Chapter 1

Question

There are many random phenomena (example?) in our real life. What is the language/mathematical structure that we use to depict them?

characteristic
★ don't know what result we will get in the future
★ the best we can do is to describe/calculate the probability of these possible results.

樂透開獎號碼
waiting time
rain tomorrow?
...

Outline

probability space

- sample space
- event
- probability measure
  - conditional probability
  - independence
- three theorems
  - multiplication law
  - law of total probability
  - Bayes' rule

Website of My Probability Course

http://www.stat.nthu.edu.tw/~swcheng/Teaching/math2810/index.php

---

Textbook page

LNp. (Lecture Note page)

Definition (sample space, TBp. 2)

A **sample space** Ω is the set of all possible outcomes in a random phenomenon.

Example 1.1 (throw a coin 3 times, TBp. 35)

$\Omega = \{hhh, hht, hth, thh, htt, tht, tth, ttt\}$ h: head
t: tail

Ω is a finite set

Example 1.2 (number of jobs in a print queue, Ex. B, TBp. 2)

$\Omega = \{0, 1, 2, \ldots\}$

Ω is an infinite, but countable, set

discrete random variable

cf.

continuous random variable

Example 1.3 (length of time between successive earthquakes, Ex. C, TBp. 2)

$\Omega = \{t | t \geq 0\} = [0, \infty)$

Ω is an infinite, but uncountable, set

Question

What are the differences between the Ω in these examples?

Definition (event, TBp. 2)

A particular subset of Ω is called an **event**.

collection of all "well-defined" events $\Rightarrow$ σ-field

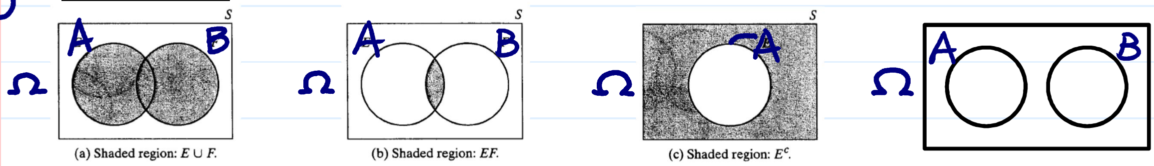
Example 1.4 (cont. Ex. 1.1)

Let A be the event that total number of heads equals 2, then $A = \{hht, hth, thh\}$.

Example 1.5 (cont. Ex. 1.2)

Let A be the event that fewer than 5 jobs in the print queue, then $A = \{0, 1, 2, 3, 4\}$.

- **union.** $C = A \cup B \Rightarrow C$: at least one of A and B occur.
- **intersection.** $C = A \cap B \Rightarrow C$: both A and B occur.
- **complement.** $C = A^c \Rightarrow C$: A does not occur.
- mutually exclusive **disjoint.** $A \cap B = \emptyset \Rightarrow A$ and B have no outcomes in common.



(a) Shaded region: $E \cup F$. (b) Shaded region: EF. (c) Shaded region: E^c.

Definition (probability measure, TBp. 4)

A **probability measure** on Ω is a function P from subsets of Ω to the real numbers that satisfies the following axioms: $\overset{\shortparallel}{\mathcal{F}}$

$P: \mathcal{F} \to [0, 1]$ $\cancel{\mathbb{R}}$

Axioms of probability

1. $P(\Omega) = 1$. ← total prob. = 1
2. If $A \subset \Omega$, then $P(A) \geq 0$. ← non-negativity
3. If A_1 and A_2 are disjoint, then ← additivity

$$P(A_1 \cup A_2) = P(A_1) + P(A_2).$$

More generally, if $A_1, A_2, \ldots$ are mutually disjoint, then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

Example 1.6 (cont. Ex. 1.1)

Suppose the coin is fair. For every outcome $\omega \in \Omega$, $P(\omega) = \frac{1}{8}$.

$$\Omega = \{hhh, hht, hth, thh, htt, tht, tth, ttt\}$$

1/8 1/8 1/8 1/8 1/8 1/8 1/8 1/8 ← $P: \Omega \to [0, 1]$

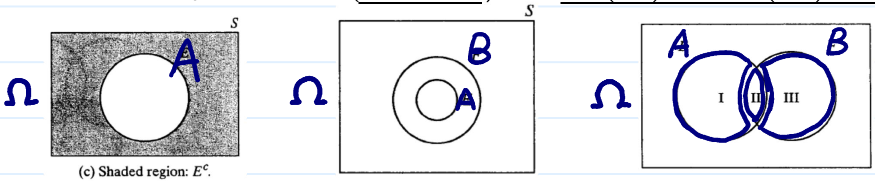
Property A. $\underline{P(A^C)} = \underline{1 - P(A)}$.

Property B. $\underline{P(\emptyset) = 0}$.

Property C. If $\underline{A \subset B}$, then $\underline{P(A) \le P(B)}$.

Property D. $P(\underline{A \cup B}) = \underline{P(A) + P(B) - P(A \cap B)}$.

generalization:
$P(A_1 \cup \cdots \cup A_n)$
$= \sum P(A_i)$
$- \sum P(A_i \cap A_j)$
$+ \sum P(A_i \cap A_j \cap A_k)$
$- \cdots$



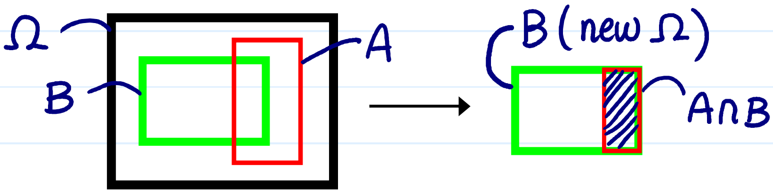
(c) Shaded region: E^c.

Definition (conditional probability, TBp. 17)

Let A and B be two events with $\underline{P(B) > 0}$. The **conditional probability** of A given B is defined to be

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

Q: Why cond. prob. important in statistics?



Ans: update information.

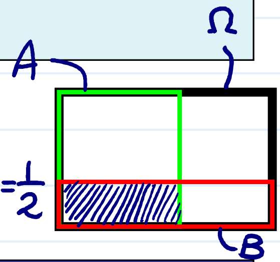
Example 1.7 (cont. Ex. 1.6)

Suppose that the first throw is h. What is the probability that we can get exact two h's in the three trials?

B
A

$\Omega = \{hhh, hht, hth, thh, htt, tht, tth, ttt\}$
$B = \{hhh, \underline{hht}, \underline{hth}, htt\}$
$A = \{\underline{hht}, \underline{hth}, thh\}$

$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{2/8}{4/8} = \frac{1}{2}$



Theorem (Multiplication Law, TBp. 17)

Let A and B be events and assume $P(B) > 0$. Then

$$P(A \cap B) = P(A|B)P(B).$$

intuition

Sometimes, this is easier to obtain ($\because \Omega \to B$)

generalization
$P(A_1 \cap A_2 \cap \cdots \cap A_n)$
$= P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_1 \cap A_2) \cdot \cdots$

Example 1.7 (Ex. B, TBp. 18)

Suppose if it is cloudy (B), the probability that it is raining (A) is 0.3, and that the probability that it is cloudy is $P(B) = 0.2$. The probability that it is cloudy and raining is $P(A \cap B) = P(A|B)P(B) = 0.3 \times 0.2 = 0.06$.

Theorem (Law of Total Probability, TBp. 18)

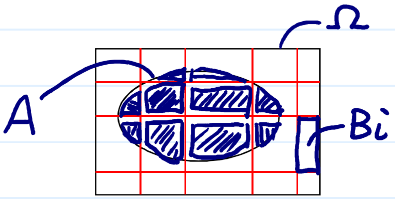
Let $B_1, B_2, \ldots, B_n$ be such that $\cup_{i=1}^n B_i = \Omega$ and $B_i \cap B_j = \emptyset$ for $i \neq j$, with $P(B_i) > 0$ for all i. Then, for any event A,

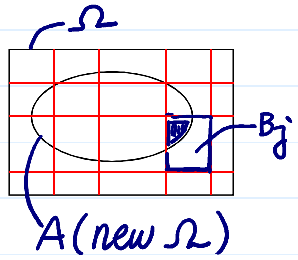
a partition of Ω

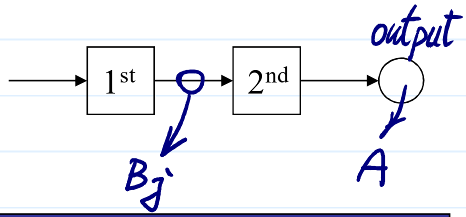
$$P(A) = \sum_{i=1}^{n} \overbrace{P(A|B_i)}^{\text{平均}}\overbrace{P(B_i)}^{\text{權重}}.$$

intuition

$P(A \cap B_i)$







Theorem (Bayes' Rule, TBp. 20)

Let A and $B_1, \ldots, B_n$ be events where the B_i are disjoint, $\cup_{i=1}^n B_i = \Omega$ and $P(B_i) > 0$ for all i. Then

update

$$\frac{P(A \cap B_j)}{P(A)} = P(B_j|A) = \frac{\boxed{P(A|B_j)P(B_j)}}{\boxed{\sum_{i=1}^n P(A|B_i)P(B_i)}}.$$

Definition (independence, TBp. 24)

Two events A and B are said to be **independent** if 獨立

definition of independence

$$P(A \cap B) = P(A)P(B).$$

A collection of events $A_1, A_2, \ldots, A_n$ are said to be **mutually independent** if for any subcollection, $A_{i_1}, \ldots A_{i_m}$,

$$P(A_{i_1} \cap \cdots \cap A_{i_m}) = P(A_{i_1}) \cdots P(A_{i_m}).$$

cf.

generalization of multiplication Law in LNp.6

cf.

When A and B are independent,

intuition of independence

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

and $P(A^c|B) = P(A^c)$.

Furthermore, $P(A|B^c) = P(A)$ and $P(A^c|B^c) = P(A^c)$.

$$\begin{tikzcd} A \arrow[r, leftrightarrow] \arrow[dr, leftrightarrow] & B \arrow[dl, leftrightarrow] \\ A^c \arrow[r, leftrightarrow] & B^c \end{tikzcd}$$

independence & complement

required

optional

Reading: textbook, Sections 1.1, 1.2, 1.3, 1.5, 1.6, 1.7

Further Reading: Roussas, Chapters 1 and 2

Chapters 2 and 3

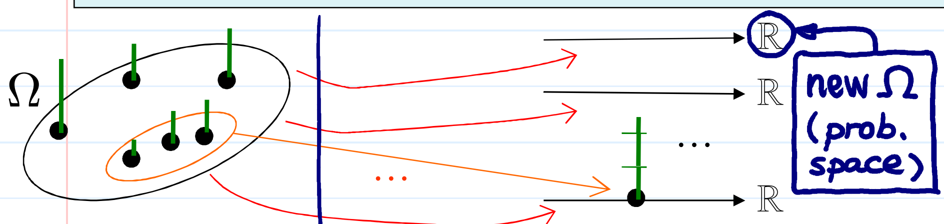
Outline

- random variables (隨機變數)
- distribution
  - discrete and continuous
  - univariate and multivariate
  - cdf, pmf, pdf
- conditional distribution
- independent random variables
- function of random variables
  - distribution of transformed r.v.
  - extrema and order statistics

• random variable

Definition 2.1 (random variable, TBp. 33)

A **random variable** is a function from Ω to the real numbers.



2 main aspects of r.v. X
- the (random) value of X
- distribution of X

digitize the world

→ statistical modeling usually starts from here.

Example 2.1 (cont. Ex. 1.1)

(1) X_1 = the total number of heads
(2) X_2 = the number of heads on the first toss
(3) X_3 = the number of heads minus the number of tails

$$\begin{array}{cccccccccc}
 & & 1/8 & 1/8 & 1/8 & 1/8 & 1/8 & 1/8 & 1/8 & 1/8 \\
\Omega & = \{ & hhh, & hht, & hth, & thh, & htt, & tht, & tth, & ttt\} \\
 & & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\
X_1 & : & \underbrace{3,}_{1/8} & \multicolumn{3}{c}{\underbrace{2, \quad 2, \quad 2,}_{3/8}} & \multicolumn{3}{c}{\underbrace{1, \quad 1, \quad 1,}_{3/8}} & \underbrace{0.}_{1/8} \\
X_2 & : & 1, & 1, & 1, & 0, & 1, & 0, & 0, & 0. \\
X_3 & : & 3, & 1, & 1, & 1, & -1, & -1, & -1, & -3.
\end{array}$$

update probability space

← new Ω

← new probability measure

Question 2.1

Why statisticians need random variables? Why they map to real line?

We need random variable because → Data Uncertainty

Data → in $\mathbb{R}^n$ space

Uncertainty → need probability measure

↳ can do "+", "−", "×", "÷", exp, log, ...

- distribution ← probability measure of r.v. — don't know what value will appear

(分配, 分布)

• For r.v., it value: random, but its distribution: fixed.

Question 2.2

A random variable have a sample space on **real line**. Does it bring some special ways to characterize its probability measure?

finite or countable infinity / uncountable

	discrete	continuous
uni-variate r.v. (one r.v.)	• pmf ⟷ • cdf • mgf/chf	• pdf • cdf • mgf/chf
multi-variate r.v.'s (at least two r.v.)	• joint pmf ⟷ • joint cdf • joint mgf/chf	• joint pdf • joint cdf • joint mgf/chf

when any of them is known, the other 2 can be obtained

pmf: probability mass function, pdf: probability density function,
cdf: cumulative distribution function

mgf (moment generating function) and chf (characteristic function) will be defined in Chapter 4

Definition 2.2 (discrete and continuous random variables, TBp. 35 and 47)

A **discrete random variable** can take on only a finite or at most a countably infinite number of values. A **continuous random variable** can take on a continuum of values. ← uncountable

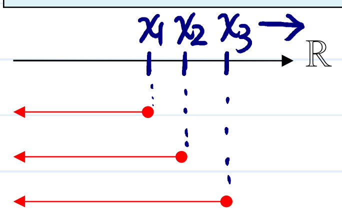
e.g.

Discrete	Continuous
$X \in \{0,1,2,3\}$	$X \in [0,1]$
$X \in \mathbb{Z}_+$	$X \in (-\infty,\infty)$

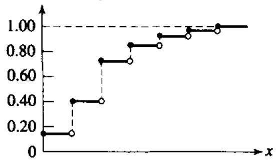
Definition 2.3 (cumulative distribution function, TBp. 36)

A function F is called the **cumulative distribution function (cdf)** of a random variable X if

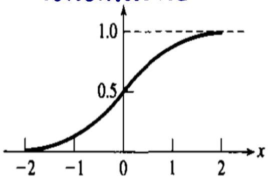
$$F(x) = P(X \leq x),\ x \in \mathbb{R}.$$



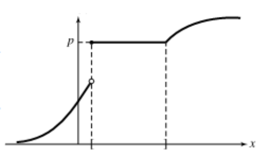
discrete



continuous



mixed.



Definition 2.4 (probability mass function/frequency function, TBp. 36)

A function $p(x)$ is called a **probability mass function (pmf)** or a **frequency function** if and only if (1) $p(x) \geq 0$ for all $x \in \mathcal{X}$, and (2) $\sum_{x\in\mathcal{X}} p(x) = 1$.

$\mathcal{X}$: a finite or countably infinite set.

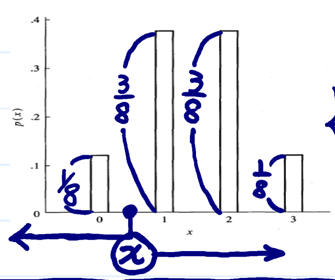
For a discrete random variable X with pmf $p(x)$,

$$P(X = x) = p(x),$$

and

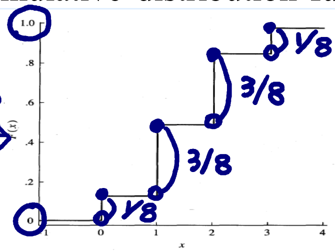
$$P(X \in A) = \sum_{x\in \underbrace{A}_{A\cap\mathcal{X}}} p(x).$$

probability mass function



relationship

cumulative distribution function



$P(X\leq 1) = \frac{1}{8} + \frac{3}{8} = \frac{4}{8} = F(1)$
$P(X<1) = \frac{1}{8} = F(1-)$
$P(X=1) = P(X\leq 1) - P(X<1) = F(1) - F(1-)$

- $F(x) = \sum_{t\leq x} P(X=t) = \sum_{t\leq x} p(t)$

  $p(x) = P(X=x) = F(x) - F(x-)$, where $F(x-) = \lim_{t\uparrow x} F(t)$

Definition 2.5 (probability density function, TBp. 46)

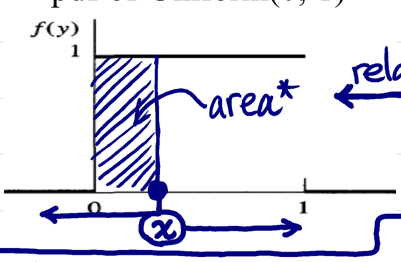
A function $f(x)$ is a **probability density function (pdf)** or **density function** if and only if (1) $f(x) \geq 0$ for all x, and (2) $\int_{-\infty}^{\infty} f(x)dx = 1$.

For a continuous random variable X with pdf f,

$$P(X \in A) = \int_A f(x)\,dx.$$

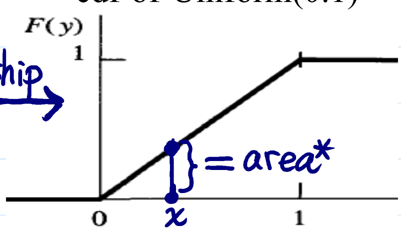
Note. pdf plays a similar role as pmf, but $\Sigma \to \int$

pdf of Uniform(0, 1)

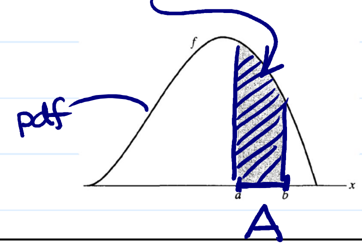


relationship

cdf of Uniform(0.1)



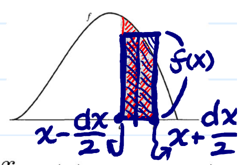
area = P(A)



- $F(x) = \int_{-\infty}^{x} f(t)\,dt$

  $f(x) = \frac{d}{dx} F(x)$

The value of a pdf can be larger than one (c.f. pmf)



(**Note.** x st $f(x) > 0$, $P(X = x) = \int_x^x f(t)dt = 0$) c.f.

Question 2.3

How to interpret $f(x)$?

For small dx, $P\left(x - \frac{dx}{2} \leq X \leq x + \frac{dx}{2}\right) = \int_{x-\frac{dx}{2}}^{x+\frac{dx}{2}} f(t)dt \approx f(x)dx$ proportional to prob.

Theorem 2.1 (properties of cdf)

If $F(x)$ is a cumulative distribution function of some random varaible X then the following properties hold.

1. $0 \le F(x) \le 1$
2. $F(x)$ is nondecreasing.
3. For any $x \in \mathbb{R}$, $F(x)$ is continuous from the right; i.e.
$$\lim_{t \downarrow x} F(t) = F(x).$$
4. $\lim_{x\to\infty} F(x) = 1$ and $\lim_{x\to-\infty} F(x) = 0$.
5. $P(X > x) = 1 - F(x)$ and $P(a < X \le b) = F(b) - F(a)$.
6. For any $x \in \mathbb{R}$, $F(x)$ has left limit. $\rightarrow F(x-) = P(X < x)$
7. There are at most countably many discontinuity points of $F(x)$.

Conversely, if a function $F(x)$ satisfies properties 2, 3, 4 then $F(x)$ is a cdf.

Question 2.4 Why need *joint* distribution for the study of multivariate r.v.'s? $\rightarrow (X_1, X_2, \cdots, X_n) \in \mathbb{R}^n$

Why several marginal distributions not enough?

Example 2.2 (cont. Ex. 2.1)

$$\Omega = \{hhh, hht, hth, thh, htt, tht, tth, ttt\}$$

$(X_1, X_2) \in \mathbb{R}^2$

X_2: # of head on 1st toss	X_1: total # of heads			
	0 (1/8)	1 (3/8)	2 (3/8)	3 (1/8)
(1/2) 0	$\frac{1}{8}\left(\frac{1}{16}\right)$	$\frac{2}{8}\left(\frac{3}{16}\right)$	$\frac{1}{8}\left(\frac{3}{16}\right)$	$0\left(\frac{1}{16}\right)$
(1/2) 1	$0\left(\frac{1}{16}\right)$	$\frac{1}{8}\left(\frac{3}{16}\right)$	$\frac{2}{8}\left(\frac{3}{16}\right)$	$\frac{1}{8}\left(\frac{1}{16}\right)$

marginal distribution

joint distribution

When $X_1 = 1$ occurs,

$P(X_2 = 0 \mid X_1 = 1) = \frac{2/8}{3/8} = \frac{2}{3}$

$P(X_2 = 1 \mid X_1 = 1) = \frac{1/8}{3/8} = \frac{1}{3}$

Note: two marginal distributions are not enough to describe their joint distribution.

Question 2.5

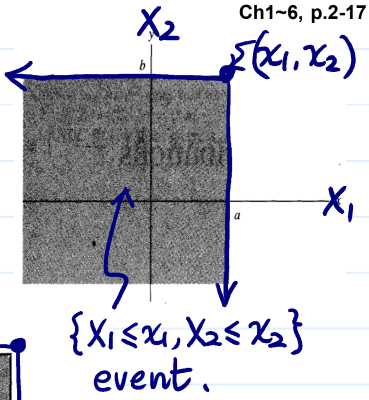
When we know the joint distribution, we can obtain every marginal distributions. Is the reverse statement true?

Definition 2.6 (joint cumulative distribution function, TBp. 71)

The **joint cdf** of $X_1, X_2, \ldots, X_n$ is

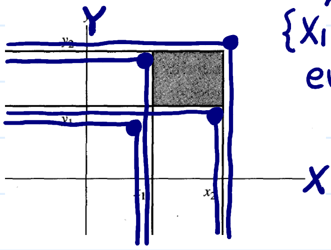
$$F(x_1, x_2, \cdots, x_n) = P(X_1 \le x_1, X_2 \le x_2, \ldots, X_n \le x_n)$$

for $x_1, x_2, \ldots, x_n \in \mathbb{R}$.



can be generalized to more than 2 r.v.'s

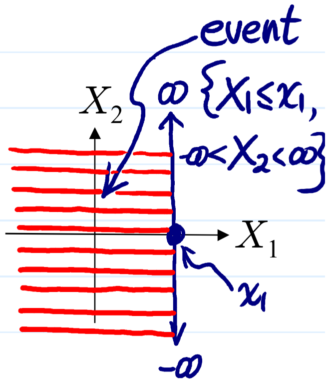
$$P(x_1 < X \le x_2, y_1 < Y \le y_2) = F(x_2, y_2) - F(x_2, y_1) - F(x_1, y_2) + F(x_1, y_1)$$



Definition 2.7 (marginal cdf, TBp. 76)

The **marginal cdf** of X_1 is

$$F_{X_1}(x_1) = P(X_1 \le x_1) = \lim_{x_2, x_3, \ldots, x_n \to \infty} F(x_1, x_2, \cdots, x_n)$$



- discrete case: marginal pmf $p_{X_1}(x) = F_{X_1}(x) - F_{X_1}(x-)$.
- continuous case: marginal pdf $f_{X_1}(x) = \frac{d}{dx} F_{X_1}(x)$.

- discrete multivariate case

cf. the similarity between pmf & pdf

$$p(x_1, x_2, \cdots, x_n) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$$

$\Rightarrow$ **joint pmf** of $X_1, X_2, \ldots, X_n$

$$P((X_1, \ldots, X_n) \in A) = \sum_{(x_1, \ldots, x_n) \in A} p(x_1, \ldots, x_n)$$

$$F(x_1, x_2, \cdots, x_n) = \sum_{t_1 \le x_1,\, t_2 \le x_2,\, \ldots,\, t_n \le x_n} p(t_1, t_2, \ldots, t_n)$$

relationship b/w joint cdf & pmf

$$p_{X_1}(x_1) = P(X_1 = x_1) = \sum_{-\infty < t_2 < \infty, \ldots, -\infty < t_n < \infty} p(x_1, t_2, \ldots, t_n)$$

relationship b/w marginal & joint pmfs

- continuous multivariate case

$$f(x_1, x_2, \cdots, x_n) = \frac{\partial^n}{\partial x_1 \cdots \partial x_n} F(x_1, x_2, \cdots, x_n)$$

$\Rightarrow$ **joint pdf** of $X_1, X_2, \ldots, X_n$

$$P((X_1, \ldots, X_n) \in A) = \int \cdots \int_A f(x_1, \ldots, x_n)\, dx_1 \cdots dx_n$$

$$F(x_1, x_2, \cdots, x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f(t_1, t_2, \ldots, t_n)\, dt_n \cdots dt_1$$

relationship b/w joint cdf & pdf

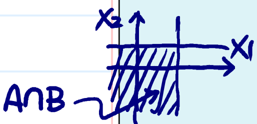
$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, t_2, \ldots, t_n)\, dt_2 \cdots dt_n$$

relationship b/w marginal & joint pdfs

• independent random variables ← Recall. independent events (LNp.8)

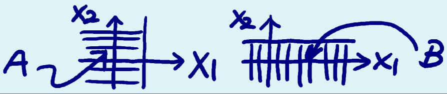
Definition 2.8 (independent random variables, TBp. 84)

Random variables $X_1, X_2, \ldots, X_n$ are said to be **independent** if their joint cdf factors into the product of their marginal cdf's



$$F(x_1, x_2, \ldots, x_n) = F_{X_1}(x_1)F_{X_2}(x_2)\cdots F_{X_n}(x_n)$$

for all $x_1, x_2, \ldots, x_n$.



$(\Rightarrow)$ $f = \frac{\partial^n}{\partial x_1 \cdots \partial x_n} F = \frac{\partial^n}{\partial x_1 \cdots \partial x_n} F_{X_1} \cdots F_{X_n} = f_{X_1} \cdots f_{X_n}$

$(\Leftarrow)$ $F = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f_{X_1} \cdots f_{X_n} = F_{X_1} \cdots F_{X_n}$

joint can be determined by marginals

Theorem 2.2 (TBp. 85-86)

1. For continuous case,

$$F(x_1, \ldots, x_n) = F_{X_1}(x_1)\cdots F_{X_n}(x_n) \Leftrightarrow f(x_1, \ldots, x_n) = f_{X_1}(x_1)\cdots f_{X_n}(x_n)$$

Note: similarity between pdf & pmf.

For discrete case,

$$F(x_1, \ldots, x_n) = F_{X_1}(x_1)\cdots F_{X_n}(x_n) \Leftrightarrow p(x_1, \ldots, x_n) = p_{X_1}(x_1)\cdots p_{X_n}(x_n)$$

2. X, Y independent
$\Leftrightarrow P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$ — For any A & B
i.e. the events $\{X \in A\}$ and $\{Y \in B\}$ are independent

For interpretation → $P(Y \in B \mid X \in A) = P(Y \in B)$

↳ No matter what data X occurs, it has no impact on the appearance probability of data Y.

3. X, Y independent $\Rightarrow$ $Z = g(X)$ and $W = h(Y)$ are independent

indep. & transformation

intuition

e.g. X:生日 $\xleftrightarrow{\text{indep.}}$ Y:身高

$\Downarrow$

$g(X)$:星座 $\xleftrightarrow{\text{indep.}}$ $h(Y)$:免費車票

> generalization
>
> $X_1, \ldots, X_n$ are independent
>
> $1 < i_0 < i_1 < \cdots < i_k = n$
>
> $$\begin{aligned} Y_1 &= g_1(X_1, \ldots, X_{i_1}), \\ Y_2 &= g_2(X_{i_1+1}, \ldots, X_{i_2}), \\ &\cdots \\ Y_k &= g_k(X_{i_{k-1}+1}, \ldots, X_{i_k}). \end{aligned}$$
>
> $Y_1, \ldots, Y_k$ are independent

☆☆ 4. marginal distributions of $X_1, X_2, \ldots, X_n$ + independence $\Rightarrow$ joint distribution of $X_1, X_2, \ldots, X_n$

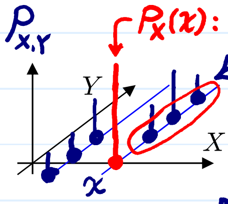
• conditional distribution ← conditional probability (LNp.5)

Definition 2.9 (conditional pmf for discrete case, TBp. 87)

X and Y are discrete random variables with joint pmf $p_{XY}(x,y)$, the **conditional pmf** of Y given X is

$$p_{Y|X}(y|x) \equiv P(\underbrace{Y=y}_{\text{event } A}|\underbrace{X=x}_{\text{event } B}) = \frac{\overbrace{P(X=x, Y=y)}^{A\cap B}}{\underbrace{P(X=x)}_{B}} = \frac{p_{XY}(x,y)}{p_X(x)} = \frac{\text{joint}}{\text{marginal}}$$

if $p_X(x) > 0$. The probability is defined to be zero if $p_X(x) = 0$.



$P_X(x)$: marginal

$\Sigma P_i = 1$ (?)

Ans: No.

LNp.16

Example 2.3 (cont. $\boxed{\text{Ex 2.2}}$)

$p_{X_2|X_1}(0|1) = 2/3$, and $p_{X_2|X_1}(1|1) = 1/3 \xleftarrow{\text{update}} \begin{cases} P_{X_2}(0) = 1/2 \\ P_{X_2}(1) = 1/2 \end{cases}$

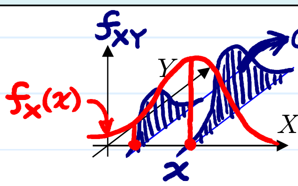
Definition 2.10 (conditional pdf for continuous case, TBp. 86)

X and Y are continuous random variables with joint pdf $f_{XY}(x,y)$, the **conditional pdf** of Y given X is defined by

$$\frac{\text{joint}}{\text{marginal}} = f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)}, \quad y \in R,$$

Notice the similarity between pmf & pdf.

if $0 < f_X(x) < \infty$ and 0 otherwise.



area $= f_X(x)$

$= \int_R f_{XY}(x,y)dy = 1$ (?)

Theorem 2.3

1. The definition of $f_{Y|X}(y|x)$ comes from

$P(a \le Y \le b, x-\frac{\Delta x}{2} \le X \le x+\frac{\Delta x}{2}) / P(x-\frac{\Delta x}{2} \le X \le x+\frac{\Delta x}{2})$

$$P(a \le Y \le b | x-\Delta x/2 \le X \le x+\Delta x/2) = \frac{\int_a^b \int_{x-\Delta x/2}^{x+\Delta x/2} f_{XY}(u,v)dudv}{\int_{x-\Delta x/2}^{x+\Delta x/2} f_X(t)dt}$$

$$\approx \frac{\int_a^b f_{XY}(x,y)\cancel{\Delta x}\, dy}{f_X(x)\cancel{\Delta x}} = \int_a^b \frac{f_{XY}(x,y)}{f_X(x)} dy$$

2. For each fixed x, $p_{Y|X}(y|x)$ is a pmf for y and $f_{Y|X}(y|x)$ is a pdf for y. ← Notice the different roles of x & y
3. $p_{XY}(x,y) = p_{Y|X}(y|x)\, p_X(x)$, and $f_{XY}(x,y) = f_{Y|X}(y|x)\, f_X(x)$
   — multiplication law ← cf. LNp.6
4. $p_Y(y) = \sum_x p_{Y|X}(y|x)p_X(x)$, and $f_Y(y) = \int_{-\infty}^{\infty} f_{Y|X}(y|x)f_X(x)dx$
   — law of total probability ← cf. LNp.7
5. $p_{X|Y}(x|y) = \dfrac{\boxed{p_{Y|X}(y|x)}\, p_X(x)}{\boxed{\sum_x p_{Y|X}(y|x)p_X(x)}}$ (update), and $f_{X|Y}(x|y) = \dfrac{\boxed{f_{Y|X}(y|x)}\, f_X(x)}{\boxed{\int_{-\infty}^{\infty} f_{Y|X}(y|x)f_X(x)dx}}$ (update),
   LNp.7 cf. → — Bayes' rule
   items 3,4,5 can be generalized to more than 2 r.v.'s
6. X, Y are independent $\Leftrightarrow$ $p_{Y|X}(y|x) = p_Y(y)$ or $f_{Y|X}(y|x) = f_Y(y)$
   $\boxed{\text{intuition (graphs in LNp.21 \& 22)}}$

- **functions of random variables**

Raw Data: $X_1, \ldots, X_n$

Transformations
$g_1(X_1, \ldots, X_n) = Y_1$
$\ldots,$
$g_k(X_1, \ldots, X_n) = Y_k$ ← r.v.

Extract Information → Θ ← unknown parameters in the statistical model

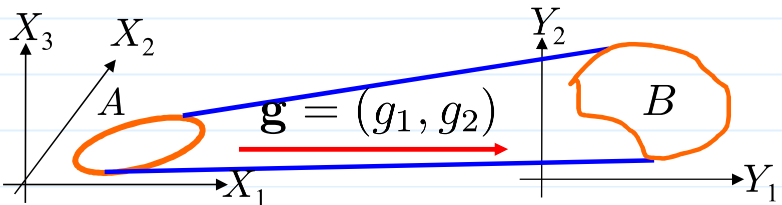
Question 2.6

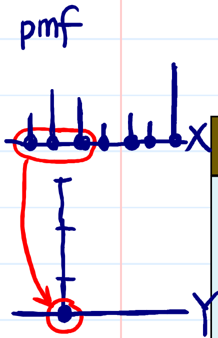
For given r.v.'s $X_1, \ldots, X_n$, how to derive the distributions of their transformations?

1. method of events → discrete r.v.'s (pmf)

Theorem 2.7

Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be random variables, and $\mathbf{Y} = \mathbf{g}(\mathbf{X})$. Then, the distribution of $\mathbf{Y}$ is determined by the distribution of $\mathbf{X}$ as follow: for any event B defined by $\mathbf{Y}$, $P(\mathbf{Y} \in B) = P(\mathbf{X} \in A)$, where $A = \mathbf{g}^{-1}(B)$.
(P ← P_Y; P ← P_X)





Example 2.4 (univariate discrete random variable)

Let X be a discrete r.v. taking the values x_i, $i = 1, 2, \ldots$, and $Y = g(X)$. Then, Y is also a discrete r.v. taking the values y_j, $j = 1, 2, \ldots$. To determine the pmf of Y, by taking $B = \{y_j\}$, we have

$$A = \{x_i : g(x_i) = y_j\} \text{ and hence}$$

$$p_Y(y_j) = P(\{y_j\}) = P(A) = \sum_{x_i \in A} p_X(x_i).$$

Example 2.5 (sum of two discrete random variables, TBp. 96)

X and Y are random variables with joint pmf $p(x, y)$. Find the distribution of $Z = X + Y$.

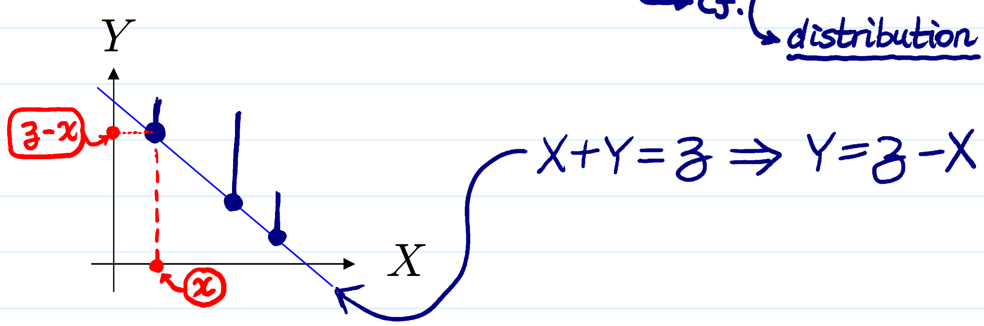
(Exercise: difference of two random variables, $Z=X-Y$) ← Ans. $p_Z(z) = \sum_y p(z+y, y)$

$$p_Z(z) = P(Z = z) = P(X + Y = z) = \sum_{x=-\infty}^{\infty} p(x, z - x)$$

When X, Y independent, $p(x, y) = p_X(x)p_Y(y)$,

$$p_Z(z) = \sum_{x=-\infty}^{\infty} p_X(x)p_Y(z - x) \quad \Rightarrow \textbf{convolution of } p_X \textbf{ and } p_Y$$

cf. value of r.v. / distribution of r.v.



2. method of cumulative distribution function (a special case of method 1)

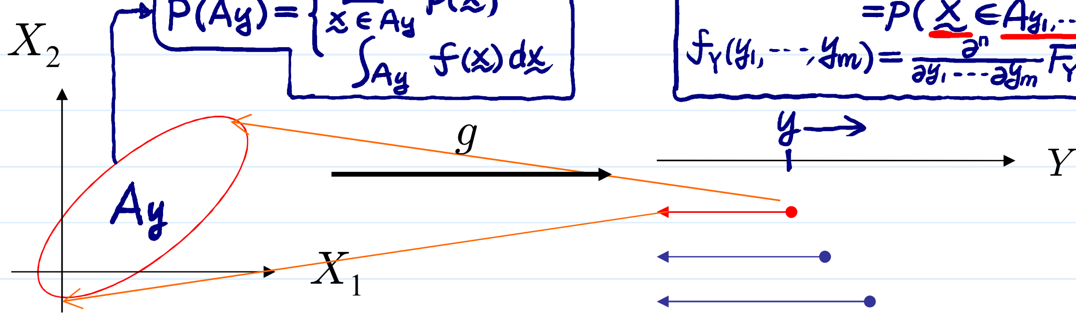
Let Y be a function of the random variables $X_1, X_2, \ldots, X_n$.

1. Find the region $Y \le y$ in the $(x_1, x_2, \ldots, x_n)$ space. ← A_y →
2. Find $F_Y(y) = P(Y \le y)$ by summing the joint pmf or integrating the joint pdf of $X_1, X_2, \ldots, X_n$ over the region $Y \le y$.
3. (for continuous case) Find the pdf of Y by differentiating $F_Y(y)$, i.e., $f_Y(y) = \frac{d}{dy}F_Y(y)$.

Note. It can be generalized to multivariate $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_m)$.

$$P(A_y) = \begin{cases} \sum_{\underline{x} \in A_y} p(\underline{x}) \\ \int_{A_y} f(\underline{x})\, d\underline{x} \end{cases}$$

$F_Y(y_1, \cdots, y_m) = P(Y_1 \le y_1, \cdots, Y_m \le y_m)$
$= P(\underline{X} \in A_{y_1, \cdots, y_m})$
$f_Y(y_1, \cdots, y_m) = \frac{\partial^n}{\partial y_1 \cdots \partial y_m} F_Y(y_1, \cdots, y_m)$



Example 2.6 (square of a random variable, similar example see TBp. 61)

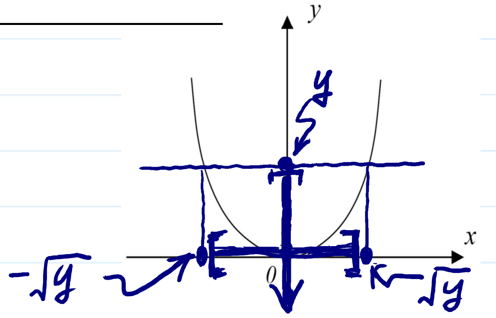
X is a random variables with pdf $f_X(x)$ and cdf $F_X(x)$. Find the distributon of $Y = X^2$. ↳ X is a continuous r.v.

For $y \geq 0$, $\{Y \leq y\} = \{-\sqrt{y} \leq X \leq \sqrt{y}\}$

$$F_Y(y) = P(Y \leq y) = P(-\sqrt{y} \cancel{\leq} X \leq \sqrt{y}) = F_X(\sqrt{y}) - F_X(-\sqrt{y})$$

$$\begin{aligned} f_Y(y) = \frac{d}{dy}F_Y(y) &= \frac{d}{dy}F_X(\sqrt{y}) - \frac{d}{dy}F_X(-\sqrt{y}) \\ &= f_X(\sqrt{y})\frac{1}{2\sqrt{y}} - f_X(-\sqrt{y})(-\frac{1}{2\sqrt{y}}) \\ &= \frac{1}{2\sqrt{y}}(f_X(\sqrt{y}) + f_X(-\sqrt{y})) \end{aligned}$$

and $f_Y(y) = 0$ for $y < 0$.



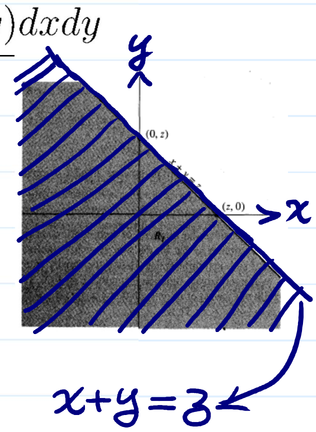
Example 2.7 (sum of two continuous random variables, TBp. 97)

X and Y are random variables with joint pdf $f(x, y)$. Find the distribution of $Z = X + Y$. ↳ X, Y: continuous r.v.'s

(Exercise: difference of two random variables, $Z = X - Y$) ← Ans. $f_Z(z) = \int_{-\infty}^{\infty} f(z+y, y)\,dy$

Let R_z be $\{(x, y) : x + y \leq z\}$. Then,

$$\begin{aligned} F_Z(z) &= P(Z \leq z) = P(X + Y \leq z) = \iint_{R_z} f(x, y)\,dx\,dy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{z-x} f(x, y)\,dy\,dx \\ &= \int_{-\infty}^{z}\int_{-\infty}^{\infty} f(x, v - x)\,dx\,dv \quad (\text{set } \begin{cases} y = v - x \\ x = x \end{cases}) \\ f_Z(z) &= \frac{d}{dz}F_Z(z) = \int_{-\infty}^{\infty} f(x, z - x)\,dx \end{aligned}$$



When X, Y independent, $f(x, y) = f_X(x)f_Y(y)$,

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x)f_Y(z-x)\,dx \quad \Rightarrow \textbf{ convolution of } f_X \textbf{ and } f_Y$$

$\boxed{\text{cf.}}$ → the convolution for discrete r.v.'s (LNp.25)

Example 2.8 (quotient of two continuous random variables, TBp. 98)

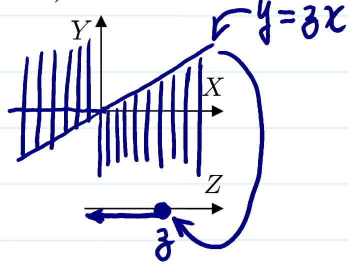
X and Y are r.v. with joint pdf $f(x,y)$. Find the distribution of $Z = Y/X$. (Exercise: product of two random variables, $Z=XY$)

$$Q_z = \{(x,y) : y/x \le z\} = \{(x,y) : x < 0, y \ge zx\} \cup \{(x,y) : x > 0, y \le zx\}$$

$P(Z \le z)$

$$\begin{aligned}
F_Z(z) &= \iint_{Q_z} f(x,y)\,dx\,dy = \int_{-\infty}^{0}\int_{xz}^{\infty} + \int_{0}^{\infty}\int_{-\infty}^{xz} f(x,y)\,dy\,dx \\
&= \int_{-\infty}^{0}\int_{z}^{-\infty} + \int_{0}^{\infty}\int_{-\infty}^{z} x f(x,xv)\,dv\,dx \quad (\text{set } y = xv) \\
&= \int_{-\infty}^{0}\int_{-\infty}^{z} (-x) f(x,xv)\,dv\,dx + \int_{0}^{\infty}\int_{-\infty}^{z} x f(x,xv)\,dv\,dx \\
&= \int_{-\infty}^{z}\int_{-\infty}^{\infty} |x| f(x,xv)\,dx\,dv
\end{aligned}$$

$\{x = x$



$$f_Z(z) = \frac{d}{dz}F_Z(z) = \int_{-\infty}^{\infty} |x| f(x,xz)\,dx$$

$$\left(= \int_{-\infty}^{\infty} |x| f_X(x) f_Y(xz)\,dx \quad \text{when } X, Y \text{ independent} \right)$$

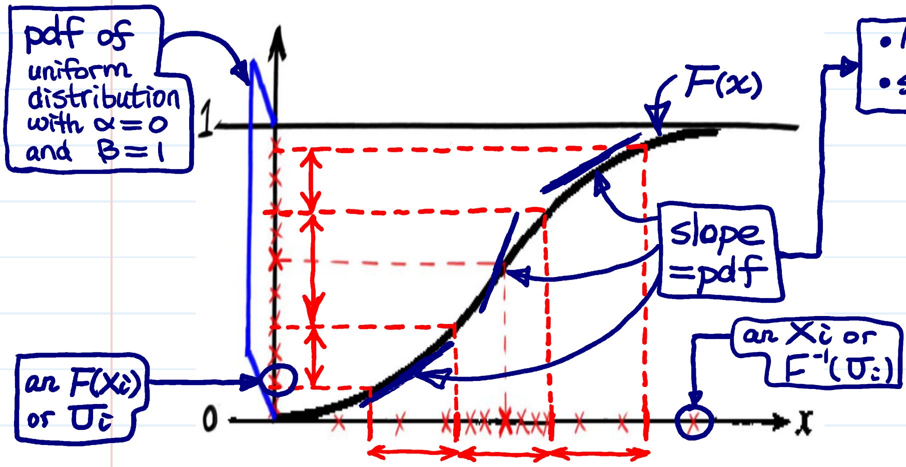
Theorem 2.4 (TBp. 63)

Let X be a random variable whose cdf F possesses a unique inverse F^{-1}. Let $Z = F(X)$, then Z has a uniform distribution on $[0,1]$.

① no jump ② strictly increasing ⇒ X : a continuous r.v.

Theorem 2.5 (TBp. 63)

Let U be a uniform random variable on $[0,1]$ and F is a cdf which possesses a unique inverse F^{-1}. Let $X = F^{-1}(U)$. Then the cdf of X is F.



• larger slope ⇒ more Xi's
• smaller slope ⇒ fewer Xi's

Note. The 2 theorems are useful for generating pseudo-random numbers in computer simulation (the concepts can be generalized to any r.v.'s).

3. method of probability density function (for continuous r.v.'s and differentiable, one-to-one transformations, a special case of method **2**) :

check its proof in textbook

Theorem 2.6 (univariate continuous case, TBp. 62)

Let X be a continuous random variable with pdf $f_X(x)$. Let $Y = g(X)$, where g is differentiable, strictly monotone. Then,

can be relaxed to piecewise strictly monotone

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{dg^{-1}(y)}{dy} \right|$$

cf. Example 2.4 in LNp.24
Q: What's the role of the term?

for y s.t. $y = g(x)$ for some x, and $f_Y(y) = 0$ otherwise.

Example 2.9

X is a random variables with pdf $f_X(x)$. Find the distributon of $Y = 1/X$.

For $x > 0$ (or $x < 0$),

$$y = 1/x \equiv g(x) \quad \Rightarrow \quad x = g^{-1}(y) = 1/y$$

$$dg^{-1}/dy = -1/y^2 \quad \text{and} \quad |dg^{-1}/dy| = 1/y^2$$

hence

$$f_Y(y) = f_X(1/y)(1/y^2)$$

Theorem 2.7 (multivariate continuous case, TBp. 102-103)

$\mathbf{X} = (X_1, X_2, \ldots, X_n)$ multivariate continuous, $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n) \equiv \mathbf{g}(\mathbf{X})$. $\mathbf{g}$ is one-to-one, so that its inverse exists and is denoted by

①

$$\mathbf{x} = \mathbf{g}^{-1}(\mathbf{y}) = \mathbf{w}(\mathbf{y}) = (\underset{x_1}{w_1(\mathbf{y})}, \underset{x_2}{w_2(\mathbf{y})}, \ldots, \underset{x_n}{w_n(\mathbf{y})}).$$

Assume $\mathbf{w}$ have continuous partial derivatives, and let

②

$$\underbrace{J}_{\text{Jacobin}} = \underbrace{\begin{vmatrix} \frac{\partial w_1(\mathbf{y})}{\partial y_1} & \frac{\partial w_1(\mathbf{y})}{\partial y_2} & \cdots & \frac{\partial w_1(\mathbf{y})}{\partial y_n} \\ \frac{\partial w_2(\mathbf{y})}{\partial y_1} & \frac{\partial w_2(\mathbf{y})}{\partial y_2} & \cdots & \frac{\partial w_2(\mathbf{y})}{\partial y_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial w_n(\mathbf{y})}{\partial y_1} & \frac{\partial w_n(\mathbf{y})}{\partial y_2} & \cdots & \frac{\partial w_n(\mathbf{y})}{\partial y_n} \end{vmatrix}}_{\text{determinant}}$$

interpretation: similar to $\left|\frac{dg^{-1}}{dy}\right|$

Then

absolute value

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}(\mathbf{g}^{-1}(\mathbf{y}))|J|.$$

for $\mathbf{y}$ s.t. $\mathbf{y} = \mathbf{g}(\mathbf{x})$ for some $\mathbf{x}$, and $f_{\mathbf{Y}}(\mathbf{y}) = 0$, otherwise.

Note. When the dimensionality of $\mathbf{Y}$, denoted by k, is less than n, we can choose another $n-k$ transformations $\mathbf{Z}$ such that $(\mathbf{Y}, \mathbf{Z})$ satisfy the above assumptions. By integrating out the last $n-k$ arguments in the pdf of $(\mathbf{Y}, \mathbf{Z})$, the pdf of $\mathbf{Y}$ can be obtained.

Example 2.10 (cont. Ex 2.8)

X_1 and X_2 are random variables with joint pdf $f_{X_1X_2}(x_1, x_2)$. Find the distribution of $Y_1 = X_2/X_1$. (Exercise: $Y_1 = X_1X_2$)

Let $Y_2 = X_1$. Then

$$\begin{aligned} x_1 &= y_2 &\equiv w_1(y_1, y_2) \\ x_2 &= y_1y_2 &\equiv w_2(y_1, y_2). \end{aligned}$$

$$\frac{\partial w_1}{\partial y_1} = 0, \quad \frac{\partial w_1}{\partial y_2} = 1, \quad \frac{\partial w_2}{\partial y_1} = y_2, \quad \frac{\partial w_2}{\partial y_2} = y_1.$$

$$J = \begin{vmatrix} 0 & 1 \\ y_2 & y_1 \end{vmatrix} = -y_2, \quad \text{and} \quad |J| = |y_2|$$

Therefore,

$$f_{Y_1Y_2}(y_1, y_2) = f_{X_1X_2}(y_2, y_1y_2)|y_2|$$

$$f_{Y_1}(y_1) = \int_{-\infty}^{\infty} f_{Y_1Y_2}(y_1, y_2)dy_2 = \int_{-\infty}^{\infty} f_{X_1X_2}(y_2, y_1y_2)|y_2|dy_2$$

cf. Ex2.8 in LNp.29

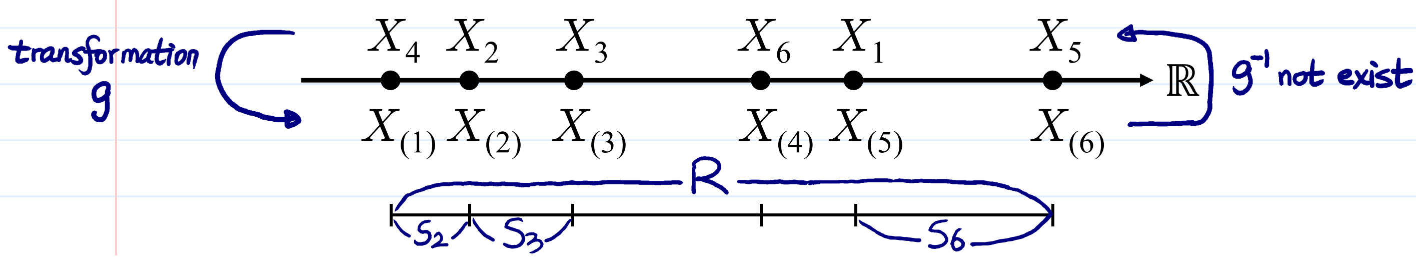
4. **method of moment generating function:** based on the *uniqueness theorem* of moment generating function. To be explained later in Chapter 4.

- extrema and order statistics 順序統計量 → quantile (分位數)

Definition 2.11 (order statistics, sec 3.7)

Let $X_1, X_2, \ldots, X_n$ be random variables. We sort the X_i's and denote by $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$ the **order statistics**. Using the notation,

$$\begin{aligned} X_{(1)} &= \min(X_1, X_2, \ldots, X_n) && \text{is the } \textbf{minimum} \\ X_{(n)} &= \max(X_1, X_2, \ldots, X_n) && \text{is the } \textbf{maximum} \\ R &\equiv X_{(n)} - X_{(1)} && \text{is called } \textbf{range} \\ S_j &\equiv X_{(j)} - X_{(j-1)}, j = 2, \ldots, n && \text{are called } j\text{th } \textbf{spacings} \end{aligned}$$



Note. In the section, we only consider the case that $X_1, X_2, \ldots, X_n$ are i.i.d continuous r.v.'s with cdf F and pdf f. Although $X_1, X_2, \ldots, X_n$ are independent, their order statistics are not independent in general. $\boxed{X_{(1)}, \cdots, X_{(n)}}$

Definition 2.12 (i.i.d.)

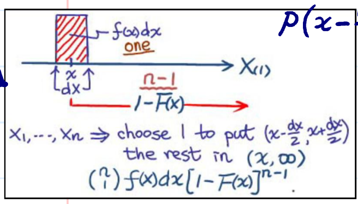
$X_1, X_2, \ldots, X_n$ are **i.i.d.** (idependent, identically distributed) with cdf F/pmf p/pdf $f \Rightarrow X_1, X_2, \ldots, X_n$ are independent and have a common marginal cdf F/pmf p/pdf f. ↳ joint $= \Pi$ marginal

$\boxed{\text{but not common value}}$

Theorem 2.8 (TBp. 104)

The cdf of $X_{(1)}$ is $1-[1-F(x)]^n$ and its pdf is $nf(x)[1-F(x)]^{n-1}$.

The cdf of $X_{(n)}$ is $[F(x)]^n$ and its pdf is $nf(x)[F(x)]^{n-1}$.



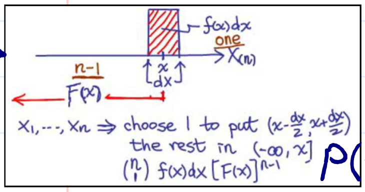
$P(x-\frac{dx}{2} < X_{(1)} < x+\frac{dx}{2}) \approx f_{X_{(1)}}(x)dx$

$$\begin{aligned} F_{X_{(n)}}(x) &= P(X_{(n)} \le x) = P(X_1 \le x, \ldots, X_n \le x) \\ &= P(X_1 \le x) \cdots P(X_n \le x) \\ &= [F(x)]^n. \end{aligned}$$

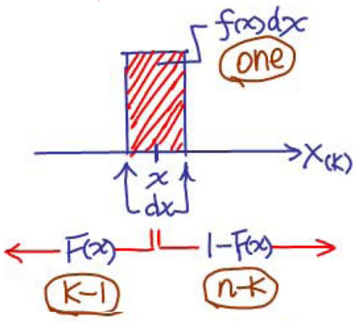
$\frac{dF_{X_{(n)}}(x)}{dx}$

$$\begin{aligned} 1 - F_{X_{(1)}}(x) &= P(X_{(1)} > x) = P(X_1 > x, \ldots, X_n > x) \\ &= P(X_1 > x) \cdots P(X_n > x) \\ &= [1-F(x)]^n. \end{aligned}$$

$\frac{dF_{X_{(1)}}(x)}{dx}$



$P(x-\frac{dx}{2} < X_{(n)} < x+\frac{dx}{2}) \approx f_{X_{(n)}}(x)\,dx$



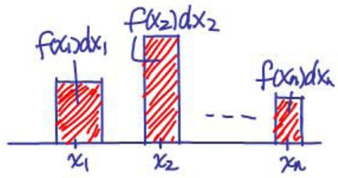
Theorem 2.9 (TBp. 105)

The pdf of the kth order statistic $X_{(k)}$ is

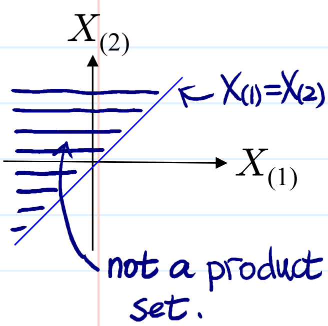
$P(x-\frac{dx}{2} < X_{(k)} < x+\frac{dx}{2}) \approx f_{X_{(k)}}(x) \cdot dx$

$$f_{X_{(k)}}(x) = \frac{n!}{(k-1)!(n-k)!} f(x)[F(x)]^{k-1}[1-F(x)]^{n-k}.$$

$X_1, \cdots, X_n \Rightarrow$ choose 1 to place in $(x-\frac{dx}{2}, x+\frac{dx}{2})$
" $k-1$ " " " $(-\infty, x)$
" $n-k$ " " " (x, ∞)
$\binom{n}{1 \;\; k-1 \;\; n-k} f(x)dx\,[F(x)]^{k-1}[1-F(x)]^{n-k}$



$\binom{n}{1 \; 1 \cdots 1} f(x_1)dx_1 \cdots f(x_n)dx_n$



Theorem 2.10 (TBp. 114, Problem 73)

The joint pdf of $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ is

$P(x_i - \frac{dx_i}{2} < X_{(i)} < x_i + \frac{dx_i}{2}, i=1, \cdots, n) \approx f_{X_{(1)} \cdots X_{(n)}}(x_1, \cdots, x_n)dx_1 \cdots dx_n$

$$f_{X_{(1)}X_{(2)}\ldots X_{(n)}}(x_1, x_2, \ldots, x_n) = n!f(x_1)f(x_2)\cdots f(x_n),$$

for $x_1 \le x_2 \le \cdots \le x_n$, and $f_{X_{(1)}X_{(2)}\ldots X_{(n)}} = 0$ otherwise.

Question: Are $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ independent, judged from the from of its joint pdf? ← C.f. Thm 2.2, item 1 (LNp.19)

Example 2.11 (range, TBp. 105-106)

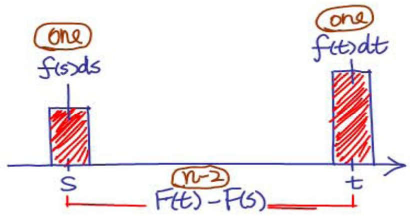
The joint pdf of $X_{(1)}$ and $X_{(n)}$ is $P(s-\frac{ds}{2}<X_{(1)}<s+\frac{ds}{2}, t-\frac{dt}{2}<X_{(n)}<t+\frac{dt}{2}) \approx f_{X_{(1)},X_{(n)}}(s,t)\,ds\,dt.$

$$f_{X_{(1)}X_{(n)}}(s,t) = n(n-1)f(s)f(t)[F(t)-F(s)]^{n-2}, \quad \text{for } s \le t,$$

and 0 otherwise. Therefore, the pdf of $R = X_{(n)} - X_{(1)}$ is

check exercise in Ex2.7(LNp.28)

$$f_R(r) = \int_{-\infty}^{\infty} f_{X_{(1)}X_{(n)}}(s, s+r)ds \quad \text{for } r > 0, \text{ and } f_R(r) = 0, \text{ otherwise.}$$



$X_1, \dots, X_n \Rightarrow$ choose one to place in $(s-\frac{ds}{2}, s+\frac{ds}{2})$
one $(t-\frac{dt}{2}, t+\frac{dt}{2})$
the rest in (s,t)
$\binom{n}{1\ 1\ n-2} f(s)ds\, f(t)dt\, [F(t)-F(s)]^{n-2}$

Exercise

1. Find the joint pdf of $X_{(i)}$ and $X_{(j)}$, where $i < j$.
2. Find the joint pdf of $X_{(j)}$ and $X_{(j-1)}$, and derive the pdf of jth spacing $S_j = X_{(j)} - X_{(j-1)}$.

❖ **Reading**: textbook, 2.1 (not including 2.1.1~5), 2.2 (not including 2.2.1~4), 2.3, 2.4, Chapter 3

❖ **Further Reading**: Roussas, 3.1, 4.1, 4.2, 7.1, 7.2, 9.1, 9.2, 9.3, 9.4, 10.1

Chapter 4

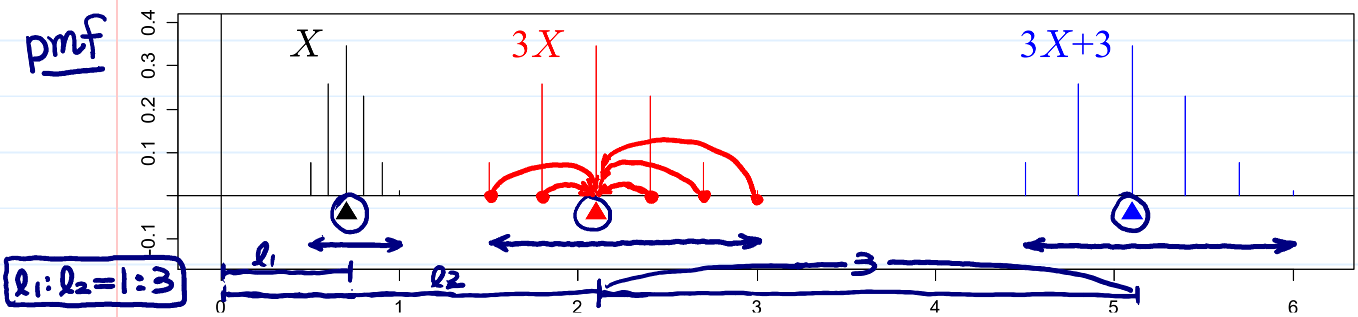
Outline

- expectation ← 期望值.
  - mean, variance, standard deviation, covariance, correlation coefficient
- moment generating function & characteristic function
- conditional expectation and prediction
- δ method

Question 3.1

Can we describe the characteristics of distributions by use of some intuitive and meaningful simple values?

pmf



• expectation

Definition 3.1 (expectation, TBp. 122, 123)

For random variables $X_1, \ldots, X_n$, the **expectation** of a univariate random variable $Y = g(X_1, \ldots, X_n)$ is defined as

$g: \mathbb{R}^n \to \mathbb{R}^1$

$$E(Y) \equiv \sum_{-\infty<y<\infty} y p_Y(y) = E[g(X_1, \ldots, X_n)]$$

$$\equiv \sum_{-\infty<x_1<\infty, \ldots, -\infty<x_n<\infty} g(x_1, \ldots, x_n) p(x_1, \ldots, x_n),$$

weighted average
加權平均
平均: y
權重: p_Y / f_Y

if $X_1, X_2, \ldots, X_n$ are discrete random variables, or

Y: random
E(Y): fixed value

$$E(Y) \equiv \int_{-\infty}^{\infty} y f_Y(y) dy = E[g(X_1, \ldots, X_n)]$$

$$\equiv \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, \ldots, x_n) f(x_1, \ldots, x_n) dx_1 \cdots dx_n,$$

if Y and $X_1, X_2, \ldots, X_n$ are continuous random variables.

Definition 3.2 (mean, variance, standard deviation, covariance, correlation coefficient)

1. (TBp.116&118) $g(x) = x \Rightarrow E[g(X)] = E(X)$ is called **mean** of X, usually denoted by $E(X)$ or μ_X.

2. (TBp.131) $g(x) = (x - \overset{\text{constant}}{\mu_X})^2 \Rightarrow E[g(X)] = E[(X - E(X))^2]$ is called **variance** of X, usually denoted by $Var(X)$ or σ_X^2. The square root of variance, i.e., σ_X, is called **standard deviation**.

   constant, not random

3. (TBp.138) $g(x, y) = (x - \mu_X)(y - \mu_Y) \Rightarrow E[g(X, Y)] = E[(X - E(X))(Y - E(Y))]$ is called **covariance** of X and Y, usually denoted by $Cov(X, Y)$ or σ_{XY}.

4. (TBp.142) The **correlation coefficient** of X, Y is defined as $\sigma_{XY}/(\sigma_X \sigma_Y)$, usually denoted by $Cor(X, Y)$ or ρ_{XY}. X and Y are called **uncorrelated** if $\rho_{XY} = 0$. $\Leftrightarrow \sigma_{XY} = 0$

Notes. (intuitive explanation of mean)

from its definition

1. Mean of a random variable parallels the notion of a weighted average.
2. It is helpful to think of the mean as the center of mass of the pmf/pdf. ← center of gravity (重心)
3. Mean can be interpreted as a long-run average. (see Chapter 5.) → LLN

Notes. (intuitive explanation of variance and standard deviation)

from its definition → how the dist. is spread out

1. variance is the average value of the squared deviation of X from μ_X.
2. If X has units, then mean and standard deviation have the same unit, and variance has unit squared.

Theorem 3.1 (properties of mean)

1. (TBp.125) For constants $a, b_1, \ldots, b_n \in \mathbb{R}$,
$$E\left(a + \sum_{i=1}^n b_i X_i\right) = a + \sum_{i=1}^n b_i E(X_i).$$
$\Rightarrow E(a+bX) = a + b \cdot E(X)$

2. (TBp.124) If X, Y are independent, then
$$E(\underbrace{g(X)}_{W}\underbrace{h(Y)}_{Z}) = E(g(X))E(h(Y)).$$
← W & Z are independent.

independent ⇒ uncorrelated

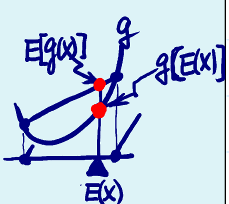
In particular, $E(XY) = E(X)E(Y)$.

(**Question 3.2**: $E(X/Y) = E(X)/E(Y)$? ← $E(\frac{X}{Y}) = E(X \cdot \frac{1}{Y}) = E(X) \cdot E(\frac{1}{Y})$, $\frac{1}{E(Y)}$?

Note. $E[g(X)] \neq g[E(X)]$ in general.

g: convex $E[g(X)] \geq g[E(X)]$

g: concave $E[g(X)] \leq g[E(X)]$



Theorem 3.2 (properties of variance and standard deviation)

1. (TBp.132) $\sigma_X^2 = Var(X) = E[(X-\mu_X)^2] = E(X^2) - \mu_X^2$. ← $[E(X)]^2$
→ for calculation purpose

2. (TBp.131) $Var(a+bX) = b^2 Var(X)$, $a, b \in \mathbb{R}$, and $\sigma_{a+bX} = |b|\sigma_X$.
→ • location shift ⇒ no impact on σ^2
• scale change ⇒ $\sigma^2 \to b^2\sigma^2$

3. (TBp.140)
$$Var\left(a + \sum_{i=1}^n b_i X_i\right) = \sum_{i=1}^n b_i^2 Var(X_i) + 2 \sum_{1 \leq i < j \leq n} b_i b_j Cov(X_i, X_j).$$
a → gone

$[b_1 \cdots b_n] \left[a_{ij} = cov(X_i, X_j) \right] \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}$ covariance matrix, $var(X_i)$

In particular, $Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$.

4. (TBp.140) If $X_1, \ldots, X_n$ are independent,
$$Var\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n Var(X_i).$$
imply → $cov(X_i, X_j) = 0$, i.e., uncorrelated, $\forall i \neq j$

cf. mean of sum item 1, Thm3.1 (LNp.41)

5. (TBp.136) $E[(X \overset{-\mu_X+\mu_X}{-} \theta)^2] = Var(X) + (\mu_X - \theta)^2$ (Mean square error = variance + bias square) $\Rightarrow = E[(X-\mu_X)^2 + (\mu_X-\theta)^2 - 2(\mu_X-\theta)(X-\mu_X)]$

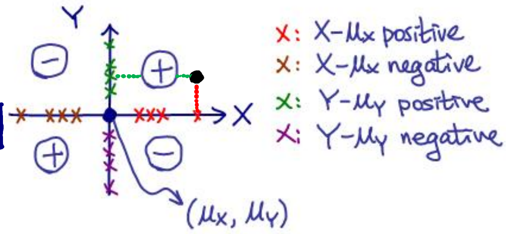
Notes. (intuitive explanation of covariance and correlation coefficient)

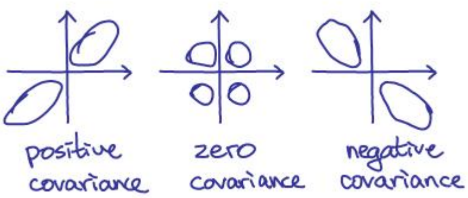
1. covariance is a measure of the joint variability of X and Y, or their degree of association. → i.e., when X (r.v.) is large (or small), will Y tend to be larger or smaller?

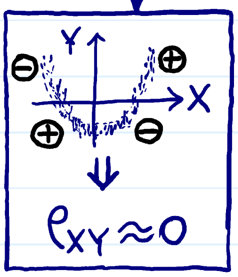
   might not be causal relation

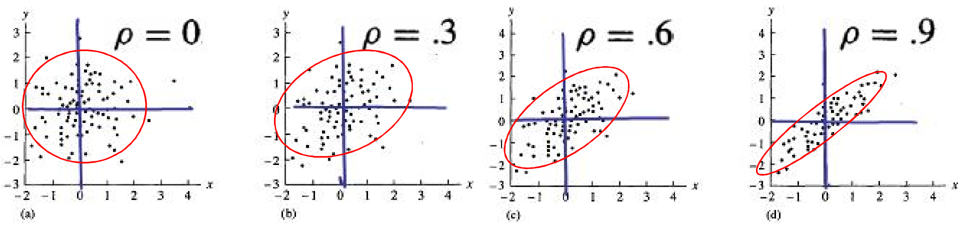
2. covariance is the average value of the product of the deviation of X from its mean and the deviation of Y from its mean. ← from its definition.
3. positive covariance and negative covariance → drawback: cov depends on the scale/unit of X & Y
4. correlation coefficient is unit free
5. correlation coefficient measures the strength of the linear relationship between X and Y.

$\mathrm{cov}(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]$









Theorem 3.4 (properties of covariance and correlation coefficient)

1. (TBp.138) $Cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] = E(XY) - \mu_X\mu_Y$
   (Note. $Cov(X,X) = Var(X)$.)
   for calculation purpose
2. (TBp.140)
   $[b_1 \cdots b_n]\left[\ \sigma_{ij} = cov(X_i, Y_j)\ \right]\begin{bmatrix} d_1 \\ \vdots \\ d_m \end{bmatrix}$
   $$Cov\left(a + \sum_{i=1}^{n} b_i X_i,\ c + \sum_{j=1}^{m} d_j Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} b_i d_j Cov(X_i, Y_j)$$
   $\boxed{\text{gone}}$
3. (TBp.140) If X, Y are independent then $Cov(X,Y) = 0$, i.e., **independent** $\Rightarrow$ **uncorrelated**. But, the converse statement is not necessarily true.
4. (TBp.143) $-1 \le \rho_{XY} \le 1$ and $\rho_{XY} = \pm 1$ if and only if $Y = aX + b$ with probability one for some a, $b \in \mathbb{R}$.
   $\begin{cases} \rho = +1 \Leftrightarrow a > 0 \\ \rho = -1 \Leftrightarrow a < 0 \end{cases}$
5. $\rho_{XY} = E\left[\left(\dfrac{X-\mu_X}{\sigma_X}\right)\left(\dfrac{Y-\mu_Y}{\sigma_Y}\right)\right]$
   standardization (標準化)
   After standardization,
   mean $= 0$
   var $= 1$
6. $|Cor(a + bX, c + dY)| = |Cor(X,Y)|$
   - location shift
   - scale change

   $\Rightarrow$ no impact on |cor|

• moment generating function & characteristics function

Definition 3.3 (moment generating function, TBp. 155)

The **moment generating function (mgf)** of a random variable X is

$$M_X(t) = E(e^{tX}), \quad t \in \mathbb{R}$$

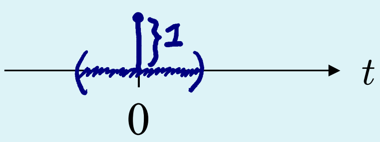
$M_X(t) = \begin{cases} \int e^{tx} f_X(x)\,dx \\ \sum e^{tx} p_X(x) \end{cases}$

Laplace transformation of

if the expectation exists.

Theorem 3.5 (properties of moment generating function)

1. The moment generating function may or may not exist for any particular value of t.
   $t = 0 \Rightarrow E(e^{0\cdot X}) = 1 \leftarrow$ always exists
   i.e., $E(e^{tX}) < \infty$
2. **uniqueness theorem** (**TBp.143**). If the moment generating function exists for t in an open interval containing zero, it uniquely determines the probability distribution.



know mgf $\Rightarrow$ know distribution

3. (TBp.156) If the moment generating function exists in an open interval containing zero, then

$$M_X^{(k)}(0) = E(X^k).$$

know all moments ⇒ know $M_X(t) = \sum_{k=0}^{\infty} \frac{M_X^{(k)}(0)}{k!} t^k$ ⇒ know dist.

the reason why it's called moment generating function.

4. (TBp.158) For any constants a, b, $M_{a+bX}(t) = e^{at} M_X(bt)$.

5. (TBp.159) X, Y independent $\Rightarrow M_{X+Y}(t) = M_X(t) M_Y(t)$.

useful for identifying the dist. of $X_1 + \cdots + X_n$

generalization: indep. $X_1, \ldots, X_n$, $M_{X_1+\cdots+X_n}(t) = \prod_{i=1}^{n} M_{X_i}(t)$

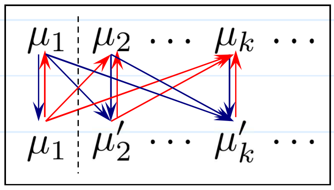
6. continuity theorem (see Chapter 5)

Definition 3.4 (moment, TBp. 155)

The kth **moment** of a random variable is $E(X^k) \equiv \mu_k$, and the kth **central moment** is $E[(X - \mu_X)^k] \equiv \mu'_k$. $(-\mu_X + \mu_k)$

➢ Some Notes.

- $\mu'_k = \sum_{i=0}^{k} \binom{k}{i} (-\mu_X)^{n-i} \mu_i$. ← μ'_k: a linear combination of $\mu_1, \cdots, \mu_k$
- $\mu_k = \sum_{i=0}^{k} \binom{k}{i} (\mu_X)^{n-i} \mu'_i$. ← μ_k: a linear combination of $\mu_1, \mu'_2, \cdots, \mu'_k$
- In particular, $E(X) = \mu_X = \mu_1$, and, $Var(X) = \sigma_X^2 = \mu_2 - \mu_1^2 = \mu'_2$.



Definition 3.5 (joint moment generating function, TBp. 161)

For random variables $X_1, X_2, \ldots, X_n$, their **joint mgf** is defined as:

$$M_{X_1X_2\cdots X_n}(t_1, t_2, \ldots, t_n) = E(e^{t_1X_1 + t_2X_2 + \cdots + t_nX_n})$$

if the expection exists.

cf. mgf of $X_1 + X_2 + \cdots + X_n = Y$ $= E(e^{tX_1 + tX_2 + \cdots + tX_n})$

$M_{X_1,\cdots,X_n}(t, \cdots, t) = M_{X_1+\cdots+X_n}(t)$

Theorem 3.6 (properties of joint mgf)

1. $M_{X_1}(t_1) = M_{X_1X_2\cdots X_n}(t_1, 0, \ldots, 0)$ ← relationship between joint mgf & marginal mgf.

2. uniqueness theorem

3. $X_1, X_2, \ldots, X_n$ are independent if and only if

$$M_{X_1X_2\cdots X_n}(t_1, t_2, \ldots, t_n) = \prod_{i=1}^{n} M_{X_i}(t_i).$$

LNp.19, joint {cdf, pmf, pdf} $= \prod_{i=1}^{n}$ marginal {cdf, pmf, pdf} cf.

cf. the mgf of the sum of indep. $X_1, \cdots, X_n$ $= \prod_{i=1}^{n} M_{X_i}(t)$

4. $$\left.\frac{\partial^{r_1+\cdots+r_n}}{\partial t_1^{r_1} \cdots \partial t_n^{r_n}} M_{X_1X_2\cdots X_n}(t_1, t_2, \ldots, t_n)\right|_{t_1=t_2=\cdots=t_n=0} = E(X_1^{r_1} X_2^{r_2} \cdots X_n^{r_n})$$

• **conditional expectation** ← Recall: conditional distribution (LNp.21~23)

Definition 3.7 (conditional expectation, TBp. 135-136)

The **conditional expectation** of $h(Y)$ given $X = x$ is

$\underbrace{h(Y)}_{\text{random}}$, $\underbrace{X = x}_{\text{fixed}}$

a function of x

平均: Y or $h(Y)$
權重: $p_{Y|X}(y|x)$, $f_{Y|X}(y|x)$

[**Discrete case**] : $E(h(Y)|X = x) = \sum_y h(y)p_{Y|X}(y|x)$

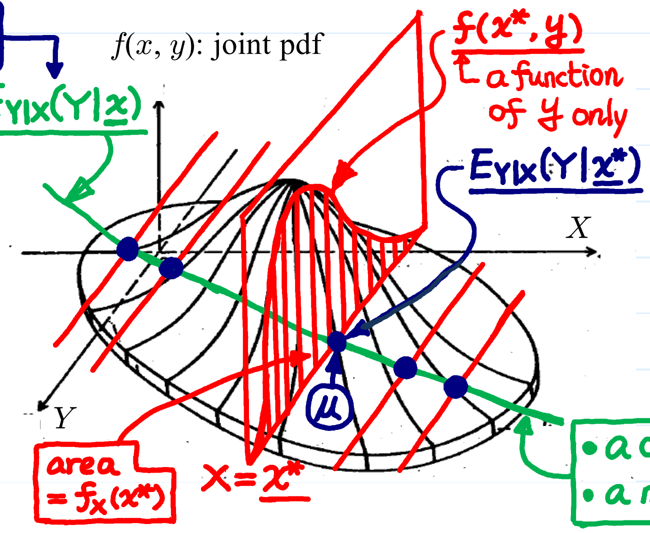
In particular, $E(Y|X = x) = \sum_y y\, p_{Y|X}(y|x)$ ← a pmf for y

[**Continuous case**]: $E(h(Y)|X = x) = \int h(y) f_{Y|X}(y|x)dy$

In particular, $E(Y|X = x) = \int y f_{Y|X}(y|x)dy$ ← a pdf for y

function of x with unit of Y

e.g., $h(Y) = Y$
X: height (cm)
Y: weight (kg)
$E(Y|X=170)$
= average weight of people whose height = 170



$$f_{Y|X}(y|x^*) = \frac{f(x^*, y)}{f_X(x^*)}$$

• a curve on (X, Y) plane
• a map from X to Y

Theorem 3.8 (properties of conditional expectation)

1. $E_{Y|X}(h(Y)|x)$ is a function of x and is free of Y.

fixed values

the Y part has been integrated or summed

2. If X and Y are independent then $E_{Y|X}(h(Y)|x) = E_Y(h(Y))$.

By Thm2.3, item 6, LNp.23, $\begin{cases} p_{Y|X}(y|x) = p_Y(y) \\ f_{Y|X}(y|x) = f_Y(y) \end{cases}$

intuition: $E_{Y|X}(h(Y)|x)$ is a constant function of x $\Rightarrow$ X offers no information of Y

cf.

3. $E(h(X)|X = x) = h(x)$

4. Let $g(x) = E_{Y|X}(h(Y)|x)$, then $g(X)$ is a random variable (transformation of X) and usually denoted by $E_{Y|X}(h(Y)|X)$.

e.g. X: height, Y: weight, $g(x)$: function that maps from height to average weight

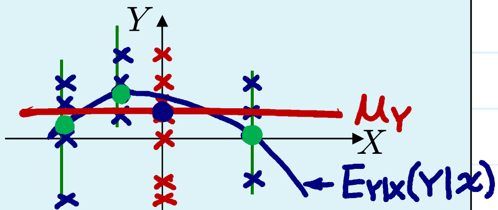
It's a function of X only. But, its random value reflects $h(Y)$

5. **law of total expection** (TBp.149)

$$E_X[E_{Y|X}(h(Y)|X)] = E_Y[h(Y)].$$

In particular,

$$E_Y[E_{X|Y}(Y|Y)] \to E_Y(Y) = E_X[E_{Y|X}(Y|X)].$$



$E_{X,Y} = E_X E_{Y|X} = E_Y E_{X|Y}$

$\sum_x \sum_y h(y) p_{XY}(x,y)$, with $p_{XY}(x,y) = p_{Y|X}(y|x) \cdot p_X(x)$; $h(x,y)$; $\updownarrow$ $\sum_y \sum_x$

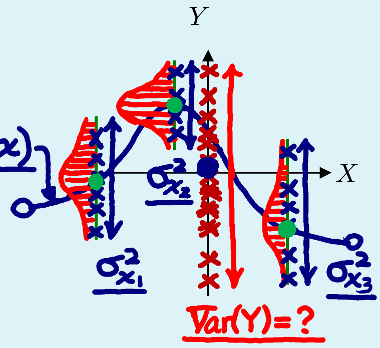
$\int\int_{xy} h(y) f_{XY}(x,y)\,dy\,dx$, with $f_{XY}(x,y) = f_{Y|X}(y|x) f_X(x)$; $h(x,y)$; $\updownarrow$ $\int_y \int_x$

*** generalization

$E_{XY}[h(X,Y)] = E_Y E_{X|Y}[h(X,Y)|Y]$
$= E_X E_{Y|X}[h(X,Y)|X]$

4. **variance decomposition** (TBp.151)

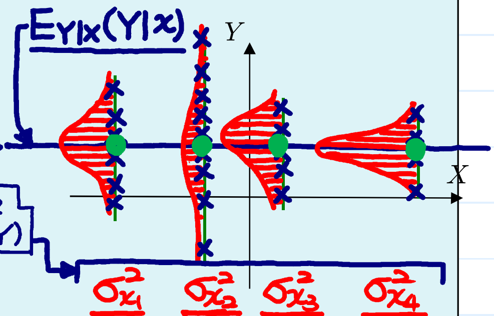
$$Var_Y(Y) = Var_X[E_{Y|X}(Y|X)] + E_X[Var_{Y|X}(Y|X)]$$



Note.

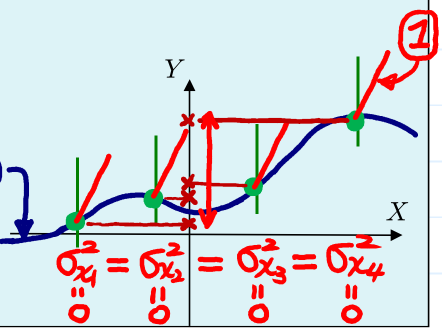
1. $Var_Y(Y) \geq E_X[Var_{Y|X}(Y|X)]$ and the equality holds if and only if $E_{Y|X}(Y|X) = E_Y(Y)$ with probability one.

$Var_X[E_{Y|X}(Y|X)] = 0$



2. $Var_Y(Y) \geq Var_X[E_{Y|X}(Y|X)]$ and the equality holds if and only if $Var_{Y|X}(Y|X) = 0$ with probability one; i.e., $Y = E_{Y|X}(Y|X)$ with probability one.

$E_X[Var_{Y|X}(Y|X)] = 0$



• prediction

Example 3.1 (predicting the value of a r.v. Y from another r.v. X, TBp. 152-154)

- **data**: X and Y (example?)

X	身高	雨量
Y	體重	米產量

- **statistical modeling**: assign (X, Y) a (known) joint distribution (cdf $F(x, y)$, pdf $f(x, y)$, or pmf $p(x, y)$)
- **objective**: Predict Y by using a function of X, i.e., $g(X)$.

  We consider the following three groups of g's:

  (i) $G_1 = \{g(x) : g(x) = c,$ where $c \in \mathbb{R}\}$ ← not use the information of X

  (ii) $G_2 = \{g(x) : g(x) = a + bx,$ where $a, b \in \mathbb{R}\}$, and

  (iii) $G_3 = \{g(x) : g$ is arbitrary$\}$.

  Note. $G_1 \subset G_2 \subset G_3$.
- **question**: Within each group, what is the "best" prediction?

  i.e. how to choose c for G_1; a, b for G_2; g for G_3
- **criterion**: minimizing mean square error:

meaning? → $\text{MSE} \equiv E_{X,Y}\{[\underbrace{Y}_{\text{true value}} - \underbrace{g(X)}_{\text{predicted value}}]^2\}$. error

G_1 → **Example 3.2** ("best" constant prediction, TBp. 153)

$$E_{X,Y}(Y - c)^2 = E_Y(Y - c)^2 \geq E_Y[Y - E_Y(Y)]^2 = Var_Y(Y) \quad \boxed{\min}$$

The equality holds if and only if $c = E_Y(Y)$. ← only need to know μ_Y

G_3 → **Example 3.3** ("best" prediction of Y using X, TBp. 153)

mean: best predictor under MSE

$$E_{X,Y}[Y - g(X)]^2 \geq E_{X,Y}[Y - E_{Y|X}(Y|X)]^2 = E_X[Var_{Y|X}(Y|X)]$$

$\boxed{\min}$ cf. check LNp.52

The equlity holds if and only if $g(x) = E_{Y|X}(Y|x)$. cf.

Notes for the best predictor in G_3.

best in G_1: $E_Y(Y)$ cf.

- $E_{Y|X}(Y|X)$ is the best predictor of Y based on X, in the mean squared prediction error sense. $\boxed{\text{intuition}}$ ← check the graph in LNp.50 | $\boxed{\text{median: best predictor under } E|Y-g(X)|}$
- need to know the joint distribution of X and Y, or at least $E_{Y|X}(Y|x)$ cf.
- $E_{Y|X}(Y|x)$ is called the regression function of Y on X. 迴歸

G_2 → **Example 3.4** ("best" linear prediction of Y using X, TBp. 153-154)

$$E_{X,Y}[Y-(a+bX)]^2 \geq E_{X,Y}\left\{Y - \left[\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(X - \mu_X)\right]\right\}^2 = \sigma_Y^2(1-\rho^2)$$

$0 \leq |\rho| \leq 1$ $\quad \boxed{\min}$

The equality holds if and only if $a = \mu_Y - b\mu_X$ and $b = \rho\frac{\sigma_Y}{\sigma_X}$. $\boxed{\text{unit=?}}$

Notes for the best predictor in G_2.

- $E_{Y|X}(Y|x) = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X)$ if (X, Y) is distributed as bivariate normal

best in G_3 linear regression analysis best in G_2 more information better predictor

- needs to know only the means, variances and covariances

c.f. the best in G_1 & G_3 ⟶ Which one require more information?

- $\sigma_Y^2(1 - \rho^2)$ is small if ρ is close to $+1$ or -1, and large if ρ is close to 0

intuition ⟵ check the plot in LNp.44

Notes.

1. $\min\limits_{a,b} E[Y - (a + bX)]^2 \le \min\limits_{c} E(Y - c)^2$ and the equality holds if and only if $\rho = 0$.

$\because G_1 \subset G_2 \subset G_3$

2. $\min\limits_{g} E(Y - g(X))^2 \le \min\limits_{a,b} E[Y - (a + bX)]^2$ and the equality holds if and only if $E_{Y|X}(Y|x) = \mu_Y + \rho(\sigma_Y/\sigma_X)(x - \mu_X)$.

Collect data of X, Y to estimate their joint dist.

Question 3.3

What if the joint distribution of X and Y is unknown?

❖ **Reading**: textbook, Chapter 4

❖ **Further Reading**: Roussas, 5.1, 5.3, 5.4, 5.5, 6.1, 6.2, 6.4, 6.5

Some Commonly Used Distributions (from Chapters 2, 3, 6)

Question 4.1

For a given random phenomenon or data, what distribution (or statistical model) is more appropriate to depict it? statistical modeling

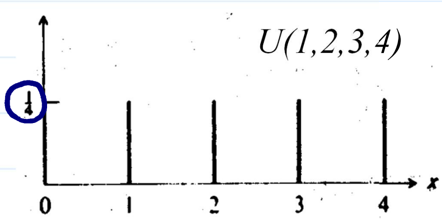
• discrete distributions

Definition 4.1 (Uniform distribution $U(a_1,\ldots,a_m)$)

Equal probability to obtain $a_1, a_2, \ldots, a_m$.

• pmf: $p(x) = \begin{cases} \frac{1}{m}, & x = a_1, \ldots, a_m \\ 0, & \text{otherwise} \end{cases}$

a pmf ? (Ec)



• mgf: $\frac{\sum_{j=1}^m e^{a_j t}}{m}$ ← by definition (Ec)

• mean: $\frac{\sum_{j=1}^m a_j}{m} \equiv \bar{a}$

• variance: $\frac{\sum_{j=1}^m (a_j-\bar{a})^2}{m}$

• parameter: $a_i \in \mathbb{R}$, $m = 1, 2, \ldots$

• example: throw a fair die once

Definition 4.2 (Bernoulli distribution $B(p)$, sec 2.1.1)

A Bernoulli distribution takes on only two values: 0 and 1, with probabilities $1-p$ and p, respectively.

- **pmf:** $p(x) = \begin{cases} p^x(1-p)^{(1-x)}, & \text{if } x=0 \text{ or } x=1 \\ 0, & \text{otherwise} \end{cases}$

  a pmf? (Ec)
- **mgf:** $pe^t + 1 - p$ — by definition (Ec)
- **mean:** p — by definition / use mgf (Ec)
- **variance:** $p(1-p)$ — $Var(X) = E(X^2) - [E(X)]^2$ ($E(X^2) = p$) / $Var(X) = E[X(X-1)] + E(X) - [E(X)]^2$ ($E[X(X-1)] = 0$) / use mgf (Ec)
- **parameter:** $p \in [0,1]$
- **example:** toss a coin once, p=probability that head occurs

Note: If A is an event, then the indicator random variable I_A follows the Bernoulli distribution.

$\hookrightarrow p = P(A)$

$I_A: \Omega \to \mathbb{R}$, $I_A(\omega) = \begin{cases} 1, & \text{if } \omega \in A \\ 0, & \text{if } \omega \notin A \end{cases}$

Definition 4.3 (Binomial distribution $B(n,p)$, sec 2.1.2)

Suppose that n independent Bernoulli trials are performed, where n is a fixed number. The total number of 1 appearing in the n trials follows a binomial distribution with parameters n and p.

$0+1+1+\cdots+0 = X$

1 2 ... n

Shape

explanation

- **pmf:** $p(x) = \begin{cases} \binom{n}{x} p^x(1-p)^{(n-x)}, & x = 0,1,\ldots,n \\ 0, & \text{otherwise} \end{cases}$

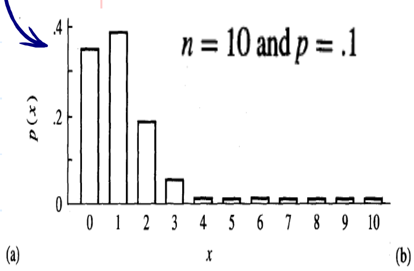
  a pmf? (Ec) ← use (*)
- **mgf:** $(pe^t + 1 - p)^n$, $t \in \mathbb{R}$. — by definition / sum of i.i.d. $B(p)$ (Ec)
- **mean:** np — use definition / use mgf / sum of iid $B(p)$ (Ec)

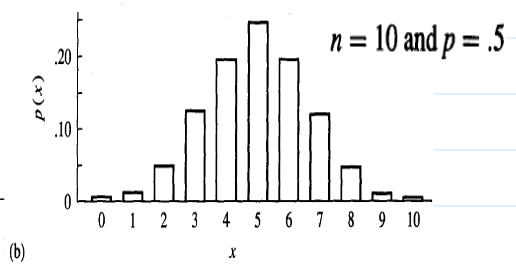
  intuition

  $E(X) = \sum_{x=0}^{n} x\binom{n}{x}p^x(1-p)^{n-x} = \left[\sum_{x=1}^{n}\binom{n-1}{x-1}p^{x-1}(1-p)^{(n-1)-(x-1)}\right]np$ — pmf of $B(n-1,p)$

  ↳ STO (sum-to-one) method
- **variance:** $np(1-p)$ ← max at $p=\frac{1}{2}$, min at $p=0$ or 1

  Find $E(X^2)$ using mgf / Find $E[X(X-1)]$ using STO / sum of i.i.d. $B(p)$ (Ec)
- **parameter:** $p \in [0,1]$, $n = 1, 2, \ldots$
- **example:** # of heads, toss a coin n times





Note: (*) $(a+b)^n = \sum_{x=0}^{n}\binom{n}{x}a^x b^{n-x}$.

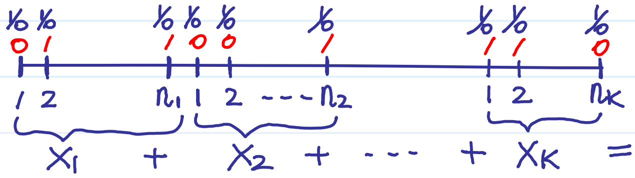
Note.

intuition

1. binomial distribution is a generalization of bernoulli distribution from 1 trial to n trials
2. Let $X_1, \ldots, X_n$ be i.i.d. $B(p)$, then $Y = X_1 + \cdots + X_n \sim B(n, p)$. – prove using ① mgf ($M_Y(t) = \prod_{i=1}^{n} M_{X_i}(t)$) ② convolution & induction (Ec)

intuition

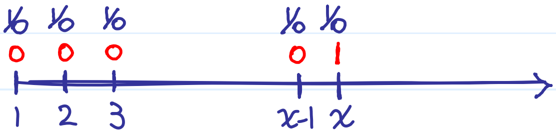
3. Let $X_i \sim B(n_i, p), i = 1, \ldots, k$, and $X_1, \ldots, X_k$ are independent. Then, $Y = X_1 + \cdots + X_k \sim B(n_1 + \cdots + n_k, p)$. – prove using mgf / prove using convolution & induction (Ec)



$= Y = \#$ of 1's in $(n_1 + \cdots + n_k)$ trial

Definition 4.4 (Geometric distribution $G(p)$, sec 2.1.3)

The geometric distribution is constructed from an infinite sequence of independent Bernoulli trials. Let X be the total number of trials up to and including the first appearance of 1. Then, X follows the geometric distribution.



- **pmf:** $p(x) = \begin{cases} (1-p)^{(x-1)}p, & x = 1, 2, 3, \ldots \\ 0, & \text{otherwise} \end{cases}$ — explanation; a pmf? (Ec) ← use (**)
- **cdf:** $F(x) = \begin{cases} 1 - (1-p)^{[x]}, & 1 \le [x] \le x < [x]+1 \\ 0, & x < 1 \end{cases}$ — Find $P(X>x)$ using (**) (Ec)
- **mgf:** $\frac{pe^t}{1-(1-p)e^t}$, $t < -\log(1-p)$. — use (**) / use STO (Ec)
- **mean:** $\frac{1}{p}$ — use $E(X) = \sum_{k=1}^{\infty} P(X \ge k)$ or use (**) / use mgf / use differentiation method (TBp.117, Example B) (Ec)

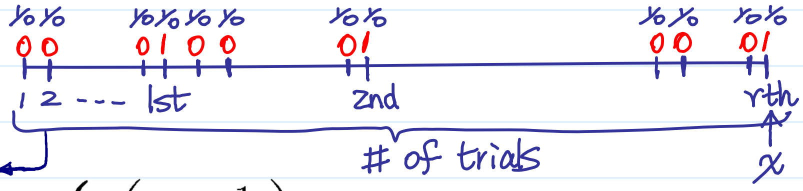
  intuition
- **variance:** $\frac{1-p}{p^2}$ — Find $E(X^2)$ using mgf / Find $E[X(X-1)]$ using differentiation method (Ec)
- **parameter:** $p \in [0, 1]$
- **example:** lottery, # of tickets a person must purchase up to and including the first winning ticket

Note: a memoryless distribution ← intuition
check its definition (LNp.74) and prove (Ec)

Note: (**) $\sum_{x=n}^{\infty} t^x = \frac{t^n}{1-t}$, for $-1 < t < 1$.

Definition 4.5 (Negative Binomial distribution $NB(r, p)$, sec 2.1.3)

An infinite sequence of independent Bernoulli trials is performed until the appearance of the rth 1. Let X denote the total number of trials. Then, X follows negative binomial distribution.



explanation

- **pmf:** $p(x) = \begin{cases} \binom{x-1}{r-1} p^r (1-p)^{(x-r)}, & x = r, r+1, \ldots \\ 0, & \text{otherwise} \end{cases}$

  a pmf? (Ec) use (***)

- **mgf:** $\frac{p^r e^{rt}}{[1-(1-p)e^t]^r}$, $t < -\log(1-p)$.

  use STO
  use (***) (Ec)
  sum of i.i.d G(p)

- **mean:** $\frac{r}{p}$

  intuition

  use mgf
  use STO
  sum of i.i.d. G(p) (Ec)

- **variance:** $\frac{r(1-p)}{p^2}$

  Find $E(X^2)$ using mgf
  Find $E(X(X+1))$ using STO (Ec)
  sum of i.i.d. G(p)

- **parameter:** $p \in [0, 1]$, $r = 1, 2, \ldots$
- **example:** lottery, # of tickets a person must purchase up to and including the rth winning ticket

Note: (***) $\sum_{x=0}^{\infty} \binom{n+x-1}{x} t^x = \frac{1}{(1-t)^n}$, for $-1 < t < 1$.

Note.

1. negative binomial distribution is a generalization of geometric distribution from 1st success to rth success

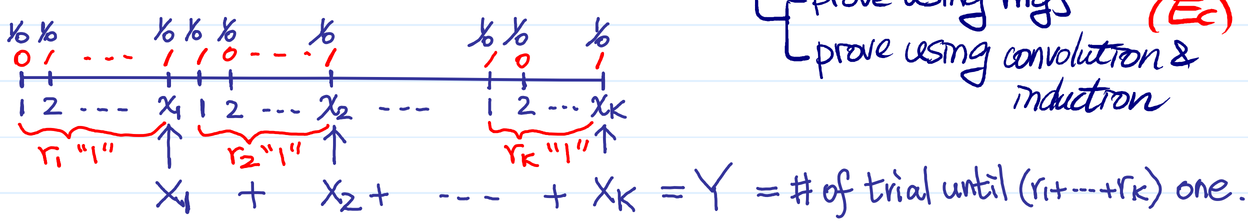
intuition

2. Let $X_1, X_2, \ldots, X_r$ be i.i.d. $G(p)$, then $Y = X_1 + \cdots + X_r \sim NB(r, p)$ — prove using ① mgf ($M_Y(t) = \prod_{i=1}^{n} M_{X_i}(t)$) ② convolution & induction (Ec)

3. Let $X_i \sim NB(r_i, p)$, $i = 1, \ldots, k$, and $X_1, \ldots, X_k$ are independent. Then, $Y = X_1 + \cdots + X_k \sim NB(r_1 + \cdots + r_k, p)$.

   prove using mgf (Ec)
   prove using convolution & induction

intuition



Definition 4.6 (Multinomial distribution $Multinomial(n, p_1, p_2, \ldots, p_r)$, TBp.73-74)

Suppose that each of n independent trials can result in one of r types of outcomes, and that on each trial the probabilities of the r outcomes are $p_1, p_2, \ldots, p_r$. Let X_i be the total number of outcomes of type i in the n trials, $i = 1, \ldots, r$. Then, $(X_1, \ldots, X_r)$ follows a multinomial distribution.

- **joint pmf:** use (****)

a joint pmf? (Ec)

$$p(x_1,\dots,x_r)=\begin{cases}\binom{n}{x_1\cdots x_r}p_1^{x_1}\cdots p_r^{x_r}, & x_i=0,1,\dots,n, \text{ and } \sum_{i=1}^r x_i=n\\ 0, & \text{otherwise}\end{cases}$$

explanation

- **joint mgf:** $(p_1e^{t_1}+\cdots+p_re^{t_r})^n$, $t_1,\dots,t_r\in\mathbb{R}$. — use (****); use STO (Ec)
- **marginal distribution:** $X_i\sim B(n,p_i)$, $i=1,\dots,r$ — intuition (Ec); prove using mgf
- **mean:** $E(X_i)=np_i$, $i=1,\dots,n$
- **variance:** $Var(X_i)=np_i(1-p_i)$, $i=1,\dots,n$ — Find $E(X_iX_j)$ using STO
- **covariance:** $Cov(X_i,X_j)=-np_ip_j$, $i\neq j$ — Find $E(X_iX_j)$ using mgf (Ec)
- **parameter:** $p_i\in[0,1]$, and $\sum_{i=1}^r p_i=1$. $n=1,2,\dots$
- **example:** randomly choose n people, record the numbers of people with different religions

why negative?

(****) **Note:** $(a_1+\cdots+a_k)^n=\sum_{x_1+\cdots+x_k=n}\binom{n}{x_1,\cdots,x_k}a_1^{x_1}\cdots a_k^{x_k}$.

Notes: multinomial distribution is a generalization of the binomial distribution from 2 outcomes to r outcomes.

Definition 4.7 (Poisson distribution $P(\lambda)$, sec 2.1.5)

Limit of binomial distributions $X_n\sim B(n,p_n)$, where $p_n\to 0$ as $n\to\infty$ in such a way that $\lambda_n\equiv np_n\to\lambda$.

$p_n=\frac{\lambda_n}{n}$

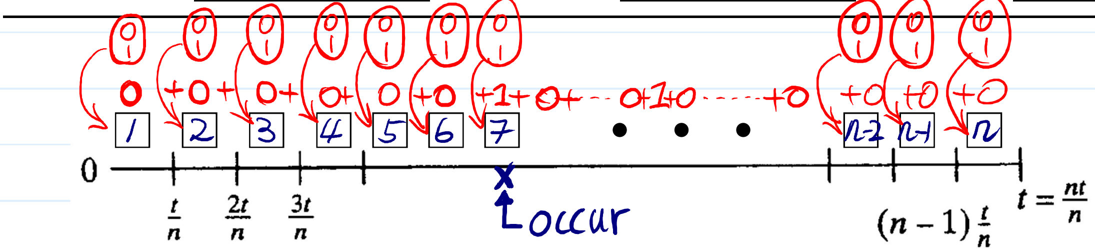
Note: if $a_n\to a$, $\left(1+\frac{a_n}{n}\right)^n\to e^a$.

$$\binom{n}{x}p_n^x(1-p_n)^{(n-x)}$$
$$=\frac{n(n-1)\cdots(n-x+1)}{x!}\left(\frac{\lambda_n}{n}\right)^x\left(1-\frac{\lambda_n}{n}\right)^{n-x}$$
$$=\frac{n(n-1)\cdots(n-x+1)}{n^x}\frac{1}{x!}\lambda_n^x\left(1-\frac{\lambda_n}{n}\right)^{n-x}$$
$$=1(1-\tfrac{1}{n})\cdots(1-\tfrac{x-1}{n})\frac{\lambda_n^x}{x!}\left(1-\frac{\lambda_n}{n}\right)^n\left(1-\frac{\lambda_n}{n}\right)^{-x}\longrightarrow 1^x\cdot\frac{\lambda^x}{x!}\cdot e^{-\lambda}\cdot 1=\frac{\lambda^x e^{-\lambda}}{x!}$$

1 1 1 $\frac{\lambda^x}{x!}$ $e^{-\lambda}$ 1

explanations.

1. if n large, the pmf of $B(n,p)$ is not easily calculated. Then, we can approximate them by pmf of $P(\lambda)$, where $\lambda=np$.



2. Let X be the number of times some event occurs in a given time interval I. Divide the interval into many small subintervals I_k, $k = 1, \ldots, n$, of equal length. Let N_k be the number of events occurring in I_k. When we can assume $N_1, \ldots, N_n$ are independent and approximately $\sim B(p)$, X has a distribution near $P(\lambda)$, where $\lambda = np$.

$X = N_1+N_2+\cdots+N_n \sim B(n,p)$ with large n & small p

shape

a pmf? (Ec) — use (*****)

- **pmf:** $p(x) = \begin{cases} \frac{\lambda^x}{x!}e^{-\lambda}, & x = 0, 1, 2, \ldots \\ 0, & \text{otherwise} \end{cases}$
- **mgf:** $e^{\lambda(e^t-1)}$, $t \in \mathbb{R}$. — use (*****) / use STO (Ec)
- **mean:** λ — use STO / use mgf (Ec)

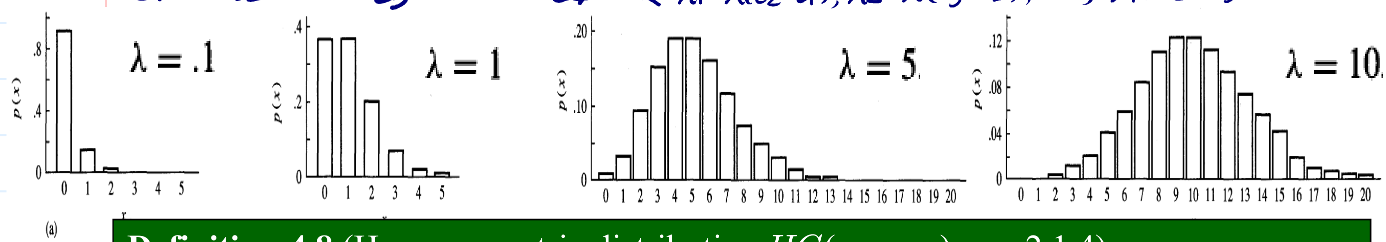
  meaning of parameter λ: average occurences
- **variance:** λ — Find $E[X(X-1)]$ using STO / Find $E(X^2)$ using mgf (Ec) / $np(1-p) \approx np$
- **parameter:** $\lambda > 0$
- **example:** number of phone calls coming into an exchange during a unit of time

Note: ~~(*****)~~ $e^\lambda = \sum_{x=0}^{\infty} \frac{\lambda^x}{x!}$.

Note: Let $X_i \sim P(\lambda_i)$, $i = 1, \ldots, k$, and $X_1, \ldots, X_k$ are independent. Then, $Y = X_1 + \cdots + X_k \sim P(\lambda_1 + \cdots + \lambda_k)$. — prove using mgf (Ec) / prove using convolution & induction / intuition

$X_1 + X_2 + X_3$

$t_1 \quad t_2 \quad t_3 \quad t_4$ $\Leftarrow \lambda_1=\lambda(t_2-t_1), \lambda_2=\lambda(t_3-t_2), \ldots, \lambda_1+\lambda_2+\lambda_3=\lambda(t_4-t_1)$



Definition 4.8 (Hypergeometric distribution $HG(r, n, m)$, sec 2.1.4)

Suppose that an urn contains n black balls and m white balls. Let X denote the number of black balls drawn when taking r balls without replacement. Then, X follows hypergeometric distribution. c.f. with replacement $\Rightarrow X \sim B(r, \frac{n}{m+n})$

explanation

- **pmf:** $p(x) = \begin{cases} \frac{\binom{n}{x}\binom{m}{r-x}}{\binom{n+m}{r}}, & x = 0, 1, \ldots, \min(r, n),\ r - x \le m \\ 0, & \text{otherwise} \end{cases}$

  a pmf? (Ec) — use (*****)

Note: ~~(******)~~ $\binom{n+m}{r} = \sum_x \binom{n}{x}\binom{m}{r-x}$.

- **mgf:** exist, but no simple expression
- **mean:** $\frac{rn}{n+m}$ ← use STO (Ec)

  intuition
- **variance:** $\frac{rnm(n+m-r)}{(n+m)^2(n+m-1)}$ ← Find E[X(X-1)] using STO (Ec)
- **parameter:** $r, n, m, = 1, 2, \ldots,\ r \le n+m$
- **example:** sampling industrial products for defect inspection

Notes. a relationship between hypergeometric and binomial distributions: Let $m,\ n \to \infty$ in such a way that

$$p_{m,n} \equiv \frac{n}{m+n} \to p,$$

where $0 < p < 1$. Then,

intuition: When m, n are large, with replacement ≈ without replacement

$$\frac{\binom{n}{x}\binom{m}{r-x}}{\binom{n+m}{r}} \to \binom{r}{x} p^x (1-p)^{r-x}.$$

• continuous distributions

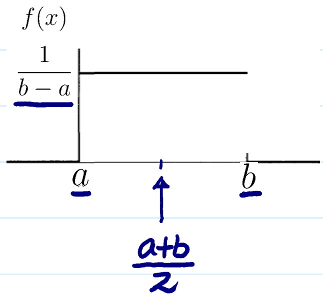
Definition 4.9 (Uniform distribution $U(a,b)$, sec 2.2)

Choose a number at random between a and b.

shape

- **pdf:** $f(x) = \begin{cases} \frac{1}{b-a}, & a \le x \le b \\ 0, & \text{otherwise} \end{cases}$

  a pdf? (Ec)



- **cdf:** $F(x) = \begin{cases} 0, & x < a \\ \frac{x-a}{b-a}, & a \le x \le b \\ 1, & x > b \end{cases}$ ← by definition (Ec)
- **mgf:** $\frac{e^{bt}-e^{at}}{t(b-a)},\ t \in \mathbb{R}$. ← by definition (Ec)
- **mean:** $\frac{a+b}{2}$ ← by definition / use mgf (Ec)

  intuition
- **variance:** $\frac{(b-a)^2}{12}$ ← Find $E(X^2)$ using definition / Find $E(X^2)$ using mgf (Ec)
- **parameter:** $a, b \in \mathbb{R},\ \ a < b$

Thm 2.4, 2.5 (LNp.30)

Note: $U(0,1)$ is useful for pseudo-random number generation

Definition 4.10 (Exponential distribution $E(\lambda)$, sec 2.2.1)

shape

- pdf: $f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$

  a pdf ? (Ec)

- cdf: $F(x) = \begin{cases} 1 - e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$ ← by definition (Ec)

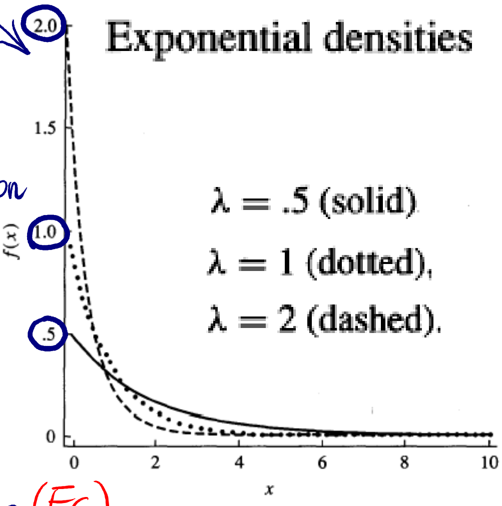
- mgf: $\frac{\lambda}{\lambda - t}$, $t < \lambda$. ← by definition / use STO (Ec)

- mean: $\frac{1}{\lambda}$ ← use STO / use mgf (Ec)

- variance: $\frac{1}{\lambda^2}$ ← Find $E(X^2)$ using STO / Find $E(X^2)$ using mgf (Ec)

- parameter: $\lambda > 0$

- example: lifetime or waiting time



meaning of parameter — $1/\lambda$: average waiting time $\left(\frac{時間}{次}\right)$; λ : average occurence rate $\left(\frac{次}{時間}\right)$

Notes:

1. **memoryless** (future independent of past): Let $T \sim E(\lambda)$, then

$$\underbrace{P(T > t+s \mid T > s)}_{T-s>t} = \frac{P(T > t+s \text{ and } T > s)}{P(T > s)} = \frac{P(T > t+s)}{P(T > s)}$$

$$= \frac{e^{-\lambda(t+s)}}{e^{-\lambda s}} = e^{-\lambda t} = P(T > t)$$

1st 2nd now 0 s t s+t

cdf of T: $F_T(t) = 1 - P(T > t)$

If discrete, then it is geometric. cf.

   - (⇐) If a continuous distribution is memoryless, it is exponential.
   - It does not mean the two events $T > s$ and $T > t+s$ are independent.

2. relationship between exponential, gamma, and Poisson distributions

   Let $T_1, T_2, T_3, \ldots$ be i.i.d. $\sim E(\lambda)$ and $S_k = T_1 + \cdots + T_k$, $k = 1, 2, \ldots$. Let X_i be the number of S_k's that falls in $[t_{i-1}, t_i]$, $i = 1, \ldots, n$, then $X_1, \ldots, X_n$ are independent, and $X_i \sim P(\lambda(t_i - t_{i-1}))$. The reverse statement is also ture.

Poisson Process

X_1: # of events occur during $[t_0, t_1] \to [0, t]$ $\sim P(\lambda(t_1 - t_0)) \to P(\lambda t)$ Binomial

$P_r(T_1 > t) = P_r(P(\lambda t) = 0) = e^{-\lambda t}(\lambda t)^0/0! = e^{-\lambda t}$

$T_i \sim E(\lambda)$ ↕ Geometric

T_1 T_2 T_3 T_4 time 0 t_0 S_1 S_2 t_1 S_3 S_4

$S_k \sim \Gamma(k, \lambda)$ ↕ Negative Binomial

Question 4.2: how to interpret λ? $\left(\frac{次}{單位時間}\right)$

3. Sometimes, the pdf is written as $\frac{1}{\lambda} e^{-\frac{x}{\lambda}}$. In the case, how to interpret λ?

Definition 4.11 (Gamma distribution $\Gamma(\alpha, \lambda)$, sec 2.2.2)

shape

- pdf: $f(x) = \begin{cases} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$

a pdf? (Ec) ← use gamma function (LNp.74)

- mgf: $(\frac{\lambda}{\lambda - t})^\alpha$, $t < \lambda$. ← use STO (Ec) / sum of i.i.d. exponential
- mean: $\frac{\alpha}{\lambda}$ ← use STO / use mgf (Ec) / sum of i.i.d exponential

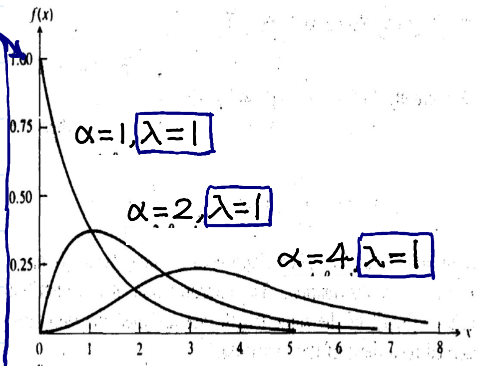
intuition

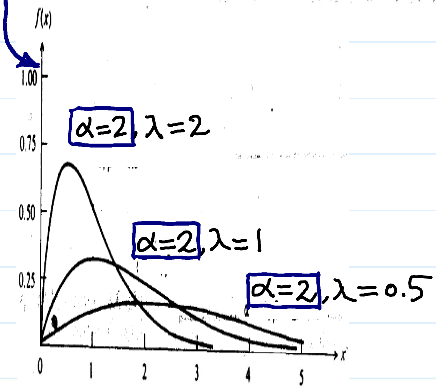
- variance: $\frac{\alpha}{\lambda^2}$
- parameter: $\alpha, \lambda > 0$

Find $E(X^2)$ using STO

Find $E(X^2)$ using mgf (Ec)

sum of i.i.d exponential





Notes.

1. α: shape parameter; λ: scale parameter (Question 4.3: how to interpret α, λ from the view point of Poisson process?)

(LNp.72) λ: occurence rate, α: # of summed exponential r.v.'s

2. properties of gamma function $\Gamma(\alpha)$:
   - $\Gamma(\alpha) \equiv \int_0^\infty y^{\alpha-1} e^{-y} dy$ (which is finite for $\alpha > 0$)
   - $\Gamma(1) = 1$ and $\Gamma(\frac{1}{2}) = \sqrt{\pi}$
   - $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$
   - $\Gamma(\alpha) = (\alpha - 1)!$ if α is an integer
   - $\Gamma(\frac{\alpha}{2}) = \frac{\sqrt{\pi}(\alpha-1)!}{2^{\alpha-1}(\frac{\alpha-1}{2})!}$ if α is an odd integer
3. gamma distribution can be viewed as a generalization of exponential distribution, i.e., $\Gamma(1, \lambda) = E(\lambda)$.

   (Ec) intuition / prove using mgf

4. Let $X_1, \ldots, X_k$ be i.i.d. $\sim E(\lambda)$, then $Y = X_1 + \cdots + X_k \sim \Gamma(k, \lambda)$.
5. Let $X_1, \ldots, X_k$ be independent, and $X_i \sim \Gamma(\alpha_i, \lambda)$, then $Y = X_1 + \cdots + X_k \sim \Gamma(\alpha_1 + \cdots + \alpha_k, \lambda)$. ← intuition / prove using mgf (Ec)
6. Let $X \sim \Gamma(\alpha, \lambda)$,, then $cX \sim \Gamma(\alpha, \lambda/c)$, where $c > 0$. ← intuition / prove using mgf (Ec)
7. $X \sim \Gamma(\alpha, \lambda) \Rightarrow E(X^k) = \frac{\Gamma(\alpha+k)}{\lambda^k \Gamma(\alpha)}$, for $0 < k$ and $E(\frac{1}{X^k}) = \frac{\lambda^k \Gamma(\alpha - k)}{\Gamma(\alpha)}$, for $0 < k < \alpha$. ← use ① STO ② mgf (Ec) integration.

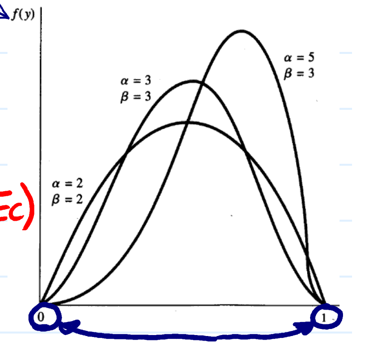
Definition 4.12 (Beta distribution $beta(\alpha, \beta)$, sec 15.3.2)

shape

- **pdf:** $f(x) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, & 0 \le x \le 1 \\ 0, & \text{otherwise} \end{cases}$

a pdf? (Ec)

cf. pmf of $B(n,p) = \binom{n}{x} p^x (1-p)^{n-x}$



- **mgf:** $1 + \sum_{k=1}^{\infty} \left(\prod_{r=0}^{k-1} \frac{\alpha+r}{\alpha+\beta+r} \right) \frac{t^k}{k!}$ ← by definition (Note: $e^{tx} = \sum_{k=0}^{\infty} \frac{(tx)^k}{k!}$) (Ec)
- **mean:** $\frac{\alpha}{\alpha+\beta}$ ← use STO / use mgf (Ec)

intuition

- **variance:** $\frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}$ ← Find $E(X^2)$ using STO / Find $E(X^2)$ using mgf (Ec)
- **parameter:** $\alpha, \beta > 0$

Notes:

1. Beta function: $B(\alpha, \beta) \equiv \int_0^1 t^{\alpha-1}(1-t)^{\beta-1} dt = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$
2. $\beta(1,1) = U(0,1)$ meaning of α & β
3. Let $X_1 \sim \Gamma(\alpha_1, \lambda)$, $X_2 \sim \Gamma(\alpha_2, \lambda)$, and X_1, X_2 independent. Then, $\frac{X_1}{X_1+X_2} \sim beta(\alpha_1, \alpha_2)$. ← $\begin{cases} Y_1 = X_1/(X_1+X_2) \\ Y_2 = X_1+X_2 \end{cases}$ → find the joint pdf of (Y_1, Y_2), then marginal pdf of Y_1 (Ec)