CS 6140

MATH REVIEW 2

09-15-2021

JORIO COCOLA
cocola.j@northeastern.edu

References

- Garrett Thomas - "Mathematics of Machine Learning"
- Deisenroth et al. - "Mathematics for Machine Learning"
- Kevin Murphy - "Probabilistic Machine Learning"
- Larsen & Marx - "An Introduction to Mathematical Statics and its Applications"

<u>Remark</u> Below sections refer to Garrett Thomas' notes

5.1 BASICS

- **Experiment** (frequentist view) is any procedure which has a well-defined set of outcomes

Experiment: "Rolling a standard 6-sided die once"

- **Sample outcomes** are the potential eventualities of an experiment

6 Possible outcomes: 1, 2, 3, 4, 5, 6

- **Sample space** the totality of the sample outcomes

  The sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$

- **Event** is a subset of the sample space

  - Event : "rolling a 2" $= \{2\}$

  - Event : "rolling an even number" $= \{2, 4, 6\}$

let $\mathcal{F}$ be the set of events of Ω

A **Probability Measure** is a function

$$\mathbb{P} : \mathcal{F} \longrightarrow [0,1]$$

that satisfies

(i) $\mathbb{P}(\Omega) = 1$

(ii) **Countable additivity**: for any countable collection of disjoint sets $\{A_i\} \subseteq \mathcal{F}$,

$$\mathbb{P}\left(\bigcup_i A_i\right) = \sum_i \mathbb{P}(A_i)$$

Note

- disjoints sets: $A_i \cap A_j = \phi$

- countable collection of sets:

  $\{A_1, A_2, \dots, A_N\}$ for some N

  or

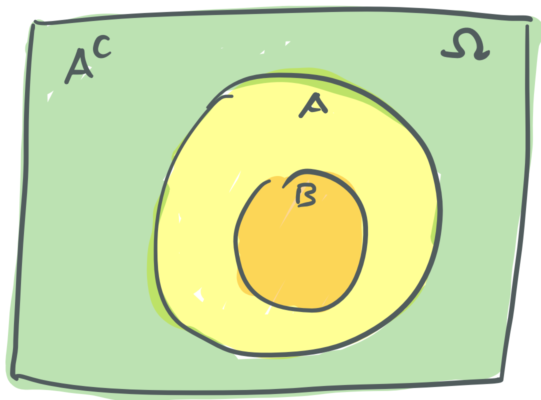
  $\{A_1, A_2, \dots, A_N, A_{N+1}, \dots\}$ (infinite)

Proposition 26. *Let A be an event. Then*

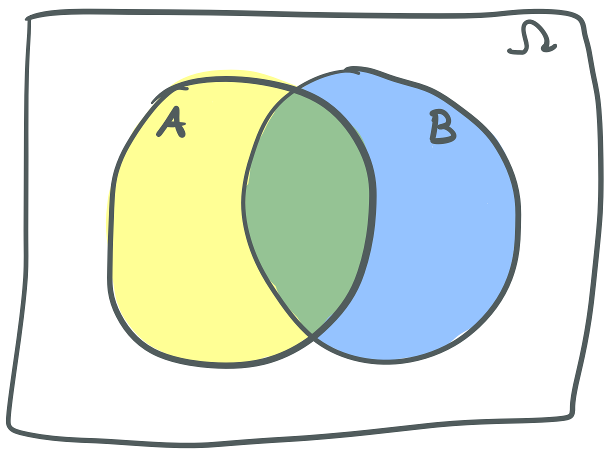
(i) $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$. (Complement of A)

(ii) *If B is an event and $B \subseteq A$, then $\mathbb{P}(B) \leq \mathbb{P}(A)$.*

(iii) $0 = \mathbb{P}(\varnothing) \leq \mathbb{P}(A) \leq \mathbb{P}(\Omega) = 1$



Proposition 27. *If A and B are events, then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.*



Proposition 28. *If $\{A_i\} \subseteq \mathcal{F}$ is a countable set of events, disjoint or not, then*

$$\mathbb{P}\left(\bigcup_i A_i\right) \leq \sum_i \mathbb{P}(A_i)$$

Q (Larsen & Marx, 2.3.2)

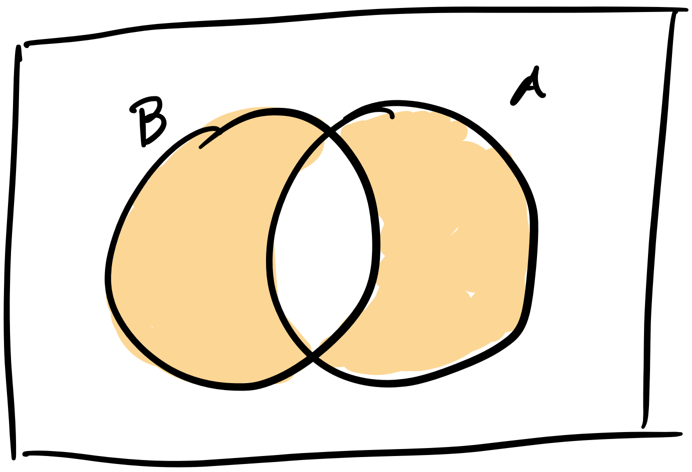
Let A and B two events in Ω.

Suppose that

$$\mathbb{P}(A) = 0.4 \; ; \quad \mathbb{P}(B) = 0.5 \quad \text{and} \quad \mathbb{P}(A \cap B) = 0.1$$

Find the probability that

E = A or B but not both occur



"E = orange"

$$E = (A \cap B^c) \cup (B \cap A^c)$$

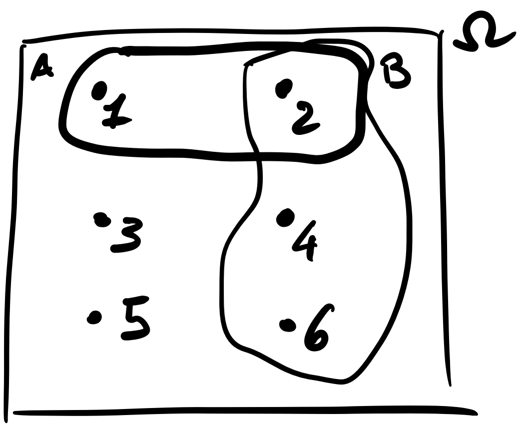
5.1.1 Conditional probability

The **conditional probability** of event A given that event B has occurred is written $\mathbb{P}(A|B)$ and defined as

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

assuming $\mathbb{P}(B) > 0$.[12]

Example "Roll a 6-sided die"



A = "Rolling a number smaller than 3"
$= \{1, 2\}$

B = "Rolling an even number"
$= \{2, 4, 6\}$

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\frac{1}{6}}{\frac{3}{6}} = \frac{1}{3}$$

5.1.2 Chain rule

Another very useful tool, the **chain rule**, follows immediately from this definition:

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

Q

Two cards are drawn from a standard deck one after the other without replacement Find the probability that the first card is a heart and the second is red

5.1.3 Bayes' rule

Taking the equality from above one step further, we arrive at the simple but crucial **Bayes' rule**:

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

5.2 RANDOM VARIABLES

A random variable X is some uncertain quantity of interest, whose value depends on the outcome of a random event:

$$X : \Omega \longrightarrow \mathbb{R}$$

where $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space.

Examples

- Flip a coin 2 times

$$\Omega = \{tt, th, ht, hh\}$$

$X =$ "Number of heads"

$$\mathbb{P}(X=2) = \mathbb{P}(\{hh\}) = \frac{1}{4}$$

$$\mathbb{P}(X=1) = \mathbb{P}(\{ht, th\}) = \frac{1}{2}$$

- 2 Fair 6-sided die are rolled

$X =$ "Sum of the outcomes"

$$\mathbb{P}(X=11) = \mathbb{P}(\{(5,6), (6,5)\}) = \frac{2}{36}$$

5.2.2 Discrete Random Variables

A random variable X is called discrete

if the set of possible outcomes

is finite or countable

Example

Flip a coin repeatedly (infinite times)

X = "Number of tosses until the first head"

$X \in \{1, 2, 3, \dots\}$

The **probability mass function** (p.m.f.)

is a function $p : X(\Omega) \to [0,1]$:

$$p(x) = \mathbb{P}(X = x)$$

such that

$$\sum_{x \in X(\Omega)} p(x) = 1$$

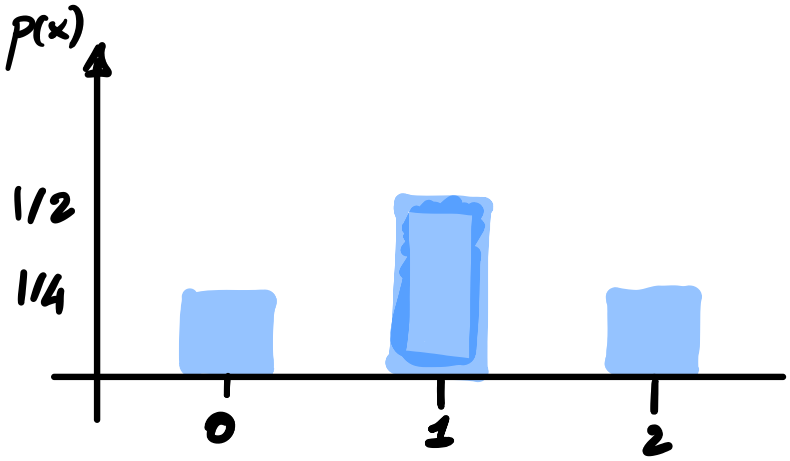
<u>Note</u>: $p(x)$ completely specify X

Example

Flip a coin twice,

$X =$ "number of heads"

	$X=0$	$X=1$	$X=2$
$p(x)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$



Q

Flip a coin repeatedly (infinite time)

X = "Number of tosses until the first head"

Example (Bernoulli)

If E is any event, then the BERNOULLI R.V. ON E is

$$X = \begin{cases} 1 & \text{if } E \text{ occurs} \\ 0 & \text{if } E \text{ does not occur} \end{cases}$$

$$\Rightarrow p(1) = \mathbb{P}(X=1) = \mathbb{P}(E) \qquad p(0) = 1 - \mathbb{P}(E)$$

5.2.3 Continuous Random Variable

X takes real values

Examples

- Temperature at the peak of Mount Everest
- The weight of a tennis ball

Def A continuous probability density function $p: \mathbb{R} \to [0, \infty)$ is a function such that

$$\int_{-\infty}^{+\infty} p(x)\, dx = 1$$

Def X is a continuous random variable if there exists a continuous probability density function such that for any

$$-\infty \leq a \leq b \leq +\infty$$

we have

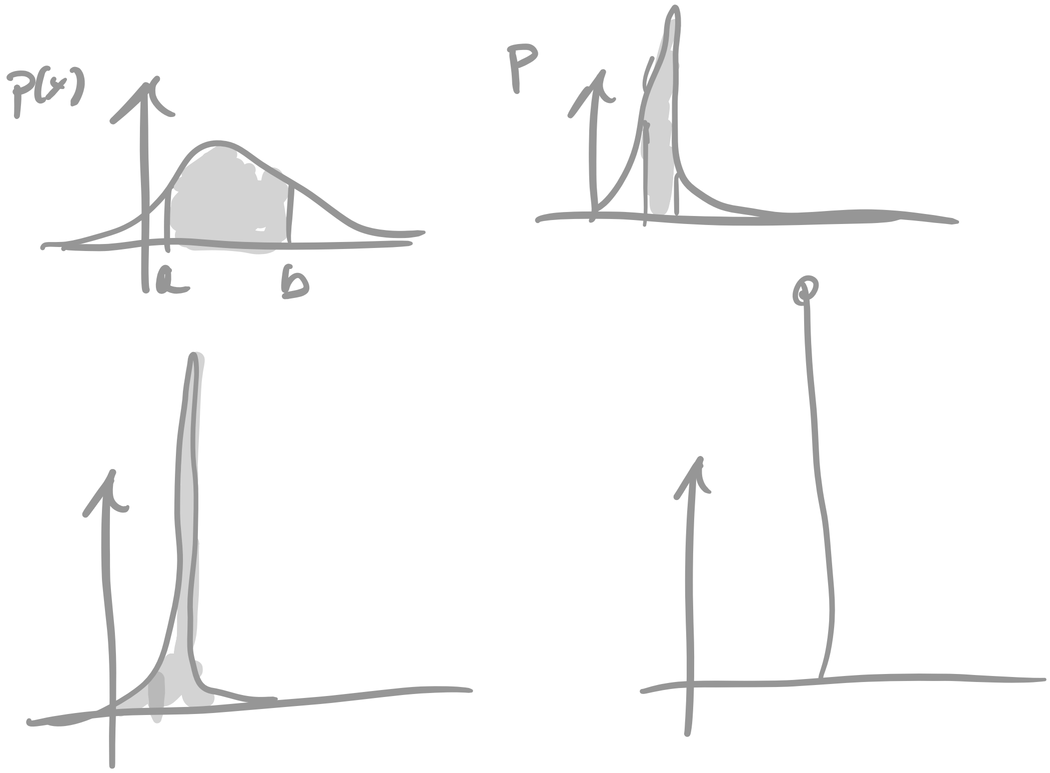
$$\mathbb{P}(a \leq X \leq b) = \int_a^b p(x)\,dx$$

Remark

$$p(a) \gg p(c)$$

$$p(a) \neq \mathbb{P}(X = a)$$

$$\mathbb{P}(X = a) = 0$$


p(x)
a
b
p

Notation

- $p(X)$ is used to denote the entire probability distribution
- $p(x)$ is used to denote p evaluated at x.

$$p : \mathbb{R} \to [0, \infty)$$

Def. The cumulative distribution of a continuous random variable X is

$$F(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} p(z)\, dz$$

It satisfies

$$\{X \leq x\} = \{X > x\}^c$$

a.) $\mathbb{P}(X > x) = 1 - F(x)$

b.) $\mathbb{P}(a < X \leq b) = F(b) - F(a)$

c.) $\lim\limits_{x \to \infty} F(x) = 1$

d.) $\lim\limits_{x \to -\infty} F(x) = 0$

e.) $F'(x) = p(x)$

Q (Larsen & Marx)

3.4.10. A continuous random variable Y has a cdf given by

$$F_Y(y) = \begin{cases} 0 & y < 0 \\ y^2 & 0 \le y < 1 \\ 1 & y \ge 1 \end{cases}$$

Find $P(\frac{1}{2} < Y \le \frac{3}{4})$ two ways—first, by using the cdf and second, by using the pdf.

5.3 JOINT DISTRIBUTIONS

X and Y are discrete R.V., then

they can be completely described by the

joint probability mass function

$$p : \mathbb{R}^2 \longrightarrow [0,1]$$

such that

$$\mathbb{P}(X = x, Y = y) = p(x,y)$$

$$\sum_{x,y} p(x,y) = 1$$

Example (Larsen & Marx)

Example 3.7.2 Suppose two fair dice are rolled. Let X be the sum of the numbers showing, and let Y be the larger of the two. So, for example,

$$p_{X,Y}(2,3) = P(X=2, Y=3) = P(\emptyset) = 0$$

$$p_{X,Y}(4,3) = P(X=4, Y=3) = P(\{(1,3)(3,1)\}) = \frac{2}{36}$$

and

$$p_{X,Y}(6,3) = P(X=6, Y=3) = P(\{(3,3)\}) = \frac{1}{36}$$

The entire joint pdf is given in Table 3.7.1.

Table 3.7.1

x \ y	1	2	3	4	5	6	Row totals
2	1/36	0	0	0	0	0	1/36
3	0	2/36	0	0	0	0	2/36
4	0	1/36	2/36	0	0	0	3/36
5	0	0	2/36	2/36	0	0	4/36
6	0	0	1/36	2/36	2/36	0	5/36
7	0	0	0	2/36	2/36	2/36	6/36
8	0	0	0	1/36	2/36	2/36	5/36
9	0	0	0	0	2/36	2/36	4/36
10	0	0	0	0	1/36	2/36	3/36
11	0	0	0	0	0	2/36	2/36
12	0	0	0	0	0	1/36	1/36
Col. totals	1/36	3/36	5/36	7/36	9/36	11/36	

X and Y are continuous R.V., then they can be completely described by the

joint probability mass function

$$p : \mathbb{R}^2 \longrightarrow [0, \infty)$$

such that

$$\mathbb{P}(a \leq X \leq b,\ c \leq Y \leq d) = \int_a^b \int_c^d p(x,y)\, dx\, dy$$

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

Example

The joint uniform distribution

$$p(x,y) = \frac{1}{(b-a)(c-d)} \qquad \text{for } \begin{cases} a \le x \le b \\ c \le x \le d \end{cases}$$

If R is a region in the rectangle

$$[a,b] \times [c,d]$$

then

$$\mathbb{P}\big((X,Y) \in R\big) = \iint_R p(x,y)\, dx\, dy = \frac{\text{Area}(R)}{(b-a)(c-d)}$$

Note this only depends on the size of the region!

Remark

All the above generalizes to

$$X_1, \dots, X_n$$

random variables, defining appropriate

Joint distributions

$$p(x_1, x_2, \dots, x_n)$$

5.3.1 Independent random variables

Two random variables X and Y are said independent if

$$P_{XY}(x,y) = P_X(x)\, P_Y(y)$$

where

- P_{XY} joint probability distrib.
- P_X probability distrib. of X
- P_Y probability distrib. of Y

A collection of random variables

$$X_1, X_2, \dots, X_n$$

are (collectively) independent when

$$P_{X_1 \dots X_n}(x_1, x_2, \dots, x_n) = \prod_{i=1}^{n} P_{X_i}(x_i) = P_{X_1}(x_1) \cdot P_{X_2}(x_2) \cdot \ldots \cdot P_{X_n}(x_n)$$

$X_1, \dots, X_n$ are independent and identically distributed (i.i.d.) when

$$P_{X_i}(x) = \hat{P}(x) \qquad \forall X_i$$

and they are independent

$$P_{X_1 \dots X_n}(x_1, \dots, x_n) = \prod_{i=1}^{n} \hat{P}(x_i)$$

5.3.2 Marginal distribution

Given the joint probability distribution of X and Y.

We can derive the probability distribution of the sigle variables

- $$P_X(x) = \sum_y p(x,y) \qquad \text{(discrete case)}$$

- $$P_X(x) = \int_{\mathbb{R}} p(x,y)\,dy \qquad \text{(continuous case)}$$

P_X is a "marginal distribution" of P_{XY}

Example (Larsen & MarX)

Example 3.7.2 Suppose two fair dice are rolled. Let X be the sum of the numbers showing, and let Y be the larger of the two. So, for example,

$$p_{X,Y}(2,3) = P(X = 2, Y = 3) = P(\emptyset) = 0$$

$$p_{X,Y}(4,3) = P(X = 4, Y = 3) = P(\{(1,3)(3,1)\}) = \frac{2}{36}$$

and

$$p_{X,Y}(6,3) = P(X = 6, Y = 3) = P(\{(3,3)\}) = \frac{1}{36}$$

The entire joint pdf is given in Table 3.7.1.

Table 3.7.1

x \ y	1	2	3	4	5	6	Row totals
2	1/36	0	0	0	0	0	1/36
3	0	2/36	0	0	0	0	2/36
4	0	1/36	2/36	0	0	0	3/36
5	0	0	2/36	2/36	0	0	4/36
6	0	0	1/36	2/36	2/36	0	5/36
7	0	0	0	2/36	2/36	2/36	6/36
8	0	0	0	1/36	2/36	2/36	5/36
9	0	0	0	0	2/36	2/36	4/36
10	0	0	0	0	1/36	2/36	3/36
11	0	0	0	0	0	2/36	2/36
12	0	0	0	0	0	1/36	1/36
Col. totals	1/36	3/36	5/36	7/36	9/36	11/36	

Q Consider $p(x,y) = \dfrac{1}{(b-a)(c-d)}$ on $\begin{cases} a \le x \le b \\ c \le x \le d \end{cases}$

- Find p_x and p_y

- Are X an Y independent?

5.4 "Expectations"

The "average value" of a random variable is described by its expected value

- $$\mathbb{E}[X] = \sum_{x \in X(\Omega)} x\, p(x)$$ (discrete case)

- $$\mathbb{E}[X] = \int_{-\infty}^{+\infty} x\, p(x)\, dx$$ (continuous case)

Properties

- $\mathbb{E}$ is a linear map on the vector space of random variables

$$\mathbb{E}\left[\sum_{i=1}^{n} \alpha X_i\right] = \sum_{i=1}^{n} \alpha_i \mathbb{E}[X_i]$$

- For $\beta \in \mathbb{R}$, $X = \beta$ (constant RV) is a RV and

$$\mathbb{E}[\beta] = \beta$$

- $X_1, \dots, X_n$ are independent

$$\mathbb{E}\left[\prod_{i=1}^{n} X_i\right] = \prod_{i=1}^{n} \mathbb{E}[X_i]$$

5.6 Covariance

For two variables X and Y, the covariance measures the linear dependence between the two

$$\mathrm{Cov}(X,Y) = \mathbb{E}_{X,Y}\left[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])\right]$$

$$= \mathbb{E}_{XY}[XY] - \mathbb{E}_X[X]\,\mathbb{E}_Y[Y]$$

When $Cov(X,Y)=0$ we say that X and Y are uncorrelated.

If X and Y are independent

$\Rightarrow Cov(X,Y) = 0$

5.5 Variance

The variance of X is

$$\mathrm{Var}(X) = \mathrm{Cov}(X, X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$$

$$= \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

The standard deviation of X is

$$\sqrt{\mathrm{Var}(X)}$$

<u>Remark</u> X and $\sqrt{\mathrm{Var}(X)}$ have the same units

Properties

- $$\mathrm{Var}(\alpha X) = \alpha^2 \mathrm{Var}(X)$$

- $X = \beta \in \mathbb{R}$

  $$\mathrm{Var}(X) = 0$$

- $$\mathrm{Var}(X_1 + X_2 + \dots + X_n) = \sum_{i=1}^{n} \mathrm{Var}(X_i)$$

  if $X_1, \dots, X_n$ are uncorrelated.

The Gaussian/Normal distribution

X has a gaussian distribution with mean μ and variance σ^2 if

$$P_X(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \quad \text{for } x \in \mathbb{R}$$

then

$$\mathbb{E}[X] = \mu \quad \text{and} \quad \mathrm{Var}[X] = \sigma^2$$

We write $X \sim N(\mu, \sigma^2)$

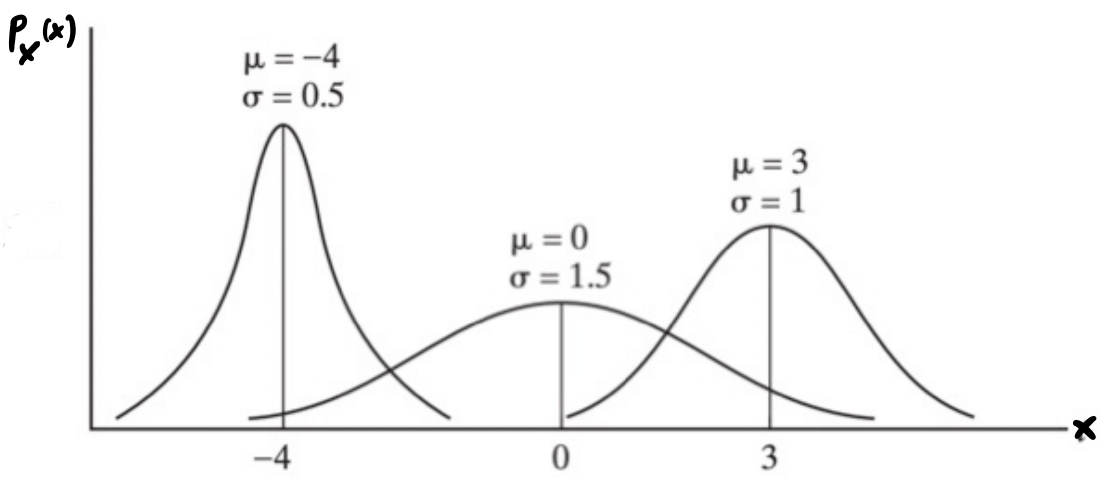


Figure 3.4.5

From Larsen & Marx

5.8 Estimation of Parameters

Probability functions provide models of random phenomena

For example a normal distribution can be used to can be used for the height Northeastern students

How to find the parameters of these models (e.g. μ, σ) ?

5.8.1 Maximum likelihood estimation

- $X_1, \dots, X_n$ are random variables with p.d.f.

$$p_{X_1 \dots X_n}(x_1, \dots, x_n; \theta)$$

where θ is an unknown parameter

- $x_1, \dots, x_n$ are the corresponding observations (real numbers)

Goal Estimate the parameter θ given data $D = \{x_1, \dots, x_n\}$

Maximum likelihood estimate

Given data $D = \{x_1, \ldots, x_n\}$ the

Likelihood function is

$$\mathcal{L}(\theta) = p(x_1, \ldots, x_n ; \theta)$$

The **Maximum likelihood estimate** $\hat{\theta}_{MLE}$ of θ is

$$\hat{\theta}_{MLE} = \underset{\theta}{\arg\max}\, \mathcal{L}(\theta)$$

Q What are possible issues with this?

When $X_1, \dots, X_n$ are i.i.d. then

$$P_{X_1, \dots, X_n}(x_1, \dots, x_n ; \theta) = \prod_{i=1}^{n} p_X(x_i ; \theta)$$

Then

$$\log \mathcal{L}(\theta) = \sum_{i=1}^{n} \log p_X(x_i ; \theta)$$

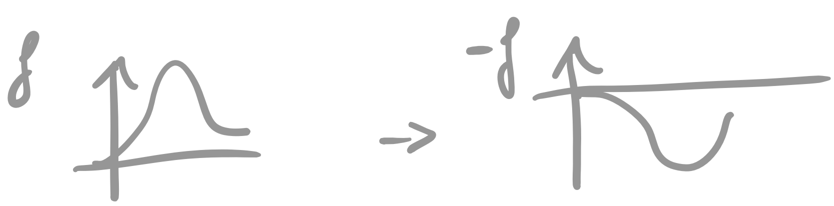
is the Log-likelihood

Then

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \log \mathcal{L}(\theta)$$

or

$$\hat{\theta}_{MLE} = \arg\min_{\theta} - \log \mathcal{L}(\theta)$$



Example

$X_1, \dots, X_n$ are i.i.d $N(\mu, \sigma^2)$ variables.

and $x_1, \dots, x_n$ their observations

Goal Estimate μ and σ

- Likelihood

$$\mathcal{L}(\mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2\right)$$

- Negative Log-likelihood

$$-\log \mathcal{L}(\mu, \theta) = \frac{n}{2} \log(2\pi\sigma^2) + \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2$$

The maximum likelihood estimates are given by

$$\hat{\mu}_{MLE}, \hat{\sigma}_{MLE} = \arg\min_{\mu,\sigma} \left[\frac{n}{2} \log(2\pi\sigma^2) + \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right]$$

Recall from calculus:

If $\hat{\mu}_{MLE}, \hat{\sigma}_{MLE}$ are minimuma of $-\log \mathcal{L}(\mu,\sigma)$

then

$$\begin{cases} \dfrac{\partial}{\partial \mu} - \mathcal{L}(\mu,\sigma) = 0 \\ \\ \dfrac{\partial}{\partial \sigma} - \mathcal{L}(\mu,\sigma) = 0 \end{cases}$$

- $$\frac{\partial}{\partial \mu} \log \mathcal{L}(\mu, \sigma) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) = 0$$

- $$\frac{\partial}{\partial \sigma} \log \mathcal{L}(\mu, \sigma) = \frac{n}{\sigma} - \frac{1}{\sigma^3} \sum_{i=1}^{n} (x_i - \mu)^2 = 0$$

solving the 2 equations we find

$$\hat{\mu}_{MLE} = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$ EMPIRICAL MEAN

$$\hat{\sigma}^2_{MLE} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$ EMPIRICAL VARIANCE

<u>Remark</u> One should check that these are indeed max of the likelihood (e.g. looking at 2nd derivatives)

- The bias of an estimator $\hat{\theta}$ of a true parameter θ_* is

$$\text{bias}(\theta) = \mathbb{E}[\hat{\theta}] - \theta_*$$

it can be shown that

$$\text{bias}(\hat{\mu}_{MLE}) = 0$$

$$\text{bias}(\hat{\sigma}_{MLE}) = -\frac{\sigma^2}{n} \neq 0 \qquad \text{(biased)}$$

In statistics the following estimator is used

$$\hat{\sigma}^2_{UHB} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \hat{\mu}_{MLE})^2$$

$$\text{bias}(\hat{\sigma}^2_{UHB}) = 0 \qquad \text{(unbasiled)}$$