Inferensteori

course page: bit.ly/2sw1dai

raazesh.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 $f^{[-1]} = \underset{\substack{\uparrow \\ \text{collection} \\ \text{of subsets of } Y}}{\wp_Y(Y)} \to \wp_X(X)$

Note

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

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

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

1) $\forall$ event a, $0 \leq P(a) \leq 1$

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

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.

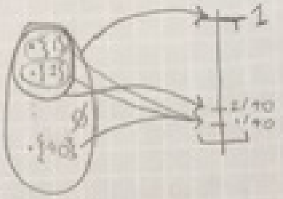
$\Omega = \{H,T\}$ $\quad \begin{Bmatrix} \{H\} \\ \{T\} \\ \{H,T\} \\ \emptyset \end{Bmatrix}$ Bernoulli random variable. Bernoulli(θ)

$2^\Omega = \mathcal{F}_\Omega$

ex2 NZ Lotto (40 balls)

Label of the 1st ball that comes out:

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



De Moivre random variable,

$2^\Omega = \mathcal{F}_\Omega$

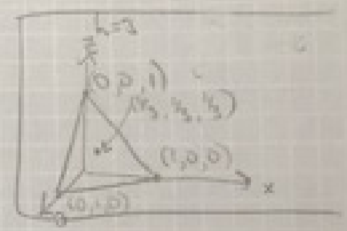
$\text{DeMoivre}(\frac{1}{L}, \frac{1}{L}, \dots, \underset{L=40}{\frac{1}{L}}) = \text{DeMoivre}(\theta_1, \theta_2, \dots, \theta_L)$

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

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

$= P(\{2\} \cup \{4\} \cup \dots \cup \{40\}) \underset{\text{axiom 3 repeatedly}}{=} P(\{2\}) + P(\{4\}) + \dots + P(\{40\}) =$

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



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) $\mathcal{F} \ni \{\emptyset, \{H\}, \{T\}, \{H,T\}\} \xrightarrow{P} [0,1]$

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

$X_1, \dots, X_n \overset{\text{iid}}{\sim} P_\theta$ (indep., ident. dist.), then $P_\theta(\{X_1 = x_1, X_2 = x_2, \dots, X_n = x_n\}) = \prod_{i=1}^{n} P_\theta(X_i = x_i)$

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 now 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

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

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

6.12

576 squares, 537 hits

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

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

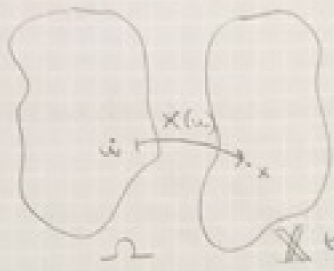
$\lambda = \frac{537}{576}$, "No. of hits" $\sim Po(\lambda)$

approx $= [576 \cdot D(Po(\lambda) = i)$ for i 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) ↘ ↙ prob. dist function (cont)

distribution

↕

Problem

Let X be a RV with values in 𝕏 and $\overset{L(X)}{Law(X)}$ 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)

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

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

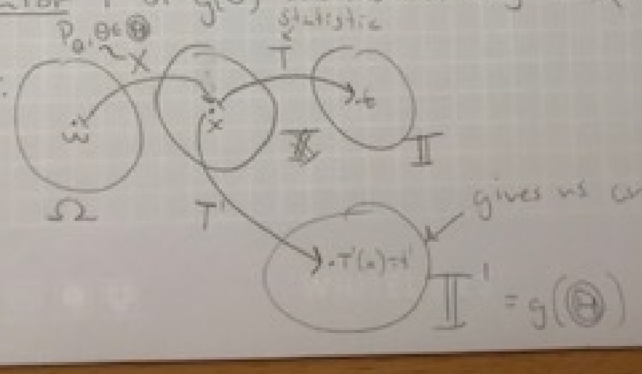
Def

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 $\underset{\text{statistic}}{T}$: $T: \mathbb{X} \to \overset{\{g(\theta) \mid \theta \in \Theta\}}{g(\Theta)}$.

Concretely:



gives us an estimate of g(θ)

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

$Law(X_1) \sim Bernoulli(\Theta) = Bin(1, \Theta)$

Suppose the data vector $x = (x_1, \ldots, x_n)$ 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 $\Theta \in \mathbf{\Theta} = [0,1]$

Note that $P_\Theta(X=x) = \prod_{i=1}^{n} P_\Theta(X_i = x_i) = \prod_{i=1}^{n} \Theta^{x_i}(1-\Theta)^{1-x_i} = \Theta^{\sum_{i=1}^{n} x_i}(1-\Theta)^{n-\sum_{i=1}^{n} x_i} \overset{\circledast}{=} \Theta^{t}(1-\Theta)^{n-t}$.

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

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

Now, consider another statistic: (sample mean)

$T(X) = \frac{1}{n}\sum_{i=1}^{n} X_i =: \overline{X}_n$.

Then $\circledast$ becomes $\Theta^{t'}(1-\Theta)^{n-t'} \overset{t'=nt}{\underset{\downarrow}{=}} \Theta^{nt}(1-\Theta)^{n(1-t)} = \Theta^{n\bar{x}_n}(1-\Theta)^{n(1-\bar{x}_n)}$

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

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

eg (Ber. 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 chose a loss function $Loss(t, \theta)$ which measures the loss (inaccuracy) w/ 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 $\Theta \in \circledS \subseteq \mathbb{R}$

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 emphasise 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

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

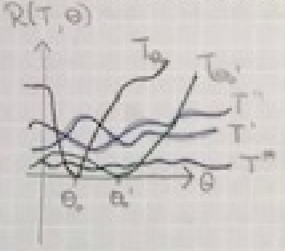
↑ Risk function of T

Note Risk might $\overset{not}{}$ exist since the expectation might not exist

Idea

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

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 estimator (in $\mathcal{T}$). (UBE)



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

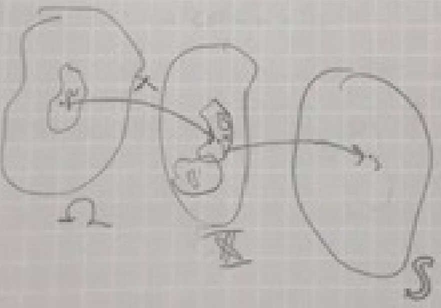
Def Consider an estimator T s.t. $E_\theta T$ exists.

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 $\overset{\text{for all}}{\vee}$ $P_\theta(X \in B \mid S(X) = s)$ is independent of θ, for all values of s and all events B.



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 $(\mathrm{Law}(X)=P_\Theta(\cdot))$.

After the value of the sufficient statistic $S(X)=s$ is known, then the conditional law $P_\Theta(\cdot\,|S(x)=t)$ is no longer dependant on Θ.

Since we are interested in $\overbrace{\text{making inferences about } \Theta}^{g(\Theta)}$, 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 $s=x$, $P_\Theta(X\in B|X=x)=\mathbb{1}_B(x)=\begin{cases}1, & \text{if } x\in B\\ 0, & \text{if } x\notin B\end{cases}$

Prop

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

Proof

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

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 $x=(x_1,\dots,x_n)$,

$$P_\Theta(X_1=x_1,X_2=x_2,\dots,X_n=x_n|n\overline{X}_n=k)=\frac{P_\Theta(X_1=x_1,\dots X_n=x_n, n\overline{X}_n=k)}{P_\Theta(n\overline{X}_n=k)} \quad \circledast$$

If $\circledast$ fulfils $\sum_{i=1}^{n} x_i=k$, then $\circledast$ 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 x, $\sum_{i=1}^{n} x_i \neq k$, 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 $\overline{X}_n$:

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

Additionally, suppose we limit ourselves further to unbiased estimators:

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

<u>Prop.</u>

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

<u>Proof</u>

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

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

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,

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

all possible values of $\overline{X}_n$ □

Def

A statistic T is called complete if, for all $h: \mathbb{T} \to \mathbb{R}$:
$[E_\theta(h(T(X)))=0$ 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

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

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 $k \in \{0,1,\dots,n\}$ 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 <u>151-156</u>

§1 - conv. of sequence of R.V.

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

- In distribution

  $X_n \rightsquigarrow X$

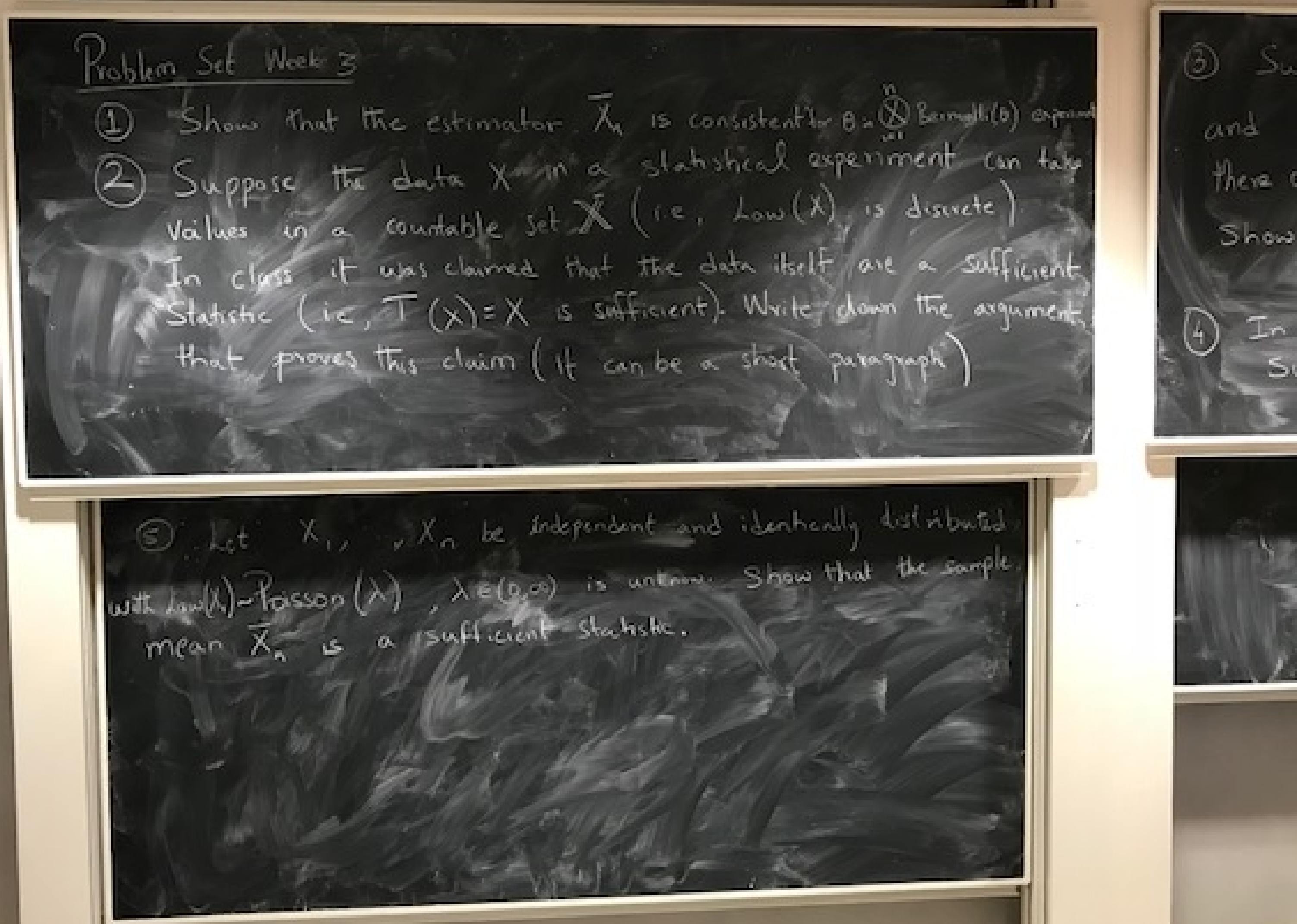
- In probability

  $X_n \xrightarrow{p} X$

-Markov's ineq., Chebyshevs 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 $\mathbb{S}^{[-1]}$ such that $\mathbb{S}^{[-1]}(S(t)) = t$ 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 $\overline{X}_n$ 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 $\mathcal{X}$ (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 $X_1, \dots, X_n$ be independent and identically distributed with Law(X_i) ~ Poisson(λ), $\lambda \in (0,\infty)$ is unknown. Show that the sample mean $\overline{X}_n$ is a sufficient statistic.

③ Su

and

there

Show

④ In

Su