta) = g(\alpha, \beta) = E(X_1; \alpha, \beta) = \alpha\beta$

Method of Moments Estimator (MOME)

Suppose the parameter θ has k components $(\theta = (\theta_1, \dots, \theta_k))$, then for $1 \le j \le k$, define the j^{th} moment:

$$\alpha_j := \alpha_j(\theta) = E(X^j; \theta) = \int x^j dF_\theta(x)$$

and the j^{th} sample moment:

$$\hat{\alpha}_j := \frac{1}{n} \sum_{i=1}^{n} X_i^j$$

Def
The MOME $T_{MOM}(X) = \hat{\theta}_n$ is defined to be the value of θ s.t.

$$\begin{cases} \alpha_1(\hat{\theta}_n) = \hat{\alpha}_1 \\ \alpha_2(\hat{\theta}_n) = \hat{\alpha}_2 \\ \vdots \\ \alpha_k(\hat{\theta}_n) = \hat{\alpha}_k \end{cases}$$

i.e. $\hat{\theta}_n$ is the MOM estimate and is the solution θ to this system of k equations in k unknowns.

Note
MOME, if they exist, are typically suboptimal but it is often easy to compute and can be used to initialise the iterative strategies used to obtain more optimal estimators (e.g. MLE)

ex 3
$X_1, \dots, X_n \overset{iid}{\sim} Be(\theta)$. Find MOME for θ.

Sol.
$$\left.\begin{aligned} \alpha_1(\hat{\theta}_n) &= E(X_1; \theta) = \hat{\alpha}_1 \\ \hat{\alpha}_1 &= \frac{1}{n} \sum_{i=1}^{n} x_i = \overline{X_n} \end{aligned}\right\} \text{By equating we solve for } \hat{\theta}_n:$$

$$\hat{\theta}_n = \overline{X_n}$$

ex 4
$X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. Find MOME.

$\theta = (\theta_1, \theta_2) = (\mu, \sigma^2)$

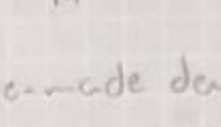
Sol
$$\alpha_1 = E(X_1; \mu, \sigma^2) = \mu$$

$$\alpha_2 = E(X_1^2; \mu, \sigma^2) = V(X_1; \mu, \sigma^2) + \left(E(X_1; \mu, \sigma^2)\right)^2 = \sigma^2 + \mu^2$$

Solve
$$\begin{cases} \hat{\mu}_n = \overline{X_n} \\ \hat{\sigma}_n^2 + \hat{\mu}_n^2 = \frac{1}{n} \sum_{i=1}^{n} X_i^2 \end{cases} \Rightarrow \sigma_n^2 = \frac{1}{n} \sum_{i=1}^{n} X_i^2 - \overline{X_n}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X_n})^2$$

ex 5

Show that MOME for parameter λ in a product $P_0(\lambda)$ exp is $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Recall $f(x;\lambda) = \mathbb{1}_{\mathbb{N}\cup\{0\}}(x)\frac{\lambda^x e^{-\lambda}}{x!}$

Exam practice: home-made densities are possible (e.g. )

Properties of MOME

Under "appropriate conditions" on the experiment, the following hold:

1) $\widehat{\theta}_n$ exists with prob $\xrightarrow[n\to\infty]{} 1$ (equations are solvable for large n)

2) $\widehat{\theta}_n \xrightarrow{p} \theta$ (asymptotic) consistency.

3) $\sqrt{n}(\widehat{\theta}_n - \theta) \rightsquigarrow N(0, \Sigma)$, $\Sigma = \underbrace{g}_{k\times k} \underbrace{E_\theta(YY^T)}_{k\times k} \underbrace{g^T}_{k\times 1}$, where $Y = [X, X^2, \dots, X^k]^T$,

$g = (g_1, \dots, g_k)$, $g_j := \frac{\partial}{\partial\theta}\alpha_j^{-1}(\theta)$

Maximum likelihood estimators (MLE)

$X_1, \dots, X_n \overset{iid}{\sim} X$, and $X \sim F(x;\theta)$ with PMF or PDF $\underbrace{f(x;\theta) = dF(x;\theta)}_{f_\theta(x) = dF_\theta(x)}$

Def

The likelihood function is defined by: if product exp

$$L_n(\theta) := L_n(\theta; X_1, \dots, X_n) = f(X_1, \dots, X_n; \theta) \overset{\text{i.i.d.}}{=} \prod_{i=1}^{n} f(X_i;\theta)$$

Note

$L_n(\theta): \Theta \to [0,\infty) =: \mathbb{R}_{\geq 0}$

Def

log-likelihood function is $\ell_n(\theta) := \ell_n(\theta; X_1, \dots, X_n) = \ln(L_n(\theta))$ ($\neq \ln$ as in logs)

Thus, the likelihood function is just the joint density of the data, except that we view this as a function of the parameter θ

random (because $x = (x_1, \dots, x_n)$ is random)

Def

The MLE $T_{ML}(x_1,\dots,x_n) = \hat{\theta}_n = \underset{\theta\in\Theta}{\operatorname{argmax}}\, L_n(\theta) = \underset{\theta\in\Theta}{\operatorname{argmax}}\, \ell_n(\theta)$

ex1

Let $x_1,\dots,x_n \overset{iid}{\sim} Be(\theta)$. Find MLE of θ.

sol

$L_n(\theta) = \prod_{i=1}^n f(x_i;\theta) = \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i} := \theta^s(1-\theta)^{n-s}$ ($f(x_i;\theta) = \mathbb{1}_{\{0,1\}}\theta^{x_i}(1-\theta)^{1-x_i}$)

$\Rightarrow \ell_n(\theta) = \ln\left(\theta^s(1-\theta)^{n-s}\right) = s\ln(\theta) + (n-s)\ln(1-\theta)$

$\frac{d}{d\theta}\ell_n(\theta) = \frac{1}{\theta}s - (n-s)\frac{1}{1-\theta} = 0 \Rightarrow \frac{1}{\theta}s = (n-s)\frac{1}{1-\theta} \Rightarrow (1-\theta)s = (n-s)\theta \Rightarrow$

$\Rightarrow s - \theta s = n\theta - s\theta \Rightarrow \theta = \frac{s}{n} = \overline{X_n}$

$\frac{d^2}{d\theta^2}\ell_n(\theta) = -\frac{s}{\theta^2} + \frac{n-s}{(1-\theta)^2} \Rightarrow \frac{\partial \ell_n}{\partial\theta^2}\left(\frac{s}{n}\right) = -\frac{s}{\left(\frac{s}{n}\right)^2} + \frac{n-s}{\left(1-\frac{s}{n}\right)^2} = -\cdots < 0 \Rightarrow$ max!

ex2

Show that MLE $\hat{\Lambda}_n$ of par. λ in a product $Exp(\lambda)$ exp is $\frac{1}{\overline{X_n}}$.

Recall $X_{iid},X_n \overset{iid}{\sim} Exp(\lambda)$ means $f(x_i;\lambda) = \lambda e^{-\lambda x}\mathbb{1}_{(0,\infty)}(x_i)$ for $\lambda\in(0,\infty)$

ex $X_1,\dots,X_n \overset{iid}{\sim} N(\mu,\sigma^2)$. Find the MLE $\hat{\Theta}_n$ for the unknown parameter $\Theta = (\mu,\sigma^2)$

sol:

Likelihood function (ignoring constants as they are irrel. for argmax)

$L_n(\Theta) = L_n(\mu,\sigma) = \prod_{i=1}^n \frac{1}{\sigma}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \sigma^{-n}e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2} = \sigma^{-n}e^{-\frac{nS_n^2}{2\sigma^2}}e^{-\frac{n(\overline{X}_n-\mu)^2}{2\sigma^2}}$

where $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n x_i$, $S_n^2 = \frac{1}{n}\sum_{i=1}^n (x_i-\overline{X}_n)^2$

The log-likelihood $\ell_n(\mu,\sigma) = -n\ln(\sigma) - \frac{nS_n^2}{2\sigma^2} - \frac{n(\overline{X}_n-\mu)^2}{2\sigma^2}$

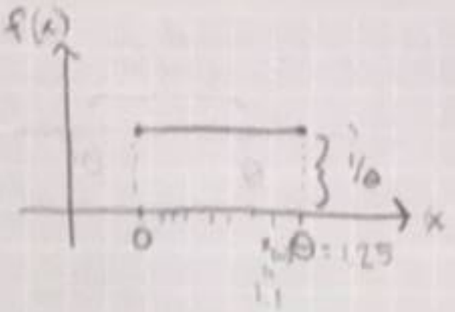
Now, solve the equations $\begin{cases} \frac{\partial}{\partial\mu}\ell_n(\mu,\sigma) = 0 \\ \frac{\partial}{\partial\sigma}\ell_n(\mu,\sigma) = 0 \end{cases}$

exercise: show that $\hat{\Theta}_n = (\hat{\Theta}_{n,1}, \hat{\Theta}_{n,2}) = (\overline{X}_n, S_n)$

($\hat{\Theta}_{n,1} \to \hat{\mu}_n$, $\hat{\Theta}_{n,2} \to \hat{\sigma}_n$)

ex3 Tricky conceptually.

$X_1, \dots, X_n \overset{iid}{\sim} \text{Uniform}(0, \theta)$



Find the MLE $\hat{\theta}_n$ of θ

$$f(x_i;\theta) = \frac{1}{\theta}\mathbb{1}_{[0,\theta]}(x_i) = \begin{cases} 1/\theta, & x_i \in [0,\theta] \\ 0, & x_i \notin [0,\theta] \end{cases}$$

Likelihood: consider θ fixed, suppose $\theta < x_i$ for some $i \in [n] = \{1, \dots, n\}$.

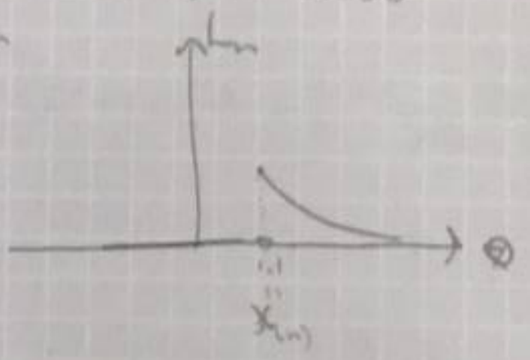
Then $f(x_i;\theta) = 0$ and thus $L_n(\theta) = 0$ if any $x_i > \theta$

so, $L_n(\theta) = 0$ if $\max(\{X_1, \dots, X_n\}) =: X_{(n)} > \theta$.

Now, consider any $\theta \geq X_{(n)} \geq 0$. Then for each x_i we have

$$f(x_i;\theta) = \frac{1}{\theta} \Rightarrow L_n(\theta) = \prod_{i=1}^{n} f(x_i;\theta) = \frac{1}{\theta^n} = \theta^{-n}$$

$$\Rightarrow L_n(\theta) = \begin{cases} \theta^{-n} & \text{if } \theta \geq X_{(n)} \\ 0 & \text{if } \theta < X_{(n)} \end{cases}$$



So, MLE is $X_{(n)}$ (nth order statistic)

Properties of MLEs

1) The MLE is consistent: $\hat{\theta}_n \xrightarrow{p} \theta_0$, where θ_0 denote the true value of param. θ if the model is "good" or "reasonable"

2) The MLE is equi-variant: if $\hat{\theta}_n$ is the MLE of θ, then $g(\hat{\theta}_n)$ is the MLE of $g(\theta)$ ("the function of interest")

3) The MLE is asymptotically normal: $\frac{(\hat{\theta}_n - \theta_0)}{\widehat{se}_n} \rightsquigarrow N(0,1)$

$se_n(\hat{\theta}_n) = \sqrt{V(\hat{\theta}_n)}$ is the standard error

4) The MLE is asymptotically optimal (efficient):

Amongst all well-behaved estimators, the MLE has the smallest variance (at least for large samples)

5) The MLE is approximately the Bayes Estimators

* These things hold when certain regularity conditions are satisfied:
Mainly smoothness conditions on the density ($f(x;\theta)$ ↳ $x = (x_1, \dots, x_n)$)

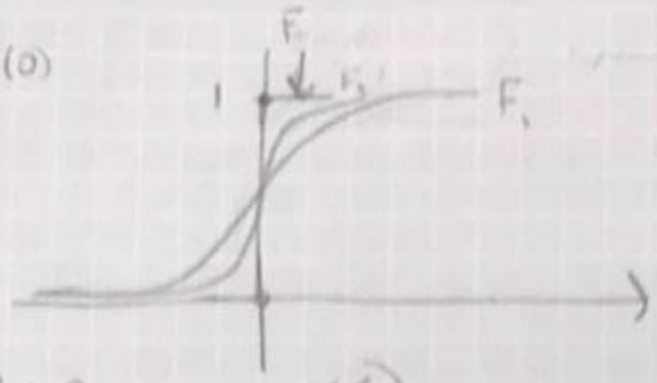
$N(0,\frac{1}{1})$ $N(0,\frac{1}{2})$ $\text{Point Mass}(0)$

$\|$ $\|$ $\|$

F_1 F_2 F

$\wr$ $\wr$ $\wr$

$X_1, \dots, X_n \rightsquigarrow X$



Def Conv. in distribution (P. 152, 154 in (SE)

Let $X_1, \dots, X_n$ have DFs $F_1, \dots, F_n$. Let X be another RV with DF F.

Then we say $X_n \rightsquigarrow X$ if $\forall t \in \mathbb{R}$ at which F is continuous, we have

$(X_n \xrightarrow{D} X)$

$\lim_{n\to\infty} F_n(t) = F(t)$ (Pointwise convergence of DFs)

Recall $\lim_{n\to\infty} P(\{w : X_n(w) \le t\}) = P(\{w : X(t) \le t\})$

Prop. CLT

Let $X_1, \dots, X_n \overset{iid}{\sim} X_1$ and $E(X_1), V(X_1)$ exist.

Then,

- $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \rightsquigarrow X \sim N(E(X_1), \frac{V(X_1)}{n})$
- $\overline{X}_n - E(X_1) \rightsquigarrow X - E(X_1) \sim N(0, \frac{V(X_1)}{n})$
- $\sqrt{n}(\overline{X}_n - E(X_1)) \rightsquigarrow \sqrt{n}(X - E(X_1)) \sim N(0, V(X_1))$
- $Z_n := \frac{\sqrt{n}}{\sqrt{V(X_1)}}(\overline{X}_n - E(X_1)) = \frac{\overline{X}_n - E(\overline{X}_n)}{\sqrt{V(\overline{X}_n)}} \rightsquigarrow Z \sim N(0,1)$
- $\lim_{n\to\infty} P\left(\frac{\overline{X}_n - E(\overline{X}_n)}{\sqrt{V(\overline{X}_n)}} \le z\right) = \lim_{n\to\infty}(Z_n \le z) = \Phi(z) := \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx$

ex Suppose collection $X_1, \ldots, X_n$ model # errors in programs ran on $1, \ldots, n$. Suppose $X_i \sim Po(\lambda = 5)$ and X_i i.i.d, $E(X_i) = 5$

Suppose $n = 125$ and we want to make prob. statement of $\overline{X}_n$

$$P(\overline{X}_n \leq 5.5) = P\left(\frac{\sqrt{n}(\overline{X}_n - E(X_i))}{\sqrt{V(X_i)}} \leq \frac{\sqrt{n}(5.5 - E(X_i))}{\sqrt{V(X_i)}}\right) = P\left(Z \leq \frac{\sqrt{125}(5.5-5)}{\sqrt{5}}\right) =$$

$$= \phi(2.5)$$

Prob. back-fill

Def

Conv. in prob

$$X_n \xrightarrow{P} X \text{ if } \forall \varepsilon > 0,\ P(|X_n - X| > \varepsilon) \xrightarrow{n \to \infty} 0 \quad (X_n \xrightarrow{P} c \text{ if } c \text{ is PointMass}(c))$$

(X_n: E_n, X: F)

Def

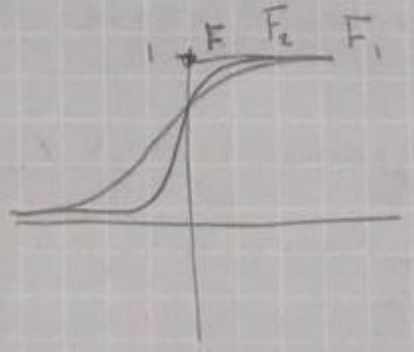
Conv. in dist.

$$X_n \rightsquigarrow X \text{ if } \forall t: F(t) \text{ continuous} : \lim_{n \to \infty} F_n(t) = F(t)$$

Def

conv. in quadratic mean (q.m.) or L_2 conv.:

$$X_n \xrightarrow{qm} X \text{ if } E((X_n - X)^2) \xrightarrow{n \to \infty} 0$$



ex $X_n \sim N(0, 1/n)$, $X \sim PointMass(0)$.

How to formalize intuition "$X_n \to 0$"?

conv. in dist.:

$$\sqrt{n} X_n \sim N(0,1) =: Z$$

for $t < 0$: $F_n(t) = P(X_n < t) = P(\sqrt{n} X_n < \sqrt{n} t) = P(Z < \sqrt{n} t) \xrightarrow{n \to \infty} 0$

for $t > 0$: $F_n(t) = P(X_n < t) = P(\sqrt{n} X_n < \sqrt{n} t) = P(Z < \sqrt{n} t) \xrightarrow{n \to \infty} 1$

Hence, $F_n(t) \to F(t)$ $\forall t \neq 0$ But this is ok since $\rightsquigarrow$ only needs to be for t where $F(t)$ cont.

Conv. in prob.: $\forall c > 0 \cdot P(|X_n - 0| > 0) = P(|X_n|^2 > \varepsilon^2) \leq \frac{E(X_n^2)}{\varepsilon^2} = \frac{\frac{1}{n}}{\varepsilon^2} = \frac{1}{n\varepsilon^2} \xrightarrow{n \to \infty} 0$

Prop. (PNN)

a) $X_n \xrightarrow{q.m.} X \Rightarrow X_n \xrightarrow{p} X$

b) $X_n \xrightarrow{p} X \Rightarrow X_n \rightsquigarrow X$

c) If $X_n \rightsquigarrow X$ and $\exists c \in \mathbb{R}: P(X_n = c) = 1$, then $X_n \xrightarrow{p} X$

So, conv. in q.m. $\Rightarrow$ conv. in prob. $\Rightarrow$ conv. in dist.

$\Leftarrow$ if $X \sim \text{PointMass}(c)$

ex conv. in prob $\not\Rightarrow$ conv. in q.m.

Let $U \sim U(0,1)$, $X_n = \sqrt{n}\,\mathbb{1}_{(0,1/n)}(U)$, $X \sim \text{PointMass}(0) = 0$

Then $P(|X_n| > \varepsilon) = P(\sqrt{n}\,\mathbb{1}_{(0,1/n)}(U) > \varepsilon) = P(0 \le U \le 1/n) = \frac{1}{n} \to 0$ so $X_n \xrightarrow{p} X$

But $E(X_n^2) = \int_0^{1/n} (\sqrt{n})^2 \cdot 1 \, du = n \int_0^{1/n} du = n[u]_0^{1/n} = n\left(\frac{1}{n} - 0\right) = 1$ for all n so $X_n \not\xrightarrow{q.m.} X$

ex conv. in dist. $\not\Rightarrow$ conv. in prob.

Let $X \sim N(0,1)$, $X_n = -X$ for $n = 1, 2, 3, \ldots$

Hence $X_n \sim N(0,1)$ (goes "backwards" $\infty \to -\infty$ but since $N(0,1)$ symmetric, still the same)

X_n has same DF as X for all n, so trivially $\lim_{n\to\infty} F_n(x) = F(x)\ \forall x \Rightarrow X_n \rightsquigarrow X$

But $P(|X_n - X| > \varepsilon) = P(|2X| > \varepsilon) = P(|X| > \varepsilon/2) \neq 0$

ex we may conjecture that if $X_n \xrightarrow{p} c$, then $E(X_n) \to c$. Not true in general.

(unless X_n is uniformly integrable)

Let X_n be s.t. $P(X_n = n^2) = \frac{1}{n}$ and $P(X_n = 0) = 1 - \frac{1}{n}$

Now, $P(|X_n - 0| < \varepsilon) = P(X_n = 0) = 1 - \frac{1}{n} \xrightarrow[n\to\infty]{} 1 \Rightarrow X_n \xrightarrow{p} 0$

but $E(X_n) = n^2 \cdot \frac{1}{n} + 0 \cdot \left(1 - \frac{1}{n}\right) = n \Rightarrow E(X_n) \to \infty$

Back to CLT

$$Z_n := \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \rightsquigarrow N(0,1) \quad \text{for } X_1, \dots, X_n \overset{iid}{\sim} X_1,$$
$$\mu := E(X_1) < \infty,\ V(X_1) := \sigma^2 < \infty$$

We typically don't know σ.
If we replace σ by its estimate $S_n := \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}$ it still works

The Berry-Esseen inequality (tells us about accuracy of Normal approx.)
suppose further that $E(|X_1|^3) < \infty$.
Then $\sup_z |P(Z_n < z) - \Phi(z)| \leq \frac{33}{4} \cdot \frac{E(|X_1 - \mu|^3)}{\sqrt{n}\,\sigma^3}$

Back to properties of MLE

Prop (consistency)
Let θ_0 denote the true (and possibly unknown) value of θ.

$$M_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} \ln\left(\frac{f(x_i;\theta)}{f(x_i;\theta_0)}\right), \quad M(\theta) = E_{\theta_0}\left(\ln\left(\frac{f(x_i;\theta)}{f(x_i;\theta_0)}\right)\right)$$

(maximizing $\Leftrightarrow$ maximizing $\lambda_n(\theta)$)

Suppose that $\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \xrightarrow{p} 0$

and that $\forall \varepsilon > 0 : \sup_{\theta : |\theta - \theta_0| \geq \varepsilon} M(\theta) < M(\theta_0)$
Let $\hat{\theta}_n$ denote the MLE. Then $\hat{\theta}_n \xrightarrow{p} \theta_0$

prop (equivariance)
Let $\tau = g(\theta)$ be a function of θ and $\hat{\theta}_n$ be MLE of θ
Then $\hat{\tau}_n = g(\hat{\theta}_n)$ is the MLE of τ

proof
Let $h = g^{-1}$ denote the inverse of g. Then $\hat{\theta}_n = h(\hat{\tau}_n)$.
For any τ, $L_n(\tau) = \prod_{i=1}^{n} f(x_i; h(\tau)) = \prod_{i=1}^{n} f(x_i;\theta) = L_n(\theta)$ where $\theta = h(\tau)$.
Hence $\forall \tau$, $L_n(\tau) = L_n(\theta) \leq L_n(\hat{\theta}_n) = L_n(\hat{\tau}_n)$

ex

Let $X_1, \dots, X_n \overset{iid}{\sim} N(\theta, 1)$. MLE of θ is $\hat{\theta}_n = \bar{X}_n$.

Let $\tau = e^\theta$. Then MLE for τ is $\hat{\tau}_n = e^{\hat{\theta}_n} = e^{\bar{X}_n}$

Def

The Score function is $s(X;\theta) = \frac{\partial}{\partial \theta} \ln(f(x;\theta))$

Def

Fisher information is $I_n(\theta) = V_\theta\left(\sum_{i=1}^n s(X_i;\theta)\right) = \sum_{i=1}^n V_\theta(s(X_i;\theta))$.

when $n=1$, $I(\theta) := I_1(\theta)$

Identifiability

$\theta \neq \bar{\theta}$

$\Rightarrow L(X;\theta) \neq L(X;\bar{\theta})$

Prop

$I_n(\theta) = nI(\theta)$. Also, $I(\theta) = -E_\theta\left(\frac{\partial^2}{\partial \theta^2} \ln(f(x;\theta))\right) =$

$= -\int_{\mathbb{R}} \left(\frac{\partial^2}{\partial \theta^2} \ln(f(x;\theta))\right) f(x;\theta)\, dx$

Prop (Asymptotic Normality)

Let $se = \sqrt{V(\hat{\theta}_n)}$. Under appropriate regularity conditions the following hold:

1. Let $se \approx \sqrt{\frac{1}{I_n(\theta)}}$, then $\frac{(\hat{\theta}_n - \theta)}{se} \rightsquigarrow N(0,1)$

2. Let $\hat{se}_n = \sqrt{\frac{1}{I_n(\hat{\theta}_n)}}$, then $\frac{(\hat{\theta}_n - \theta)}{\hat{se}_n} \rightsquigarrow N(0,1)$

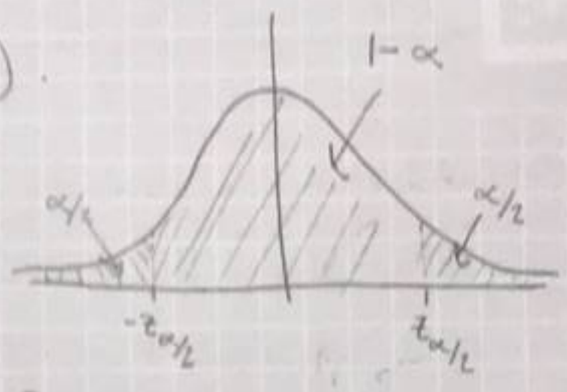
So, $\hat{\theta}_n \rightsquigarrow N(\theta, \hat{se}_n^2)$

We can thus construct asymptotic (normal-based) confidence intervals for θ.

Prop

Let $C_n = [\underline{C}_n, \overline{C}_n] = (\hat{\theta}_n - z_{\alpha/2}\widehat{se}_n, \hat{\theta}_n + z_{\alpha/2}\widehat{se}_n)$.

Then, $P_\theta(\theta \in C_n) \to 1-\alpha$ as $n \to \infty$



Proof

Let Z be standard normal RV. Then,

$$P_\theta(\theta \in C_n) = P_\theta(\hat{\theta}_n - z_{\alpha/2}\widehat{se}_n \le \theta \le \hat{\theta}_n + z_{\alpha/2}\widehat{se}_n) = P_\theta\left(-z_{\alpha/2} \le \frac{\hat{\theta}_n - \theta}{\widehat{se}_n} \le z_{\alpha/2}\right) \to P(-z_{\alpha/2} \le Z \le z_{\alpha/2}) = 1-\alpha$$

For $\alpha = 0.05$, $z_{\alpha/2} = 1.96 \approx 2$, so: $\hat{\theta}_n \pm 2\widehat{se}_n$ is an approx. 95% confidence interval

ex/ When you read an opinion poll in a newspaper, you see statements like "the poll is accurate to within one point 95% of the time". By this, they are giving 95% confidence interval of the form $\hat{\theta}_n \pm 2\widehat{se}_n$

ex/ Let $X_1, \dots, X_n \overset{iid}{\sim} Be(\theta)$. The MLE is $\hat{\theta}_n = \overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$.

$f(x;\theta) = \theta^x(1-\theta)^{1-x}$, $\ell n(f(x;\theta)) = x\ell n\theta + (1-x)\ell n(1-\theta)$

$s(X;\theta) = \frac{x}{\theta} - \frac{1-x}{1-\theta}$ and $s'(x;\theta) = \frac{x}{\theta^2} + \frac{1-x}{(1-\theta)^2} =$ ___

Thus $I(\theta) = E_\theta(-s'(x;\theta)) = \frac{\theta}{\theta^2} + \frac{(1-\theta)}{(1-\theta)^2} = \frac{1}{\theta(1-\theta)} = \frac{1}{V(X)}$

Hence, $\widehat{se}_n = \frac{1}{\sqrt{I_n(\hat{\theta}_n)}} = \frac{1}{\sqrt{nI(\hat{\theta}_n)}} = \left(\frac{\hat{\theta}(1-\hat{\theta})}{n}\right)^{1/2}$

An approx. 95% conf. interval is: $\hat{\theta}_n \pm 2\left(\frac{\hat{\theta}_n(1-\hat{\theta}_n)}{n}\right)^{1/2}$

ex/ Let $X_1, \dots, X_n \overset{iid}{\sim} N(\theta, \sigma^2)$ where σ^2 is known.

$s(X;\theta) = (X-\theta)/\sigma^2$ and $s'(x;\theta) = \frac{-1}{\sigma^2}$, so $I_1(\theta) = \frac{1}{\sigma^2}$.

The MLE of θ is $\hat{\theta}_n = \overline{X}_n$. $\overline{X}_n \sim N(\theta, \sigma^2/n)$.

In this case the Normal approx. is actually exact.

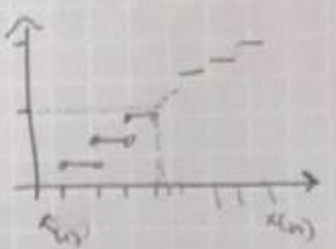
Exercise

Find 95% conf. int. for (and MLE) $X_1, \dots, X_n \overset{iid}{\sim} P_o(\lambda)$

Optimality of MLE:

Suppose $X_1, \dots, X_n \overset{iid}{\sim} N(\theta, \sigma^2)$. The MLE of $\hat{\theta}_n$ is $\overline{X}_n$.

Another reasonable estimator is sample median $\tilde{\theta}_n$.



We know that MLE satisfies $\sqrt{n}(\hat{\theta}_n - \theta) \rightsquigarrow N(0, \sigma^2)$

One can show that $\tilde{\theta}_n$ satisfies $\sqrt{n}(\tilde{\theta}_n - \theta) \rightsquigarrow N(0, \sigma^2 \frac{\pi}{2})$

So, the median converges to the right value θ but has a larger variance than MLE.

More generally, consider two estimators T_n and U_n.

Suppose $\sqrt{n}(T_n - \theta) \rightsquigarrow N(0, \sigma_t^2)$ and $\sqrt{n}(U_n - \theta) \rightsquigarrow N(0, \sigma_u^2)$

Def

The asymptotic relative efficiency (ARE) is given by

$$ARE(T, U) = \frac{\sigma_t^2}{\sigma_u^2}$$

So, $ARE(\tilde{\theta}_n, \hat{\theta}_n) = \frac{2}{\pi} = 0.63$, in words: if you use the median-based estimate $\tilde{\theta}_n$, then you are effectively using only a fraction of the samples ($\approx 0.63n$)

Prop.

If $\hat{\theta}_n$ is the MLE and $\tilde{\theta}_n$ is any other estimator, then

$ARE(\tilde{\theta}_n, \hat{\theta}_n) \leq 1$, so MLE is asymptotically optimal (almost, +details)

All of this MLE stuff is only as good as the $\mathcal{Q}$-parametric family of models!

Delta Method

Let $\tau = g(\theta)$, g smooth $g \in C^1$.
Due to equivariance of MLE we see that $\hat{\tau} = g(\hat{\theta}_n)$ but what about the distribution of $\hat{\tau}_n$?

Prop (The Delta Method)
If $\tau = g(\theta)$ where g is differentiable and $g'(\theta) \neq 0$, then

$$\frac{(\hat{\tau}_n - \tau)}{\widehat{se}(\hat{\tau})} \rightsquigarrow N(0,1) \quad \text{where } \hat{\tau}_n = g(\hat{\theta}_n) \text{ and } \widehat{se}(\hat{\tau}_n) = |g'(\theta_n)| \widehat{se}(\hat{\theta}_n)$$

Hence, if $C_n = \left(\hat{\tau}_n - z_{\alpha/2} \cdot \widehat{se}(\hat{\tau}_n), \hat{\tau} - z_{\alpha/2} \cdot \widehat{se}(\hat{\tau}_n)\right)$, then

$$P(\tau \in C_n) \xrightarrow[n \to \infty]{} 1 - \alpha$$

ex/ Let $X_1, \dots, X_n \overset{iid}{\sim} Bd(\theta)$ and let $\Psi = g(\theta) = \ln\left(\frac{\theta}{1-\theta}\right)$

The Fisher information function is $I(\theta) = \frac{1}{\theta(1-\theta)}$ so the estimated se of the MLE $\hat{\theta}_n$ is $\widehat{se} = \sqrt{\frac{\hat{\theta}_n(1-\hat{\theta}_n)}{n}}$. The MLE of Ψ is $\hat{\Psi} = \ln\left(\frac{\hat{\theta}_n}{1-\hat{\theta}_n}\right)$.

since $g'(\theta) = \frac{1}{\theta(1-\theta)}$, according to the delta method,

$\widehat{se}(\hat{\Psi}_n) = |g'(\hat{\theta}_n)| \widehat{se}(\hat{\theta}_n) = \frac{1}{\sqrt{n\hat{\theta}_n(1-\hat{\theta}_n)}}$ and an approx. 95% ci. is $\hat{\Psi}_n \pm \frac{2}{\sqrt{n\hat{\theta}_n(1-\hat{\theta}_n)}}$

Exercise

$X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. Suppose μ is known but not $\sigma > 0$.
We want to estimate $\Psi = \ln(\sigma)$.

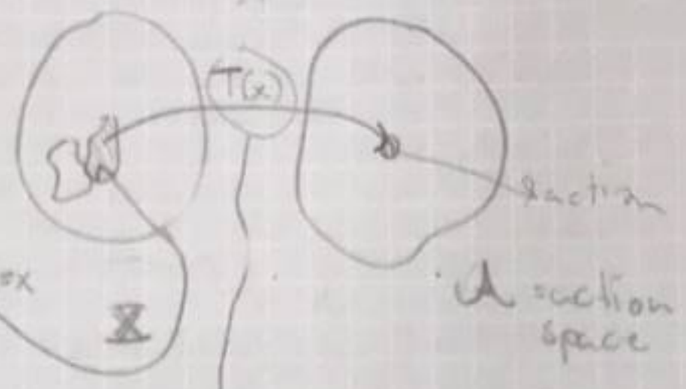
hints
- Find MLE (log-likelihood etc.)
- $\hat{\Psi} = \ln(\hat{\sigma}_n)$
- find $\widehat{se}(\sigma)$
- find $\widehat{se}(\hat{\Psi})$

cons
$\hat{\sigma}_n = \sqrt{\frac{\sum (x_i - \mu)^2}{n}}$

$\widehat{se}(\hat{\Psi}_n) = \frac{1}{\hat{\sigma}_n} \cdot \frac{\hat{\sigma}_n}{\sqrt{2n}} = \frac{1}{\sqrt{2n}}$

95% ci: $\hat{\Psi}_n \pm \frac{2}{\sqrt{2n}}$

Generally in decision theory:



Decision problems

estimation problem

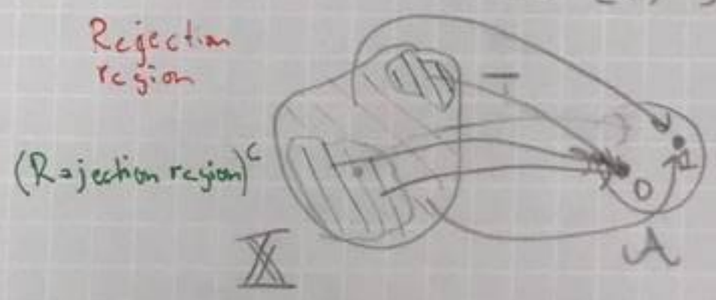
$A = \Theta$, T is $\hat{\theta}_n$ (estimator)

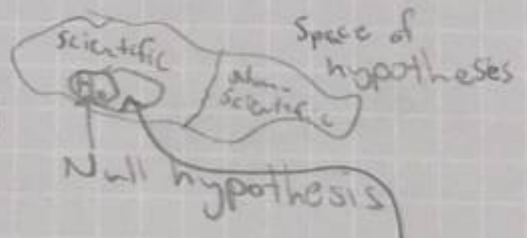
$|\Theta| < \infty$

	Parametric	non-parametric
	$X_1, \dots, X_n \sim L_\theta$	$X_1, \dots, X_n \sim F \in \{\text{All DF's}\}$
Point estimators:	• MoME • MLE $\hat{\theta}_n$	• $\hat{F}_n \to F_n$ (Glivenko-Cantelli Lemma)
Set estimators:	$[\underline{C}_n, \overline{C}_n]$, C_n (conf. intervals)	• $[\hat{\underline{F}}_n, \hat{\overline{F}}_n]$ (interval in function space) (Dvoretzky-Kiefer-Wolfowitz ineq.)

Another Decision Problem: (hypothesis testing)

$A = \{0, 1\}$, T is Test statistic





H_A - alternative hypothesis (or H_1)

Karl Popper

"A scientific hyp. is one that is falsifiable"

Hyp. testing Problem

Should H_0 be rejected? (don't know)

Data $(X_1, \dots, X_n) \sim \text{Law}(H_0)$ or $\text{Law}(H_A)$

So, the summary of outcomes of hyp. testing:

	(fail to reject) Retain Null	Reject Null (false pos.)
H_0 true	✓	Type I error $= \alpha$
H_A true	Type II error $= \beta$ (false neg.)	✓ (Power $= 1-\beta$)

We want to minimize α (size) at some "significance level", typically $\alpha = 0.05$

we also want to maximize Power for given size α test

exe The Null hypothesis: "The disease rate is the same in two groups".

Alt. hypothesis: "The disease rate is not the same".

Formally

We partition index set Θ into two disjoint sets Θ_0, Θ_1

and we want to test

$H_0: \theta \in \Theta_0$ versus $H_1: \theta \in \Theta_1$

↑ Null hyp. ↑ Alt. hyp.

e.g. $\{\bar{\theta}\}$

Def Rejection region

$R \subseteq \mathbb{X}$ s.t. if data in R, we reject H_0

optimality of MLE ; Delta Method ; Hypothesis Testing ; Computer Exs ?

Falsifiability

Karl Popper

Science? Vs. Nonsuence?

Demarcation Problem

Decision Problems

Estimation Problem

$\mathbb{A} = \Theta$, $T = \hat{\Theta}_n$ is called an estimator

parametric $|\Theta| < \infty$ / non-parametric

$\mathbb{A}$:= Action Space

$T(x)$

X data = x

$\mathbb{A}$ = Action Space

Another Decision Problem is Hypothesis Testing Problem

$\mathbb{A} = \{0, 1\}$, T is called a Test Statistic

Possable Hypotheses

Alternative Hypothesis H_a

Null Hypothesis H_0

non-scientific

A scientific hypothesis is one that is falsifiable

Rejection Region

"Acceptance Region"

1

0

$X_1, \ldots, X_n \sim F \in \{\text{all DFs}\}$

$X_1, \ldots, X_n \sim f_\theta$

Point Estimator

- MOME
- MLE $\hat{\theta}_n$

" $\hat{F}_n$ "

$\hat{F}_n \to F_n$

Glivenko-Cantelli Lemma

Hyp Testing Problem is

Should H_0 be rejected?

Set Estimators

$[\underline{C}_n, \bar{C}_n]$, C_n

$[\hat{\underline{F}}_n, \hat{\bar{F}}]$

Dvoretzky–Kiefer–Wolfowitz ≠

ex. The Null hypothesis: "The disease rate is the same in two groups".
Alt. hypothesis: "The disease rate is not the same".

Formally

We partition index set Θ into two disjoint sets Θ_0, Θ_1
and we want to test

$H_0: \theta \in \Theta_0$ versus $H_1: \theta \in \Theta_1$

(H_0: Null hyp. Θ_0: e.g. $\{\bar{\theta}\}$ H_1: Alt. hyp)

Let X be data,
$X: \Omega \to \mathbb{X}$

Def Rejection region

$R \subseteq \mathbb{X}$ s.t. if data in R, we reject H_0.
Then R is the rejection region

Outcomes of test:	Fail to reject H_0	Reject H_0
H_0 true	✓	Type I error
H_0 false $\Leftrightarrow$ H_1 true	Type II error	✓

So, if $x \in R$, we reject H_0, otherwise we fail to reject H_0

Typically, the rejection region is of the form

$$R = \{x : T(x) > c\}$$

where T is a test statistic and c is a critical value.
The problem of hyp. testing is to find appropriate T and c.

Def

The <u>power function</u> of a test with rejection region R is

$\beta(\theta) := P_\theta(X \in R)$.

The <u>size</u> of a test is $\alpha := \sup_{\theta \in \Theta_0} \beta(\theta)$

A test is said to have <u>level</u> α if its size is $\leq \alpha$

Types of hypothesis:

- Simple hypothesis: $H_0: \theta = \theta_0$, $H_1: \theta \neq \theta_0$
- Composite hypothesis: $H_0: \theta < \theta_0$, $H_1: \theta > \theta_0$ or $H_0: \theta > \theta_0$, $H_1: \theta < \theta_0$

Types of tests:

- Two-sided test: $H_0: \theta = \theta_0$, $H_1: \theta \neq \theta_0$
- One-sided test: $H_0: \theta < \theta_0$, $H_1: \theta > \theta_0$ or $H_0: \theta > \theta_0$, $H_1: \theta < \theta_0$

Typically, tests are two-sided under a simple hypothesis.

ex Let $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ where σ is known.

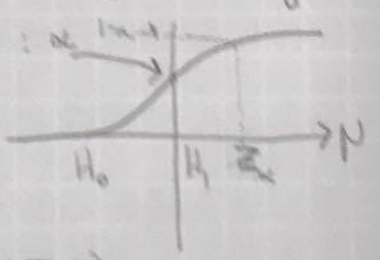
We want to test $H_0: \mu \leq 0$ vs $H_1: \mu > 0$. $(\Theta_0 = (-\infty, 0]$, $\Theta_1 = (0, \infty))$

consider the test: reject H_0 if $T > c$ where $T = \bar{X}_n$,

so the rejection region is $R = \{(x_1, \dots, x_n) : T(x_1, \dots, x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i > c\}$

Let $Z \sim N(0,1)$. Then the power function is $\beta(\mu) = P_\mu(\bar{X}_n > c) = P_\mu\left(\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} > \frac{\sqrt{n}(c-\mu)}{\sigma}\right) =$

$= P\left(Z > \frac{\sqrt{n}(c-\mu)}{\sigma}\right) = 1 - \phi\left(\frac{\sqrt{n}(c-\mu)}{\sigma}\right)$. Thus $\beta(\mu)$ is increasing in μ :



Hence, size is $\sup_{\substack{\mu \in \Theta \\ (\mu \leq 0)}} \beta(\mu) = \beta(0) = 1 - \phi\left(\frac{\sqrt{n}c}{\sigma}\right)$

For a size α test, we set size $= \alpha$ and solve for μ : $\alpha = 1 - \phi\left(\frac{\sqrt{n}c}{\sigma}\right) \Rightarrow \dots \Rightarrow c = \phi^{-1}(1-\alpha)\frac{\sigma}{\sqrt{n}}$

So, we reject the null hypothesis if $\bar{X}_n > \phi^{-1}(1-\alpha)\frac{\sigma}{\sqrt{n}}$ for a size α test

or equivalently, when $\frac{\sqrt{n}(\bar{X}_n - 0)}{\sigma} > \phi^{-1}(1-\alpha) = z_\alpha$

It is desirable to find the test with the highest power under H_1 ($\theta \in H_1$) among all size α tests. Such a test (if it exists) is called the most powerful size α test. (Generally very hard to find)

A common test is the so-called <u>Wald Test</u>:

Let $\theta \in \mathbb{R}$, $\hat{\theta}_n$ be an estimator of θ with realization (estimate) $\hat{\theta}_n$, and let $\hat{se}_n$ be the estimated standard error of $\hat{\theta}_n$.

<u>Def</u>

Consider testing $H_0: \theta = \theta_0$ vs $H_1: \theta \neq \theta_0$.

Assume $\hat{\theta}_n$ is asymptotically normal, $\frac{\hat{\theta}_n - \theta_0}{\hat{se}_n} \rightsquigarrow N(0,1)$.

Then the <u>size α Wald test</u> is:

reject H_0 when $|W| > z_{\alpha/2}$, $W = \frac{\hat{\theta}_n - \theta_0}{\hat{se}_n}$

<u>prop.</u>

Asymptotically, the Wald test has size α, i.e.

$P_{\theta_0}(|W| > z_{\alpha/2}) \to \alpha$ as $n \to \infty$

<u>Proof</u>

Under H_0 (if H_0 is true): $\theta = \theta_0$; $\frac{(\hat{\theta}_n - \theta_0)}{\hat{se}_n} \rightsquigarrow N(0,1)$ (by assumption), hence

the probability of rejecting H_0 when H_0 is true is $P_{\theta_0}(|W| > z_{\alpha/2}) = P_{\theta_0}\left(\frac{|\hat{\theta}_n - \theta_0|}{\hat{se}_n} > z_{\alpha/2}\right) \to$

$\underset{n \to \infty}{\to} P(|Z| > z_{\alpha/2}) = \alpha$

<u>Remark</u>

Alternatively, $W = \frac{(\hat{\theta}_n - \theta_0)}{se_0}$ where se_0 is the standard error computed with $\theta = \theta_0$ directly. So, one can use $\hat{se}_n$ or se_0 for a valid Wald test.

Let's consider the power of the Wald test when H_0 is false.

<u>Prop.</u>

Suppose the true value of θ is $\theta^* \neq \theta_0$. The power $\beta(\theta^*)$, i.e. prob. of rejecting H_0, is given approximately by $\beta(\theta^*) = 1 - \Phi\left(\frac{\theta_0 - \theta^*}{\hat{se}_n} + z_{\alpha/2}\right) + \Phi\left(\frac{\theta_0 - \theta^*}{\hat{se}_n} - z_{\alpha/2}\right)$

<u>Note</u>

$\hat{se}_n \to 0$ as $n \to \infty$ so power is large if i) $|\theta_0 - \theta^*|$ is large
ii) n is large

Prop (Wald test & $1-\alpha$ asymp. conf. interval: $\hat{\theta}_n \pm z_{\alpha/2}\hat{se}_n$)
The size α Wald test rejects $H_0: \theta = \theta_0$ vs. $H_1: \theta \neq \theta_0$
iff $\theta_0 \notin C_n = [\underline{C_n}, \overline{C_n}] = [\hat{\theta}_n - \hat{se}_n z_{\alpha/2}, \hat{\theta}_n + \hat{se}_n z_{\alpha/2}]$.

Thus, testing the hypothesis is equivalent to checking if the null value θ_0 is in the conf. interval.

ex Compare two predictive algorithms (A/B testing).
Suppose the first algorithm gives X many incorrect predictions out of m tricks.
And the second — " — Y — " — " — n tricks.
Assuming X, Y independent binomially distributed R.V.s, test the null hypothesis that their probabilities for incorrect predictions are the same.

sol
Let $X \sim Bin(m, \theta)$, $Y \sim Bin(n, \theta)$. We want to test $H_0: \theta = \theta$ or $\delta = 0$ where $\delta := \theta - \theta$. So, $H_0: \delta = 0$, $H_1: \delta \neq 0$ ← can do Wald test.
The MLE of δ is $\hat{\delta} = \hat{\delta}_{m,n} = \hat{\theta}_m - \hat{\theta}_n$ by equi-variance prop of MLE.
The estimated std. error of $\hat{\delta} = \hat{\theta}_m - \hat{\theta}_n$ is

$$\hat{se}_n = \sqrt{\frac{\hat{\theta}_m(1-\hat{\theta}_m)}{m} + \frac{\hat{\theta}_n(1-\hat{\theta}_n)}{n}}$$

$X \sim Bin(m, \theta) \Rightarrow V(X) = m\theta(1-\theta)$
$\hat{\theta}_m = \bar{X}_m \Rightarrow V(\hat{\theta}_m) = \frac{1}{m^2} V(X) = \frac{\theta(1-\theta)}{m}$

The size α Wald test is to reject H_0 when
$|W| > z_{\alpha/2}$, where $W = \frac{\hat{\delta} - 0}{\hat{se}_n} = \frac{\hat{\theta}_m - \hat{\theta}_n}{\hat{se}_n}$

Suppose you observe 18 bad prediction out of 117 tricks for alg. 1
and 21 — " — " — 111 — " — 2.
Does the size $\alpha = 0.05$ Wald test reject the null hypothesis that $\delta = 0$

exercise
Suppose $X_1, \ldots, X_n \overset{iid}{\sim} exp(\lambda)$ model waiting times at bus stop and you have values 7.6, 9.2, 11.8, 6.3, 13.2, 10.6, 6.3, 7.8, 8.9, 10.2
First, obtain MLE $\hat{\lambda}_n$ of λ, $\hat{se}_n$ of $\hat{\lambda}_n$, $(1-\alpha)$ c.i. for λ.
Then give a size $\alpha = 0.05$ Wald Test to reject $H_0: \lambda = 1/10$

p-values

Instead of just reporting "reject H_0" or "fail to reject H_0" at a given size α, we could make a more informed report of the test.
If a test rejects at level α, it will also reject at $\alpha' \geq \alpha$.
But there will be a smallest α at which the test rejects H_0.

Def Suppose for every $\alpha \in (0,1)$ we have a test of size α with rejection region R_α. Then,

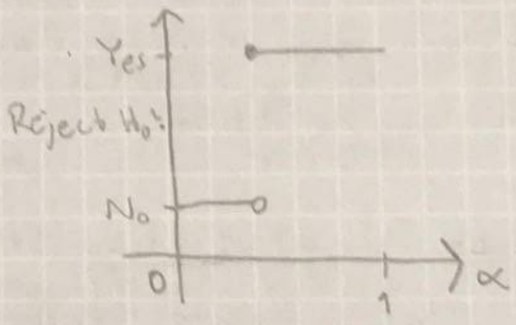
$$\text{p-value} := \inf_{\alpha \in (0,1)} \{\alpha : T(X_1, \dots X_n) \in R_\alpha\}.$$

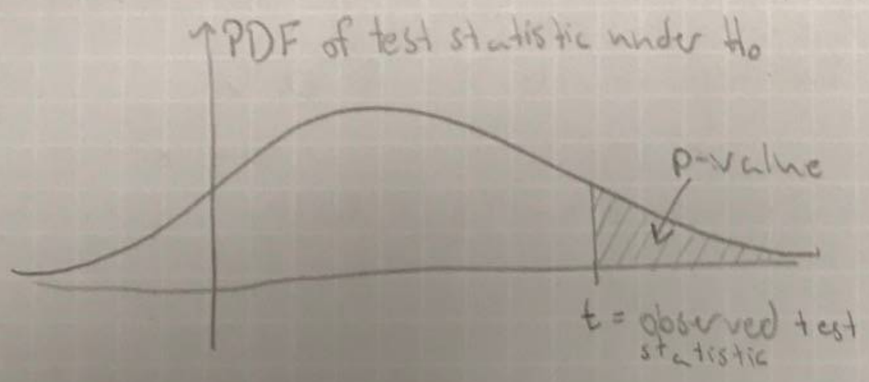
i.e. p-value is the smallest α at which we reject H_0.

Remarks

p-value is a measure of evidence against H_0; the smaller the p-value, the stronger the evidence against H_0.

p-value	evidence against H_0
<0.01	very strong
$0.01-0.05$	strong
$0.05-0.10$	weak
>0.1	little or no evidence





To find p-value for a particular test statistic T, we find α s.t. the observed statistic t is just at the boundary of rejection region R
$\Rightarrow$ p-value is tail area $P(T>t)$
(if T is of the form $T>c$)

Pearson's χ^2 test for multinomial data

Recall, if $X=(X_1,\dots,X_k)$ has Multinomial(n,θ) distribution.

$\mathcal{X}=\{1,2,\dots,k\}$

$k=2 \Rightarrow$ binomial

$$P((X_1,X_2)=(x_1,x_2);n,\theta_1,\theta_2)=\binom{n}{x_1}\theta_1^{x_1}\theta_2^{x_2}=\frac{n!}{\underbrace{(n-x_1)!}_{x_2!}\cdot x_1!}\theta_1^{x_1}\theta_2^{x_2}$$

Multinomial:

$2\le k<\infty$

$$P((X_1,\dots,X_k)=(x_1,\dots,x_k);n,\theta_1,\theta_2,\dots,\theta_k)=\binom{n}{x_1\ x_2\ \dots\ x_k}\theta_1^{x_1}\theta_2^{x_2}\cdots\theta_k^{x_k}=\frac{n!}{x_1!x_2!\cdots x_k!}\theta_1^{x_1}\theta_2^{x_2}\cdots\theta_k^{x_k}$$

Let $\theta\in\Delta^{k-1}:=\{(\theta_1,\dots,\theta_k)\in\mathbb{R}^k:\sum_{i=1}^k\theta_i=1,\ \theta_i\ge 0\ \forall i\}$

We can check that the MLE of $\theta=(\theta_1,\dots,\theta_k)$, $\hat{\theta}=(\hat{\theta}_1,\dots,\hat{\theta}_k)=\left(\frac{x_1}{n},\frac{x_2}{n},\dots,\frac{x_k}{n}\right)$

Let $\theta_0=(\theta_{01},\dots,\theta_{0k})$ be some fixed vector and suppose we want to test

$H_0:\theta=\theta_0$ vs. $H_1:\theta\neq\theta_0$.

Def Pearson χ^2-statistic is

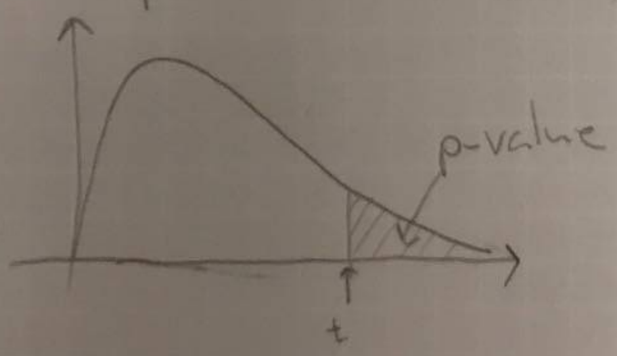
$$T=\sum_{j=1}^k\frac{(x_j-n\theta_{0j})^2}{n\theta_{0j}}=\sum_{j=1}^k\frac{(x_j-E_{\theta_0}(x_j))^2}{E_{\theta_0}(x_j)}$$

where $E_{\theta_0}(x_j)$ is the expected value of x_j under H_0.

Prop.

Under H_0, $T\rightsquigarrow\chi^2_{k-1}$. Hence the test rejects H_0 if $T>\chi^2_{k-1,\alpha}$ has asymptotic level α.

The p-value is $P(\chi^2_{k-1}>t)$, where t is observed test statistic.



ex (Mendel's peas)
Mendel bred peas with round yellow seeds and green wrinkled seeds.
There are 4 types of progeny: rY, rg, wY, wg. The number of each type is modeled as a multinomial RV with probs $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)$

Mendel's theory of inheritance predicts that $\theta = \theta_0 = (\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16})$.
In $n = 556$ trials, Mendel observed $X = (315, 101, 108, 32)$.
We will test $H_0: \theta = \theta_0$ vs $H_1: \theta \neq \theta_0$.

Since $n\theta_1 = 312.75$, $n\theta_2 = n\theta_3 = 104.25$, $n\theta_4 = 34.75$,
the χ^2 test statistic is $\frac{(315-312.75)^2}{312.75} + \frac{(101-104.25)^2}{104.25} + \frac{(108-104.25)^2}{104.25} + \frac{(32-34.75)^2}{34.75} =$
0.47.

The $\alpha = 0.05$ value for a χ^2_3 is 7.815. Since $0.47 < 7.815$ we do not reject the null. The p-value is $P(\chi^2 > 0.47) = 0.93$ ← table lookup
which is not evidence against H_0.

The χ^2-distribution (Prelude to Pearson's χ^2-test)

Def Let $Z_1, \dots, Z_k \overset{iid}{\sim} N(0,1)$ and $Y = \sum_{i=1}^{k} Z_i^2$. Then Y has χ^2 distr. with k degrees of freedom (d.f.) and we write $Y \sim \chi^2_k$.

The PDF of Y is $f(y; k) = \frac{y^{\frac{k}{2}-1} e^{-y/2}}{2^{k/2}\,\Gamma(k/2)}$ for $y > 0$.

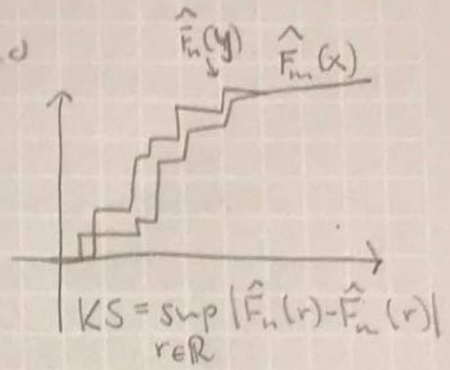
Facts
$E(Y) = k$, $V(Y) = 2k$

Def Upper α quantile is $\chi^2_{k,\alpha} := F^{-1}(1-\alpha; k)$ where F is CDF of Y.
i.e. $P(Y > \chi^2_{k,\alpha}) = \alpha$

Permutation test

This is a non-parametric test (in A/B testing context).

Say we want to test if two sets of samples come from the same distribution or not. $X_1, \dots, X_m \overset{iid}{\sim} F_X \in \{\text{all DFs}\}$ and $Y_1, \dots, Y_n \overset{iid}{\sim} F_Y \in \{\text{all DFs}\}$



$H_0: F_X = F_Y$ vs $H_1: F_X \neq F_Y$

Let $T(X_1, \dots, X_m, Y_1, \dots, Y_n)$ be a given test statistic

examples of T: $T = |\bar{X}_m - \bar{Y}_n|$ or $T = KS$

Let $N = m+n$ and consider all $N!$ permutations of the data $X_1, \dots, X_m, Y_1, \dots, Y_n$.

Under H_0 (distributions equal), all permutations should have same prob!

For each permutation, compute T and denote them $T_1, \dots, T_{N!}$

Under H_0, all the T_i's have equal prob. So the distribution P_0, under H_0, puts mass $\frac{1}{N!}$ on each T_i where $1 \leq i \leq N!$

This P_0 is called the permutation distribution of T.

Let t_{obs} be the observed statistic. Assuming T is such that we reject for large values of T, the p-value is:

$$\text{p-value} = P_0(T > t_{obs}) = \frac{1}{N!} \sum_{j=1}^{N!} \mathbb{1}(T_j > t_{obs})$$

ex (Toy example)

Suppose data are $(X_1, X_2, Y) = (1, 9, 3)$

Let $T(X_1, X_2, Y) = |\bar{X}_2 - \bar{Y}| = 2$

permutation	T	prob under H_0
(1,9,3)	$2 = t_{obs}$	$1/6 = 1/3!$
(9,1,3)	2	1/6
(1,3,9)	7	1/6
(3,1,9)	7	1/6
(3,9,1)	5	1/6
(9,3,1)	5	1/6

p-value $= P_0(T > 2) =$

$= \frac{4}{6} = \frac{2}{3} > 0.1$ ⇒

It's not practical to do N! permutations. However, we can approximate the p-value by sampling uniformly from the set of all permutations of $\{1, 2, \dots, N\}$ due LLN.

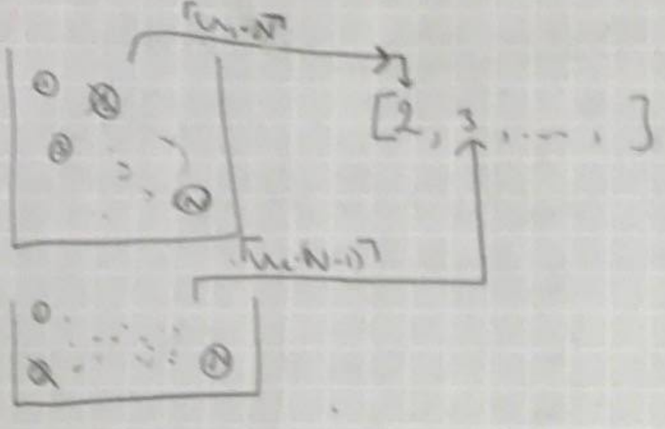
Alg. for Permutation test

1. Compute observed value of test statistic.
$$t_{obs} = T(X_1, \dots, X_m, Y_1, \dots, Y_n)$$
2. Randomly permute the data. Compute statistic again using permuted data.
3. Repeat 2. B times and let $T_1, \dots, T_B$ denote the results
4. The approx. p-value is $\frac{1}{B} \sum_{j=1}^{B} \mathbb{1}(T_j > t_{obs})$

Uniform sampling:

$U_1, \dots, U_N \overset{iid}{\sim} \text{Uniform}(0,1)$

$\text{index}(X) = [1, 2, \dots, N]$



Likelihood Ratio Test Statistic (LRTS)

A more general test when $|\Theta_0| = r \ll n$ (much smaller than)

($|\Theta_0|$ = dim Θ_0)

Def

Consider testing the following:

$H_0: \theta \in \Theta_0$ vs. $H_1: \theta \notin \Theta_0$ (or $\theta \in \Theta \setminus \Theta_0 =: \Theta_1$)

The LRTS is

$$\Lambda_n := 2 \ln \left(\frac{\sup_{\theta \in \Theta} L_n(\theta)}{\sup_{\theta \in \Theta_0} L_n(\theta)} \right) = 2 \ln \left(\frac{L_n(\hat{\theta}_n)}{L_n(\hat{\theta}_{0,n})} \right)$$

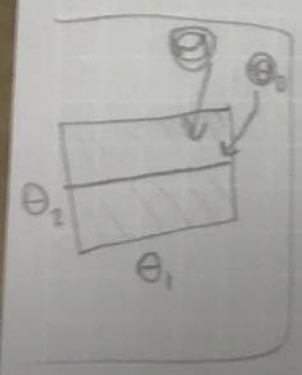
(Λ: uppercase λ)

where $\hat{\theta}_n$ is MLE and $\hat{\theta}_{0,n}$ is MLE for θ restricted to Θ_0

Prop.

Suppose $\theta = (\theta_1, \dots, \theta_r)$ and let $\Theta_0 = \{\theta : (\theta_{q+1}, \dots, \theta_r) = (\theta_{0,q+1}, \dots, \theta_{0,r})\}$

$|\Theta| = r$, $|\Theta_0| = q$



Let Λ_n be the LRTS under $H_0: \theta \in \Theta_0$.

$$\Lambda_n(X_1, \dots, X_n) \overset{n \to \infty}{\rightsquigarrow} \chi^2_{r-q}$$

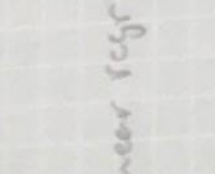
The p-value of the test is $P(\chi^2_{r-q} > \lambda_{n,obs})$

($\lambda_{n,obs}$: observed LRTS)

For instance, if $\theta = (\theta_1, \dots, \theta_6)$ and we want to test $\theta_4 = \theta_5 = \theta_6 = 0$, then Λ_n has limiting distr. given by $\chi^2_{6-3} = \chi^2_3$ RV.

Consider a model $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, where $\varepsilon_i \overset{iid}{\sim} N(0,1)$
for data of the form $((X_i, Y_i))_{i=1}^n$
We are interested in testing $H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$

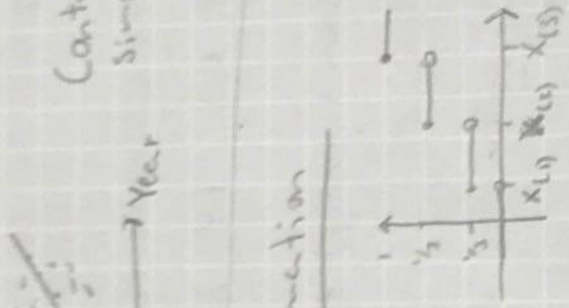
Recall
Suppose with notebook from day 1:



Can't use simple linear regression

Non-parametric Estimation

Recall: Let $X_1, \dots, X_n \overset{iid}{\sim}$ cdf F
EDF: $\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}(X_i \leq x)$



Prop
At any fixed x, $E(\hat{F}_n(x)) = F(x)$
$V(\hat{F}_n(x)) = \frac{F(x)(1-F(x))}{n} \to 0$

So, MSE (Mean Squared Error) $= \frac{F(x)(1-F(x))}{n} \to 0$
$\hat{F}_n(x) \xrightarrow{p} F(x)$

Prop. (Glivenko-Cantelli Lemma)
Let $X_1, \dots, X_n \overset{iid}{\sim} F$. Then $\sup_x |\hat{F}_n(x) - F(x)| \xrightarrow{p} 0$ as $n \to \infty$

Prop (The Dvoretzky-Kiefer-Wolfowitz (DKW) ineq.)
Let $X_1, \dots, X_n \overset{iid}{\sim} F$. Then, for any $\varepsilon > 0$,

$$P(\sup_x |\hat{F}_n(x) - F(x)| > \varepsilon) \leq 2e^{-2n\varepsilon^2}$$

A non-parametric $1-\alpha$ confidence band for F:

$$L(x) = \max\{\hat{F}_n(x) - \varepsilon_n, 0\}, \quad U(x) = \min\{\hat{F}_n(x) + \varepsilon_n, 1\} \text{ where } \varepsilon_n = \sqrt{\frac{1}{2n} \ln\left(\frac{2}{\alpha}\right)}$$

Due to DKW $\neq$, for any fixed but unknown $F \in \{\text{all DFs}\}$,

$$P(L(x) \leq F(x) \leq U(x) \ \forall x) \geq 1 - \alpha$$

Simple linear regression

Regression is a method of studying the relationship between a response RV Y and a covariate (predictor/feature) RV X.

[Galton (1822-1911)].

We summarize using regression function: $r(x) := E(Y|X=x) =$

$= \int x f(y|x)\,dy$.

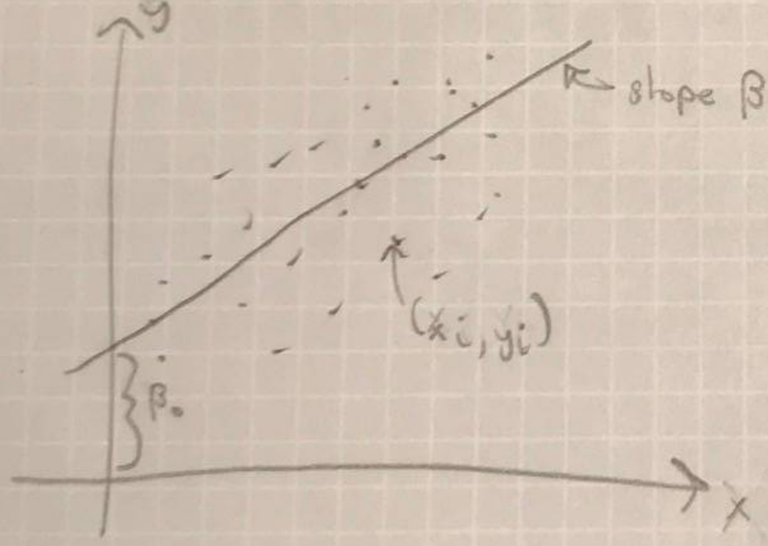
Our goal is to estimate $r(x)$ from the data of of the form

$(Y_1, X_1), (Y_2, X_2), \ldots, (Y_n, X_n) \overset{iid}{\sim} F_{X,Y}(x,y;\theta)$

Simplest version is called simple linear regression:

X_i, Y_i real-valued.

$r(x) = \beta_0 + \beta_1 x$, assume $V(Y|X=x) = \sigma^2$ does not depend on X.



Def

SLR Model:

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ where $E(\varepsilon_i|X_i) = 0$ and $V(\varepsilon_i|X_i) = \sigma^2$

We want estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ for β_0, β_1 from the data.

That gives us the fitted line: $\hat{r}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$

and the residuals: $\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$

$\hat{y}_i = \hat{r}(x_i)$ is the predicted value.

Def

Residual sum of squares (RSS) is given by $\sum_{i=1}^{n} \hat{\varepsilon}_i^2$

Def

Least squares estimate (LSE) is the value of $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimizes RSS.

$$\hat{\beta}_1 = \frac{\sum (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\sum (X_i - \bar{X}_n)^2}, \quad \hat{\beta}_0 = \bar{Y}_n - \beta_1 \bar{X}_n$$

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum \hat{\varepsilon}_i^2$$

Prop.

If $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, then MLE = LSE