Introduction to Data Science

(AoS)

Chapter 1: Probability (mathematical language to quantify uncertainty)

sample space (Ω) is the set of possible outcomes of an experiment. Subsets of Ω are called events. Points of ω in Ω are called sample outcomes / elements

ex 1 If we toss a coin forever, then the sample space is the infinite set

$\Omega = \{\omega = (\omega_1, \omega_2, \omega_3, \ldots) : \omega_i \in \{H, T\}\}$

Let E be the event that the first head appears on the third toss. Then

$E = \{(\omega_1, \omega_2, \omega_3, \ldots) : \omega_1 = T, \omega_2 = T, \omega_3 = H, \omega_i \in \{H, T\} \text{ for } i > 3\}$

Given an event A let $A^c = \{\omega \in \Omega : \omega \notin A\}$ complement of A (not A)

true event (always true) $\leftarrow$ $\Omega^c = \emptyset$ (empty set) $\rightarrow$ null event - always false

union of events A and B: $A \cup B = \{\omega \in \Omega : \omega \in A \text{ or } \omega \in B \text{ or } \omega \in \text{both}\}$

"A or B". If A is a sequence of ~~ere~~ sets then $\bigcup_{i=1}^{\infty} A_i = \{\omega \in \Omega : \omega \in A_i \text{ for at least one } i\}$

Intersection "A and B": $AB = A \cap B = \{\omega \in \Omega : \omega \in A \text{ and } \omega \in B\}$

if $A_1, A_2, \ldots$ is a sequence of sets: $\bigcap_{i=1}^{\infty} A_i = \{\omega \in \Omega : \omega \in A_i \text{ for all } i\}$

The set difference is defined by $A - B = \{\omega : \omega \in A, \omega \notin B\}$

If every element of A contained in B: $A \subset B$ or $B \supset A$

$A_1, A_2, \ldots$ are disjoint / mutually exclusive if $A_i \cap A_j = \emptyset$ whenever $i \neq j$

eg. $A_1 = [0, 1)$, $A_2 = [1, 2)$, $A_3 = [2, 3)$

A partition of Ω is a sequence of disjoint sets $A_1, A_2, \ldots$ such that $\bigcup_{i=1}^{\infty} A_i = \Omega$

Event A: the indicator function of A

$$I_A(\omega) = I(\omega \in A) = \begin{cases} 1, & \text{if } \omega \in A \\ 0, & \text{if } \omega \notin A \end{cases}$$

A sequence sets $A_1, A_2, \ldots$ is monotone increasing if $A_1 \subset A_2 \subset \ldots$ and we define $\lim_{n\to\infty} A_n = \bigcup_{i=1}^{\infty} A_i$, if —"— decreasing if $A_1 \supset A_2 \supset \ldots$ and we define $\lim_{n\to\infty} A_n = \bigcap_{i=1}^{\infty} A_i$

A function $\mathbb{P}$ that assigns a real number $\mathbb{P}(A)$ to each event A is a probability distribution or a probability measure if it satisfies the following axioms:

Axiom 1: $\mathbb{P}(A) \geq 0$ for every A

Axiom 2: $\mathbb{P}(\Omega) = 1$

Axiom 3: If $A_1, A_2, \ldots$ are disjoint the $\mathbb{P}(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$

Derived Properties from Axioms:

$\mathbb{P}(\emptyset) = 0$ (use Axiom 3) $\qquad$ $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$ (use Axiom 3)

$A \subset B \Rightarrow \mathbb{P}(A) \leq \mathbb{P}(B)$ $\qquad$ $A \cap B = \emptyset \Rightarrow \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$

$0 \leq \mathbb{P}(A) \leq 1$

$\{x \mid x \in A \cap B\}$ → definition 2 $\iff$ → implications $\Rightarrow$ conditionals needs ...

Proof $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. We can start from the following events: $= P((AB^c) \cup (AB) \cup (A^cB)) = P(AB^c) + P(AB) + P(A^cB)$

$= P(AB^c) + P(AB) + P(A^cB) + P(AB) - P(AB)$

$= P((AB^c) \cup (AB)) + P((A^cB) \cup (AB)) - P(AB)$

$= P(A) + P(B) - P(AB)$

Theorem. Continuity of Probabilities. If $A_n \to A$ then $P(A_n) \to P(A)$ as $n \to \infty$

Proof. Suppose A_n is monotone increasing (→ the sequence increases and → ...) so that $A_1 \subset A_2 \subset$. Let $A = \lim_{n\to\infty} A_n = \bigcup_{i=1}^{\infty} A_i$. Define $B_1 = A_1$, $B_2 = \{\omega \in \Omega : \omega \in A_2, \omega \notin A_1\}$, $B_3 = \{\omega \in \Omega : \omega \in A_3, \omega \notin A_2, \omega \notin A_1\}$ ...

$B_1, B_2, B_3, \ldots$ disjoint. $A_n = \bigcup_{i=1}^{n} A_i = \bigcup_{i=1}^{n} B_i$ (→ set of all outcomes) for each n and $\bigcup_{i=1}^{\infty} B_i = \bigcup_{i=1}^{\infty} A_i$

From Axiom 3: $P(A_n) = P\left(\bigcup_{i=1}^{n} B_i\right) = \sum_{i=1}^{n} P(B_i)$

$\Rightarrow \lim_{n\to\infty} P(A_n) = \lim_{n\to\infty} \sum_{i=1}^{n} P(B_i) = \sum_{i=1}^{\infty} P(B_i) = P\left(\bigcup_{i=1}^{\infty} B_i\right) = P(A)$

→ If Ω is finite and if each outcome is equally likely then $P(A) = \frac{|A|}{|\Omega|}$ and it's called uniform probability distribution.

Given n objects, the number of ways of ordering these is $n! = (n-1)(n-2) \cdot n \ldots 3 \cdot 2 \cdot 1$

$0! = 1$ $\quad \binom{n}{k} = \frac{n!}{k!(n-k)!}$ "n choose k" $\quad \binom{n}{0} = \binom{n}{n} = 1$

Independent Events. Two events A and B are independent if $P(AB) = P(A)P(B)$

$A \perp\!\!\!\perp B$, when are not we write $A \not\perp\!\!\!\perp B$

Suppose A and B are ~~independent~~ disjoint each with +tive probabilities. Can they be independent?

No. $P(A) \cdot P(B) > 0$ yet $P(AB) = P(\phi) = 0$

Conditional Probability: If $P(B) > 0$ then $P(A|B) = \frac{P(AB)}{P(B)}$

* If A and B independent events $\Rightarrow P(A|B) = P(A)$

$P(AB) = P(A|B) \cdot P(B) = P(B|A) P(A)$

1) Bayes' Theorem: Let $A_1, \ldots A_k$ be a partition of Ω such that $P(A_i) > 0$ for each i. If $P(B) > 0$ then for each $i = 1, \ldots, k$.

posterior prob. of A: $P(A_i|B) = \frac{P(B|A_i)\, P(A_i)}{\sum_j P(B|A_j) \cdot P(A_j)}$ → prior prob of A

2) The Law of Total Probability. Let $A_1, \ldots, A_k$ be a partition of Ω (→ mutually exclusive & exhaustive events). Then for any event B

$P(B) = \sum_{i=1}^{k} P(B|A_i) P(A_i)$

Proof 2: $C_j = BA_j$ and $C_1 \ldots C_k$: disjoint and $B = \bigcup_{j=1}^{k} C_j$

$P(B) = \sum_j P(C_j) = \sum_j P(BA_j) = \sum_j P(B|A_j) P(A_j)$

Proof 1: $P(A_i|B) = \frac{P(A_iB)}{P(B)} = \frac{P(B|A_i)P(A_i)}{P(B)} \underset{\text{Proof 2}}{=} \frac{P(B|A_i)P(A_i)}{\sum_j P(B|A_j) \cdot P(A_j)}$

$P(A) = P(A \cap B_1) + P \ldots + P(A \cap B_k)$

$P(A) = P(B_1) P(A|B_1) + \ldots + P(B_k) P(A|B_k)$

(AUS)

Chapter 2 : Random Variables (RV)

A RV is a mapping $X: \Omega \to \mathbb{R}$ that assigns a real number $X(\omega)$ to each outcome ω.

Let $\Omega = \{(x,y); x^2+y^2 \leq 1\}$, ω typical outcome $= (x,y)$

RV : $X(\omega) = x$, $Y(\omega) = y$ $Z(\omega) = x+y$ $W(\omega) = \sqrt{x^2+y^2}$

X is a subset A define $X^{-1}(A) = \{\omega \in \Omega \; X(\omega) \in A\}$ and

$\mathbb{P}(X \in A) = \mathbb{P}(X^{-1}(A)) = \mathbb{P}(\{\omega \in \Omega ; X(\omega) \in A\})$

$\mathbb{P}(X = x) = \mathbb{P}(X^{-1}(x)) = \mathbb{P}(\{\omega \in \Omega ; X(\omega) = x\})$

RV ← ↳ particular value

Ex. Flip a coin twice and let X be the number of heads. Then $\mathbb{P}(X=0) = \mathbb{P}(\{TT\}) = 1/4$, $\mathbb{P}(X=1) = \mathbb{P}(\{HT, TH\}) = 1/2$ and $\mathbb{P}(X=2) = \mathbb{P}(\{HH\}) = 1/4$

ω	$\mathbb{P}(\{\omega\})$	$X(\omega)$
TT	1/4	0
TH	1/4	1
HT	1/4	1
HH	1/4	2

x	$P(X=x)$
0	1/4
1	1/2
2	1/4

Discrete : Probab Mass Function (PMF)

Continuous : Probability density funct (PDF)

cumulative distribut function (CDF)

* The cumulative distribution function (CDF) is the function $F_X : \mathbb{R} \to [0,1]$ defined by $F_X(x) = \mathbb{P}(X \leq x)$ } right continuous non-decreasing

~~Ex. Flip a coin twice and X~~

$$F_X(x) = \begin{cases} 0 & , x<0 \\ 1/4 & ; 0 \leq x < 1 \\ 3/4 & ; 1 \leq x < 2 \\ 1 & , x \geq 2 \end{cases}$$

Let X have CDF F and Y have CDF G. If $F(x) = G(x)$ for all x then $\mathbb{P}(X \in A) = \mathbb{P}(Y \in A)$ for all A.

A function F mapping to $[0,1]$ is a CDF for some probability $\mathbb{P}$ if and only if F satisfies : i) F is non-decreasing $x_1 < x_2$ implies $F(x_1) \leq F(x_2)$

ii) F is normalised $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$

iii) F is right-continuous $F(x) = F(x^+)$ for all x where $F(x^+) = \lim_{\substack{y \to x \\ y > x}} F(y)$

X is discrete if it takes countably many values $\{x_1, x_2 \dots\}$ We define the probability function or probability mass function for X by $f_X(x) = \mathbb{P}(X = x)$

Thus, $f_X(x) \geq 0$ for all $x \in \mathbb{R}$ and $\sum_i f_X(x_i) = 1$ The CDF of X :

$$F_X(x) = \mathbb{P}(X \leq x) = \sum_{x_i \leq x} f_X(x_i)$$

$$f_X(x) = \begin{cases} 1/4 &, x=0 \\ 1/2 &, x=1 \\ 1/4 &, x=2 \\ 0 &, \text{otherwise} \end{cases}$$

A RV X is continuous if there exists a function f_X such that $f_X(x) \geq 0$ for all x $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$ and for every $a \leq b$

$\mathbb{P}(a < X < b) = \int_a^b f_X(x)\,dx$

The function f_X is called the probability density function (PDF):

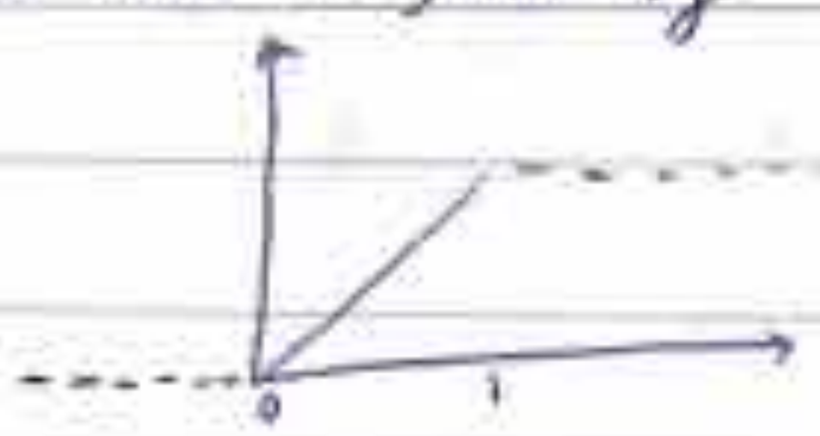
$F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt$ and $f_X(x) = F_X'(x)$ at all points x at which F_X is differentiable

Ex.1 Let say X has PDF:

$$f_X(x) = \begin{cases} 1 & \text{for } 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

$f_X(x) \geq 0$ and $\int f_X(x)\,dx = 1$. A RV with this density is said to have a Uniform(0,1) distribution. The CDF is given by

$$F_X(x) = \begin{cases} 0 &, x < 0 \\ x &, 0 \leq x \leq 1 \\ 1 &, x > 1 \end{cases}$$



Ex.2: X has PDF:

$$f_X(x) = \begin{cases} 0 & \text{for } x < 0 \\ \frac{1}{(1+x)^2} & \text{otherwise} \end{cases}$$

since $\int_{-\infty}^{+\infty} f_X(x)\,dx = 1 \to$ well-defined PDF

* If X continuous $\to \mathbb{P}(X = x) = 0$ for every x $\neq$ $f(x) = \mathbb{P}(X = x)$
* A PDF can be bigger than 1 $\neq$ mass function. A PDF can be unbounded

Ex.3: Let $f(X) = \begin{cases} 0 & , \text{for } x < 0 \\ \frac{1}{1+x} & , \text{otherwise} \end{cases}$ this is not a PDF since $\int f(x)\,dx = \int_0^\infty dx/(1+x) = \int_1^\infty du/u = \log(\infty) = \infty$

Lemma 2.15 Let F be the CDF for a RV X. Then:

1. $\mathbb{P}(X = x) = F(x) - F(x^-)$ where $F(x^-) = \lim_{y \uparrow x} F(y)$
2. $\mathbb{P}(x < X \leq y) = F(y) - F(x)$
3. $\mathbb{P}(X > x) = 1 - F(x)$
4. If X continuous: $F(b) - F(a) = \mathbb{P}(a < X < b) = \mathbb{P}(a \leq X < b) = \mathbb{P}(a < X \leq b) = \mathbb{P}(a \leq X \leq b)$

Let X be a RV with CDF F. The inverse CDF (quantile function) is denoted by:

$F^{-1}(q) = \inf\{x : F(x) > q\}$ for $q \in [0,1]$. If F is strictly increasing & continuous then $F^{-1}(q)$ is the unique real number x such that $F(x) = q$

1st quartile $F^{-1}(1/4)$ 3rd quartile $F^{-1}(3/4)$

2nd —"— (median) : $F^{-1}(1/2)$

* Two variables are equal in distribution $X \overset{d}{=} Y$ if $F_X(x) = F_Y(x)$ for all x
but X and Y not equal eg. $\mathbb{P}(Y=1) = \mathbb{P}(Y=-1) = 1/2$.
↳ $P(X=Y) = 0$

Some Important Discrete RV.

1. Point Mass Distribution: $X \sim \delta_a$ if $\mathbb{P}(X=a) = 1$ $F_X(x) = \begin{cases} 0, & x<a \\ 1, & x \geq a \end{cases}$
PMF: $f(x) = 1$ for $x=a$ and 0 otherwise.
2. Discrete Uniform Distribution: $k > 1$ (given integer), then X has PDF:
$f(x) = \begin{cases} 1/k, & \text{for } x = 1 \dots k \\ 0, & \text{otherwise} \end{cases}$
3. Bernoulli Distribution: $\mathbb{P}(X=1) = p$ and $\mathbb{P}(X=0) = 1-p$, $p \in [0,1]$
$X \sim \text{Bernoulli}(p)$. Probability function: $f(x) = p^x (1-p)^{1-x}$, $x \in \{0,1\}$
4. Binomial Distribution: $P(=\text{heads}) = p$ ∴ Flip it n times, $X = \#$ of heads
$f(x) = \mathbb{P}(X=x)$ mass function
$f(x) = \begin{cases} \binom{n}{x} \cdot p^x (1-p)^{n-x}, & \text{For } x = 0 \dots n \\ 0, & \text{otherwise} \end{cases}$ $X \sim \text{Binomial}(n,p)$
and you can add binomials
5. Geometric Distribution: $X \sim \text{Geom}(p)$, $p \in (0,1)$
$\mathbb{P}(X=k) = p(1-p)^{k-1}$, $k \geq 1$
$\sum_{k=1}^{\infty} \mathbb{P}(X=k) = p \sum_{k=1}^{\infty} (1-p)^k = \frac{p}{1-(1-p)} = 1$
$X =$ "# of flips until the first success of heads"
6. Poisson Distribution $X \sim \text{Poisson}(\lambda)$
$f(x) = e^{-\lambda} \cdot \frac{\lambda^x}{x!}$, $x \geq 0$ $\sum_{x=0}^{\infty} f(x) = e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^{-\lambda} e = 1$
eg. use for rare events like radioactive decay or traffic accidents → you can add

Some important ~~Dis~~ Continuous RV:

1. Uniform Distribution $X \sim \text{uniform}(a,b)$ if:
$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } x \in [a,b] \\ 0 & \text{otherwise} \end{cases}$ when $a<b$ Distribution function is:
$F(x) = \begin{cases} 0, & x < a \\ \frac{x-a}{b-a} & x \in [a,b] \\ 1 & x > b \end{cases}$
2. Normal (Gaussian) $X \sim N(\mu, \sigma^2)$ $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}$, $x \in \mathbb{R}$
standard normal distribution $\sim N(0,1^2)$ PDF → $\phi(z)$ CDF → $\Phi(z)$
i) If $X \sim N(\mu, \sigma^2)$ then $Z = (X-\mu)/\sigma \sim N(0,1)$
ii) If $Z \sim N(0,1)$ then $X = \mu + \sigma Z \sim N(\mu, \sigma^2)$
iii) if $X_i \sim N(\mu_i, \sigma_i^2)$, $i = 1 \dots n$ independent then $\sum_{i=1}^{n} X_i \sim N\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$

$P(a<X<b) = P\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$

Ex. $X \sim N(3,5)$ find $P(X>1)$

$1 - P(X<1) = 1 - P\left(Z < \frac{1-3}{\sqrt{5}}\right) = 1 - \Phi(-0.8944) = 0.81$

now find $q = \Phi^{-1}(0.2)$ $\quad P(X<q) = 0.2$

$= P\left(Z < \frac{q-\mu}{\sigma}\right) = \Phi\left(\frac{q-\mu}{\sigma}\right) = \Phi(-0.8416) = 0.2$

$= -0.8416 = \frac{q-\mu}{\sigma} = \frac{q-3}{\sqrt{5}} \Leftrightarrow q = 3 - 0.8416 \cdot \sqrt{5} = 1.118$

3. Exponential Distribution. $X \sim Exp(\beta)$ $\quad f(x) = \frac{1}{\beta} e^{-x/\beta}, x>0, \quad \beta > 0$

↳ used for the lifetimes of electronic components and the waiting times between rare events

4. Gamma Distribution: $\alpha > 0$ $\quad \Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y}\, dy$

$X \sim Gamma(\alpha, \beta)$ if $f(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} \cdot e^{-x/\beta}, x>0, \quad \alpha, \beta > 0$

exponential distribution is $Gamma(1, \beta)$

5. Beta Distribution: $\alpha > 0, \beta > 0$ $\quad X \sim Beta(\alpha, \beta)$ if $f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$ $\quad 0<x<1$

6. t and Cauchy Distribution $X \sim t_\nu$ (similar to normal but with thicker tails)

normal corresponds to t with $\nu = \infty$

special case when $\nu = 1$ $\quad f(x) = \frac{1}{\pi(1+x^2)}$

5. χ^2 Distribution. $X \sim \chi^2_p$

if $Z_1, \dots Z_p$ independent standard Normal RV then $\sum_{i=1}^p Z_i^2 \sim \chi^2_p$

Bivariate Distribution.

X, Y discrete - Joint Mass function $f(x,y) = P(X=x \,\&\, Y=y)$

→ X, Y continuous we call $f(x,y)$ a PDF if

i) $f(x,y) \geq 0$ for all (x,y)

ii) $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y)\, dx\, dy = 1$

iii) for any set $A \subset \mathbb{R} \times \mathbb{R}$, $P((X,Y) \in A) = \iint_A f(x,y)\, dx\, dy$

In both discrete & continuous the Joint CDF $F_{XY}(x,y) = \mathbb{P}(X \leq x, Y \leq y)$

Ex. Let (X,Y) uniform

$f(x,y) = \begin{cases} 1 & \text{if } 0 \leq x \leq 1, \ 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$

$P(X < 1/2, Y < 1/2)$ event $A = \{X < 1/2, Y = 1/2\}$

integrating f over this subset (compute area of A) $= 1/4$, $P(X<\frac{1}{2}, Y<\frac{1}{2}) = 1/4$

Ex. Let (X,Y) have a density

$f(x,y) = \begin{cases} x+y & \text{if } 0 \leq x \leq 1, \ 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$ then

$\iint_0^1 (x+y)\, dx\, dy = \int_0^1 \left[\int_0^1 x\, dx\right] dy + \int_0^1 \left[\int_0^1 y\, dx\right] dy$

$= \int_0^1 \frac{1}{2}\, dy + \int_0^1 y\, dy = \frac{1}{2} + \frac{1}{2} = 1 \Rightarrow$ verifies it is a PDF

If $X_1, X_2 \ldots X_n$ continuous RV then the marginal pdf of X_1 is found by integrating $x_2, x_3 \ldots x_n$ out of the joint pdf.

Marginal Distributions →

If (X, Y) have joint distribution with mass function $f_{X,Y}$ then marginal mass function for X is:

$f_X(x) = \mathbb{P}(X = x) = \sum_y P(X = x, Y = y) = \sum_y f(x,y)$ and the same for y.

Ex. The marginal distributions for X corresponds to the row totals and for Y the column totals

	$y=0$	$y=1$	
$X=0$	$1/10$	$2/10$	$3/10$
$X=1$	$3/10$	$4/10$	$7/10$
	$4/10$	$6/10$	1

$f_X(0) = 3/10$, $f_X(1) = 7/10$

For continuous RV the marginal densities are (same x, y):

$f_X(x) = \int f(x,y)\, dy$, $f_Y(y) = \int f(x,y)\, dx$

Ex. ① $f(x,y) = \begin{cases} x+y & \text{if } 0 \le x \le 1,\ 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$

then

$$f_Y(y) = \int_0^1 (x+y)\, dx = \int_0^1 x\, dx + \int_0^1 y\, dx = \frac{1}{2} + y$$

② $f(x,y) = \begin{cases} \frac{21}{4} x^2 y & \text{if } x^2 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$

then $f_X(x) = \int f(x,y)\, dy = \frac{21}{4} x^2 \int_{x^2}^{1} y\, dy = \frac{21}{8} x^2 (1 - x^4)$ for $-1 \le x \le 1$

and $f_X(x) = 0$ otherwise

Independent RV. Two RVs X and Y are independent if for every A and B $P(X \in A, Y \in B) = P(X \in A)\, P(Y \in B)$ → $X \perp\!\!\!\perp Y$ otherwise $X \not\perp\!\!\!\perp Y$

Ex.

	$y=0$	$y=1$	
$X=0$	$1/4$	$1/4$	$1/2$
$X=1$	$1/4$	$1/4$	$1/2$
	$1/2$	$1/2$	1

$f_X(0) = f_X(1) = 1/2$ $f_Y(0) = f_Y(1) = 1/2$

$f_X(0)\, f_Y(0) = f(0,0)$, $f_X(0)\, f_Y(1) = f(0,1)$

$f_X(1)\, f_Y(0) = f(1,0)$, $f_X(1) \cdot f_Y(1) = f(1,1)$

X, Y independent.

Ex. 2 X, Y independent and have same density.

$f(x) = \begin{cases} 2x & \text{if } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$

$P(X + Y \le 1)$ using independence, the joint density is

$f(x,y) = f_X(x)\, f_Y(y) = \begin{cases} 4xy & \text{if } 0 \le x \le 1,\ 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$

$$P(X + Y \le 1) = \iint_{x+y \le 1} f(x,y)\, dy\, dx$$

$$= 4 \int_0^1 x \left[\int_0^{1-x} y\, dy \right] dx$$

$$= 4 \int_0^1 x \frac{(1-x)^2}{2}\, dx = 1/6$$

Conditional Distributions

for discrete: The conditional probability mass function: PDF

$$f_{X|Y}(x|y) = \mathbb{P}(X=x|Y=y) = \frac{P(X=x, Y=y)}{P(Y=y)} = \frac{f_{XY}(x,y)}{f_Y(y)}$$

→ joint

↳ Marginal PDF of y

for continuous: $f_Y(y) > 0$: $P(X \in A | Y=y) = \int_A f_{X|Y}(x|y)\,dx$

Ex 1. $f(x,y) = \begin{cases} x+y & \text{if } 0 \le x \le 1, 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$ $\quad \mathbb{P}(X < 1/4 | Y = 1/3)$

$$f_y(y) = \int_0^1 x+y\,dx = \frac{x^2}{2} + xy \Big]_0^1 = \frac{1}{2} + y$$

$$f_{X|Y}(x|y) = \frac{f_{XY}(x,y)}{f_Y(y)} = \frac{x+y}{y+1/2}$$

$$\Rightarrow \int_0^{1/4} f_{X|Y}(x|\tfrac{1}{3})\,dx = \int_0^{1/4} \frac{x+\frac{1}{3}}{\frac{1}{3}+\frac{1}{2}}\,dx = \frac{\frac{1}{32}+\frac{1}{12}}{\frac{1}{3}+\frac{1}{2}} = \frac{11}{80}$$

Ex 2. Suppose that $X \sim \text{Uniform}(0,1)$. $Y|X=x \sim \text{Uniform}(x,1)$ what is the marginal distribution of y

$f_X(x) = \begin{cases} 1 & \text{if } 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$ and $f_{Y|X}(y|x) = \begin{cases} \frac{1}{1-x} & \text{if } 0 < x < y < 1 \\ 0 & \text{otherwise} \end{cases}$

$f_{XY}(x,y) = f_{Y|X}(y|x)\, f_X(x) = \begin{cases} \frac{1}{1-x} & \text{if } 0 < x < y < 1 \\ 0 & \text{otherwise} \end{cases}$

the marginal for y is $f_Y(y) = \int_0^y f_{X,Y}(x,y)\,dx = \int_0^y \frac{dx}{1-x} = -\int_1^{1-y} \frac{du}{u} = -\log(1-y)$

$0 < y < 1$

Multivariate Distributions and IID Samples

Let $X = (X_1 \ldots X_n)$, where $X_1 \ldots X_n$ RVs. X random vector. Let $f(x_1, \ldots x_n)$ denote the PDF

we say $X_1 \ldots X_n$ independent for every $A_1, \ldots A_n$,

$P(X_1 \in A_1, \ldots, X_n \in A_n) = \prod_{i=1}^n P(X_i \in A_i)$. It suffices to check $f(x_1, \ldots x_n) = \prod_{i=1}^n f_{X_i}(x_i)$

→ If $X_1 \ldots X_n$ independent and each has the same marginal distribution with CDF F we say that $X_1 \ldots X_n$ are independent and identically distributed (iid).

$\underbrace{X_1, X_2 \ldots X_n}_{\text{random sample of size } n \text{ from } F} \sim F$. If F has density f we also write $X_1 \ldots X_n \sim f$

with replacement — Multinomial: is the multivariate version of Binomial. eg Draw a coloured ball from an urn with k different colours. Let $p = (p_1 \ldots p_k)$ where $p_j \ge 0$ and $\sum_{j=1}^k p_j = 1$

p_j probability of drawing a ball of ~~one~~ colour j. Draw n times (independent with replacement) and let $X = (X_1 \ldots X_k)$ where X_j: number of times that colour j appears.

$X \sim \text{Multinomial}(n, p)$ with a prob function: $f(x) = \binom{n}{x_1 \ldots x_k} p_1^{x_1} \ldots p_k^{x_k}$

⊛ Suppose that $X \sim \text{Multinomial}(n, p)$ where $X = (X_1 \ldots X_k)$ and $p = (p_1 \ldots p_k)$. The marginal distribution of X_j is $\text{Binomial}(n, p_j)$

Multivariate Normal has two parameters μ and Σ

$Z = \begin{pmatrix} z_1 \\ z_k \end{pmatrix}$ where $z_1, \dots, z_k \sim N(0,1)$ independent. The density z is

$f(z) = \prod_{i=1}^{k} f(z_i) = \frac{1}{(2\pi)^{k/2}} \exp\{-\frac{1}{2}\sum_{j=1}^{k} z_j^2\} = \frac{1}{(2\pi)^{k/2}} \exp\{-\frac{1}{2} z^T z\}$ Σ : variance-covariance matrix

Transformations of RVs Let X: RV with PDF f_X and CDF F_X. Let $Y = r(X)$

eg $Y = X^2$, $y = e^x$. Y is a function of X and $r(X)$ transformation of X

Compute the PDF and CDF of Y. In the discrete case is easy and mass function is $f_Y(y) = P(Y=y) = P(r(X)=y) = P(\{x; r(x)=y\}) = P(X \in r^{-1}(y))$

Ex. Suppose $P(X=-1) = P(X=1) = 1/4$ and $P(X=0) = 1/2$. Let $Y = X^2$

Then $P(Y=0) = P(X=0) = 1/2$ and $P(Y=1) = P(X=1) + P(X=-1) = 1/2$

x	$f_X(x)$
-1	1/4
0	1/2
1	1/4

y	$f_Y(y)$
0	1/2
1	1/2

Y takes fewer values than X because the transformation is not one-to-one

continuous (3 steps to find f_Y) 1. For each y, find the set $A_y = \{x : r(x) \le y\}$

2. find CDF $F_Y(y) = P(Y \le y) = P(r(X) \le y) = P(\{x, r(x) \le y\})$
$= \int_{A_y} f_X(x)\, dx$

3. PDF is $f_Y(y) = F_Y'(y)$

Ex. Let $f_X(x) = e^{-x}$, $x > 0$. Hence $F_X(x) = \int_0^x f_X(s)\, ds = 1 - e^{-x}$

Let $Y = r(X) = \log(X)$. Then $A_y = \{x : x \le e^y\}$

$F_Y(y) = P(Y \le y) = P(\log X \le y) = P(X \le e^y) = F_X(e^y) = 1 - e^{-e^y}$ for $y \in \mathbb{R}$

Ex 2: Let $X \sim$ Uniform $(-1, 3)$ find the PDF of $Y = X^2$. The density of X is

$f_X(x) = \begin{cases} 1/4 & \text{if } -1 < x < 3 \\ 0 & \text{otherwise} \end{cases}$ Y can only take $(0, 9)$

Two cases. i) $0 < y < 1$ and ii) $1 \le y < 9$. For (i) $A_y = [-\sqrt{y}, \sqrt{y}]$ and $F_Y(y) = \int_{A_y} f_X(x)\, dx = 1/2 \cdot \sqrt{y}$. For (ii) $A_y = [-1, \sqrt{y}]$ and $F_Y(y) = \int_{A_y} f_X(x)\, dx = (1/4)(\sqrt{y} + 1)$. if we differentiate:

$f_Y(y) = \begin{cases} \frac{1}{4\sqrt{y}} & \text{if } 0 < y < 1 \\ \frac{1}{8\sqrt{y}} & \text{if } 1 < y < 9 \\ 0 & \text{otherwise} \end{cases}$

* when r strictly monotone increasing or decreasing then r has an inverse $s = r^{-1}$ $f_Y(y) = f_X(s(y)) \left| \frac{ds(y)}{dy} \right|$

Transformations of several RVs. if X, Y RVs we might want to know the distribution of X/Y, $X+Y$, $\max\{X,Y\}$, $\min\{X,Y\}$. Let $Z = r(X,Y)$ the steps to find f_Z.

independence, then:
1) $P(A|B) = P(A)$
2) $P(A \cap B) = P(A) \cdot P(B)$

1. For each z, find the set $A_z = \{(x,y) : r(x,y) \le z\}$
2. Find the CDF $F_Z(z) = P(Z \le z) = P(r(X,Y) \le z)$
$= P(\{(x,y); r(x,y) \le z\}) = \iint_{A_z} f_{X,Y}(x,y)$ ...
3. Then $f_Z(z) = F'_Z(z)$

Ex. Let $X_1, X_2 \sim$ Uniform $(0,1)$ be independent. Find density $Y = X_1 + X_2$. The joint density of (X_1, X_2)

$f(x_1, x_2) = \begin{cases} 1, & 0 < x_1 < 1,\ 0 < x_2 < 1 \\ 0, & \text{otherwise} \end{cases}$

Let $r(x_1, x_2) = x_1 + x_2$

$F_Y(y) = P(Y \le y) = P(r(X_1, X_2) \le y)$
$= P(\{(x_1, x_2) : r(x_1, x_2) \le y\}) = \iint_{A_y} f(x_1, x_2) \, dx_1 \, dx_2$

Find A_y: Suppose $0 < y \le 1$. Then A_y is the triangle with vertices $(0,0)$, $(y,0)$ and $(0,y)$. Then $\iint_{A_y} f(x_1, x_2) \, dx_1 \, dx_2$ is the area of the triangle with vertices $(1, y-1)$, $(1,1)$, $(y-1, 1)$. This set has area $1 - (2-y)^2/2$

$F_Y(y) = \begin{cases} 0, & y < 0 \\ y^2/2, & 0 \le y < 1 \\ 1 - \left[\frac{(2-y)^2}{2}\right], & 1 \le y < 2 \\ 1, & y \ge 2 \end{cases}$ if we differentiate the PDF:

$f_Y(y) = \begin{cases} y & 0 \le y \le 1 \\ 2-y & 1 \le y \le 2 \\ 0 & \text{otherwise} \end{cases}$

(**) helps you ~~find the probability~~ calculate the Expected value of a function of a RV without having to find the probab distribution of a function itself.

Chapter 3: Expectation

Expected value, mean, first moment of X is defined to be

$E(X) = \int x \, dF(x) = \begin{cases} \sum_x x \cdot f(x), & \text{if } X \text{ discrete} \\ \int x \cdot f(x) \, dx & \text{if } X \text{ continuous} \end{cases}$ assuming integral is well defined

$= \mu = \mu_X$

$E(X)$ well defined if $\int_x |x| \, dF_X(x) < \infty$

$X \sim$ Bernoulli (p). Then $E(X) = \sum_{x=0}^{1} x \cdot f(x) = 0(1-p) + (1p) = p$

$X \sim$ Uniform $(-1, 3)$. Then $E(X) = \int x \, dF_X(x) = \int x f_X(x) \, dx = \frac{1}{4} \int_{-1}^{3} x \, dx = 1$

* Cauchy distribution: $f_X(x) = \{\pi(1+x^2)\}^{-1}$ using $u = x$, $v = \tan^{-1} x$

$\int |x| \, dF(x) = \frac{2}{\pi} \int_0^\infty \frac{x \, dx}{1+x^2} = [x \tan^{-1}(x)]_0^\infty - \int_0^\infty \tan^{-1} x \, dx = \infty$

Mean does not exist. If you simulate many times the average never settles down
↳ because has thick tails + extreme observations are common

(**) The Rule of the Lazy Statistician: Let $Y = r(X)$. Then (function/transformation)

$E(Y) = E(r(X)) = \int r(x) \, dF_X(x)$ or $\sum_{i=1}^{k} r(x_i) \cdot P(X = x_i)$

* The k-th moment of X ~~of~~ is defined $E(X^k)$ assuming $E(|X|^k) < \infty$

If the k-th moment exists and if $j < k$ then the j-th moment exists.

Proof $E|X|^j = \int_{-\infty}^{\infty} |x|^j f_X(x) dx$

k-th moment is defined to be $E((X-\mu)^k)$

Properties of Expectations

1. If $X_1 \dots X_n$ RVs and $a_1 \dots a_n$ constants

$$E\left(\sum_i a_i X_i\right) = \sum_i a_i E(X_i)$$

Let X RV with mean μ. The variance of X (σ^2 or σ_X^2 or $V(X)$)

is $\sigma^2 = E(X-\mu)^2 = \int (x-\mu)^2 dF(x)$

Properties: 1) $V(X) = E(X^2) - \mu^2$

2. If a, b constants $V(aX+b) = a^2 V(X)$

3. If $X_1 \dots X_n$ independent and $a_1 \dots a_n$ constants then:

$$V\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 V(X_i)$$

Let X and Y RVs with means μ_X and μ_Y and s.d σ_X, σ_Y

covariance: $Cov(X,Y) = E((X-\mu_X)(Y-\mu_Y))$

correlation $\rho = \rho_{X,Y} = \rho(X,Y) = \dfrac{Cov(X,Y)}{\sigma_X \sigma_Y}$

Covariance satisfies. $Cov(X,Y) = E(XY) - E(X)E(Y)$

correlation — // — $-1 \leq \rho(X,Y) \leq 1$

* If X, Y independent $Cov(X,Y) = \rho = 0$

$V(X+Y) = V(X) + V(Y) + 2\,Cov(X,Y)$

$V(X-Y) = V(X) + V(Y) - 2\,Cov(X,Y)$

Variance - Covariance matrix.

Conditional Expectation

X, Y: RVs, the conditional expectation of X given $Y=y$ is:

$$E(X|Y=y) = \begin{cases} \sum x\, f_{X|Y}(x|y)\, dx & \text{discrete} \\ \int x\, f_{X|Y}(x|y)\, dx & \text{continuous} \end{cases}$$

If $r(x,y)$ is a function of x and y then:

$$E(r(X,Y)|Y=y) = \begin{cases} \sum r(x,y)\, f_{X|Y}(x|y)\, dx & \text{discrete} \\ \int r(x,y)\, f_{X|Y}(x|y)\, dx & \text{continuous} \end{cases}$$

Ex. $X \sim Unif(0,1)$ After we observe $X=x$ we draw $Y|X=x \sim Unif(x,1)$

intuitively we expect: $E(Y|X=x) = (1+x)/2$ In fact:

$f_{Y|X}(y|x) = 1/(1-x)$ for $x<y<1$ and

$E(Y|X=x) = \int_x^1 y\, f_{Y|X}(y|x)\, dy = \frac{1}{1-x}\int_x^1 y\, dy = \frac{1+x}{2}$ thus $E(Y|X=x) =$

$E(Y|X) = (1+X)/2$ is a RV whose value is the number $(1+x)/2$ once $X=x$ is observed

The Rule of Iterated Expectations: for RV X and Y we have:

$E[E(Y|X)] = E(Y)$ and $E[E(X|Y)] = E(X)$

for any function $r(x,y)$ $E[E(r(X,Y)|X)] = E(r(X,Y))$

compute $E(Y)$? we know $E(Y|X) = (1+X)/2$

$\Rightarrow E(Y) = EE(Y|X) = E\left(\frac{(1+X)}{2}\right) = \frac{(1+E(X))}{2} = \frac{1+1/2}{2} = 3/4$

Conditional Variance

$V(Y|X=x) = \int (y-\mu(x))^2 f(y|x)\, dy$, $\mu(x) = E(Y|X=x)$

Moment Generating Functions: (MGF) or Laplace transformation of X is:

$\psi_X(t) = E(e^{tX}) = \int e^{tx}\, dF(x)$ t = real number

Properties 1) If $Y = aX + b$ $\Rightarrow \psi_Y(t) = e^{bt}\cdot \psi_X(at)$

2) If $X_1 \dots X_n$ independent and $Y = \sum_i X_i$ then $\psi_Y(t) = \prod_i \psi_i(t)$, where ψ_i is the MGF of X_i

Moment Generating Function For Some Common Distributions.

Distribution	MGF $\psi(t)$
Bernoulli (p)	$pe^t + (1-p)$
Binomial (n,p)	$(p\cdot e^t + (1-p))^n$
Poisson (λ)	$e^{\lambda(e^t-1)}$
Normal (μ,σ)	$\exp\{\mu t + \frac{\sigma^2 t^2}{2}\}$
Gamma (α,β)	$\left(\frac{1}{1-\beta t}\right)^\alpha$ for $t < 1/\beta$

$E(X)$ first moment
$Var(X)$ - second moment

Chapter 4 - AOS: Inequalities → useful for bounding quantities that are otherwise difficult to compute. For Bernoulli vars:

1. Theorem 4.1: Markov's inequality: Let X be a non-negative and suppose that $E(X)$ exists. For any $t>0$ $P(X>t) \le \frac{E(X)}{t}$ You need to know the first moment (you know only the expected value)

Proof: Since $X \ge 0$ $E(X) = \int_0^\infty x\, f(x)\, dx = \int_0^t x f(x)\, dx + \int_t^\infty x f(x)\, dx$
$\ge \int_t^\infty x f(x)\, dx \ge t \int_t^\infty f(x)\, dx = t \cdot P(X>t)$

Markov's inequality is an application of the Chebyshev

2. Theorem 4.2: Chebyshev's inequality: Let $\mu = E(X)$ and $\sigma^2 = V(X)$. Then
$P(|X-\mu| \ge t) \le \frac{\sigma^2}{t^2}$ ($t = \sigma k$) and $P(|Z| \ge k) \le \frac{1}{k^2}$, where $Z = (X-\mu)/\sigma$
In particular, $P(|Z|>2) \le 1/4$ and $P(|Z|>3) \le 1/9$ You need to know the second moment
Proof: (use Markov's inequality) symmetrized non-negative → power of 2 $\mathbb{P}(|X-\mu| \ge t) = P(|X-\mu|^2 \ge t^2) \le$
$\frac{E(X-\mu)^2}{t^2} = \frac{\sigma^2}{t^2}$, to proof second part: $t = k \cdot \sigma$

3. Theorem 4.4: Hoeffding's Inequality: Let $Y_1, \ldots, Y_n$ be independent observations such that $E(Y_i) = 0$ and $a_i \le Y_i \le b_i$. Let $\varepsilon > 0$. Then for any $t>0$, $P\left(\sum_{i=1}^n Y_i \ge \varepsilon\right) \le e^{-t\varepsilon} \prod_{i=1}^n e^{t^2(b_i-a_i)^2/8}$

Theorem 4.5: Let $X_1, \ldots, X_n \sim$ Bernoulli(p). Then, for any $\varepsilon>0$
$\mathbb{P}(|\bar{X}_n - p| > \varepsilon) \le 2e^{-2n\varepsilon^2}$ where $\bar{X} = n^{-1}\sum_{i=1}^n X_i$

eg. Let $X_1, \ldots, X_n \sim$ Bernoulli(p). Let $n=100$, $\varepsilon = 0{,}2$. Using Chebyshev's inequality yielded $\mathbb{P}(|\bar{X}_n - p| > \varepsilon) \le 0.0625$. According to Hoeffding's inequality $\mathbb{P}(|\bar{X}_n - p| > 0{,}2) \le 2\, e^{-2 \cdot 100 \cdot 0{,}2^2} = 0.00067$ (much smaller)

Fix $\alpha > 0$ and let $\varepsilon_n = \sqrt{\frac{1}{2n} \log\left(\frac{2}{\alpha}\right)}$. By Hoeffding's inequality:
$\mathbb{P}(|\bar{X}_n - p| > \varepsilon_n) \le 2e^{-2n\varepsilon_n^2} = \alpha$

Let $C = (\bar{X}_n - \varepsilon_n, \bar{X}_n + \varepsilon_n)$. Then $P(p \notin C) = P(|\bar{X}_n - p| > \varepsilon_n) \le \alpha$
$P(p \in C) \ge 1-\alpha$ → this interval C traps the true parameter p with prob $1-\alpha$ so C is a $1-\alpha$ confidence interval.

For normal RV:

Theorem: Mill's inequality: Let $Z \sim N(0,1)$ then
$\mathbb{P}(|Z| > t) \le \sqrt{\frac{2}{\pi}} \frac{e^{-t^2/2}}{t}$

Inequalities for Expectations

1) Theorem 4.8: Cauchy - Schwartz inequality (cauchy - schwarz): If X and Y have finite variances then $E|XY| \le \sqrt{E(X^2) \cdot E(Y^2)}$
$|\vec{x}|\,|\vec{y}| \le \|\vec{x}\|\,\|\vec{y}\|$
the only way $|\vec{x}|\,|\vec{y}| = \|\vec{x}\|\,\|\vec{y}\|$ when one is a multiple scalar of the other $\vec{x} = c\,\vec{y}$

* A function g is convex if for each x, y and each $a \in [0,1]$
$g(ax + (1-a)y) \le a\,g(x) + (1-a)\,g(y)$

If g twice differentiable and $g''(x) \geq 0$ for all $x \Rightarrow g$: convex

convex then g lies above any line that touches g at some point (tangent)

* A function is concave if $-g$ is convex.

Convex functions: x^2, e^x , Concave functions: $-x^2$, $\log(x)$

Theorem 4.9 Jensen's Inequality: If g is convex then $\mathbb{E}g(X) \geq g(\mathbb{E}X)$

If g is concave then $\mathbb{E}g(X) \leq g(\mathbb{E}X)$

Proof: Let $L(x) = a + bx$ line tangent to $g(x)$ at the point $\mathbb{E}[X]$ Since

g convex $\rightarrow$ lies above the line $L(x)$.

$\mathbb{E}g(X) \geq \mathbb{E}L(X) = \mathbb{E}(a+bX) = a + b\,\mathbb{E}(X) = L(\mathbb{E}(X)) = g(\mathbb{E}X)$

* From Jensen's inequality we see $\mathbb{E}(X^2) \geq (\mathbb{E}X)^2$ if X positive,

then $\mathbb{E}(1/X) \geq 1/\mathbb{E}(X)$ $\Rightarrow$ Since $\log$: concave $\mathbb{E}(\log X) \leq \log \mathbb{E}(X)$

Chapter 5: (AoS) Convergence of RVs:

Behaviour of sequences of RVs: Large sample theory, limit theory and asymptotic theory. What happen's if we gather more data?

A sequence of real numbers x_n converges to a limit x if for every $\epsilon > 0$, $|x_n - x| < \epsilon$ for all large n. Suppose $x_n = x$ for all n. Then $\lim_{n \to \infty} x_n = x$.

Suppose $X_1, X_2, \dots$ sequence of RV independent and each $\sim N(0,1)$.

X_n "converges" to $X \sim N(0,1) \rightsquigarrow$ this not quite right since $P(X_n = X) = 0$ for all n (two continuous RVs are equal with prob. zero)

1) The Law of Large Numbers says that the sample average $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ converges in probability to the expectation $\mu = \mathbb{E}(X_i)$. This means that $\bar{X}_n$ "close to μ with high prob"

2) The Central Limit Theorem says that $\sqrt{n}(\bar{X}_n - \mu)$ converges in distribution to a Normal distribution. This means that the average has approximately a Normal distr for large n.

Types of Convergence: (2 main types). Let $X_1, X_2, \dots$ be a sequence of RV and let X be another random variable. Let F_n denote the CDF of X_n and let F denote the CDF of X.

1. X_n converges to X in probability written $X_n \xrightarrow{P} X$ if for every $\epsilon > 0$: $\mathbb{P}(|X_n - X| > \epsilon) \to 0$ as $n \to \infty$

2. X_n converges to X in distribution written $X_n \rightsquigarrow X$ if $\lim_{n \to \infty} F_n(t) = F(t)$ at all t for which F continuous

X_n converges to X in quadratic mean (convergence in L_2), written $X_n \xrightarrow{qm} X$ if $\mathbb{E}(X_n - X)^2 \to 0$ as $n \to \infty$

Ex. Let $X_n \sim N(0, 1/n)$. Let F - distribution function for a point mass at 0. Note $\sqrt{n} \cdot X_n \sim N(0,1)$. Z - standard normal RV. For $t<0$: $F_n(t) = \mathbb{P}(X_n < t) = \mathbb{P}(\sqrt{n} X_n < \sqrt{n} t) = P(Z < \sqrt{n} \cdot t) \to 0$ since $\sqrt{n} t \to -\infty$

For $t>0$. $F_n(t) = P(X_n < t) = P(\sqrt{n} X_n < \sqrt{n} t) = P(Z < \sqrt{n} t) \to 1$ since $\sqrt{n} t \to \infty$. Hence $F_n(t) \to F(t)$ for all $t \neq 0$ and so $X_n \rightsquigarrow 0$.

$F_n(0) = 1/2 \neq F(1/2) = 1 \Rightarrow$ convergence fails at $t=0$. That it doesn't matter because is a continuity point of F and definition of convergence in distribution requires convergence at continuity points. Now consider convergence in probability. For any $\varepsilon > 0$ using Markov's inequality,

$\mathbb{P}(|X_n| > \varepsilon) = \mathbb{P}(|X_n|^2 > \varepsilon^2) \le \frac{E[X_n^2]}{\varepsilon^2} = \frac{\frac{1}{n}}{\varepsilon^2} \to 0$ as $n \to \infty$

$X_n \xrightarrow{P} 0$

Theorem 5.4. The following relationships hold:

(a) $X_n \xrightarrow{qm} X$ implies that $X_n \xrightarrow{P} X$

(b) $X_n \xrightarrow{P} X$ implies that $X_n \rightsquigarrow X$

} Proofs page 74

(c) If $X_n \rightsquigarrow X$ and if $\mathbb{P}(X=c) = 1$ for some real number c, then $X_n \xrightarrow{P} X$

* Inverse implication don't hold except the special case in c).

Theorem 5.5: Let X_n, X, Y_n, Y. RVs. Let g continuous function

(a) If $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$ then $X_n + Y_n \xrightarrow{P} X + Y$

(b) If $X_n \xrightarrow{qm} X$ and $Y_n \xrightarrow{qm} Y$ then $X_n + Y_n \xrightarrow{qm} X + Y$

(c) If $X_n \rightsquigarrow X$ and $Y_n \rightsquigarrow c$ then $X_n + Y_n \rightsquigarrow X + c$

(d) If $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$ then $X_n Y_n \xrightarrow{P} XY$

(e) If $X_n \rightsquigarrow X$ and $Y_n \rightsquigarrow c$ then $X_n Y_n \rightsquigarrow cX$

(f) If $X_n \xrightarrow{P} X$ then $g(X_n) \xrightarrow{P} g(X)$

(g) If $X_n \rightsquigarrow X$ then $g(X_n) \rightsquigarrow g(X)$

Parts c & e are known as Slutzky's theorem. Note $X_n \rightsquigarrow X$ and $Y_n \rightsquigarrow Y$ doesn't in general imply $X_n + Y_n \rightsquigarrow X + Y$

The Law of Large Numbers: the mean of a large sample is close to the mean of the distribution. Let $X_1, X_2 \ldots$ be an IID sample. Let $\mu = \mathbb{E}(X_1)$ and $\sigma^2 = \mathbb{V}(X_1)$. $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$ $E(\bar{X}_n) = \mu$, $V(\bar{X}_n) = \frac{\sigma^2}{n}$

The Weak Law of Large Numbers (WLLN): If $X_1, \ldots X_n$ IID then $\bar{X}_n \xrightarrow{P} \mu$

↳ The distribution of $\bar{X}_n$ becomes more concentrated around μ as n gets large

Ex. Flip a coin. Le X_i : outcome of a single toss (0 or 1) Hence p $= E(X_i)$ After n times $\bar{X}_n$ According to the Law of Large Numbers $\bar{X}_n$ to p in probability. This doesn't mean that $\bar{X}_n$ will be numerically equal to p means when n = large the distribution of $\bar{X}_n$ is tightly concentrated around p

Let $p = 1/2$. How large should be n so that $P(0.4 \le \bar{X}_n \le 0.6) \ge 0.7$

first, $E(\bar{X}_n) = p = 1/2$ and $V(\bar{X}_n) = \sigma^2/n = p(1-p)/n = 1/4n$

From Chebyshev's inequality: $\mathbb{P}(0.4 \le \bar{X}_n \le 0.6) = P(|\bar{X}_n - \mu| \le 0.1)$

$= 1 - P(|\bar{X}_n - \mu| > 0.1) \ge 1 - \frac{1}{4n(0.1)^2} = 1 - \frac{25}{n}$

$1 - \frac{25}{n} \ge 0.7 \Rightarrow n = 84$

The Central Limit Theorem (CLT): Let $X_1, \dots X_n$ be IID with mean μ and variance σ^2. Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ then $Z_n = \frac{\bar{X}_n - \mu}{\sqrt{V(\bar{X}_n)}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \rightsquigarrow Z$

where $Z \sim N(0,1)$. In other words, $\lim_{n\to\infty} P(Z_n \le z) = \Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$

Ex. Suppose the number of errors per computer program has a Poisson distribution ~~of Z_n converging to~~ with mean 5 we get 125 programs. Let $X_1 \dots X_{125}$ be number of errors in the programs. We want to approximately $P(\bar{X}_n < 5.5)$. Let $\mu = E(X_1) = \lambda = 5$ and $\sigma^2 = V(X_1) = \lambda = 5$.

$\mathbb{P}(\bar{X}_n < 5.5) = P\left(\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} < \frac{\sqrt{n}(5.5 - \mu)}{\sigma}\right) \approx P(Z < 2.5) = 0.9938$.

* We rarely know σ → we can estimate σ^2 from $X_1, \dots X_n$ by

$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ if replace σ with S_n → CLT still holds.

Theorem 5.10: Assume the same conditions as the CLT. Then:

$\frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} \rightsquigarrow N(0,1)$

Theorem 5.11: The Berry-Esséen Inequality: Suppose that $E|X_1|^3 < \infty$. Then

$\sup_z |P(Z_n \le z) - \Phi(z)| \le \frac{33}{4} \frac{E|X_1 - \mu|^3}{\sqrt{n}\,\sigma^3}$

Theorem 5.12: Multivariate central limit theorem. Let $X_1, \dots X_n$ be IID random vectors where

$X_i = \begin{pmatrix} X_{1i} \\ X_{2i} \\ \vdots \\ X_{ki} \end{pmatrix}$ with mean $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{pmatrix} = \begin{pmatrix} E(X_{1i}) \\ E(X_{2i}) \\ \vdots \\ E(X_{ki}) \end{pmatrix}$ and variance matrix Σ.

Let $\bar{X} = \begin{pmatrix} \bar{X_1} \\ \bar{X_2} \\ \vdots \\ \bar{X_k} \end{pmatrix}$ where $\bar{X_j} = \frac{1}{n}\sum_{i=1}^{n} X_{ji}$ Then. $\sqrt{n}(\bar{X} - \mu) \rightsquigarrow N(0, \Sigma)$

The Delta Method: If Y_n has a limiting Normal distribution then the delta method allows us to find the limiting distribution of $g(Y_n)$ where g any smooth function. Suppose that $\frac{\sqrt{n}(Y_n - \mu)}{\sigma} \rightsquigarrow N(0,1)$ and that g is a differentiable function such that $g'(\mu) \neq 0$. Then

$$\frac{\sqrt{n}(g(Y_n) - g(\mu))}{|g'(\mu)|\sigma} \rightsquigarrow N(0,1)$$ In other words:

$Y_n \approx N(\mu, \frac{\sigma^2}{n})$ implies that $g(Y_n) \approx N(g(\mu), (g'(\mu))^2 \cdot \frac{\sigma^2}{n})$

Ex. Let $X_1 \dots X_n$ be IID with finite mean μ and finite variance σ^2. By the central limit theorem $\sqrt{n}(\bar{X_n} - \mu)/\sigma \rightsquigarrow N(0,1)$. Let $W_n = e^{\bar{X}_n}$

Thus $W_n = g(\bar{X_n})$. where $g(s) = e^s$. Since $g'(s) = e^s$ the delta method implies that $W_n \approx N(e^{\mu}, e^{2\mu} \cdot \sigma^2/n)$

Theorem 5.15: The Multivariate Delta Method. Suppose that $Y_n = (Y_{n1} \dots Y_{nk})$ is a sequence of Random vectors such that $\sqrt{n}(Y_n - \mu) \rightsquigarrow N(0, \Sigma)$

Let $g: R^k \to \mathbb{R}$ and let $\nabla g(y) = \begin{pmatrix} \frac{\partial g}{\partial y_1} \\ \vdots \\ \frac{\partial g}{\partial y_k} \end{pmatrix}$ Let ∇_μ denote $\nabla g(y)$ evaluated at $y = \mu$ and assume that the elements of ∇_μ are nonzero

Then $\sqrt{n}(g(Y_n) - g(\mu)) \rightsquigarrow N(0, \nabla_\mu^T \Sigma \nabla_\mu)$

Ex. Let $\begin{pmatrix} X_{11} \\ X_{21} \end{pmatrix}, \begin{pmatrix} X_{12} \\ X_{22} \end{pmatrix} \dots \begin{pmatrix} X_{1n} \\ X_{2n} \end{pmatrix}$ be IID random vectors with mean $\mu = (\mu_1, \mu_2)^T$ and variance Σ. Let $\bar{X_1} = \frac{1}{n}\sum_{i=1}^{n} X_{1i}$, $\bar{X_2} = \frac{1}{n}\sum_{i=1}^{n} X_{2i}$ and define $Y_n = \bar{X_1}\bar{X_2}$. Thus, $Y_n = g(\bar{X_1}, \bar{X_2})$ where $g(s_1, s_2) = s_1 s_2$. By the CLT:

$\sqrt{n}\begin{pmatrix} \bar{X_1} - \mu_1 \\ \bar{X_2} - \mu_2 \end{pmatrix} \rightsquigarrow N(0, \Sigma)$ Now: $\nabla g(s) = \begin{pmatrix} \frac{\partial g}{\partial s_1} \\ \frac{\partial g}{\partial s_2} \end{pmatrix} = \begin{pmatrix} s_2 \\ s_1 \end{pmatrix}$ and so

$$\nabla_\mu^T \Sigma \nabla_\mu = (\mu_2 \ \ \mu_1)\begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}\begin{pmatrix} \mu_2 \\ \mu_1 \end{pmatrix} = \mu_2^2 \cdot \sigma_{11} + 2\mu_1 \cdot \mu_2 \cdot \sigma_{12} + \mu_1^2 \cdot \sigma_{22}$$

Therefore, $\sqrt{n}(\bar{X_1}\bar{X_2} - \mu_1\mu_2) \rightsquigarrow N(0, \mu_2^2 \cdot \sigma_{11} + 2\mu_1 \cdot \mu_2 \cdot \sigma_{12} + \mu_1^2 \cdot \sigma_{22})$

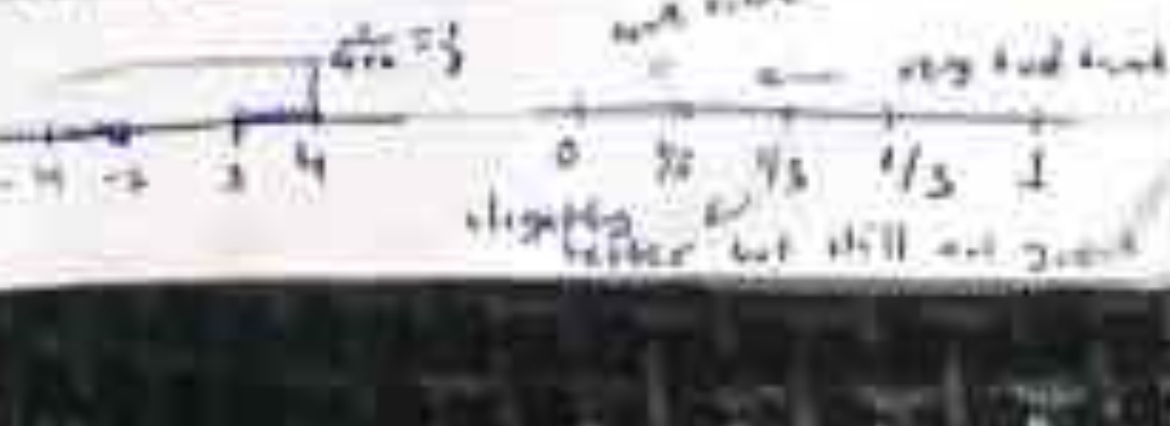
Markov's Inequality: let's say a RV, is non-negative, the probability that the RV exceeds a particular number (a) is bounded by the ratio $\frac{E[X]}{a}$. If $E(X)$ very small $\Rightarrow$ the probability of exceeding that value of a will also be small, if a very large $\Rightarrow$ prob of exceeding that large value drops.

ex1) X exponential $(\lambda = 1)$. $P(X \geq a) \leq \frac{E[X]}{a} = \frac{1}{a}$

ex2) X uniform $[-4, 4]$ $P(X \geq 3) \leq \frac{E[X]}{3}$

non-negative so we cannot apply it directly (violated)

$P(X \geq 3) \leq P(|X| \geq 3) \leq \frac{E[|X|]}{3} = \frac{2}{3}$ Because of symmetry



Jensen's Inequality
e.g. convex

2) The Chebyshev's Inequality: RV x with finite mean μ and variance σ^2 (c = positive)

↳ "If the variance is small, then X is unlikely to be too far from the mean"

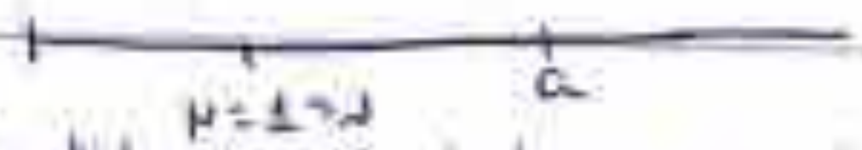
$P(|X-\mu| \geq c) \leq \frac{\sigma^2}{c^2}$ if the variance is small the prob of falling far from the mean will be small

~~If c large ⇒ prob falls~~ at $Ex.1$: $P(|X-\mu| \geq k\sigma) \leq \frac{\sigma^2}{(k\sigma)^2} = \frac{1}{k^2}$

if $k=3$: the probability that you fall 3 standard deviations from the mean or more will be less than or equal to $1/9$ (this is true for all distributions)

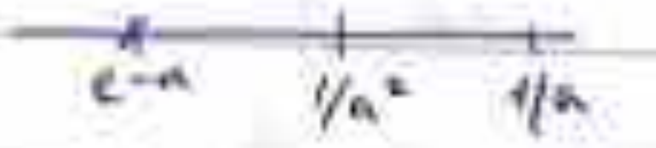
Ex.2: X: $\exp(\lambda=1)$ $P(X \geq a) \leq \frac{1}{a}$ ⇝ Markov's , real value $= e^{-a}$

$P(X \geq a) = P(X-1 \geq a-1) \leq P(|X-1| \geq a-1)$



using Chebyshev's: $\leq \frac{1}{(a-1)^2}$ ⇝ if a: large number this quantity will behave as $\frac{1}{a^2}$

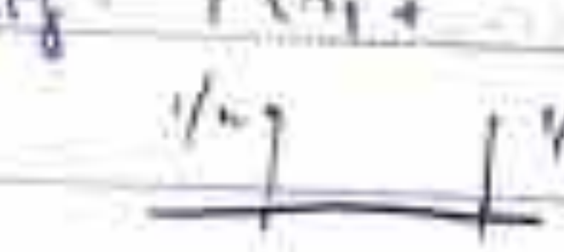
for large a: the Chebyshev bound



more informative & strong compared to Markov's because it exploits more info mean & variance

3) Hoeffding's Inequality: $P(X_1 + \dots + X_n \geq na)$ X_i = iid the RV are equally likely to take the values -1 and 1 with prob $=\frac{1}{2}$

$Y = X_1 + \dots + X_n$



$E[Y] = 0$, $Var(Y) = n$

$E[X_i] = 0$ bc symmetric

What we know about the RV Y? Using the CLT ⇒ Y

$Var(X_i) = 1$

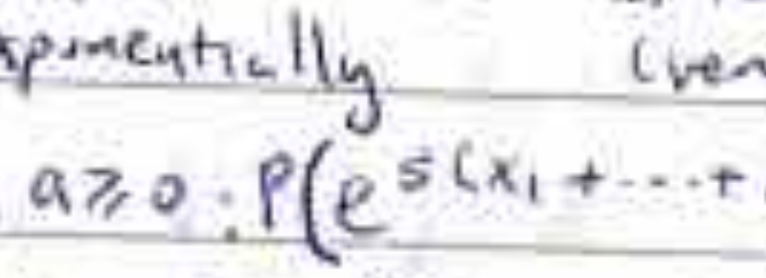
has approximately Normal distr ($\mu=0$): $P\left(\frac{Y}{\sqrt{n}} \geq a\right) \approx 1 - Q(a)$ (std normal CDF)

using Chebyshev's: $P(Y \geq na) \leq \frac{Var(Y)}{n^2a^2} = \frac{1}{na^2}$ → this prob goes to zero as fast as $\frac{1}{n}$ goes to 0 (very conservative)

Hoeffding's Inequality says that this prob falls exponentially with n: $P(X_1 + \dots + X_n \geq na)$. let $s>0$; $a>0$: $P(e^{s(X_1+\dots+X_n)} \geq e^{s\,na})$

using Markov's inequality: $\leq E[e^{s(X_1+\dots+X_n)}]/e^{s\,na} = E[e^{sX_1}\cdots e^{sX_n}]/e^{s\,na}$ since X_i independent

$= \underbrace{E[e^{sX_1}]\cdots E[e^{sX_n}]}_{\text{since identically distributed → the term is the same}} = [E[e^{sX_1}]]^n/e^{s\,na} = \left(\frac{E[e^{sX_1}]}{e^{sa}}\right)^n$ (*) $= \rho^n$: if $\rho<1$: this bound falls exponentially with n and this prob will fall the same. You can choose s to be as informative for the bound: $\left[\frac{\frac{1}{2}(e^{s}+e^{-s})}{e^{sa}}\right]$ where -1 & 1



you need an exponentially decaying bound

at 0 the denominator (e^{sa}) will be larger than the numerator ⇒ the fraction will be less than 1 ⇒ exponentially decaying bound

The denominator has a positive derivative at 0

For "small" s: $\rho<1$. If $s=a$ using the Hoeffding's: $\leq e^{-na^2/2}$

we Taylor's series: $e^{s} = 1 + s + \frac{s^2}{2!} + \dots = \sum_{i=0}^{\infty} \frac{s^i}{i!}$

$\frac{1}{2}(e^{s}+e^{-s}) = \frac{1}{2}\left(1+s+\frac{s^2}{2!}+\frac{s^3}{3!}+\dots\right) + \frac{1}{2}\left(1-s+\frac{s^2}{2!}-\frac{s^3}{3!}\dots\right) = \sum_{i=0}^{\infty}\frac{s^{2i}}{(2i)!} \leq \sum_{i=0}^{\infty}\frac{s^{2i}}{i!\,2^i}$

$(2i)! = 1\cdot2\cdot3\cdots i\cdot(i+1)(\dots)(2i) \geq i!\,2^i$

$= \sum_{i=0}^{\infty}\frac{(s^2/2)^i}{i!} = e^{s^2/2}$ (*) $\leq \left(\frac{e^{s^2/2}}{e^{sa}}\right)^n = \left(e^{s^2/2 - sa}\right)^n \overset{\text{if } s=a}{=} e^{-na^2/2}$

Jensen's Inequality: Comparing $E[\overset{\text{function}}{g}(\overset{\text{RV}}{X})]$ to $g(E[X])$ if g: linear these two are equal s

g: convex if $0 \le p \le 1$ then $g(px+(1-p)y) \le pg(x)+(1-p)g(y)$

or $g''(x) \geq 0$

or $g(x) \geq g(c) + g'(c)(x-c)$ c: any number

$$g(x) \geq \underbrace{g(E[X])}_{\text{number}} + \underbrace{g'(E[X])}_{\text{number}}\underbrace{(X-E[X])}_{\text{expected value of}\ = \text{number} - \text{number}}$$

$Eg(x) \geq g(E[X]) + 0$

eg. $g(x)=x^4$ $(E[X])^4 \le E[X^4]$ $g(x) = -\log(x)$: $-\log(E[X]) \le E[-\log(x)]$

$\log(E[X]) \geq E[\log(x)]$

Probability Distributions

1. Binomial Distribution: $P(x) = \binom{n}{x} \cdot p^x q^{n-x}$ getting x successes in n tries, p: success, q = fail $= 1-p$

$p(x) = \frac{n!}{(n-x)!\,x!} \cdot p^x \cdot q^{n-x}$ $\quad \mu = n \cdot p \quad \sigma = \sqrt{npq}$

2. Geometric Distribution: The probability that the first success will happen on the n^{th} time - the n^{th} event will succeed

$P(x > n) = q^n \quad P(x \geq n) = q^{n-1}$

~~P(x)~~ $P(x = n) = q^{n-1} p$

$P(x \leq n) = 1 - q^n \quad P(x < n) = 1 - q^{n-1}$

$\sigma^2 = \frac{1}{p}\left(\frac{1}{p} - 1\right) = m(m-1)$

$m = \frac{1}{p}$

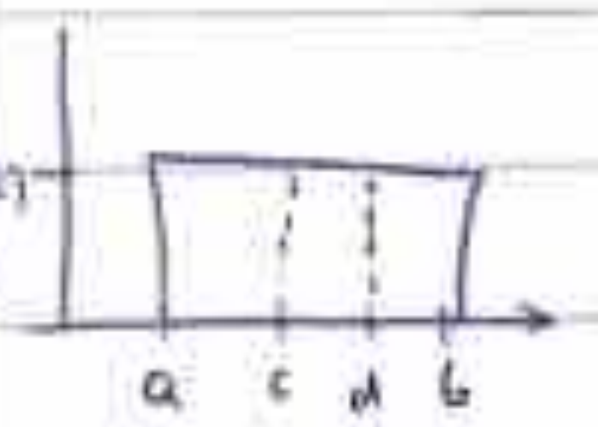
3. Poisson Distribution: Number of times an event occurs in a given interval.

$P(X = n) = \frac{m^n e^{-m}}{n!}$ or $\frac{\lambda^n \cdot e^{-\lambda}}{n!}$ $\quad m = \lambda = n \cdot p \quad \sigma^2 = np = \lambda$

$P(x > n) = 1 - e^{-m}\left[\sum_{x=0}^{n} \frac{m^n}{n!}\right]$

$P(x \leq n) = e^{-m}\left[\sum_{x=0}^{n} \frac{m^n}{n!}\right]$

4. Uniform Distribution

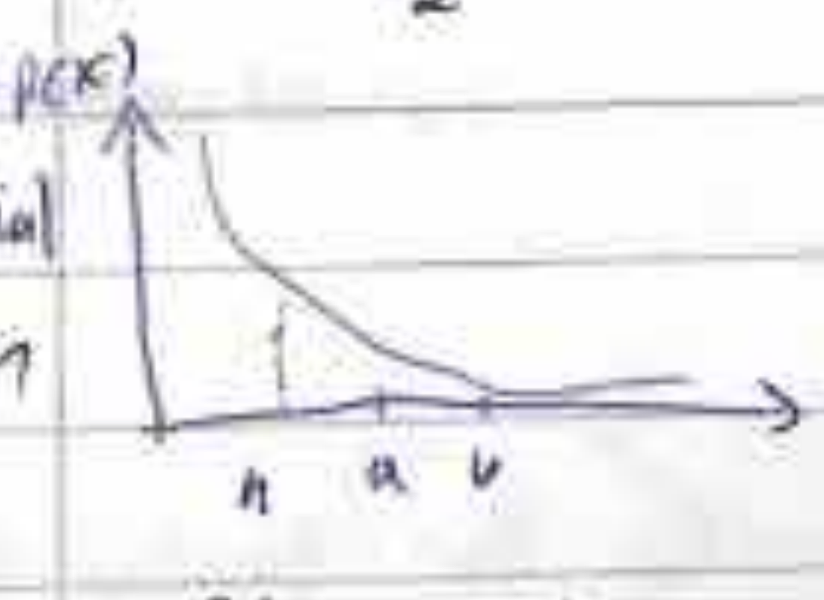


Area = 1 = Base x Height $= (b-a) \cdot f(x) \iff f(x) = \frac{1}{b-a}$

$P(a \leq x \leq c) = B \times H = \frac{c-a}{b-a}$

$m = \frac{a+b}{2} \quad \sigma = \frac{b-a}{\sqrt{12}}$

5. Exponential Distribution



λ: rate parameter

$m = \frac{1}{\lambda} \quad \sigma^2 = \frac{1}{\lambda^2}$

Area $= \int_a^b f(x)\,dx$

left: $A_L = P(x \leq n) = 1 - e^{-\lambda n}$

$f(x) = \lambda \cdot e^{-\lambda x}$

$A_R = P(x \geq n) = \int_n^\infty \lambda e^{-\lambda x} dx = e^{-\lambda n}$

$\int_0^n \lambda \cdot e^{-\lambda x} dx = 1 - e^{-\lambda n}$

6. Bernoulli Distribution: One trial with two possible discrete outcomes

$p(x) = \begin{cases} p, & \text{for } x = 1 \\ 1-p, & \text{for } x = 0 \end{cases}$ $\quad P(x) = p^x (1-p)^{1-x} \quad m = p, \quad \sigma^2 = p(1-p)$

Mutually exclusive events: they cannot happen the same time eg. get head & tails by flipping one coin $\quad P(A \cap B) = 0$

Independent: the occurrence of one event doesn't affect the occurrence of another

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

eg. Two independent fair coin tosses. H_1: ~~proba~~ 1st toss is H, H_2: 2nd toss is H

$P(H_1) = P(H_2) = 1/2$ $\quad$ Let C: 2 tosses with same result. $\{HH\}, \{TT\} = \frac{1}{2}$

Check for Pairwise independence: $P(H_1, C) = P(H_1 \cap H_2) = 1/4 \quad P(H_1)\,P(C) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$ $\}$ pairwise independent

(*) if mutually independent → always pairwise independent (not the other way round)

Student Passing Exam:

10 yes/no questions $\quad N \sim (10, 1/10)$ — known answers

guess the answer with equal probability $\quad Z \sim (10 - N, 1/2)$ # of correct answers

X

$Y = N + Z \quad$ Y: number of total correct answers

Deterministic threshold T for passing such that $Y \geq T$, $T \in \{0, 1, \dots, 10\}$

for each T compute the probability that the student knows less than 5 correct answers given that he passed $(N < 5)$

Check for independence: $P(H_1 \cap H_2 \cap C) = P(HH) = 1/4$

$P(H_1) \cdot P(H_2) \cdot P(C) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8}$ } they are not jointly independent

Ex.2 Let 4 balls and events. $A = \{1,2\}$, $B = \{1,3\}$ and $C = \{1,4\}$

To be pairwise independent should: $P(A \cap B) = P(\{1\})$ because $\quad P(\text{any ball is}) = \frac{1}{4}$

$A = \{1,2\}$ and $B = \{1,3\}$ $\quad P(\{1\}) = \frac{1}{4} \quad P(A) = \frac{2}{4} = \frac{1}{2}, \quad P(B) = \frac{2}{4} = \frac{1}{2}$

$P(A \cap B) = P(A)\, P(B) \Rightarrow$ Holds

same holds for $P(A \cap C) = P(A) \cdot P(C)$ and $P(C \cap B) = P(C) \cdot P(B)$

Joint Independence is a stronger condition: $P(A \cap B \cap C) = P(A) \cdot P(B) \cdot P(C)$

$P(A \cap B \cap C) = P(\{1\}) = \frac{1}{4}$

$P(A) \cdot P(B) \cdot P(C) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8}$ } $\neq \Rightarrow$ not jointly independent

Multivariate Distributions: Ex.1) Find $P(X_1 < X_2 < X_3)$ for X_1, X_2, X_3 with joint pdf

$f(x_1, x_2, x_3) = 1 \qquad 0 < x_1 < 1, \; 0 < x_2 < 1, \; 0 < x_3 < 1$

$P(X_1 < X_2 < X_3) = \int_0^1 \int_{x_1}^1 \int_{x_2}^1 1 \, dx_3 \, dx_2 \, dx_1 = \frac{1}{6}$

Ex.2) Find $F(x_1, x_2, x_3)$ for X_1, X_2, X_3 with joint pdf $f(x_1, x_2, x_3) = 6e^{-x_1 - 2x_2 - 3x_3}$

$x_1 > 0, \; x_2 > 0, \; x_3 > 0$

CDF → $F(x_1, x_2, x_3) = P(X_1 \leq x_1, X_2 \leq x_2, X_3 \leq x_3) = \int_0^{x_1} \int_0^{x_2} \int_0^{x_3} 6e^{-w_1 - 2w_2 - 3w_3} \, dw_3 \, dw_2 \, dw_1$

$= \int_0^{x_1} 6e^{-w_1} \, dw_1 \cdot \int_0^{x_2} 6e^{-2w_2} \, dw_2 \cdot \int_0^{x_3} 6e^{-3w_3} \, dw_3$

$= (1 - e^{-x_1})(1 - e^{-2x_2})(1 - e^{-3x_3})$

$$F(x_1, x_2, x_3) = \begin{cases} 0, & x_1 \leq 0, \; x_2 \leq 0, \; x_3 \leq 0 \\ (1 - e^{-x_1}) \cdot (1 - e^{-2x_2})(1 - e^{-3x_3}) & x_1 > 0, \; x_2 > 0, \; x_3 > 0 \end{cases}$$

Ex. Show that X_1, X_2, X_3 are pairwise independent but not mutually independent

joint ~~pdf~~ pmf: $f(x_1, x_2, x_3) = \frac{1}{4} \quad (x_1, x_2, x_3) \in \{(1,0,0), (0,1,0), (0,0,1), (1,1,1)\}$

Let consider X_1, X_2: steps: find $f(x_1, x_2)$, $f(x_1, x_2) = \frac{1}{4} \quad x_1 = 0,1 \; ; \; x_2 = 0,1$

marginal distr. $f_{X_1}(x_1) = \frac{1}{2}, \; x_1 = 0,1 \quad$ since $f(x_1, x_2) = f_{X_1}(x_1)\, f_{X_2}(x_2) \Rightarrow X_1, X_2$: pairwise independent

$f_{X_2}(x_2) = \frac{1}{2} \quad x_2 = 0,1$

Mutually independent? No, since $\frac{1}{4} \neq \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2}$

[Ch]apter 6: Models, Statistical Inference and Learning:

[Parametr]ic & Non-parametric Models: A statistical model $\mathfrak{F}$ is a set of distributions ([de]nsities or regression functions). A parametric model is a set $\mathfrak{F}$ that can be [pa]rameterised by a finite number of parameters.

If data come from Normal distribution then the model is:

$$\mathfrak{F} = \left\{ f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\},\ \mu \in \mathbb{R}, \sigma > 0 \right\} \quad (6.1)$$

↳ this is a two-parameter model → we have written the density as $f(x; \mu, \sigma)$ → x is a value of the RV and μ, σ: parameters [2-parameter model]

Parametric model takes the form: $\mathfrak{F} = \{ f(x, \theta)\ \theta \in \Theta \}$ where θ is an unknown parameter that takes values from the parameter space Θ. If θ is a vector but we are interested in one component of θ we call the remaining parameters nuisance parameters.

A nonparametric model is a set $\mathfrak{F}$ that cannot be parameterized by a finite number of parameters. eg $\mathfrak{F}_{all}$ = {all CDF's} is nonparametric.

Ex 1 (One-dimensional Parametric Estimation): Let $X_1 \dots X_n$ be independent Bernoulli(p) observations. The problem is to estimate the parameter p.

Ex 2 (Two-dimensional Parametric Estimation): Suppose that $X_1 \dots X_n \sim F$ and we assume that the PDF $f \in \mathfrak{F}$ where $\mathfrak{F}$ is given in (6.1) → there are two parameters μ and σ. The goal is to estimate the parameters from the data.

If we are only interested in estimating μ then μ is the parameter of interest and σ is a nuisance parameter.

Ex 3 Nonparametric estimation of functionals. Let $X_1 \dots X_n \sim F$. Suppose we want to estimate $\mu = E(X_1) = \int x\, dF(x)$ assuming only that μ exists. The mean μ may be thought of as a function of F called statistical functional: $\mu = T(F) = \int x\, dF(x)$

variance $T(F) = \int x^2 dF(x) - \left(\int x\, dF(x)\right)^2$

Ex 4 Suppose we observe pairs of data $(X_1, Y_1) \dots (X_n, Y_n)$. Perhaps X_i is the blood pressure of subject i and Y_i how long they live. X is called a predictor / regressor / feature / independent va[riable]. $r(x) = E(Y \mid X = x)$ regression function.

Y is the outcome, response, dependent var. The goal of predicting Y for a new patient based on their X value is prediction. If Y is discrete ⇒ classification. We need to estimate the function r → this is called regressi[on] or curve estimation. $Y = r(X) + \epsilon$ where $E(\epsilon) = 0$. Moreover $E(\epsilon) = E(E(\epsilon \mid X))$

let $\epsilon = Y - r(X)$ and hence $Y = Y + r(X) - r(X) = r(X) + \epsilon$

$E(E(Y - r(X)) \mid X) = E[E(Y \mid X) - r(X)] = E(r(X) - r(X)) = 0$

Point Estimation refers to providing a single "best guess" of some quantity.

$\hat{\theta}$ or $\hat{\theta}_n$: point estimate of θ → fixed unknown quantity and $\hat{\theta}$ depends on the data so $\hat{\theta}$ is random variable.

Let $X_1 \dots X_n$ be n iid data points from some distribution F.

→ A point estimator $\hat{\theta}_n$ of a parameter θ is some function of $X_1 \dots X_n$:

$\hat{\theta}_n = g(X_1 \dots X_n)$. The bias of an estimator is defined by

$\text{bias}(\hat{\theta}_n) = E_\theta(\hat{\theta}_n) - \theta$, $\hat{\theta}_n$ unbiased if $E_\theta(\hat{\theta}_n) = \theta$

* Requirement of an estimator: it should converge to the true parameter value as we collect more & more data.

* A point estimator $\hat{\theta}_n$ of a parameter θ is consistent if $\hat{\theta}_n \xrightarrow{P} \theta$

→ The distribution of $\hat{\theta}_n$: sampling distribution

→ standard error: $se = se(\hat{\theta}_n) = \sqrt{V(\hat{\theta}_n)}$ → estimation $\widehat{se}$

Ex 5 Let $X_1 \dots X_n \sim \text{Bernoulli}(p)$ and let $\hat{p}_n = n^{-1}\sum_i X_i$ Then $E(\hat{p}_n) = p$ so $\hat{p}_n$: unbiased. The $se = \sqrt{V(\hat{p}_n)} = \sqrt{p(1-p)/n} \Rightarrow \hat{se} = \sqrt{\hat{p}(1-\hat{p})/n}$

The quality of a point estimate is assessed by the $MSE = E_\theta(\hat{\theta}_n - \theta)^2$

$MSE = bias^2(\hat{\theta}_n) + V_\theta(\hat{\theta}_n)$

Theorem: If bias $\to 0$ and se $= 0$ as $n \to \infty$ then $\hat{\theta}_n$: consistent that $\hat{\theta}_n \xrightarrow{P} \theta$

An estimator is asymptotically Normal if $\frac{\hat{\theta}_n - \theta}{se} \rightsquigarrow N(0,1)$

Confidence sets: A $1-\alpha$ confidence interval for a parameter θ is an interval $C_n = (a, b)$ where $a = a(X_1 \dots X_n)$ are functions of the data st
$P_\theta(\theta \in C_n) \geq 1-\alpha$ for all $\theta \in \Theta$. In other words, (a,b) traps θ with probability $1-\alpha$: coverage of confidence interval.

* If θ : vector $\Rightarrow$ use of confidence set (such as a sphere/ellipse) instead of an interval.

Hypothesis Testing: Null hypothesis: check if the data provide sufficient evidence to reject the theory

Ex 6 Let $X_1 \dots X_n \sim \text{Bernoulli}(p)$ be n independent coin flips. Test if a coin is fair. Let H_0 hypothesis that the coin is fair and let H_1 : hypothesis that the coin is not fair. H_1 : alternative hypothesis. $H_0 : p = 1/2$ Vs $H_1 : p \neq 1/2$. It is reasonable to reject H_0 if $T = |\hat{p}_n - (1/2)|$ is large

$MSE = bias^2(\hat{\theta}_n) + V_\theta(\hat{\theta}_n)$

proof $E_\theta(\hat{\theta}_n - \theta)^2 = E_\theta(\hat{\theta}_n - \bar{\theta}_n + \bar{\theta}_n - \theta)^2$

$= E_\theta(\hat{\theta}_n - \bar{\theta}_n)^2 + 2(\bar{\theta}_n - \theta)E_\theta(\hat{\theta}_n - \bar{\theta}_n) + E_\theta(\bar{\theta}_n - \theta)^2$

$= (\bar{\theta}_n - \theta)^2 + E_\theta(\hat{\theta}_n - \bar{\theta}_n)^2$

$= bias^2(\hat{\theta}_n) + V(\hat{\theta}_n)$

where : $E_\theta(\hat{\theta}_n - \bar{\theta}_n) = \bar{\theta}_n - \bar{\theta}_n = 0$

$1-\alpha$: coverage of the Confidence Interval

CI : if you repeat the experiment over & over the interval will contain the parameter 95% of the time

[Chap]ter 9 Parametric Inference: parametric models $\mathfrak{F} = \{f(x;\theta) : \theta \in \Theta\}$

[We are] interested in a function $T(\theta)$ eg. $X \sim N(\mu, \sigma^2)$ then the parameter is $\theta = (\mu, \sigma)$
[the] goal is to estimate $\mu = T(\theta)$: parameter of interest, σ nuisance parameter
[can] be a complicated function of θ

[Ex] Let $X_1 \ldots X_n \sim$ Normal (μ, σ^2). the parameter is $\theta = (\mu, \sigma)$ and the parameter space is $\Theta = \{(\mu, \sigma) : \mu \in \mathbb{R}, \sigma > 0\}$ suppose that X_i is the outcome of a blood test and suppose we are interested in τ (the fraction of the population whose test score is larger ~~than~~ than 1. Let Z denote a standard Normal RV Then:

$$\tau = \mathbb{P}(X > 1) = 1 - P(X < 1) = 1 - P\left(\frac{X-\mu}{\sigma} < \frac{1-\mu}{\sigma}\right) = 1 - P\left(Z < \frac{1-\mu}{\sigma}\right) = 1 - \Phi\left(\frac{1-\mu}{\sigma}\right)$$

The parameter of interest is $\tau = T(\mu, \sigma) = 1 - \Phi((1-\mu)/\sigma)$

<u>Method of Moments</u> for generating parametric estimators (usually are not optimal but they are easy to compute)

Suppose, parameter $\theta = (\theta_1, \ldots \theta_k)$ has k components. For $1 \le j \le k$, the j^{th} moment

$$\alpha_j = \alpha_j(\theta) = \mathbb{E}_\theta(X^j) = \int x^j \, dF_\theta(x) \quad \text{and the } j\text{th sample moment: } \hat{\alpha}_j = \frac{1}{n}\sum_{i=1}^{n} X_i^j$$

The method of moments estimator $\hat{\theta}_n$ is defined to be the value of θ such that:

$$\alpha_1(\hat{\theta}_n) = \hat{\alpha}_1, \quad \alpha_2(\hat{\theta}_n) = \hat{\alpha}_2, \quad \ldots, \quad \alpha_k(\hat{\theta}_n) = \hat{\alpha}_k$$

Ex.1 Let $X_1, X_2 \ldots X_n \sim$ Bernoulli (p). Then $\alpha_1 = \mathbb{E}_p(X) = p$ and $\hat{\alpha}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i$

$$\Rightarrow \hat{p}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Ex.2 Let $X_1 \ldots X_n \sim$ Normal (μ, σ^2). Then $\alpha_1 = \mathbb{E}_\theta(X_1) = \mu$ and $\alpha_2 = \mathbb{E}_\theta(X_1^2) = \mathbb{V}(X_1) + (\mathbb{E}_\theta(X_1))^2 = \sigma^2 + \mu^2$

we need to solve 2 equations: $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $\hat{\sigma}^2 + \hat{\mu}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$

$$\Rightarrow \text{solution } \hat{\mu} = \bar{X}_n$$
$$\hat{\sigma}^2 = \frac{1}{n}\sum (X_i - \bar{X}_n)^2$$

Theorem: Let $\hat{\theta}_n$ denote the method of moments estimator. Under appropriate conditions on the model the following statements hold

1. The estimate $\hat{\theta}_n$ exists with prob tending to 1
2. The estimate is consistent $\hat{\theta}_n \xrightarrow{P} \theta$
3. The estimate is asymptotically Normal $\sqrt{n}(\hat{\theta}_n - \theta) \rightsquigarrow N(0, \Sigma)$

where $\Sigma = g\,\mathbb{E}_\theta(YY^T)g^T$, $Y = (X, X^2, \ldots X^k)^T$, $g = (g_1 \ldots g_k)$ and $g_j = \partial\alpha_j^{-1}(\theta)/\partial\theta$

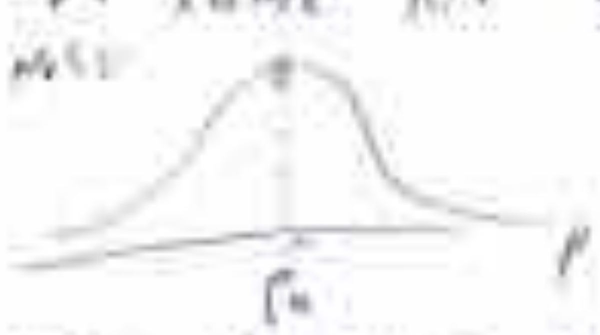
<u>Maximum Likelihood</u> method for estimating parameters in a parametric model. Let $X_1 \ldots X_n$ be IID with PDF $f(x;\theta)$. The Likelihood function is defined by

$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f(X_i;\theta)$. The Log-likelihood function is defined by $\ell_n(\theta) = \log \mathcal{L}_n(\theta)$

The likelihood function is the joint density of the data except that we treat it as a function of the parameter θ. Thus $\mathcal{L}_n : \Theta \to [0, \infty)$. This function is not a density function so $\mathcal{L}_n(\theta)$ doesn't integrate to 1 (w.r.t. θ)

The Maximum Likelihood Estimator [MLE] $\hat{\theta}_n$ is the value that max. $\mathcal{L}_n(\theta)$ (will be the same for $\ell_n(\theta)$) * If we multiply $\mathcal{L}_n(\theta)$ by any positive constant c (not depending on θ) then this will not change the MLE so we drop these constants in the function

MLE



Ex. Suppose $X_1, X_2 \ldots X_n \sim$ Bernoulli (p). The probability function is $f(x;p) = p^x(1-p)^{1-x}$ for $x = 0, 1$

$\mathcal{L}_n(p) = \prod_{i=1}^{n} f(X_i; p) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i} = p^S(1-p)^{n-S}$, $S = \sum X_i$

Hence $\ell_n(p) = S\log p + (n-S)\log(1-p)$ * take derivative and set to zero, MLE is $\hat{p}_n = S/n$

Ex. 2) Let $X_1 \dots X_n \sim N(\mu, \sigma^2)$. The parameter is $\vartheta = (\mu, \sigma)$ and the (ignoring some constants) is:

$$\mathcal{L}_n(\mu,\sigma) = \prod \frac{1}{\sigma}\exp\left\{-\frac{1}{2\sigma^2}(X_i-\mu)^2\right\} = \sigma^{-n}\exp\left\{-\frac{1}{2\sigma^2}\sum(X_i-\mu)^2\right\}$$

$$= \sigma^{-n}\exp\left\{-\frac{nS^2}{2\sigma^2}\right\}\exp\left\{\frac{n(\bar{X}-\mu)^2}{2\sigma^2}\right\} \quad \text{where } \bar{X} = n^{-1}\sum X_i, \; S^2 = n^{-1}\sum(X_i-\bar{X})^2$$

$$\sum(X_i-\mu)^2 = nS^2 + n(\bar{X}-\mu)^2 = \sum(X_i-\bar{X}+\bar{X}-\mu)^2$$

The log-likelihood is $\ell(\mu,\sigma) = -n\log\sigma - \frac{nS^2}{2\sigma^2} - \frac{n(\bar{X}-\mu)^2}{2\sigma^2}$

$\frac{\partial \ell(\mu,\sigma)}{\partial\mu} = 0$ and $\frac{\partial \ell(\mu,\sigma)}{\partial\sigma} = 0$, $\hat{\mu} = \bar{X}$ and $\hat{\sigma} = S$ global maximum of likelihood

Properties of MLE:

1) MLE consistent: $\hat{\vartheta}_n \xrightarrow{P} \vartheta_*$ where ϑ_* true value of parameter ϑ

2) MLE equivariant: if $\hat{\vartheta}_n$ is the MLE of ϑ then $g(\hat{\vartheta}_n)$ is the MLE of $g(\vartheta)$

3) MLE asymptotically Normal $(\hat{\vartheta}-\vartheta_*)/\hat{se} \rightsquigarrow N(0,1)$ also the estimated $\hat{se}$ can often be computed analytically

4) MLE asymptotically optimal/efficient: this means that among all well-behaved estimators the MLE has the smallest variance (at least for large samples)

5) MLE is approximately the Bayes estimator

* Only hold if the model satisfies certain regularity conditions (smoothness conditions) on $f(x;\vartheta)$

· Consistency of MLE: If f and g are PDF's define the Kullback-Leibler distance between them: $D(f,g) = \int f(x)\log\left(\frac{f(x)}{g(x)}\right)dx$. It can be shown that $D(f,g) \geq 0$ and $D(f,f) = 0$

· Equivariance of the MLE: Theorem: Let $\tau = g(\vartheta)$ be a function of ϑ. Let $\hat{\vartheta}_n$ be the MLE of ϑ. The $\hat{\tau}_n = g(\hat{\vartheta}_n)$ is the MLE of τ

Ex. Let $X_1 \dots X_n \sim N(\vartheta, 1)$. The MLE for ϑ is $\hat{\vartheta}_n = \bar{X}_n$. Let $\tau = e^{\vartheta}$ then the MLE for $\hat{\tau} = e^{\hat{\vartheta}} = e^{\bar{X}}$

· Asymptotic Normality: The score function is defined to be $s(X;\vartheta) = \frac{\partial \log f(X;\vartheta)}{\partial\vartheta}$

The Fisher information is $I_n(\vartheta) = V_\vartheta\left(\sum_{i}^{n} s(X_i;\vartheta)\right) = \sum^n V_\vartheta(s(X_i;\vartheta))$

Ex. Let $X_1 \dots X_n \sim \text{Poisson}(\lambda)$. Then $\hat{\lambda}_n = \bar{X}_n$ and some calculations show that $I_1(\lambda) = \frac{1}{\lambda}$

so $\hat{se} = \frac{1}{\sqrt{n I(\hat{\lambda}_n)}} = \sqrt{\frac{\hat{\lambda}_n}{n}}$ Approximate $1-\alpha$ CI for λ: $\hat{\lambda}_n \pm z_{\alpha/2}\sqrt{\hat{\lambda}_n/n}$

· Optimality: Suppose that $X_1 \dots X_n \sim N(\vartheta,\sigma^2)$. The MLE is $\hat{\vartheta}_n = \bar{X}_n$. Another reasonable estimator of ϑ is the sample median $\tilde{\vartheta}_n$. The MLE satisfies $\sqrt{n}(\hat{\vartheta}_n - \vartheta) \rightsquigarrow N(0,\sigma^2)$

The Delta Method: If $\tau = g(\vartheta)$ where g differentiable and $g'(\vartheta) \neq 0$ then $\frac{(\hat{\tau}_n - \tau)}{\hat{se}(\hat{\tau})} \rightsquigarrow N(0,1)$ where $\hat{\tau}_n = g(\hat{\vartheta}_n)$ and $\hat{se}(\hat{\tau}_n) = |g'(\hat{\vartheta})|\,\hat{se}(\hat{\vartheta}_n)$. Hence if

$C_n = (\hat{\tau}_n - z_{\alpha/2}\hat{se}(\hat{\tau}_n),\ \hat{\tau}_n + z_{\alpha/2}\hat{se}(\hat{\tau}_n))$ then $P_\vartheta(\tau \in C_n) \to 1-\alpha$ as $n \to \infty$

Ex. Let $X_1 \dots X_n \sim \text{Bernoulli}(p)$ and let $\psi = g(p) = \log(p/(1-p))$. The Fisher information function is $I(p) = 1/(p(1-p))$ so the estimated s.e. of MLE $\hat{p}_n$ is $\hat{se} = \sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}$

The MLE of ψ is $\hat{\psi} = \log \hat{p}/(1-\hat{p})$. Since $g'(p) = 1/(p(1-p))$ according to the delta method:

$\hat{se}(\hat{\psi}_n) = |g'(\hat{p}_n)|\hat{se}(\hat{p}_n) = \frac{1}{\sqrt{n\hat{p}_n(1-\hat{p}_n)}}$ An approximate 95% CI:

$\hat{\psi}_n \pm \frac{2}{\sqrt{n\hat{p}_n(1-\hat{p}_n)}}$

$(\vartheta_1 \ldots \vartheta_K)$ and $\hat{\vartheta} = (\hat{\vartheta}_1 \ldots \hat{\vartheta}_K)$ be the MLE. Let $l_n = \sum^n \log f(x_i; \vartheta)$, $\frac{\partial^2 l_n}{\partial \vartheta_j^2}$ and $H_{jk} = \frac{\partial^2 l_n}{\partial \vartheta_j \partial \vartheta_k}$

...ine the Fisher Information Matrix by:

$$I_n(\vartheta) = - \begin{bmatrix} E_\vartheta(H_{11}) & E_\vartheta(H_{12}) & \cdots & E_\vartheta(H_{1K}) \\ E_\vartheta(H_{21}) & E_\vartheta(H_{22}) & \cdots & E_\vartheta(H_{2K}) \\ \vdots & & & \\ E_\vartheta(H_{K1}) & E_\vartheta(H_{K2}) & \cdots & E_\vartheta(H_{KK}) \end{bmatrix}$$

Let $J_n(\vartheta) = I_n^{-1}(\vartheta)$ be the inverse of I_n.

Under appropriate regularity conditions $(\hat{\vartheta} - \vartheta) \approx N(0, J_n)$

Also if $\hat{\vartheta}_j$ is the j-th component of $\hat{\vartheta}$ then: $\frac{(\hat{\vartheta}_j - \vartheta_j)}{\hat{se}_j} \rightsquigarrow N(0,1)$

where $\hat{se}_j^2 = J_n(j,j)$ is the jth diagonal element of J_n. The approximate covariance of $\hat{\vartheta}_j$ and $\hat{\vartheta}_K$ is $Cov(\hat{\vartheta}_j, \hat{\vartheta}_K) \approx J_n(j,K)$

→ There is also a multiparameter delta method. Let $\tau = g(\vartheta_1 \ldots \vartheta_K)$ be a function and let

$$\nabla g = \begin{pmatrix} \frac{\partial g}{\partial \vartheta_1} \\ \vdots \\ \frac{\partial g}{\partial \vartheta_K} \end{pmatrix}$$ be the gradient of g.

Suppose that ∇g evaluated at $\hat{\vartheta}$ is not 0. Let $\hat{\tau} = g(\hat{\vartheta})$. Then: $\frac{(\hat{\tau} - \tau)}{\hat{se}(\hat{\tau})} \rightsquigarrow N(0,1)$

where $\hat{se}(\hat{\tau}) = \sqrt{(\hat{\nabla} g)^T \hat{J}_n (\hat{\nabla} g)}$, $\hat{J}_n = J_n(\hat{\vartheta}_n)$ and $\hat{\nabla} g$ is ∇g evaluated at $\vartheta = \hat{\vartheta}$.

Ex. Let $X_1 \ldots X_n \sim N(\mu, \sigma^2)$. Let $\tau = g(\mu, \sigma) = \sigma/\mu$. In Exercise 8 you will show that $I_n(\mu, \sigma) = \begin{bmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{2n}{\sigma^2} \end{bmatrix}$. Hence, $J_n = I_n^{-1}(\mu, \sigma) = \frac{1}{n} \begin{bmatrix} \sigma^2 & 0 \\ 0 & \frac{\sigma^2}{2} \end{bmatrix}$. The gradient of g

$$\nabla g = \begin{pmatrix} -\frac{\sigma}{\mu^2} \\ \frac{1}{\mu} \end{pmatrix}$$ Thus, $\hat{se}(\hat{\tau}) = \sqrt{(\hat{\nabla} g)^T \hat{J}_n (\hat{\nabla} g)} = \frac{1}{\sqrt{n}} \sqrt{\frac{1}{\hat{\mu}^4} + \frac{\hat{\sigma}^2}{2\hat{\mu}^2}}$

Parametric Bootstrap: For parametric models we can use this technique to estimate se and CI. In the nonparametric bootstrap we sampled $X_1^* \ldots X_n^*$ from the empirical distribution $\hat{F}_n$. In the parametric bootstrap we sample instead from $f(x, \hat{\vartheta}_n)$. Hence $\hat{\vartheta}_n$ could be MLE or method of moments estimator.

Ex. Consider previous example. To get the bootstrap standard error, simulate $X_1 \ldots X_n^* \sim N(\hat{\mu}, \hat{\sigma}^2)$, compute $\hat{\mu}^* = n^{-1} \sum X_i^*$ and $\hat{\sigma}^{2*} = n^{-1} \sum (X_i^* - \hat{\mu})^2$. Then compute $\hat{\tau}^* = g(\hat{\mu}^*, \hat{\sigma}^*) = \hat{\sigma}^*/\hat{\mu}^*$. Repeat this B times yields bootstrap replications $\hat{\tau}_1^*, \ldots \hat{\tau}_B^*$ and:

$\hat{se}_{boot} = \sqrt{\frac{\sum^B (\hat{\tau}_b^* - \hat{\tau})^2}{B}}$. Bootstrap easier than delta method. Delta's method advantage is the closed form expression for the standard error.

Checking Assumptions: Good idea to check that your data come from a parametric model, by inspecting the data. Eg. if a histogram of the data looks very bimodal then the assumption of Normality might be questionable. A formal way to test a parametric model is to use a goodness-of-fit test.

AOS Chapter 22: Classification

Consider IID data $(X_1, Y_1) \ldots (X_n, Y_n)$ where $X_i = (X_{i1}, \ldots X_{id}) \in \mathcal{X} \subset \mathbb{R}^d$ and Y_i takes values in some finite set $\mathcal{Y}$. A classification rule is a function $h: \mathcal{X} \to \mathcal{Y}$. When we observe a new X we predict Y to be $h(X)$.
Our goal is to find a classification rule h that makes accurate predictions.
The true error rate of a classifier h is $L(h) = P(\{h(X) \neq Y\})$ and the empirical error rate is $\hat{L}_n(h) = \frac{1}{n}\sum_{i=1}^{n} I(h(X_i) \neq Y_i)$

First we consider the special case when $\mathcal{Y} = \{0, 1\}$. Let $r(x) = \mathbb{E}(Y|X=x) = P(Y=1|X=x)$ ↳ regression function

From Bayes' theorem we have that $r(x) = P(Y=1|X=x)$
$= \frac{f(x|Y=1)P(Y=1)}{f(x|Y=1)P(Y=1) + f(x|Y=0)P(Y=0)} = \frac{\pi f_1(x)}{\pi f_1(x) + (1-\pi) f_0(x)}$, where $f_0(x) = f(x|Y=0)$, $f_1(x) = f(x|Y=1)$, $\pi = P(Y=1)$

The Bayes' classification rule h^* is
$h^*(x) = \begin{cases} 1 & \text{if } r(x) > \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$
The set $D(h) = \{x : P(Y=1|X=x) = P(Y=0|X=x)\}$ is called the decision boundary

The Bayes rule is optimal, that is if h is any other classification rule then $L(h^*) \leq L(h)$
3 main approaches for approximation to the Bayes rule:
1. Empirical Risk Minimization: choose a set of classifier $\mathcal{H}$ and find $\hat{h} \in \mathcal{H}$ that min. some estimate of $L(h)$
2. Regression: find an estimate $\hat{r}$ of the regression function r and define
$\hat{h}(x) = \begin{cases} 1 & \text{if } \hat{r}(x) > \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$
3. Density Estimation: Estimate f_0 from the X_i's for which $Y_i = 0$, estimate f_1 from the X_i's for which $Y_i = 1$ and let $\hat{\pi} = n^{-1}\sum Y_i$. Define:
$\hat{r}(x) = \hat{P}(Y=1|X=x) = \frac{\hat{\pi}\hat{f}_1(x)}{\hat{\pi}\hat{f}_1(x) + (1-\hat{\pi})\hat{f}_0(x)}$ and $\hat{h}(x) = \begin{cases} 1 & \text{if } \hat{r}(x) > \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$

Suppose that $Y \in \mathcal{Y} = \{1, \ldots K\}$. The optimal rule is:
$h(x) = \arg\max_k P(Y=k|X=x) = \arg\max_k \pi_k f_k(x)$
where $P(Y=k|X=x) = \frac{f_k(x)\pi_k}{\sum_r f_r(x)\pi_r}$, $\pi_r = P(Y=r)$, $f_r(x) = f(x|Y=r)$ and $\arg\max_k$ means the value of k that max. the expression

Gaussian & Linear Classifiers: Suppose that $\mathcal{Y} = \{0,1\}$ and that $f_0(x) = f(x|Y=0)$, $f_1(x) = f(x|Y=1)$ are both multivariate Gaussians: $f_k(x) = \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} \exp\{-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)\}$, $k = 0, 1$
Thus $X|Y=0 \sim N(\mu_0, \Sigma_0)$ and $X|Y=1 \sim N(\mu_1, \Sigma_1)$
then the Bayes rule is: $h^*(x) = \begin{cases} 1 & \text{if } r_1^2 < r_0^2 + \log\left(\frac{\pi_1}{\pi_0}\right) + \log\left(\frac{|\Sigma_0|}{|\Sigma_1|}\right) \\ 0 & \text{otherwise} \end{cases}$
where $r_i^2 = (x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)$, $i = 1, 2$
is the Mahalanobis distance
equivalent of Bayes rule: $h^*(x) = \arg\max_k \delta_k(x)$
where $\delta_k(x) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k) + \log\pi_k$
and $|A|$ denotes the determinant of a matrix A

...ion boundary of the classifier is quadratic → Quadratic Discriminant Analysis
... sample estimates of π, $\mu_0, \mu_1, \Sigma_0, \Sigma_1$ in place of the true

$\hat{\pi}_0 = \frac{1}{n}\sum_{i=1}^{n}(1-Y_i)$, $\hat{\pi}_1 = \frac{1}{n}\sum Y_i$

$\hat{\mu}_0 = \frac{1}{n_0}\sum_{Y_i=0} X_i$, $\hat{\mu}_1 = \frac{1}{n_1}\sum_{Y_i=1} X_i$

$S_0 = \frac{1}{n_0}\sum_{Y_i=0}(X_i-\hat{\mu}_0)(X_i-\hat{\mu}_0)^T$, $S_1 = \frac{1}{n_1}\sum_{Y_i=1}(X_i-\hat{\mu}_1)(X_i-\hat{\mu}_1)^T$

where $n_0 = \sum_i (1-Y_i)$ and $n_1 = \sum_i Y_i$

A simplification occurs if we assume that $\Sigma_0 = \Sigma$

$h^*(x) = \arg\max_k \delta_k(x)$, $\delta_k(x) = x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1} + \log \pi_k$

The MLE of Σ is $S = \frac{n_0 S_0 + n_1 S_1}{n_0+n_1}$ $\quad h^*(x) = \begin{cases} 1 & \text{if } \delta_1(x) > \delta_0(x) \\ 0 & \text{otherwise} \end{cases}$

where $\delta_j(x) = x^T S^{-1}\hat{\mu}_j - \frac{1}{2}\hat{\mu}_j^T S^{-1}\hat{\mu}_j + \log\hat{\pi}_j$

This is the discriminant function. The decision boundary $\{x : \delta_0(x) = \delta_1(x)\}$ is linear and called Linear Discrimination Analysis

Linear & Logistic Regression: assume we have an estimator $\hat{r}(x)$

classification rule: $\hat{h}(x) = \begin{cases} 1 & \text{if } \hat{r}(x) > \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$

↳ Regression model:

$Y = r(x) + \epsilon = \beta_0 + \sum_{j=1}^{d}\beta_j X_j + \epsilon$ where $E[\epsilon] = 0$

Least Squares estimate of $\beta = (\beta_0, \beta_1 \dots \beta_d)^T$ minimizes the residual sums of squares:

$RSS(\beta) = \sum^{n}(Y_i - \beta_0 - \sum_{j=1}^{d} X_{ij}\beta_j)^2$

X is a $N\times(d+1)$ matrix: $X = \begin{bmatrix} 1 & X_{11} & X_{1d} \\ 1 & X_{12} & X_{2d} \\ 1 & X_{n1} & X_{nd} \end{bmatrix}$ $\quad Y = (Y_1 \dots Y_n)^T$

$RSS(\beta) = (Y - X\beta)^T(Y - X\beta)$

and model is $Y = X\beta + \epsilon$, $\epsilon = (\epsilon_1 \dots \epsilon_n)^T$

$\hat{\beta} = (X^TX)^{-1}X^TY \Rightarrow \hat{Y} = X\cdot\hat{\beta}$

for logistic regression: $r(x) = P(Y=1 \mid X=x) = \dfrac{e^{\beta_0 + \sum_j \beta_j x_j}}{1 + e^{\beta_0 + \sum_j \beta_j x_j}}$

Relationship Between Logistic Regression & LDA

The difference is in how we estimate the parameters. The joint density of a single observation is $f(x,y) = f(x|y)f(y) = f(y|x)f(x)$. In LDA we estimated the whole joint distribution by maximizing the likelihood:

$\prod f(x_i, y_i) = \underbrace{\prod f(x_i|y_i)}_{\text{Gaussian}} \underbrace{\prod f(y_i)}_{\text{Bernoulli}}$

In logistic regression we maximized the conditional likelihood $\prod f(y_i|x_i)$ but we ignored the second term:

$\prod f(x_i, y_i) = \underbrace{\prod f(y_i|x_i)}_{\text{Logistic}} \underbrace{\prod f(x_i)}_{\text{ignored}}$

Since classification only requires knowing $f(y|x)$ we don't need to estimate the whole joint distribution. Logistic regression leaves the marginal distribution $f(x)$ unspecified so it is more nonparametric than LDA. (advantage over LDA)

* The naive Bayes classifier is popular when x is high-dimensional & discrete

Trees:

Let $\hat{p}_k(j) = \dfrac{\sum I(Y_i = j,\ X_i \in A_k)}{\sum I(X_i \in A_k)}$ for $k = 1, 2$ and $j = 0, 1$. The impurity of the split t is: $I(t) = \sum_{k=1}^{2} \gamma_k$

where $\gamma_k = 1 - \sum_{j=0}^{1} \hat{p}_k(j)^2$. This is known as the Gini index (split to min. impurity)

* Usually the training error rate $\hat{L}_n(h)$ as an estimate of the true error rate because biased downward.

2 ways to estimate the error rate: 1) cross validation and 2) probability inequalities

CV: splitting the data into: Training set T ($\to \hat{h}$) and validation set V ($\to \hat{L}$)

we estimate $\hat{L}(h) = \frac{1}{m} \sum_{X_i \in V} I(h(X_i) \neq Y_i)$

or you can use K-fold CV: 1) Randomly divide data into k chunks of approx. equal size. A common choice is $K = 10$. 2) for $k = 1$ to K: a) delete chunk k from the data b) compute the classifier $\hat{h}_k$ from the rest of the data c) use $\hat{h}_k$ to the predict the data in chunk K. Let $\hat{L}_k$ denote the observed error rate:

$$\hat{L}(h) = \frac{1}{K} \sum^{K} \hat{L}_k$$

Probabilities Inequalities: This method is useful in the context of empirical risk minimisation

Let $\mathcal{H}$: set of linear classifier.

$$\hat{h} = \underset{h \in \mathcal{H}}{\arg\min}\ \hat{L}_n(h) = \arg\min_{h \in \mathcal{H}} \left(\frac{1}{n} \sum I(h(X_i) \neq Y_i) \right)$$

use Hoeffding's Inequality: if $X_1 \ldots X_n \sim \text{Bernoulli}(p)$ for any $\varepsilon > 0$ $P(|\hat{p} - p| > \varepsilon) \leq 2e^{-2n\varepsilon^2}$

Theorem: Uniform convergence: Assume $\mathcal{H}$ is finite and has m elements. Then $P(\max_{h \in \mathcal{H}} |\hat{L}_n(h) - L(h)| > \varepsilon) \leq 2m\, e^{-2n\varepsilon^2}$

Theorem: Let

$\varepsilon = \sqrt{\frac{2}{n} \log\left(\frac{2m}{\alpha}\right)}$ Then $\hat{L}_n(\hat{h}) \pm \varepsilon$ is a $1-\alpha$ CI for $L(\hat{h})$

* The larger is $\mathcal{H}$, the larger the CI for $L(\hat{h})$

↳ more likely to overfit but compensate with larger CI.

Support Vector Machines: (SVM) = class of linear classifiers.

↳ Assume Y: binary $\in \{-1, +1\}$

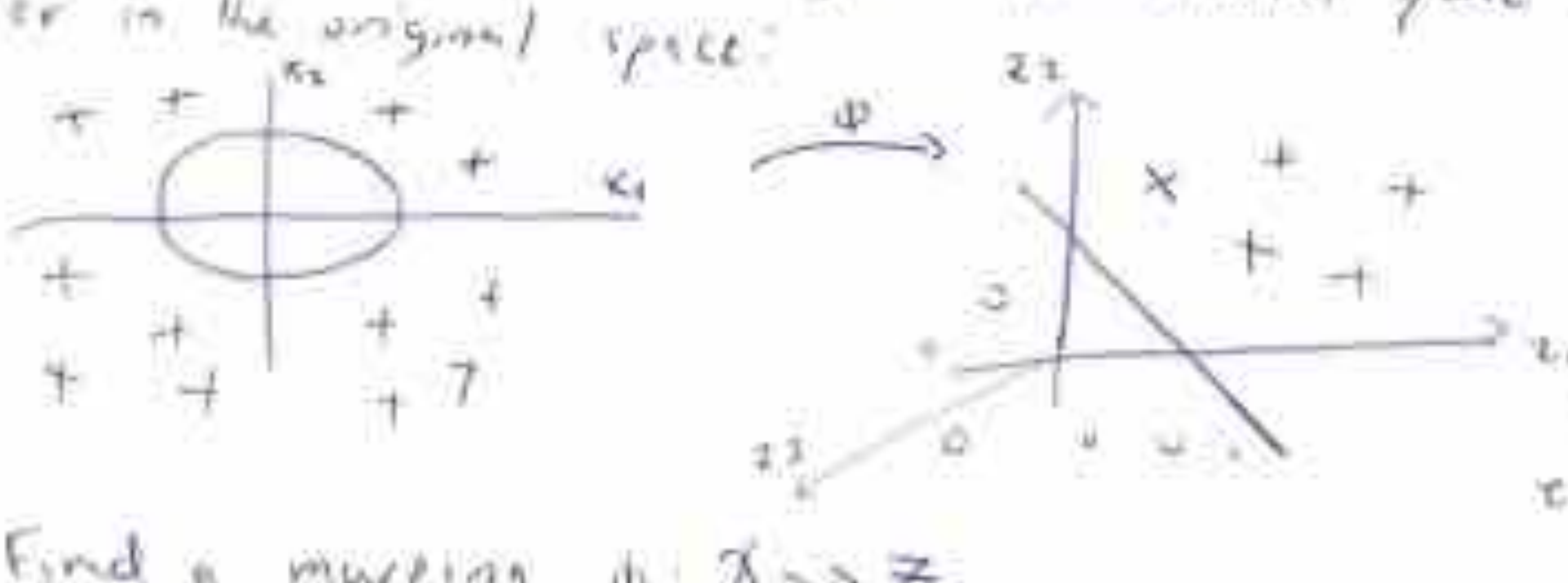
Linear classifier: $h(x) = \text{sign}(H(x))$ where $x = (x_1 \ldots x_d)$

$H(x) = a_0 + \sum^{d} a_i x_i$ and $\text{sign}(z) = \begin{cases} -1 & \text{if } z < 0 \\ 0 & \text{if } z = 0 \\ 1 & \text{if } z > 0 \end{cases}$

* If data are linearly seperable there exists a hyperplane that perfectly separates the 2 classes: $H(x) = x_0 + \sum^{d} a_i x_i$ s.t. $Y_i H(X_i) \geq 1$, $i = 1, \ldots n$. The separate data and max the margin is given by minimizing $\frac{1}{2} \sum^{d} a_i^2$

Kernelisation: use to improve computationally a simple classifier h.

A linear classifier in a higher-dimensional space corresponds to a non-linear classifier in the original space:



$\langle z, \tilde{z} \rangle = \langle \phi(x), \phi(\tilde{x}) \rangle$
$= x_1^2 \tilde{x}_1^2 + 2 x_1 \tilde{x}_1 x_2 \tilde{x}_2 + x_2^2 \tilde{x}_2^2$
$= (\langle x, \tilde{x} \rangle)^2 = K(x, \tilde{x})$

* we can compute $\langle z, \tilde{z} \rangle$ without ever computing $Z_i = \phi(X_i)$

Find a mapping $\phi: \mathcal{X} \rightsquigarrow \mathcal{Z}$

...vied kernels:
...mial $K(x,\tilde{x}) = (\langle x,\tilde{x}\rangle + a)^r$
...moid $K(x,\tilde{x}) = \tanh(a\langle x,\tilde{x}\rangle + b)$
...ussian $K(x,\tilde{x}) = \exp(-\|x-\tilde{x}\|^2/(2\sigma^2))$

...classifier:
...-NN very simple.
Bagging: method for reducing the variability of a classifier. Useful for highly non-linear classifier such as trees.
Boosting: start with a simple classifier and gradually improving it by refitting the data giving higher weight to misclassified samples. eg. AdaBoost.

Chapter 23: Probability Redux: Stochastic Processes:

↳ Imagine a sequence of DEPENDENT Random variables. (eg temperature between days) A stochastic process $\{X_t : t \in T\}$ is a collection of RV. The variables X_t take values in some set $\mathcal{X}$ called state space. The set T is called the index set (can be thought as time) and it can be discrete or continuous. $\{0,1,2,3 ...\}$ or $[0,+\infty)$

Markov Chains: A stochastic process for which the distribution of X_t depends only on X_{t-1}. Assuming that the state space is discrete: $\mathcal{X} = \{1, ..., N\}$ or $\mathcal{X} = \{1,2,...\}$ and the index set is $T = \{0,1,2 ...\}$

→ The process $\{X_n : n \in T\}$ is a Markov chain if $P(X_n = x | X_0, ..., X_{n-1}) = P(X_n = x | X_{n-1})$ for all n and for all $x \in \mathcal{X}$.

$f(x_1, ..., x_n) = f(x_1)\, f(x_2|x_1)\, f(x_3|x_2) \dots f(x_n|x_{n-1})$

$X_0 \to X_1 \to X_2 \to \dots \to X_n \to$

Each variable has a single parent (the previous observation)

Transition Probabilities: jumping from one state to another. A MC is <u>homogeneous</u> if $P(X_{n+1} = j | X_n = i)$ does not change with time. Thus, a homogeneous MC: $P(X_{n+1} = j | X_n = i) = P(X_1 = j | X_0 = i)$

$p_{ij} = P(X_{n+1} = j | X_n = i)$: transition probabilities. The matrix P whose (i,j) element is p_{ij} is called Transition matrix.

Properties of P: a) $p_{ij} \geq 0$ and $\sum_j p_{ij} = 1$ (each row is a PDF)

$P_{m+n} = P_m P_n$, $P_1 = P$, $P_2 = P_{1+1} = P_1 P_1 = P \cdot P = P^2$ so
$P_n = P^n = \underbrace{P\, P\, P \dots P}_{n \text{ times}}$

* We say that i reaches j if $p_{ij}(n) > 0$ for some n and we write $i \to j$. If $i \to j$ and $j \to i$ then we write $i \leftrightarrow j$ and say i and j <u>communicate</u>.
* If all states communicate with each other then the chain is called <u>irreducible</u>. A set of states is <u>closed</u> if once you enter that set of states you never leave. A <u>closed</u> set consisting of a single state is called an <u>absorbing state</u>.

eg.
$$P = \begin{array}{c|c} & \begin{matrix} 1 & 2 & 3 & 4 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \end{matrix} & \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{3} & \frac{2}{3} & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$
state 4 : is an absorbing state.

State i is recurrent / persistent if $P(X_n = i$ for some $n \geq 1 \mid X_0 = i) = 1$ → $\sum_n p_{ii}(n) = \infty$

otherwise state i is transient. → $\sum_n p_{ii}(n) < \infty$

* If i is recurrent $a_i = 1$. The chain will eventually return to i and once it does, argue again that since $a_i = 1$ the chain will return to state i again $[E[Y \mid X_0 = i) = \infty]$

* If i is transient $a_i < 1$. When i is in state i there is a probability $1 - a_i > 0$ that will never return to state i. Thus the prob that the chain is in state i exactly n times: $a_i^{n-1}(1-a_i)$ geometric distribution with finite mean.

Facts about recurrence

1. If state i is recurrent and $i \leftrightarrow j$ then j is recurrent.
2. If state i is transient and $i \leftrightarrow j$ then j is transient.
3. A finite MC must have at least one recurrent state.
4. The states of a finite, irreducible MC are all recurrent.

Convergence of MC: Suppose that $X_0 = i$. Define the recurrence time: $T_{ij} = \min\{n > 0 : X_n = j\}$ assuming X_n ever returns to state i, otherwise define $T_{ij} = \infty$. The mean recurrence time of a recurrent state i is: $m_i = E(T_{ii}) = \sum_n n f_{ii}(n)$

where $f_{ij}(n) = P(X_1 \neq j, X_2 \neq j, \ldots, X_{n-1} \neq j, X_n = j \mid X_0 = i)$

A recurrent state is null if $m_i = \infty$, otherwise it's called non-null / positive.

Formally the period of state i is d if $p_{ii}(n) = 0$ whenever n is not divisible by d and d is the largest integer with this property. Thus $d = \gcd\{n : p_{ii}(n) > 0\}$ where gcd: "greatest common divisor". State i is periodic if $d(i) > 1$ and aperiodic if $d(i) = 1$.

→ Let $\pi = (\pi_i ; i \in \mathcal{X})$ be a vector of non-negative numbers that sum up to one (can be thought as PDF). We say π is a stationary distribution if $\pi = \pi P$

* Just because a chain has a stationary distribution it doesn't mean it converges.

Poisson Processes:

$$P(X = x) = p(x; \lambda) = \frac{e^{-\lambda} \cdot \lambda^x}{x!}, \quad x = 0, 1, 2, \ldots \qquad E(X) = \lambda = V(X)$$

A poisson process is a stochastic process $\{X_t : t \in [0, \infty)\}$ with state space $\mathcal{X} = \{0, 1, \ldots\}$ s.t. 1) $X(0) = 0$, 2) For any $0 = t_0 < t_1 < t_2 \ldots < t_n$ the increments $X(t_1) - X(t_0), X(t_2) - X(t_1), \ldots, X(t_n) - X(t_{n-1})$ are independent.

There is a function $\lambda(t)$ s.t.

$$P(X(t+h) - X(t) = 1) = \lambda(t) \cdot h + o(h) \qquad ; \lambda(t) : \text{intensity function}$$
$$P(X(t+h) - X(t) \geq 2) = o(h).$$

→ A poisson process with intensity function $\lambda(t) = \lambda$ for some $\lambda > 0$ is called a homogeneous Poisson process with rate λ. In this case, $X(t) \sim \text{Poisson}(\lambda t)$

ITDS: Chapter 6 - Random Variable Generation

Pseudorandom: Consider a finite set $\mathcal{M} = \{0, 1, \ldots, M-1\}$ and consider the sequence $u_0, u_1, \ldots \in \mathcal{M}$. For every $a \in \mathcal{M}$ define $N_n(a)$ as the number of $u_i = a$ for $i = 0, 1, 2, \ldots, n-1$. We call the sequence $u_0, u_1, \ldots$ pseudorandom on $\mathcal{M}$ if and only if for every $a \in \mathcal{M}$

$$\frac{N_n(a)}{n} \to \frac{1}{M}$$

...ical Generators start with a number v_0 (that we call a random seed) and ...ly a map D to get the next number in the sequence

... is fixed and let D be a map define the dynamical system $u_i = D(u_{i-1}), i=1...$
... To the period of D started at v_0 the smallest positive integers s.t. $v_{i+T_0} = v_i$
...smallest period T for all admissible starting points v_0 is called the period of D

modulus $(m) = 16$
multiplier $(a) = 3$ $\qquad U_{n+1} (3U_n + 0) \text{ mod } 16$
increment $(c) = 0$

1. $V_0 = 2$
2. $V_1 = 3 \cdot 2 \text{ mod } 16 = 6 \text{ mod } 16 = 6$
3. $V_2 = 3 \cdot 6 \text{ mod } 16 = 18 \text{ mod } 16 = 2$
4. $V_3 = 3 \cdot 2 \ldots$

sequence 2, 6, 2, 6 ...

Sampling

Algorithm 1: Accept-Reject Sampler
- Target Density $f(x)$: PDF from which you want to generate samples
- Sampling Density $g(x)$: A simpler distribution that is easy to sample from & satisfies: $f(x) \le M g(x) \; \forall x$ where $M > 0$

steps:
1. Initialisation - initial state x_0 from the sampling density $g(x)$
2. Main Iteration Loop (Repeat until desired # of samples)
   - 2.1 Draw a candidate x from sampling distr. $g(x)$
   - 2.2 Compute the acceptance rate $r(x) = \frac{f(x)}{Mg(x)}$
   - 2.3 Generate a Uniform RV $u \sim \text{Uniform}(0,1)$
   - 2.4 Accept/Reject the proposal x: if $u < r(x)$ accept and let $X_{t+1} = z$ otherwise reject and return to step 2.1, resampling z from $g(x)$
3. Repeat until desired # of samples obtained

ITDS: Chapter 7 Finite MC

Let $\{X_n, n \in N\}$ be a homogeneous MC. Let the statespace $X = \{s_1, \ldots s_N\}$ be enumerated and let p_0 be the PMF of X_0. Then the PMF p_n for X_n is: $p_n = p_0 \cdot P^n$

ITDS Chapter 8 Pattern Recognition:

In many ML text books that are practically oriented you will see the recommendation that the training/testing split should be 70/30. In pattern recognition problem this doesn't make much sense, it is better to ~~use that to choose the number of~~ determine with how high probability you want the bound to hold and use that to choose the number of samples to reach a (high) confidence interval.

Precision $P(Y=1 \mid g(X)=1)$ $\qquad$ Recall $P(g(X)=1 \mid Y=1)$

often used in medical testing and are then called sensitivity

ITDS Chapter 11 Dimensionality Reduction:

SVD:
consider a line given by the unit vector v and consider a point x then the projection of x onto v is given by $(v \cdot x) v$. Let v be a unit vector. Consider the projection of each X_i onto v, but only consider the proportion $Y_i = (X_i \cdot v)$

$\bar{Y}_n = \frac{1}{n} \sum Y_i = \frac{1}{n} \sum X v = 0$ (assumed zero empirical mean)

$v_1 = \arg\max_{\|v\|=1} \frac{1}{n} \Sigma (Y_i - \bar{Y}_n)^2 = \arg\max_{\|v\|=1} \Sigma |X_i \cdot v|^2$

let A n×m matrix with the rows x_i : $\sum_{j=1}^{n} |X_i \cdot v|^2 = |Av|^2$

$\Rightarrow \arg\max_{\|v\|=1} |Av|$

* The singular vectors are not necessarily unique, in fact if v is a singular vector then so is -v. We can also have ties, we arbitrarily pick one

$A = UDV^T$

left singular vectors, D ↳ diagonal matrix with σ_i

The power method:

$A^TA = (UDV^T)^T(UDV^T) = (VDU^TUDV^T) = VD^2V^T \quad U^TU = I$ (columns orthonormal)

Principal Component Analysis:

it is a coordinate transformation from the original coordinates to the coordinate system given by the singular vectors.

$PCA(A) = AV = UDV^TV = UD$

Explained Variance: is how much % of the total variance is captured by our singular vectors.