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

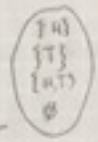
ii) $\frac{N(\Omega,n)}{n} = \frac{n}{n} = 1$ ← "something" happens every time

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

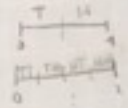
iv) If this is ignored the mathematics you need is much harder

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

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



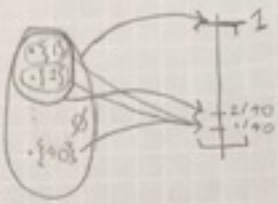
Bernoulli random variable. Bernoulli(θ)



ex2 NZ Lotto (40 balls)

Label of the 1st ball that comes out:

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



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

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

$L = 40$

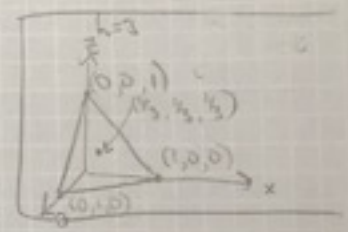
Here we have $\theta_1 = \theta_2 = \dots = \theta_L = \frac{1}{40}$, an Equi-probable DeMoivre variable

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. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

—

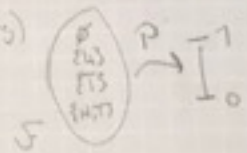
The domain of P is called a sigma field or sigma algebra,

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

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

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

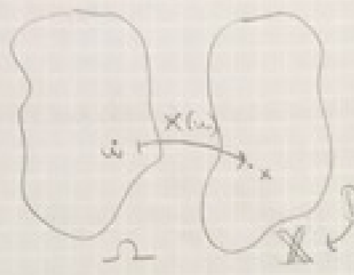
$\lambda = \frac{537}{576}$, "No. of hits" $\sim P_o(\lambda)$

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.)

prob. mass function (disc) ↔ distribution ↔ prob. dist function (cont)

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 $\subseteq \mathbb{R}^n$.

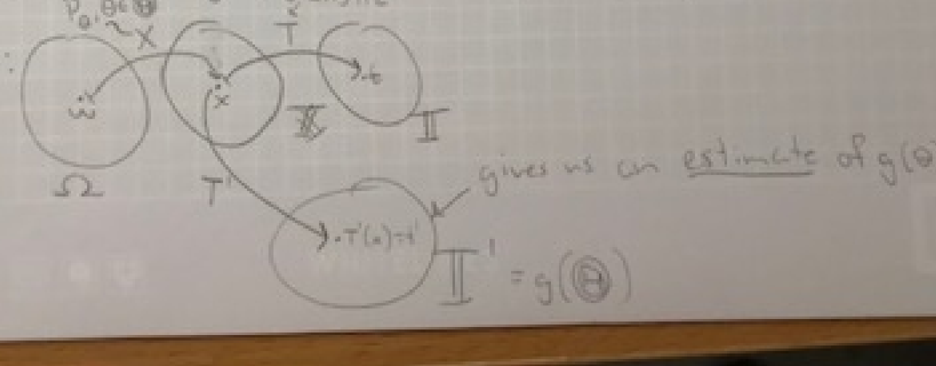
Def

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:



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

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

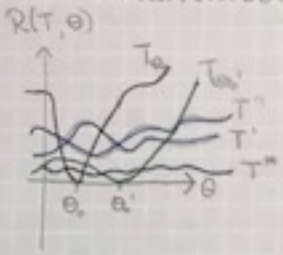
↑ Risk function of T

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

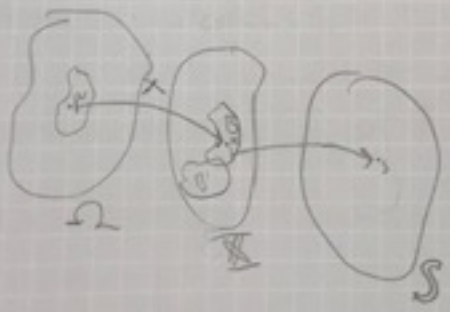
Consider an estimator T s.t. $E_\theta T$ exists.

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

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

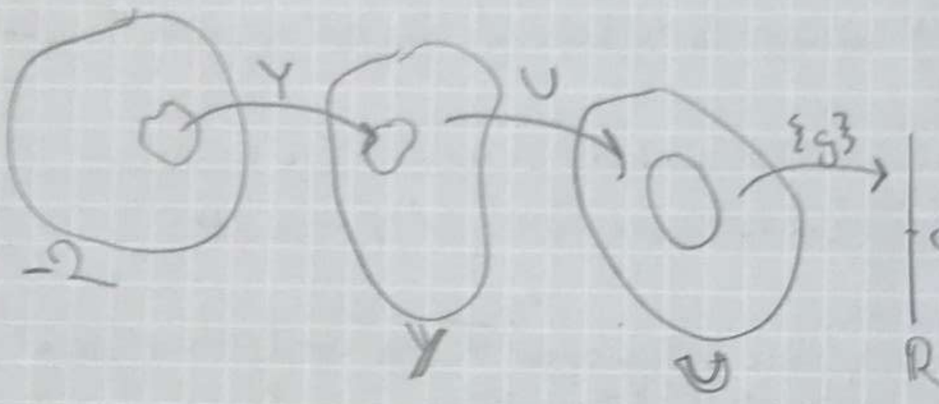
Def

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))

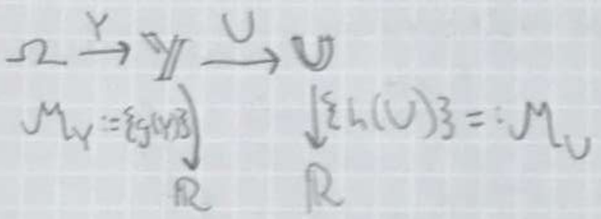
hint: add and subtract $E(h(Y)|U)$ in ⊛

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.