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?

Outline

- 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

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)

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

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

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

$$\Omega = \{t | t \geq 0\}$$

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**.

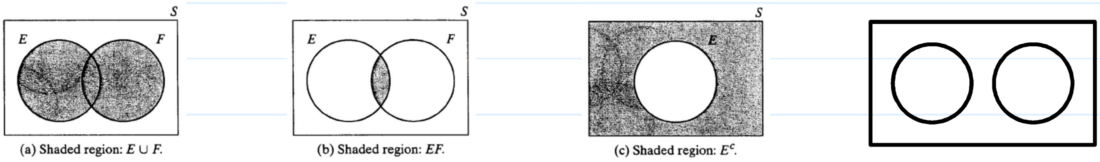
Example 1.4 (cont. Ex. 1.1)

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.
- **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:

1. $P(\Omega) = 1$.
2. If $A \subset \Omega$, then $P(A) \geq 0$.
3. If A_1 and A_2 are disjoint, then
$$P(A_1 \cup A_2) = P(A_1) + P(A_2).$$
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 $\omega \in \Omega$, $P(\omega) = \frac{1}{8}$.

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

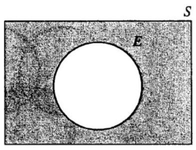
1/8　1/8　1/8　1/8　1/8　1/8　1/8　1/8

Property A. $P(A^C) = 1 - P(A)$.

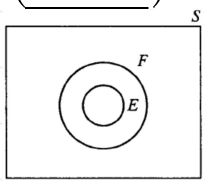
Property B. $P(\emptyset) = 0$.

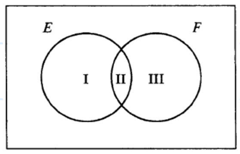
Property C. If $A \subset B$, then $P(A) \leq P(B)$.

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



(c) Shaded region: E^c.

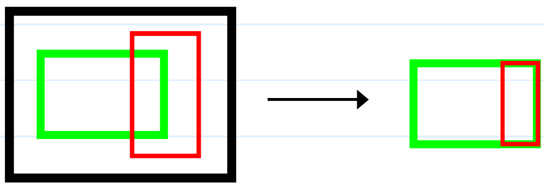




Definition (conditional probability, TBp. 17)

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

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



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?

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

Theorem (Multiplication Law, TBp. 17)

Let A and B be events and assume $P(B) > 0$. Then
$$P(A \cap B) = P(A|B)P(B).$$

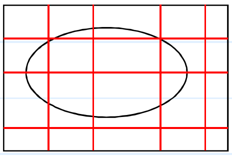
Example 1.7 (Ex. B, TBp. 18)

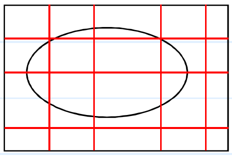
Suppose if it is cloudy (B), the probability that it is raining (A) is 0.3, and that the probability that it is cloudy is $P(B) = 0.2$. The probability that it is cloudy and raining is
$P(A \cap B) = P(A|B)P(B) = 0.3 \times 0.2 = 0.06$.

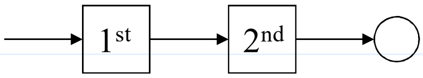
Theorem (Law of Total Probability, TBp. 18)

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,

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







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

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

Definition (independence, TBp. 24)

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

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

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

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

When A and B are independent,

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

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

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

❖ **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

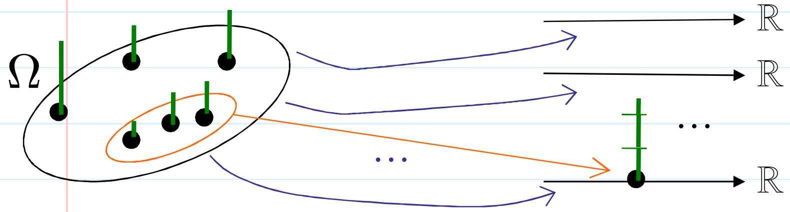
Outline

- random variables
- 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.



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

	1/8	1/8	1/8	1/8	1/8	1/8	1/8	1/8
$\Omega =$	$\{hhh,$	$hht,$	$hth,$	$thh,$	$htt,$	$tht,$	$tth,$	$ttt\}$
	↓	↓	↓	↓	↓	↓	↓	↓
X_1 :	3,	2,	2,	2,	1,	1,	1,	0.
X_2 :	1,	1,	1,	0,	1,	0,	0,	0.
X_3 :	3,	1,	1,	1,	−1,	−1,	−1,	−3.

Question 2.1

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

• distribution

Question 2.2

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

	discrete	continuous
uni-variate r.v.	• pmf • cdf • mgf/chf	• pdf • cdf • mgf/chf
multi-variate r.v.'s	• joint pmf • joint cdf • joint mgf/chf	• joint pdf • joint cdf • joint mgf/chf

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

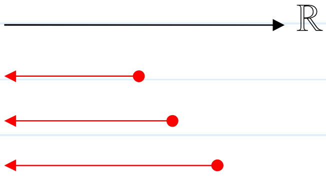
Definition 2.2 (discrete and continuous random variables, TBp. 35 and 47)

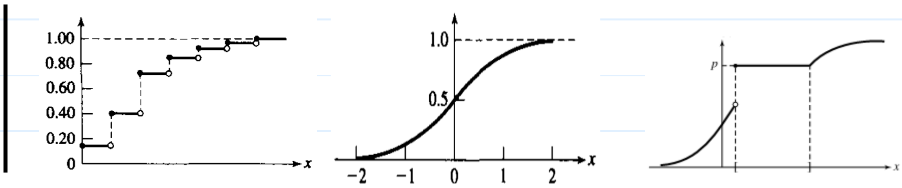
A **discrete random variable** can take on only a finite or at most a countably infinite number of values. A **continuous random variable** can take on a continuum of values.

Definition 2.3 (cumulative distribution function, TBp. 36)

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

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





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

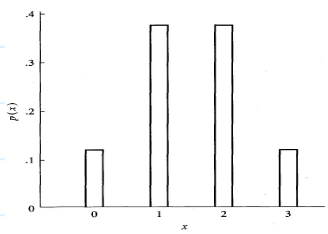
For a discrete random variable X with pmf $p(x)$,

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

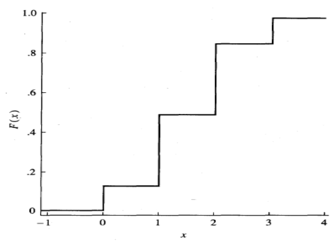
and

$$P(X \in A) = \sum_{x \in A} p(x).$$

probability mass function



cumulative distribution function



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

$$p(x) = P(X = x) = F(x) - F(x-)$$

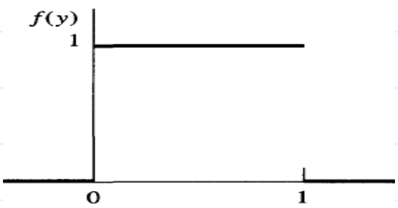
Definition 2.5 (probability density function, TBp. 46)

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) $\int_{-\infty}^{\infty} f(x)dx = 1$.

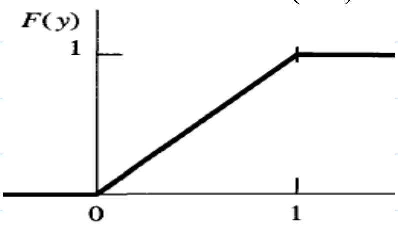
For a continuous random variable X with pdf f,

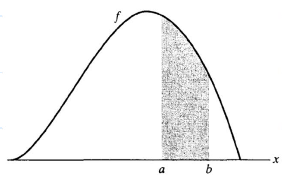
$$P(X \in A) = \int_A f(x)\, dx.$$

pdf of Uniform(0, 1)



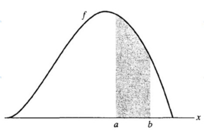
cdf of Uniform(0.1)





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

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



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

Question 2.3

How to interpret *f(x)*?

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

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.
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?

Example 2.2 (cont. Ex. 2.1)

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

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	$0\left(\frac{1}{16}\right)$	$\frac{1}{8}\left(\frac{3}{16}\right)$	$\frac{2}{8}\left(\frac{3}{16}\right)$	$\frac{1}{8}\left(\frac{1}{16}\right)$

Question 2.5

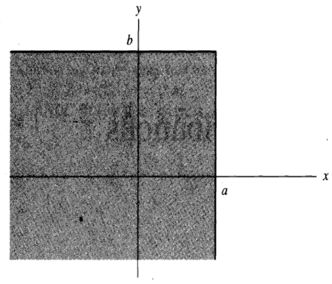
When we know the joint distribution, we can obtain every marginal distributions. Is the reverse statement true?

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

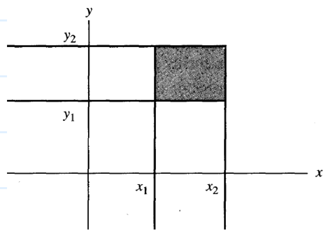
The **joint cdf** of $X_1, X_2, \ldots, X_n$ is

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

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



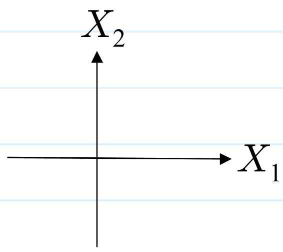
$$\begin{aligned} P(x_1 < X \le x_2, y_1 < Y \le y_2) &= F(x_2, y_2) - F(x_2, y_1) \\ &\quad - F(x_1, y_2) + F(x_1, y_1) \end{aligned}$$



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

- discrete multivariate case

$$\begin{aligned} p(x_1, x_2, \cdots, x_n) &= P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) \\ &\Rightarrow \textbf{joint pmf} \text{ of } X_1, X_2, \ldots, X_n \\ P((X_1, \ldots, X_n) \in A) &= \sum_{(x_1, \ldots, x_n) \in A} p(x_1, \ldots, x_n) \\ F(x_1, x_2, \cdots, x_n) &= \sum_{t_1 \le x_1,\, t_2 \le x_2,\, \ldots,\, t_n \le x_n} p(t_1, t_2, \ldots, t_n) \\ p_{X_1}(x_1) = P(X_1 = x_1) &= \sum_{-\infty < t_2 < \infty, \ldots, -\infty < t_n < \infty} p(x_1, t_2, \ldots, t_n) \end{aligned}$$

- continuous multivariate case

$$\begin{aligned} f(x_1, x_2, \cdots, x_n) &= \frac{\partial^n}{\partial x_1 \cdots \partial x_n} F(x_1, x_2, \cdots, x_n) \\ &\Rightarrow \textbf{joint pdf} \text{ of } X_1, X_2, \ldots, X_n \\ P((X_1, \ldots, X_n) \in A) &= \int \cdots \int_A f(x_1, \ldots, x_n)\, dx_1 \cdots dx_n \\ F(x_1, x_2, \cdots, x_n) &= \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f(t_1, t_2, \ldots, t_n)\, dt_n \cdots dt_1 \\ f_{X_1}(x_1) &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, t_2, \ldots, t_n)\, dt_2 \cdots dt_n \end{aligned}$$

• independent random variables

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 $x_1, x_2, \ldots, x_n$.

Theorem 2.2 (TBp. 85-86)

1. For continuous case,

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

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

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

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 $X_1, X_2, \ldots, X_n$

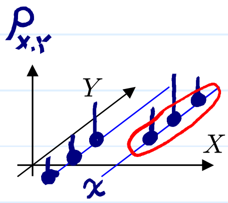
• conditional distribution

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

$$p_{Y|X}(y|x) \equiv P(Y = y|X = x) = \frac{P(X = x, Y = y)}{P(X = x)} = \frac{p_{XY}(x, y)}{p_X(x)}$$

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



Example 2.3 (cont. Ex 2.2)

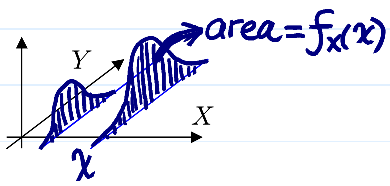
$p_{X_2|X_1}(0|1) = 2/3$, and $p_{X_2|X_1}(1|1) = 1/3$

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

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

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



Theorem 2.3

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

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

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

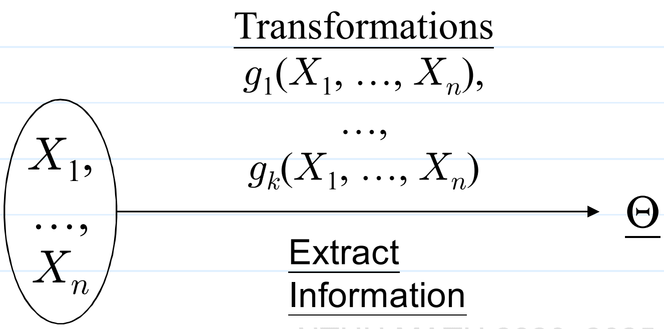
2. For each fixed x, $p_{Y|X}(y|x)$ is a pmf for y and $f_{Y|X}(y|x)$ is a pdf for y.
3. $p_{XY}(x,y) = p_{Y|X}(y|x)\ p_X(x)$, and $f_{XY}(x,y) = f_{Y|X}(y|x)\ f_X(x)$

   — multiplication law
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
5. $p_{X|Y}(x|y) = \frac{p_{Y|X}(y|x)\ p_X(x)}{\sum_x p_{Y|X}(y|x)p_X(x)}$, and $f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)\ f_X(x)}{\int_{-\infty}^{\infty} f_{Y|X}(y|x)f_X(x)dx}$,

   — Bayes' rule
6. X, Y are independent $\Leftrightarrow$ $p_{Y|X}(y|x) = p_Y(y)$ or $f_{Y|X}(y|x) = f_Y(y)$

• functions of random variables



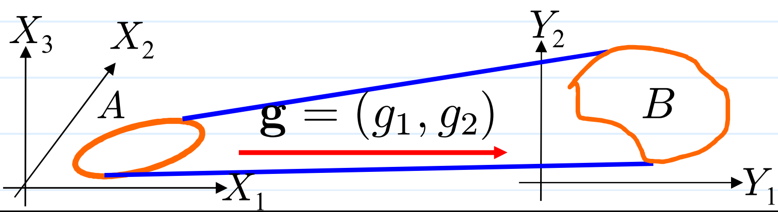
Question 2.6

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

1. method of events

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 $A = \mathbf{g}^{-1}(B)$.



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

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

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

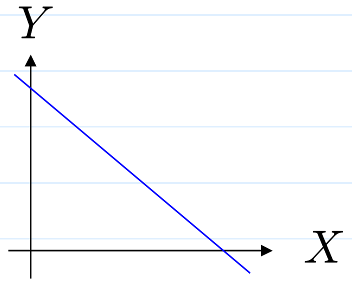
Example 2.5 (sum of two discrete random variables, TBp. 96)

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, $Z=X-Y$)

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

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

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

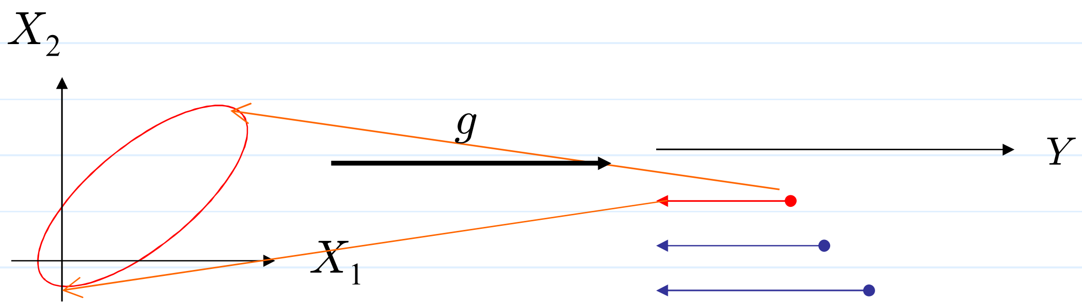


2. method of cumulative distribution function (a special case of method **1**)

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.
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 $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_m)$.



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

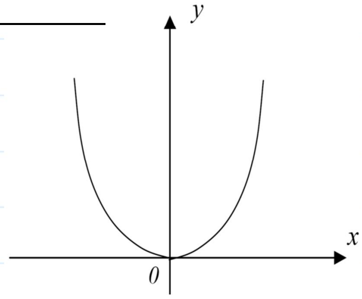
X is a random variables with pdf $f_X(x)$ and cdf $F_X(x)$. Find the distributon of $Y = X^2$.

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

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

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

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



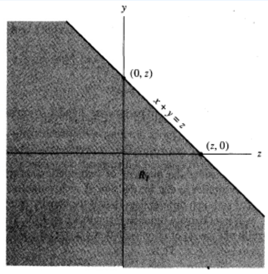
Example 2.7 (sum of two continuous random variables, TBp. 97)

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

(Exercise: difference of two random variables, $Z=X-Y$)

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

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



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

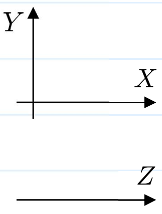
$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x)f_Y(z-x)dx \quad \Rightarrow \textbf{convolution of } f_X \textbf{ and } f_Y$$

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, $Z=XY$)

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

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



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

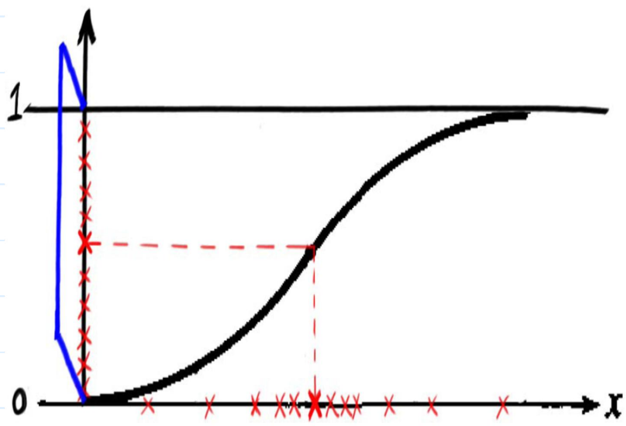
$$\left(= \int_{-\infty}^{\infty} |x| f_X(x) f_Y(xz)\,dx \quad \text{when } X, Y \text{ independent} \right)$$

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 $[0, 1]$.

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.



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**) :

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,

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

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

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

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

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

Then

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

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

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

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

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

Therefore,

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

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

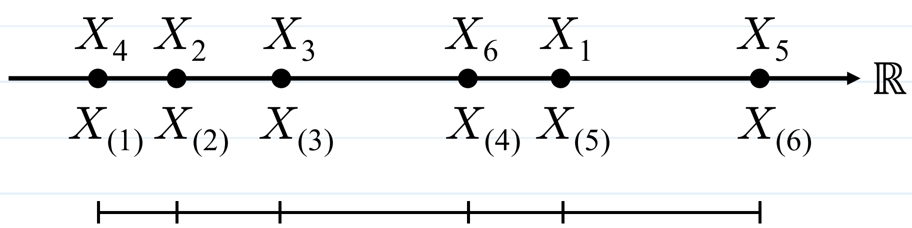
4. **method of moment generating function:** based on the *uniqueness theorem* of moment generating function. To be explained later in Chapter 4.

- extrema and order statistics

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 $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ 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.

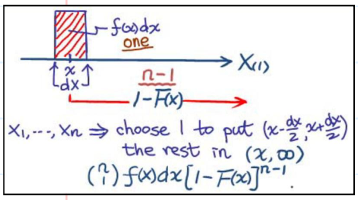
Definition 2.12 (i.i.d.)

$X_1, X_2, \ldots, X_n$ 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.

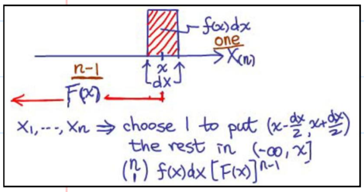
Theorem 2.8 (TBp. 104)

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 $nf(x)[F(x)]^{n-1}$.



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

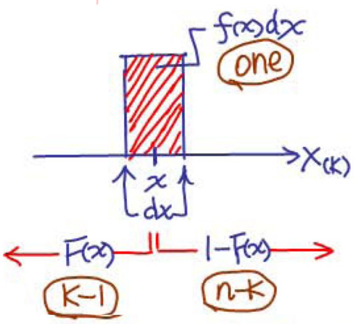


$$
\begin{aligned}
1 - F_{X_{(1)}}(x) &= P(X_{(1)} > x) = P(X_1 > x, \ldots, X_n > x) \\
&= P(X_1 > x) \cdots P(X_n > x) \\
&= [1-F(x)]^n.
\end{aligned}
$$

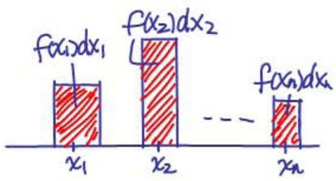
Theorem 2.9 (TBp. 105)

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

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



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



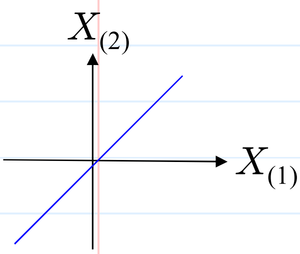
$\binom{n}{1\ 1 \cdots 1} f(x_1)dx_1 \cdots f(x_n)dx_n$

Theorem 2.10 (TBp. 114, Problem 73)

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

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

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?

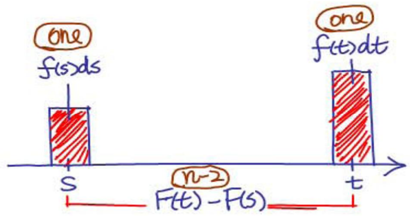
Example 2.11 (range, TBp. 105-106)

The joint pdf of $X_{(1)}$ and $X_{(n)}$ is

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

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

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



$X_1, \ldots, X_n \Rightarrow$ 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

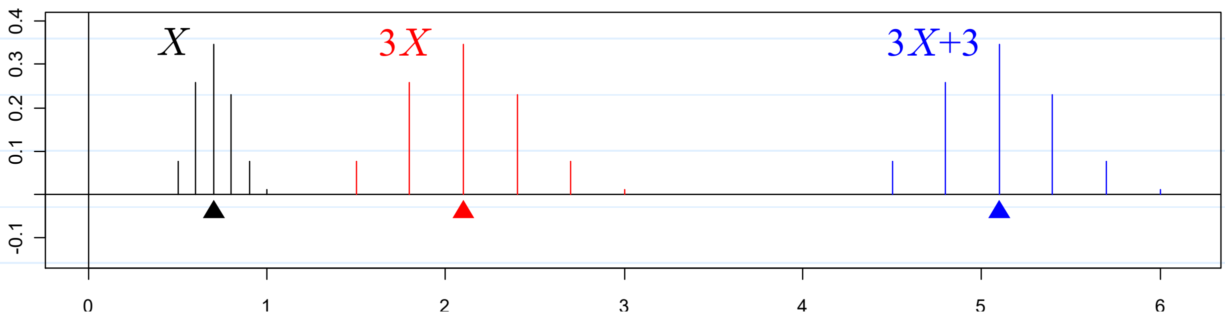
Chapter 4

Outline

- expectation
  - 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?



• 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

$$\begin{aligned} E(Y) &\equiv \sum_{-\infty<y<\infty} y p_Y(y) = E[g(X_1, \ldots, X_n)] \\ &\equiv \sum_{-\infty<x_1<\infty, \ldots, -\infty<x_n<\infty} g(x_1, \ldots, x_n) p(x_1, \ldots, x_n), \end{aligned}$$

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

$$\begin{aligned} E(Y) &\equiv \int_{-\infty}^{\infty} y f_Y(y) dy = E[g(X_1, \ldots, X_n)] \\ &\equiv \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, \ldots, x_n) f(x_1, \ldots, x_n) dx_1 \cdots dx_n, \end{aligned}$$

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) $g(x) = (x - \mu_X)^2 \Rightarrow E[g(X)] = E[(X - E(X))^2]$ 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**.

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 $\rho_{XY} = 0$.

Notes. (intuitive explanation of mean)

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.
3. Mean can be interpreted as a long-run average. (see Chapter 5.)

Notes. (intuitive explanation of variance and standard deviation)

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 $a, b_1, \ldots, b_n \in \mathbb{R}$,
$$E\left(a + \sum_{i=1}^n b_i X_i\right) = a + \sum_{i=1}^n b_i E(X_i).$$
2. (TBp.124) If X, Y are independent, then
$$E(g(X)h(Y)) = E(g(X))E(h(Y)).$$
In particular, $E(XY) = E(X)E(Y)$.
(Question 3.2: $E(X/Y) = E(X)/E(Y)$?)
Note. $E[g(X)] \neq g[E(X)]$ in general.

Theorem 3.2 (properties of variance and standard deviation)

1. (TBp.132) $\sigma_X^2 = Var(X) = E[(X-\mu_X)^2] = E(X^2) - \mu_X^2$.
2. (TBp.131) $Var(a + bX) = b^2 Var(X)$, $a, b \in \mathbb{R}$, and $\sigma_{a+bX} = |b|\sigma_X$.
3. (TBp.140)
$$Var\left(a + \sum_{i=1}^n b_i X_i\right) = \sum_{i=1}^n b_i^2 Var(X_i) + 2 \sum_{1 \le i < j \le n} b_i b_j Cov(X_i, X_j).$$
In particular, $Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$.
4. (TBp.140) If $X_1, \ldots, X_n$ are independent,
$$Var\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n Var(X_i).$$
5. (TBp.136) $E[(X-\theta)^2] = Var(X) + (\mu_X - \theta)^2$ (Mean square error = variance + bias square)

Theorem 3.3 (Chebyshev's inequality, TBp. 133)

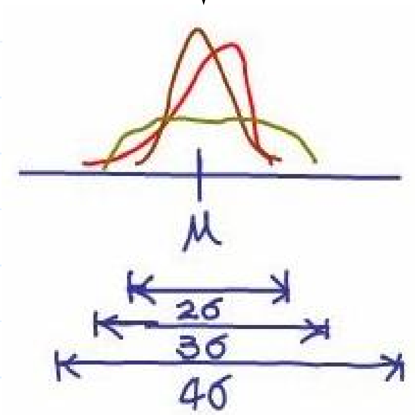
Let X be a random variable with mean μ and variance σ^2. Then for any $t > 0$,

$$P\big(|X - \mu| > t\big) \leq \frac{\sigma^2}{t^2}.$$

Proof. Let $f(x)$ be the pdf of X. Let $R = \{x : |x - \mu| > t\}$. Then

$$\begin{aligned} P\big(|X-\mu|>t\big) &= \int_R f(x)\,dx \leq \int_R \frac{(x-\mu)^2}{t^2} f(x)\,dx \\ &\leq \int_{-\infty}^{\infty} \frac{(x-\mu)^2}{t^2} f(x)\,dx = \frac{\sigma^2}{t^2}. \end{aligned}$$

No other restriction on the functional form of pdf/pmf



Note.

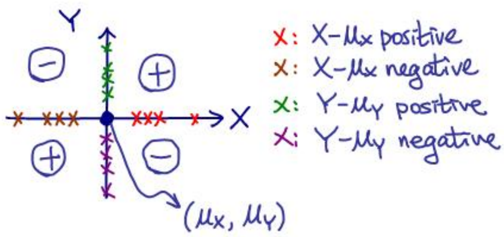
1. Setting $t = k\sigma$ we have

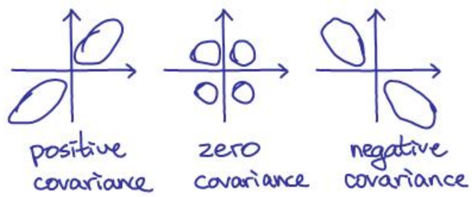
$$P\big(|X - \mu| > k\sigma\big) \leq \frac{1}{k^2}.$$

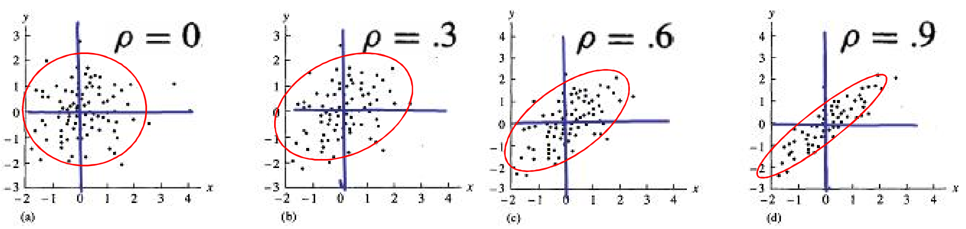
2. (TBp. 134) $\quad Var(X) = 0 \Rightarrow P(X = \mu_X) = 1.$

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.
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.
3. positive covariance and negative covariance
4. correlation coefficient is unit free
5. correlation coefficient measures the strength of the linear relationship between X and 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. $Cov(X,X) = Var(X)$.)

2. (TBp.140)

$$Cov\left(a + \sum_{i=1}^{n} b_i X_i, c + \sum_{j=1}^{m} d_j Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} b_i d_j Cov(X_i, Y_j)$$

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, $b \in \mathbb{R}$.

5. $\rho_{XY} = E\left[\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right]$

6. $|Cor(a+bX, c+dY)| = |Cor(X,Y)|$

• moment generating function & characteristics function

Definition 3.3 (moment generating function, TBp. 155)

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

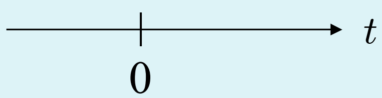
$$M_X(t) = E(e^{tX}), \quad t \in \mathbb{R}$$

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.

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.



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

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

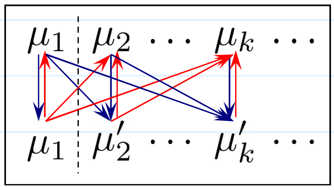
4. (TBp.158) For any constants a, b, $M_{a+bX}(t) = e^{at} M_X(bt)$.
5. (TBp.159) X, Y independent $\Rightarrow M_{X+Y}(t) = M_X(t) M_Y(t)$.
6. continuity theorem (see Chapter 5)

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 $E[(X - \mu_X)^k] \equiv \mu'_k$.

➢ Some Notes.

- $\mu'_k = \sum_{i=0}^{k} \binom{k}{i} (-\mu_X)^{n-i} \mu_i$.
- $\mu_k = \sum_{i=0}^{k} \binom{k}{i} (\mu_X)^{n-i} \mu'_i$.
- 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:

$$M_{X_1 X_2 \cdots X_n}(t_1, t_2, \ldots, t_n) = E(e^{t_1 X_1 + t_2 X_2 + \cdots + t_n X_n})$$

if the expection exists.

Theorem 3.6 (properties of joint mgf)

1. $M_{X_1}(t_1) = M_{X_1 X_2 \cdots X_n}(t_1, 0, \ldots, 0)$
2. uniqueness theorem
3. $X_1, X_2, \ldots, X_n$ are independent if and only if

$$M_{X_1 X_2 \cdots X_n}(t_1, t_2, \ldots, t_n) = \prod_{i=1}^{n} M_{X_i}(t_i).$$

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

Definition 3.6 (characteristic function, TBp. 161)

The **characteristic function (chf)** of a random variable X is

$$\phi_X(t) = E(e^{itX}) = E[\cos(tX)] + i \cdot E[\sin(tX)],$$

where $i = \sqrt{-1}$, and the joint characteristic function of $X_1, X_2, \ldots, X_n$ is

$$\phi_{X_1X_2\cdots X_n}(t_1, t_2, \ldots, t_n) = E(e^{it_1X_1+it_2X_2+\cdots+it_nX_n}).$$

Theorem 3.7 (properties of characteristic function)

1. The characteristic function always exists.
2. If $M_X(t)$ exists, then $\phi_X(t) = M_X(it)$.
3. uniqueness theorem
4. (FYI) inversion theorem:
   - discrete case: $p_X(x) = \lim_{T\to\infty} \int_{-T}^{T} e^{-itx}\phi_X(t)\, dt$
   - continuous case: $f_X(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx}\phi_X(t)\, dt$
5. The properties of characteristic function are similar to those of moment generating function.

• conditional expectation

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

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

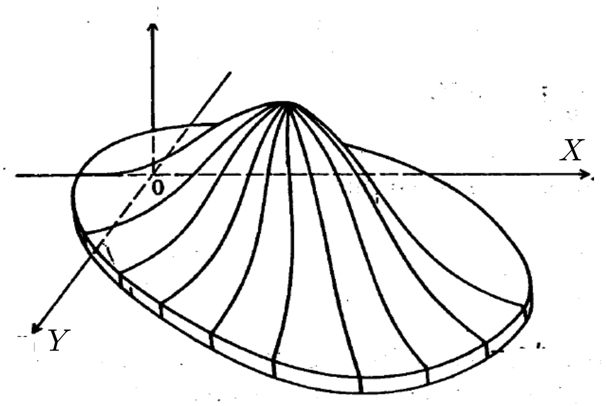
[**Discrete case**] : $E(h(Y)|X = x) = \sum_y h(y)p_{Y|X}(y|x)$

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

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

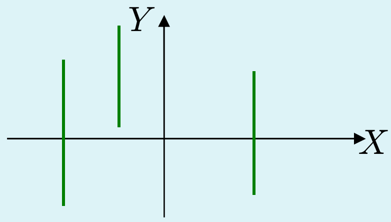
In particular, $E(Y|X = x) = \int y f_{Y|X}(y|x)dy$

$f(x, y)$: joint pdf

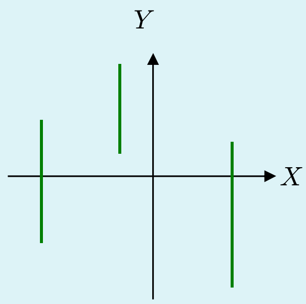


Theorem 3.8 (properties of conditional expectation)

1. $E_{Y|X}(h(Y)|x)$ is a function of x and is free of Y.
2. If X and Y are independent then $E_{Y|X}(h(Y)|x) = E_Y(h(Y))$.
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 $E_{Y|X}(h(Y)|X)$.
5. **law of total expection** (TBp.149)
   $$E_X[E_{Y|X}(h(Y)|X)] = E_Y[h(Y)].$$
   In particular,
   $$E_Y(Y) = E_X[E_{Y|X}(Y|X)].$$

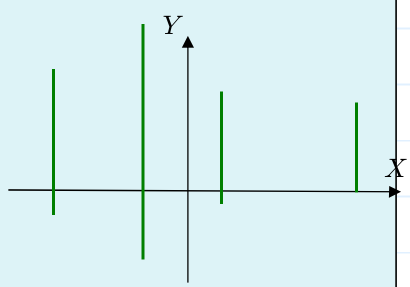


4. **variance decomposition** (TBp.151)
$$Var_Y(Y) = Var_X[E_{Y|X}(Y|X)] + E_X[Var_{Y|X}(Y|X)]$$

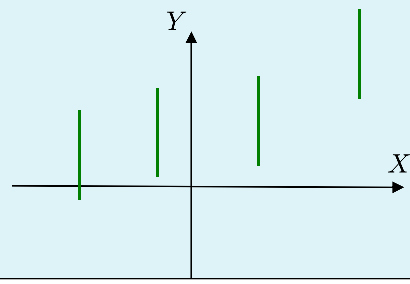


Note.

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.



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.



• 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?)
- **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) $G_1 = \{g(x) : g(x) = c, \text{ where } c \in \mathbb{R}\}$

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

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

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

$$\text{MSE} \equiv E_{X,Y}\{[Y - g(X)]^2\}.$$

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

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

The equality holds if and only if $c = E_Y(Y)$.

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

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

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

Notes for the best predictor in G_3.

- $E_{Y|X}(Y|X)$ is the best predictor of Y based on X, in the mean squared prediction error sense.
- need to know the joint distribution of X and Y, or at least $E_{Y|X}(Y|x)$
- $E_{Y|X}(Y|x)$ is called the regression function of Y on X.

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

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

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

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
- needs to know only the means, variances and covariances
- $\sigma_Y^2(1-\rho^2)$ is small if ρ is close to $+1$ or -1, and large if ρ is close to 0

Notes.

1. $\min_{a,b} E[Y-(a+bX)]^2 \le \min_c E(Y-c)^2$ and the equality holds if and only if $\rho = 0$.
2. $\min_g E(Y-g(X))^2 \le \min_{a,b} E[Y-(a+bX)]^2$ and the equality holds if and only if $E_{Y|X}(Y|x) = \mu_Y + \rho(\sigma_Y/\sigma_X)(x-\mu_X)$.

Question 3.3

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

- δ method

Question 3.3

Let $Y = g(X)$. Suppose we only know the mean μ_X and variance σ_X^2 of X, but **not** the entire distribution (i.e., do not know cdf, pdf/pmf of X). Can we derive the distribution of Y? If not, can we "roughly" describe the mean and variance of Y? (**Note.** $E[g(X)] \ne g[E(X)]$.)

Theorem 3.9 (δ method for univariate case, TBp. 162)

$$\begin{aligned}
Y = g(X) &\approx g(\mu_X) + (X-\mu_X)g'(\mu_X) \quad \text{(by Taylor expansion)}\\
\Rightarrow \quad E[g(X)] &\approx g(\mu_X)\\
Var[g(X)] &\approx Var(X)[g'(\mu_X)]^2\\
\text{or} \qquad Y = g(X) &\approx g(\mu_X) + (X-\mu_X)g'(\mu_X) + \frac{1}{2}(X-\mu_X)^2 g''(\mu_X)\\
\Rightarrow \quad E[g(X)] &\approx g(\mu_X) + \frac{1}{2}\sigma_X^2 g''(\mu_X)
\end{aligned}$$

Note. How good these approximations are depends on whether g can be reasonably well approximated by the 1st or 2nd order polynomials in a neighborhood of μ_X and on the size of σ_X.

Theorem 3.10 (δ method for multivariate case, TBp. 165)

Function of two univariate random variables $Z = g(X, Y)$:
Let $\mu = (\mu_X, \mu_Y)$.

$$Z = g(X,Y) \approx g(\mu) + (X-\mu_X)\frac{\partial g(\mu)}{\partial x} + (Y-\mu_Y)\frac{\partial g(\mu)}{\partial y}$$

$$\Rightarrow \quad \mathrm{E}(Z) \approx g(\mu)$$

$$\mathrm{Var}(Z) \approx \sigma_X^2 \left[\frac{\partial g(\mu)}{\partial x}\right]^2 + \sigma_Y^2 \left[\frac{\partial g(\mu)}{\partial y}\right]^2 + 2\sigma_{XY}\left[\frac{\partial g(\mu)}{\partial x}\right]\left[\frac{\partial g(\mu)}{\partial y}\right]$$

or

$$g(X,Y) \approx g(\mu) + (X-\mu_X)\frac{\partial g(\mu)}{\partial x} + (Y-\mu_Y)\frac{\partial g(\mu)}{\partial y}$$
$$+\frac{1}{2}(X-\mu_X)^2\frac{\partial^2 g(\mu)}{\partial x^2} + (X-\mu_X)(Y-\mu_Y)\frac{\partial^2 g(\mu)}{\partial x \partial y}$$
$$+\frac{1}{2}(Y-\mu_Y)^2\frac{\partial^2 g(\mu)}{\partial y^2}$$

$$\Rightarrow \quad E[g(X,Y)] \approx g(\mu) + \frac{1}{2}\sigma_X^2\frac{\partial^2 g(\mu)}{\partial x^2} + \sigma_{XY}\frac{\partial^2 g(\mu)}{\partial x \partial y} + \frac{1}{2}\sigma_Y^2\frac{\partial^2 g(\mu)}{\partial y^2}$$

Note. The general case of a function of n random variables can be worked out similarly.

- ❖ **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

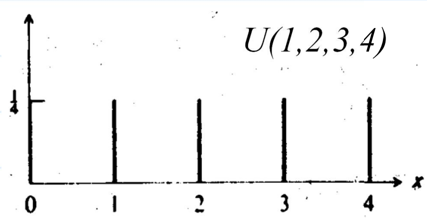
For a given random phenomenon or data, what distribution (or statistical model) is more appropriate to depict it?

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



- **mgf:** $\frac{\sum_{j=1}^m e^{a_j t}}{m}$
- **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.

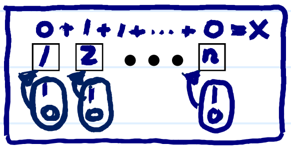
- **pmf:** $p(x) = \begin{cases} p^x(1-p)^{(1-x)}, & \text{if } x = 0 \text{ or } x = 1 \\ 0, & \text{otherwise} \end{cases}$
- **mgf:** $pe^t + 1 - p$
- **mean:** p
- **variance:** $p(1-p)$
- **parameter:** $p \in [0, 1]$
- **example:** toss a coin once, p=probability that head occurs

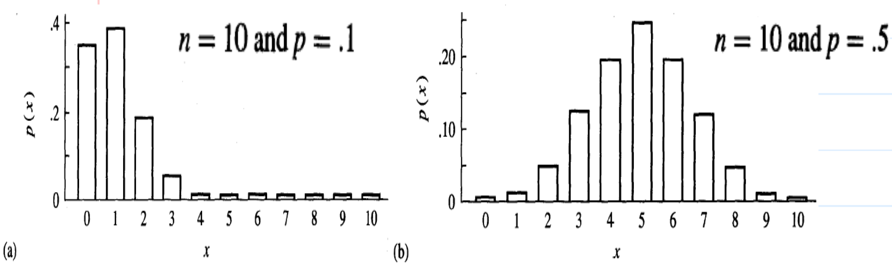
Note: If A is an event, then the indicator random variable I_A follows the Bernoulli distribution.

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.

- **pmf:** $p(x) = \begin{cases} \binom{n}{x} p^x(1-p)^{(n-x)}, & x = 0, 1, \ldots, n \\ 0, & \text{otherwise} \end{cases}$
- **mgf:** $(pe^t + 1 - p)^n$, $t \in \mathbb{R}$.
- **mean:** np
- **variance:** $np(1-p)$
- **parameter:** $p \in [0, 1]$, $n = 1, 2, \ldots$
- **example:** # of heads, toss a coin n times

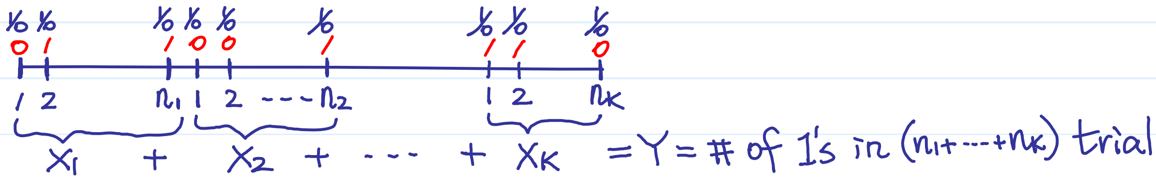




Note:
$(a+b)^n = \sum_{x=0}^{n} \binom{n}{x} a^x b^{n-x}$.

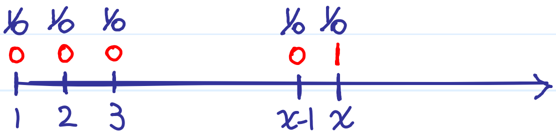
Note.

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 $Y = X_1 + \cdots + X_n \sim B(n, p)$.
3. Let $X_i \sim B(n_i, p), i = 1, \ldots, k$, and $X_1, \ldots, X_k$ are independent. Then, $Y = X_1 + \cdots + X_k \sim B(n_1 + \cdots + n_k, p)$.



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.



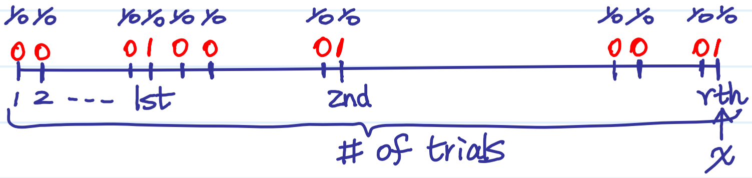
- **pmf:** $p(x) = \begin{cases} (1-p)^{(x-1)}p, & x = 1, 2, 3, \ldots \\ 0, & \text{otherwise} \end{cases}$
- **cdf:** $F(x) = \begin{cases} 1 - (1-p)^{[x]}, & 1 \le [x] \le x < [x] + 1 \\ 0, & x < 1 \end{cases}$
- **mgf:** $\frac{pe^t}{1-(1-p)e^t}$, $t < -\log(1-p)$.
- **mean:** $\frac{1}{p}$
- **variance:** $\frac{1-p}{p^2}$
- **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

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.

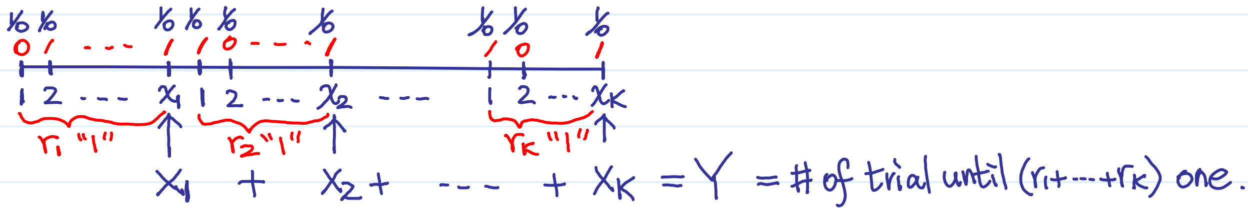


- **pmf:** $p(x) = \begin{cases} \binom{x-1}{r-1} p^r (1-p)^{(x-r)}, & x = r, r+1, \ldots \\ 0, & \text{otherwise} \end{cases}$
- **mgf:** $\frac{p^r e^{rt}}{[1-(1-p)e^t]^r}$, $t < -\log(1-p)$.
- **mean:** $\frac{r}{p}$
- **variance:** $\frac{r(1-p)}{p^2}$
- **parameter:** $p \in [0, 1]$, $r = 1, 2, \ldots$
- **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
2. Let $X_1, X_2, \ldots, X_r$ be i.i.d. $G(p)$, then $Y = X_1 + \cdots + X_r \sim NB(r, p)$.
3. Let $X_i \sim NB(r_i, p), i = 1, \ldots, k$, and $X_1, \ldots, X_k$ are independent. Then, $Y = X_1 + \cdots + X_k \sim NB(r_1 + \cdots + r_k, p)$.



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:**

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

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

$$\boxed{\textbf{Note: } (a_1+\cdots+a_k)^n=\sum_{x_1+\cdots+x_k=n}\begin{pmatrix} n \\ x_1,\cdots,x_k\end{pmatrix}a_1^{x_1}\cdots a_k^{x_k}.}$$

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)

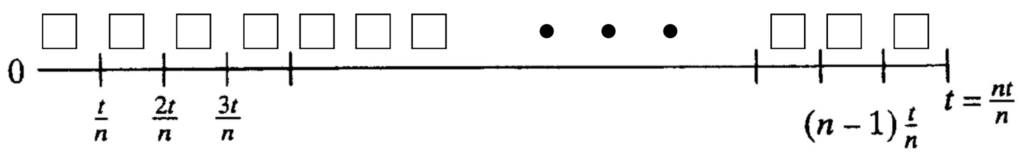
Limit of binomial distributions $X_n\sim B(n,p_n)$, where $p_n\to 0$ as $n\to\infty$ in such a way that $\lambda_n\equiv np_n\to\lambda$.

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

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

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 $\lambda=np$.

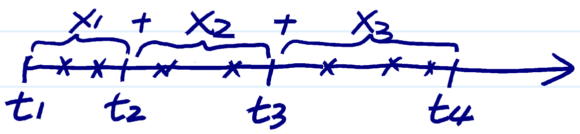


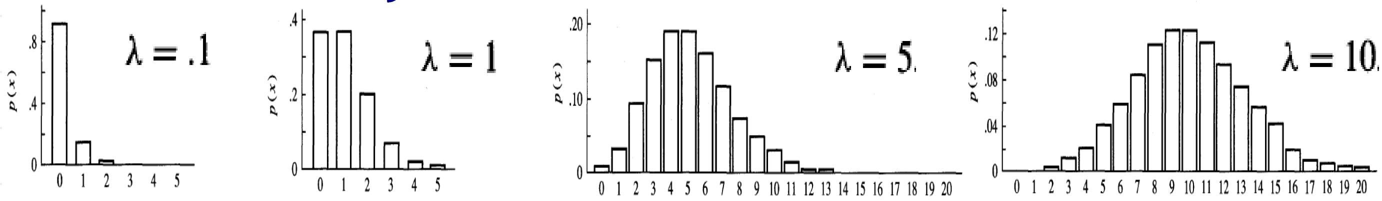
2. Let X be the number of times some event occurs in a given time interval I. Divide the interval into many small subintervals I_k, $k = 1, \ldots, n$, of equal length. Let N_k be the number of events occurring in I_k. When we can assume $N_1, \ldots, N_n$ are independent and approximately $\sim B(p)$,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

- **pmf:** $p(x) = \begin{cases} \frac{\lambda^x}{x!} e^{-\lambda}, & x = 0, 1, 2, \ldots \\ 0, & \text{otherwise} \end{cases}$
- **mgf:** $e^{\lambda(e^t - 1)}$, $t \in \mathbb{R}$.
- **mean:** λ
- **variance:** λ
- **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, $Y = X_1 + \cdots + X_k \sim P(\lambda_1 + \cdots + \lambda_k)$.





(a)

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.

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

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

- **mgf:** exist, but no simple expression
- **mean:** $\frac{rn}{n+m}$
- **variance:** $\frac{rnm(n+m-r)}{(n+m)^2(n+m-1)}$
- **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,

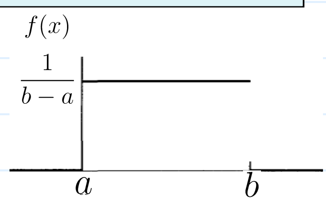
$$\frac{\binom{n}{x}\binom{m}{r-x}}{\binom{n+m}{r}} \to \binom{r}{x} p^x (1-p)^{r-x}.$$

- continuous distributions

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

Choose a number at random between a and b.

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

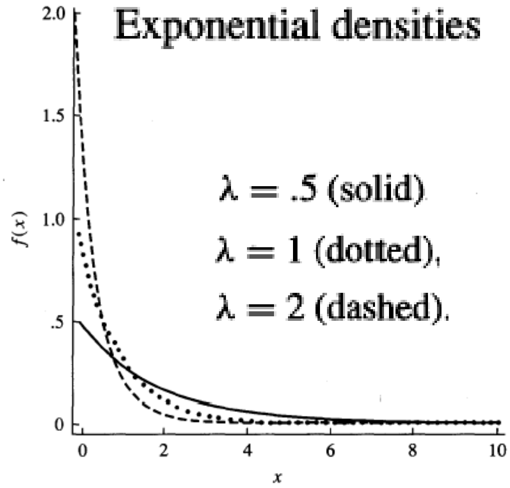


- **cdf:** $F(x) = \begin{cases} 0, & x < a \\ \frac{x-a}{b-a}, & a \le x \le b \\ 1, & x > b \end{cases}$
- **mgf:** $\frac{e^{bt}-e^{at}}{t(b-a)}, \; t \in \mathbb{R}.$
- **mean:** $\frac{a+b}{2}$
- **variance:** $\frac{(b-a)^2}{12}$
- **parameter:** $a, b \in \mathbb{R}, \quad a < b$

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

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

- **pdf:** $f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$
- **cdf:** $F(x) = \begin{cases} 1 - e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$
- **mgf:** $\frac{\lambda}{\lambda - t}$, $t < \lambda$.
- **mean:** $\frac{1}{\lambda}$
- **variance:** $\frac{1}{\lambda^2}$
- **parameter:** $\lambda > 0$
- **example:** lifetime or waiting time

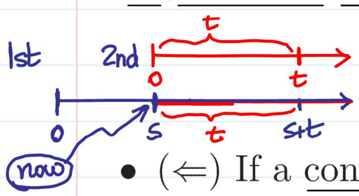


Notes:

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

$$P(T > t + s | T > s) = \frac{P(T > t + s \text{ and } T > s)}{P(T > s)} = \frac{P(T > t + s)}{P(T > s)}$$

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



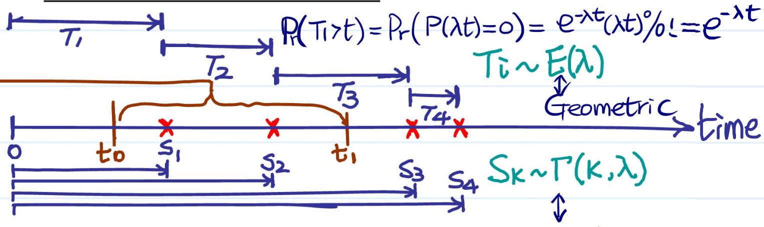
- ($\Leftarrow$) If a continuous distribution is memoryless, it is exponential.
- It does not mean the two events $T > s$ and $T > t+s$ are independent.

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 $[t_0, t_1]$ $\sim P(\lambda(t_1 - t_0))$ Binomial



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

$T_i \sim E(\lambda)$ Geometric

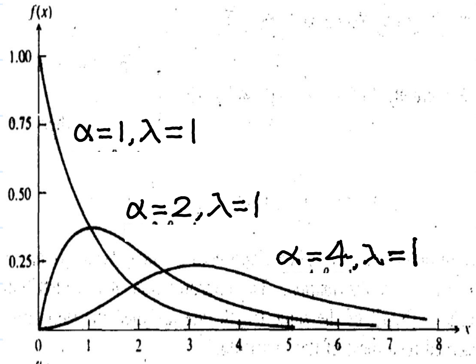
$S_k \sim \Gamma(k, \lambda)$ Negative Binomial

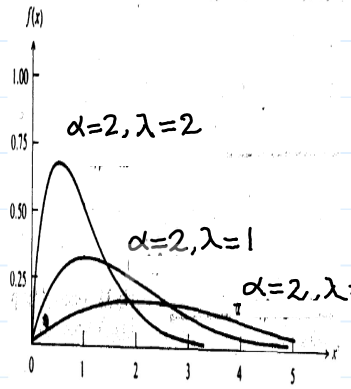
Question 4.2: how to interpret λ?

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)

- **pdf:** $f(x) = \begin{cases} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$
- **mgf:** $(\frac{\lambda}{\lambda - t})^\alpha,\ t < \lambda.$
- **mean:** $\frac{\alpha}{\lambda}$
- **variance:** $\frac{\alpha}{\lambda^2}$
- **parameter:** $\alpha, \lambda > 0$





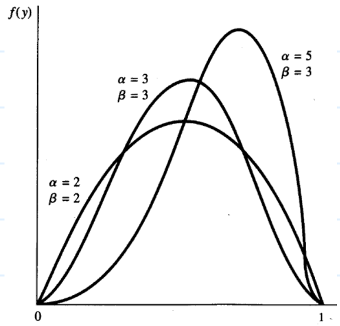
Notes.

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

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., $\Gamma(1, \lambda) = E(\lambda)$.
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 $Y = X_1 + \cdots + X_k \sim \Gamma(\alpha_1 + \cdots + \alpha_k, \lambda)$.
6. Let $X \sim \Gamma(\alpha, \lambda)$,, then $cX \sim \Gamma(\alpha, \lambda/c)$, where $c > 0$.
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 $0 < k < \alpha$.

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

- **pdf:** $f(x) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, & 0 \le x \le 1 \\ 0, & \text{otherwise} \end{cases}$
- **mgf:** $1 + \sum_{k=1}^{\infty} \left(\prod_{r=0}^{k-1} \frac{\alpha+r}{\alpha+\beta+r} \right) \frac{t^k}{k!}$
- **mean:** $\frac{\alpha}{\alpha+\beta}$
- **variance:** $\frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}$
- **parameter:** $\alpha, \beta > 0$

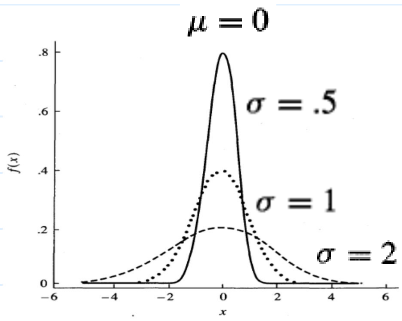


Notes:

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. $\beta(1,1) = U(0,1)$
3. Let $X_1 \sim \Gamma(\alpha_1, \lambda)$, $X_2 \sim \Gamma(\alpha_2, \lambda)$, and X_1, X_2 independent. Then, $\frac{X_1}{X_1+X_2} \sim beta(\alpha_1, \alpha_2)$.

Definition 4.13 (Normal distribution $N(\mu, \sigma^2)$, sec. 2.2.3)

- **pdf:** $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}.$
- **mgf:** $e^{\mu t + \frac{\sigma^2 t^2}{2}}$, $t \in \mathbb{R}$
- **mean:** μ
- **variance:** σ^2
- **parameter:** $\mu \in \mathbb{R}, \sigma > 0$

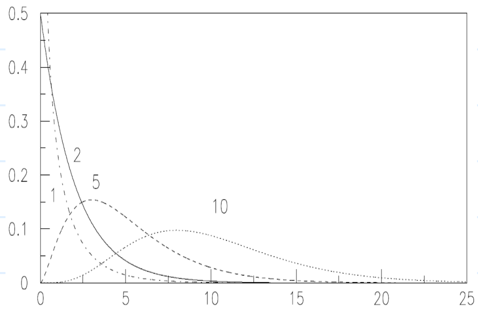


Notes:

1. bell-shaped pdf (symmetric about μ, where it has maximum, and falls off in the rate determined by σ)
2. play a central role in probability and statistics (e.g., CLT, Chapter 5)
3. $X \sim N(\mu, \sigma^2) \Rightarrow$ for $a, b \in \mathbb{R}$, $aX + b \sim N(a\mu + b, a^2\sigma^2)$. In particular, $\frac{X-u}{\sigma} \sim N(0,1)$.
4. Let $X_1, \ldots, X_k$ be independent, and $X_i \sim N(\mu_i, \sigma_i^2)$. Then $Y = X_1 + \cdots + X_k \sim N(\sum_{i=1}^k \mu_i, \sum_{i=1}^k \sigma_i^2)$
5. by 3 and 4, let $X_1, \ldots, X_k$ be i.i.d $\sim N(\mu, \sigma^2)$, then $\overline{X}_k \sim N(\mu, \frac{\sigma^2}{k})$.

Definition 4.14 (Chi-square distribution χ^2_n, TBp.177)

- **pdf:** $f(x) = \begin{cases} \frac{1}{\Gamma(\frac{n}{2})2^{\frac{n}{2}}} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}, & x \geq 0 \\ 0, & x \leq 0 \end{cases}$
- **mgf:** $(\frac{1}{1-2t})^{\frac{n}{2}}$
- **mean:** n
- **variance:** $2n$
- **parameter:** $n = 1, 2, 3, \ldots$



Notes:

1. n is called *degree of freedom*
2. $\chi^2_n = \Gamma(\frac{n}{2}, \frac{1}{2})$
3. Let $X_1, \ldots, X_k$ be independent and $X_i \sim \chi^2_{n_i}$, then $Y = X_1 + \ldots + X_k \sim \chi^2_{n_1+\cdots+n_k}$
4. Let $Z \sim N(0,1)$, then $X = Z^2 \sim \chi^2_1$.
5. By 3 and 4, let $Z_1, \ldots, Z_n$ be i.i.d. $\sim N(0,1)$, then $Y = Z_1^2 + \cdots + Z_n^2 \sim \chi^2_n$.

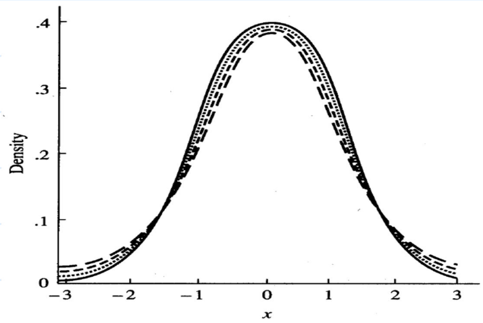
Definition 4.15 (t distribution t_n, TBp.178)

- **pdf:** $f(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}, \quad x \in \mathbb{R}.$
- **mgf:** not exist, except at $t = 0$
- **mean:** 0, $(n > 1)$
- **variance:** $\frac{n}{n-2}$, $(n > 2)$
- **moments:**

$$E(X^k) = \begin{cases} \frac{\Gamma(\frac{k+1}{2})\Gamma(\frac{n-k}{2})}{\sqrt{\pi}\Gamma(\frac{n}{2})} n^{\frac{k}{2}}, & k < n \text{ and even} \\ 0, & k < n \text{ and odd} \end{cases}$$

- **parameter:** $n = 1, 2, 3, \ldots$

t_5 (long dash), t_{10} (short dash), t_{30} (dot), N(0,1) (solid)

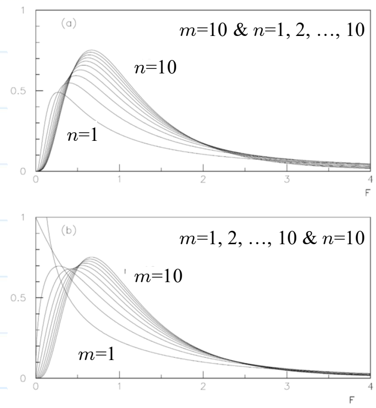


Notes:

1. Let $Z \sim N(0,1)$ and $U \sim \chi^2_n$ be independent, then $\frac{Z}{\sqrt{U/n}} \sim t_n$.
2. $f(x) = f(-x)$, i.e., t_n distribution is symmetric about zero
3. as $n \to \infty$, t_n tends to $N(0,1)$. (by LLN, Chapter 5)
4. t_n has heavier tail than $N(0,1)$

Definition 4.16 ($\underline{F}$ distribution $\underline{F_{m,n}}$, TBp.179)

- **pdf:** $f(x) = \begin{cases} \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \left(\frac{m}{n}\right)^{-\frac{m}{2}} x^{\frac{m}{2}-1} \left(1+\frac{m}{n}x\right)^{-\frac{m+n}{2}}, & x \geq 0 \\ 0, & x < 0 \end{cases}$
- **mgf:** not exist, except at $t = 0$
- **mean:** $\frac{n}{n-2}$, $(n > 2)$
- **variance:** $\frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}$, $(n > 4)$
- **moments:** $E(X^k) = \frac{\Gamma(\frac{m+2k}{2})\Gamma(\frac{n-2k}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \left(\frac{n}{m}\right)^k$, $k < \frac{n}{2}$
- **parameter:** $m, n = 1, 2, 3, \ldots$



Notes:

1. Let $U \sim \chi^2_m$ and $V \sim \chi^2_n$ be independent, then $\frac{U/m}{V/n} \sim F_{m,n}$.

2. Let $X \sim t_n$, then $Y = X^2 \sim F_{1,n}$.

3. $X \sim F_{m,n} \Rightarrow X^{-1} \sim F_{n,m}$

Theorem 4.1 (distributions of sample mean and sample variance of i.i.d. normal, sec. 6.3)

Let $X_1, X_2, \ldots, X_n$ be i.i.d. $\sim N(\mu, \sigma^2)$. Define

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{(called } \textit{sample mean}\text{)}$$

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \quad \text{(called } \textit{sample variance}\text{)}$$

Then,

1. $\overline{X}_n \sim N(\mu, \frac{\sigma^2}{n})$ and $\sqrt{n}(\overline{X}_n - \mu)/\sigma \sim N(0, 1)$.
2. (TBp.195) The random variable $\overline{X}_n$ and the random vector $(X_1 - \overline{X}_n, X_2 - \overline{X}_n, \ldots, X_n - \overline{X}_n)$ are independent.
3. (TBp.196) $\overline{X}_n$ and S_n^2 are independently.
4. (TBp.197) The distribution of $(n-1)S_n^2/\sigma^2$ is the chi-square distribution with $n-1$ degrees of freedom.
5. (TBp.198) $$\frac{\sqrt{n}(\overline{X}_n - \mu)}{S_n} \sim t_{n-1}.$$

Proof of 2. The joint mgf of $\overline{X}_n$ and $(X_1-\overline{X}_n, X_2-\overline{X}_n, \ldots, X_n-\overline{X}_n)$ is

$$M(s,t_1,t_2,\ldots,t_n) = E\left\{e^{\left[s\overline{X}_n+\sum_{i=1}^n t_i(X_i-\overline{X}_n)\right]}\right\} = E\left\{e^{\left[\sum_{i=1}^n\left(\frac{s}{n}+t_i-\bar{t}\right)X_i\right]}\right\}.$$

Let $a_i = \frac{s}{n}+t_i-\bar{t}$, $i=1,2,\ldots,n$. Then

$$\sum_{i=1}^n a_i = s, \quad \sum_{i=1}^n a_i^2 = \frac{s^2}{n}+\sum_{i=1}^n(t_i-\bar{t})^2.$$

Now we have

$$\begin{aligned} M(s,t_1,t_2,\ldots,t_n) &= \prod_{i=1}^n M_{X_i}(a_i) = \prod_{i=1}^n \exp\left(\mu a_i+\frac{\sigma^2}{2}a_i^2\right) \\ &= \exp\left(\mu\sum_{i=1}^n a_i+\frac{\sigma^2}{2}\sum_{i=1}^n a_i^2\right) = \exp\left[\mu s+\frac{\sigma^2}{2}\frac{s^2}{n}+\frac{\sigma^2}{2}\sum_{i=1}^n(t_i-\bar{t})^2\right] \\ &= \exp\left(\mu s+\frac{\sigma^2}{2n}s^2\right)\exp\left[\frac{\sigma^2}{2}\sum_{i=1}^n(t_i-\bar{t})^2\right] \end{aligned}$$

Thus, the joint mgf factorizes into product of the mgf of $\overline{X}_n$ and the mgf of $(X_1-\overline{X}_n, X_2-\overline{X}_n, \ldots, X_n-\overline{X}_n)$.

Question 4.4: Are $(X_1-\overline{X}_n, \ldots, X_n-\overline{X}_n)$ independent? [Hint: $\sum_{i=1}^n(X_i-\overline{X}_n)=0$]

Proof of 4. First note that

$$\frac{1}{\sigma^2}\sum_{i=1}^n(X_i-\mu)^2 = \sum_{i=1}^n\left(\frac{X_i-\mu}{\sigma}\right)^2 \sim \chi_n^2.$$

Also,

$$\underbrace{\frac{1}{\sigma^2}\sum_{i=1}^n(X_i-\mu)^2}_{W} = \underbrace{\frac{1}{\sigma^2}\sum_{i=1}^n(X_i-\overline{X}_n)^2}_{U} + \underbrace{\left(\frac{\sqrt{n}(\overline{X}_n-\mu)}{\sigma}\right)^2}_{V}.$$

$$\boxed{+\frac{2}{\sigma^2}\sum_{i=1}^n(X_i-\overline{X}_n)(\overline{X}_n-\mu)=0}$$

Since V and U are independent,

$$M_W(t) = M_U(t)M_V(t)$$

and then

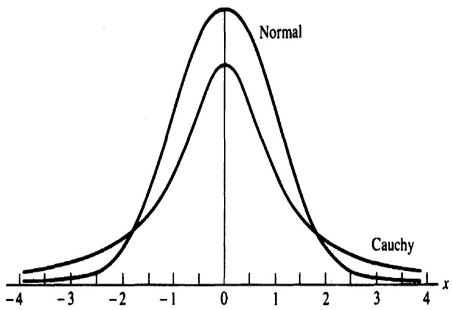
$$M_U(t) = \frac{M_W(t)}{M_V(t)} = \frac{(1-2t)^{-\frac{n}{2}}}{(1-2t)^{-\frac{1}{2}}} = (1-2t)^{-\frac{n-1}{2}},$$

which is the mgf of a χ_{n-1}^2 distribution. Thus $\frac{(n-1)S^2}{\sigma^2}\sim\chi_{n-1}^2$

Question 4.5: Why degree of freedom $= n-1$, rather than n?

Definition 4.17 (Cauchy distribution $C(\mu, \sigma)$, TBp.95)

- **pdf:** $f(x) = \frac{\sigma}{\pi} \frac{1}{\sigma^2+(x-\mu)^2}$, $x \in \mathbb{R}$.
- **cdf:** $\frac{1}{2} + \frac{1}{\pi}\tan^{-1}(\mu + \sigma x)$, $x \in \mathbb{R}$.
- **mgf:** not exist, except at $t = 0$
- **chf:** $e^{i\mu t - \sigma|t|}$
- **mean:** not exist
- **variance:** not exist
- **parameter:** $\mu \in \mathbb{R}, \sigma > 0$



Notes:

1. a heavy tail distribution
2. $C(0, 1) = t_1$
3. Let X, Y be i.i.d. $\sim N(0, 1)$. Then, $X/Y \sim C(0, 1)$.
4. $X \sim C(\mu, \sigma) \Rightarrow$ for $a, b \in \mathbb{R}$, $aX + b \sim C(a\mu + b, |a|\sigma)$
5. Let $X_1, \ldots, X_k$ be independent, and $X_i \sim C(\mu_i, \sigma_i)$. Then $Y = X_1 + \cdots + X_k \sim C(\sum_{i=1}^k \mu_i, \sum_{i=1}^k \sigma_i)$
6. by 4 and 5, let $X_1, \ldots, X_k$ be i.i.d $\sim C(\mu, \sigma)$, then $\overline{X}_k \sim C(\mu, \sigma)$.

Some other distributions

Log-normal (TBp. 69)
Weibull (TBp. 69)
Double exponential (TBp. 111)
Logistic
Pareto (TBp. 323)
Maxwell (TBp. 121)

In this chapter, you should learn

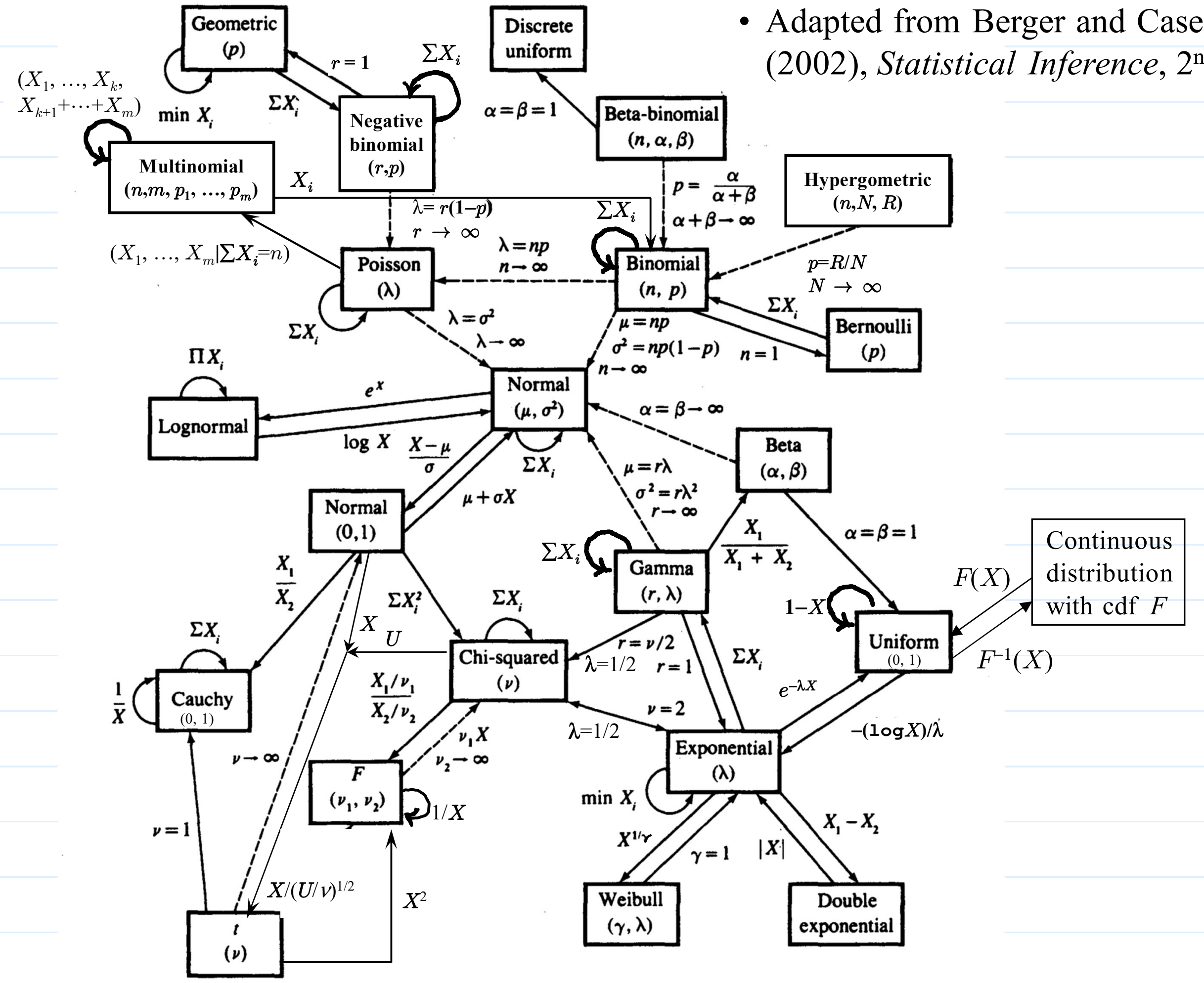
1. random phenomenon behind each distribution
2. statistical modeling (assigning a distribution) of data
3. relationship among distributions
4. meaning of parameters in each distribution
5. HOW to derive cdf/mgf/chf/mean/variance from pdf/pmf (optional)

but you are not necessary to

1. memorize their pdf/pmf/cdf/mgf/chf/mean/variance/…

❖ **Reading**: textbook, 2.1.1~2.1.5, 2.2.1~2.2.4, chapter 6
Roussas, 3.2, 3.3, 3.4, 5.2, 6.3, 7.3

- Adapted from Berger and Casella (2002), *Statistical Inference*, 2nd Ed., p.627



Relationships among common distributions. Solid lines represent transformations and special cases, dashed lines represent limits. Adapted from Leemis (1986).

Chapter 5

Outline

- 3 types of convergence
  - a.s., in prob., in dist.
- law of large number
- central limit theorem

Question 5.1

1. repeat flipping a fair coin **2 or 3 times**. Can you accurately predict the average appearance of heads?
2. repeat flipping a fair coin **many many times**. What will you predict the average appearance of heads?

Note.

1. Some *deterministic* patterns emerge from *random* phenomena when more and more data are collected, i.e., more and more information is gathered.
2. In the following, $n \to \infty$ can be interpreted as *sample size of data is large enough*.

- three types of convergence

Definition 5.1 (converge almost surely, TBp. 178)

A sequence of random variables $\{Z_n : \Omega \to \mathbb{R}\}$ is said to **converge almost surely** to a random variable $Z : \Omega \to \mathbb{R}$, and denoted as $Z_n \xrightarrow{\text{a.s.}} Z$, if for any $\epsilon > 0$,

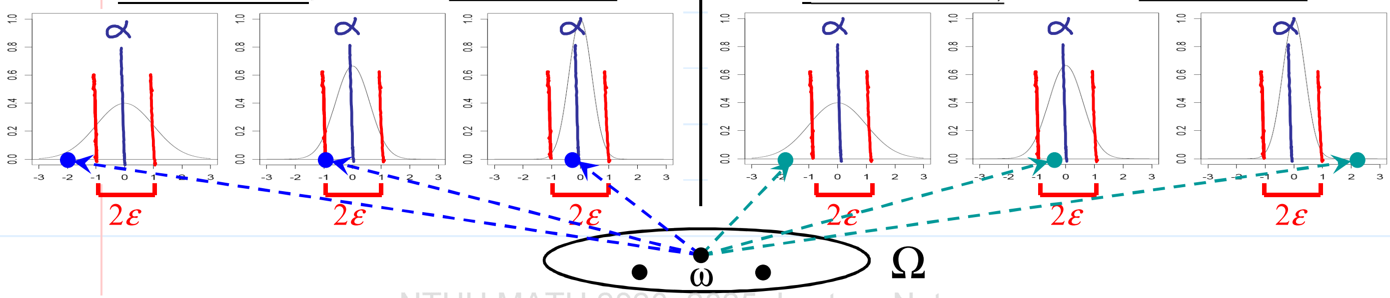
$$P\left(\left\{\omega \in \Omega : \lim_{n \longrightarrow \infty} |Z_n(\omega) - Z(\omega)| < \epsilon\right\}\right) = 1.$$

Definition 5.2 (converge in probability, TBp. 178)

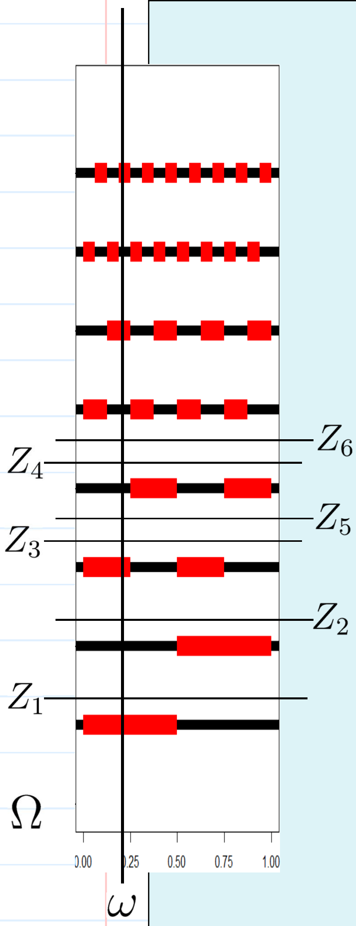
A sequence of random variables $\{Z_n : \Omega \to \mathbb{R}\}$ is said to **converge in probability** to a random variable $Z : \Omega \to \mathbb{R}$, and denoted as $Z_n \xrightarrow{P} Z$, if for any $\epsilon > 0$,

$$\lim_{n \to \infty} P\left(\{\omega \in \Omega : |Z_n(\omega) - Z(\omega)| < \epsilon\}\right) = 1.$$

$Z_n \xrightarrow{a.s.} \alpha$, α : a constant. | $Z_n \xrightarrow{P} \alpha$, α : a constant.



Example 5.1 (convergence in probability need not imply convergence a.s.)



- $\Omega = (0, 1]$
- P: uniform probability measure on Ω
- For $k = 1, 2, \ldots$, divide $(0, 1]$ into 2^k subintervals of equal length. These intervals are given by
$$I_{k,j} = \left(\tfrac{j-1}{2^k}, \tfrac{j}{2^k}\right]$$
for $j = 1, 2, \ldots, 2^k$.
- Let $Z_1, Z_2, \cdots : \Omega \to \mathbb{R}$ be a sequence of r.v.'s defined as follows:
$$Z_n(\omega) = \begin{cases} 1, & \text{if } \omega \in I_{k,j}, \\ 0, & \text{if } \omega \notin I_{k,j}, \end{cases}$$
where $n = 2^k + j - 2$.
- $Z : \Omega \to \mathbb{R}$ such that $Z(\omega) = 0$ for $\omega \in \Omega$
- $Z_n \xrightarrow{P} Z$, because for $0 < \epsilon < 1$,
$$P\left(\{\omega \in \Omega : |Z_n(\omega) - Z(\omega)| < \epsilon\}\right) = 1 - \frac{1}{2^k} \to 1.$$
- Z_n not converge to Z a.s. because
$$P\left(\{\omega \in \Omega : Z_n(\omega) \to Z(\omega)\}\right) = P(\emptyset) = 0.$$

Definition 5.3 (converge in distribution, TBp. 181)

Let $Z_1, Z_2, \ldots$ be a sequence of random variables with cdf's $F_1, F_2, \ldots$ and let Z be a random variable with cdf F. Then Z_n **converges in distribution** to Z, denoted as $Z_n \xrightarrow{d} Z$, if

$$\lim_{n\to\infty} F_n(z) = F(z)$$

at every point z where F is continuous.

Theorem 5.1 (some properties about the 3 types of convergence)

1. $Z_n \xrightarrow{\text{a.s.}} Z \quad \Rightarrow \quad Z_n \xrightarrow{P} Z$
2. $Z_n \xrightarrow{P} Z \quad \Rightarrow \quad Z_n \xrightarrow{d} Z$
3. $Z_n \xrightarrow{d} c$, c: a constant $\quad \Rightarrow \quad Z_n \xrightarrow{P} c$
4. **(convergence of transformation)** Let $g : \mathbb{R}^k \mapsto \mathbb{R}$ be a continuous function.
   (a) $Z_n^{(j)} \xrightarrow{\text{a.s.}} Z^{(j)}, j = 1, \ldots, k \;\Rightarrow\; g(Z_n^{(1)}, \ldots, Z_n^{(k)}) \xrightarrow{\text{a.s.}} g(Z^{(1)}, \ldots, Z^{(k)})$.
   (b) $Z_n^{(j)} \xrightarrow{P} Z^{(j)}, j = 1, \ldots, k \;\Rightarrow\; g(Z_n^{(1)}, \ldots, Z_n^{(k)}) \xrightarrow{P} g(Z^{(1)}, \ldots, Z^{(k)})$.
   (c) $(Z_n^{(1)}, \ldots, Z_n^{(k)}) \xrightarrow{d} (Z^{(1)}, \ldots, Z^{(k)}) \Rightarrow g(Z_n^{(1)}, \ldots, Z_n^{(k)}) \xrightarrow{d} g(Z^{(1)}, \ldots, Z^{(k)})$.

5. **(Slutsky's theorem)** If $X_n \xrightarrow{d} X$, $Y_n \xrightarrow{P} a$, where a is a constant, then

   (a) $Y_n X_n \xrightarrow{d} aX$

   (b) $X_n + Y_n \xrightarrow{d} X + a$

   (c) $\dfrac{X_n}{Y_n} \xrightarrow{d} \dfrac{X}{a}$, provided that $P(Y_n \neq 0) = 1$ for all n and $a \neq 0$.

6. **(limit theorem for δ method)** Suppose

$$\frac{\sqrt{n}(X_n - \theta)}{\sigma} \xrightarrow{d} N(0,1).$$

   For a given function g, suppose that $g'(\theta) \neq 0$ exists. Then

$$\frac{\sqrt{n}[g(X_n) - g(\theta)]}{\sigma|g'(\theta)|} \xrightarrow{d} N(0,1).$$

Theorem 5.2 (Continuity Theorem, TBp. 181)

Let $F_n(x)$ be a sequence of cdfs with the corresponding mgfs $M_n(t)$. Let $F(x)$ be a cdf with the mgf $M(t)$. If $M_n(t) \to M(t)$ as $n \to \infty$ for all t in an open interval containing zero, then $F_n(x) \to F(x)$ at all continuity point of F.

Notes.

1. The reverse of the continuity theorem also holds.
2. The continuity theorem still holds when the moment generating function is replaced by characteristics function (chf always exists).

(for your information)

F_n, F : cdf; f_n, f : pdf; p_n, p : pmf;

Q: $F_n \xrightarrow{d} F$ implies $\lim_{n\to\infty} f_n(x) = f(x)$?

or $\lim_{n\to\infty} p_n(x) = p(x)$?

Ans: In general, **NO.**

Example 5.2 (Convergence of Poisson to Normal, TBp. 181-182)

Let $X_n \sim P(\lambda_n)$, $n = 1, 2, \ldots$ with $\lambda_n \to \infty$. We know that $E(X_n) = Var(X_n) = \lambda_n$ and $M_{X_n}(t) = e^{\lambda_n(e^t - 1)}$. Let

$$Z_n = (X_n - \lambda_n)/\sqrt{\lambda_n},$$

Then $M_{Z_n}(t) = e^{-t\sqrt{\lambda_n}} M_{X_n}\left(\frac{t}{\sqrt{\lambda_n}}\right) = e^{-t\sqrt{\lambda_n}} e^{\lambda_n(e^{t/\sqrt{\lambda_n}} - 1)}$. Because

$$\lim_{n\to\infty} \log M_{Z_n}(t) = \lim_{n\to\infty} -t\sqrt{\lambda_n} + \lambda_n(e^{t/\sqrt{\lambda_n}} - 1) = \frac{t^2}{2},$$

$M_{Z_n}(t) \to e^{t^2/2}$, which is the mgf of $N(0,1)$. By continuity theorem, $Z_n \xrightarrow{d} N(0,1)$, i.e., when λ is large, we can approximate the distribution of $P(\lambda)$ by $N(\lambda, \lambda)$.

• LLN and CLT

Theorem 5.3 (Weak Law of Large Numbers (WLLN), TBp. 178)

Let $X_1, X_2, \ldots, X_n, \ldots$ be a sequence of independent random variables with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$.
Let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Then $\overline{X}_n \xrightarrow{P} \mu$.

Proof: $E(\overline{X}_n) = \mu, \quad Var(\overline{X}_n) = \sigma^2/n$

By Chebyshev's inequality,

$$P(|\overline{X}_n - \mu| > \epsilon) \leq \frac{Var(\overline{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \longrightarrow 0 \qquad \text{as } n \to \infty$$

Notes. Under the same assumptions, a **strong law of large numbers (SLLN)**, which asserts that $\overline{X}_n \xrightarrow{\text{a.s.}} \mu$, can be proved.

Example 5.3 (Monte Carlo integration, TBp. 179)

To calculate $I(f) = \int_0^1 f(x)dx$, we can generate $X_1, X_2, \ldots, X_n$ i.i.d. $\sim U(0,1)$ and compute $\hat{I}(f) = \frac{1}{n}\sum_{i=1}^n f(X_i)$. By the LLN, $\hat{I}(f)$ will be close to $E[f(X_i)] = \int_0^1 f(x) \times 1\, dx = I(f)$ as n is large.

Example 5.4 (Repeated Measurements, TBp. 179-180)

Let $X_1, \ldots, X_n$ be i.i.d. with mean μ and variance σ^2, then

$$\overline{X}_n \xrightarrow{P} \mu.$$

Let

$$S_n^2 \equiv \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2 = \left(\frac{1}{n}\sum_{i=1}^n X_i^2\right) - \overline{X}_n^2.$$

Because $g(x) = x^2$ is continuous, $\overline{X}_n^2 \xrightarrow{P} \mu^2$. Next, the r.v.'s $X_1^2, \ldots, X_n^2$ are i.i.d. with mean $\sigma^2 + \mu^2$. By WLLN

$$\frac{1}{n}\sum_{i=1}^n X_i^2 \xrightarrow{P} \sigma^2 + \mu^2.$$

Therefore,

$$S_n^2 \xrightarrow{P} (\sigma^2 + \mu^2) - \mu^2 = \sigma^2.$$

(**Note.** $\frac{1}{n}$ in S_n^2 can be replaced by $\frac{1}{n-1}$.)

Example 5.5

If $X_n \sim t_n$, then $X_n \xrightarrow{d} N(0,1)$.

(Ec) If $X_n \sim F_{m,n}$, then $mX_n \xrightarrow{d} \chi^2_m$ as $n \to \infty$.

- Let $X_1, \ldots, X_n$ be i.i.d. $\sim$ Exponential(1), then $\mu_X=E(X_i)=1$, $\sigma_X^2=Var(X_i)=1$, for $i=1, \ldots, n$, and

$$n\overline{X}_n = \sum_{i=1}^n X_i \sim \mathrm{Gamma}(n,1) \Rightarrow \overline{X}_n \sim \tfrac{1}{n}\mathrm{Gamma}(n,1).$$

$$\Rightarrow \quad \mathrm{E}(\overline{X}_n) = \mu_X = 1, \mathrm{Var}(\overline{X}_n) = \sigma_X^2/n = 1/n$$

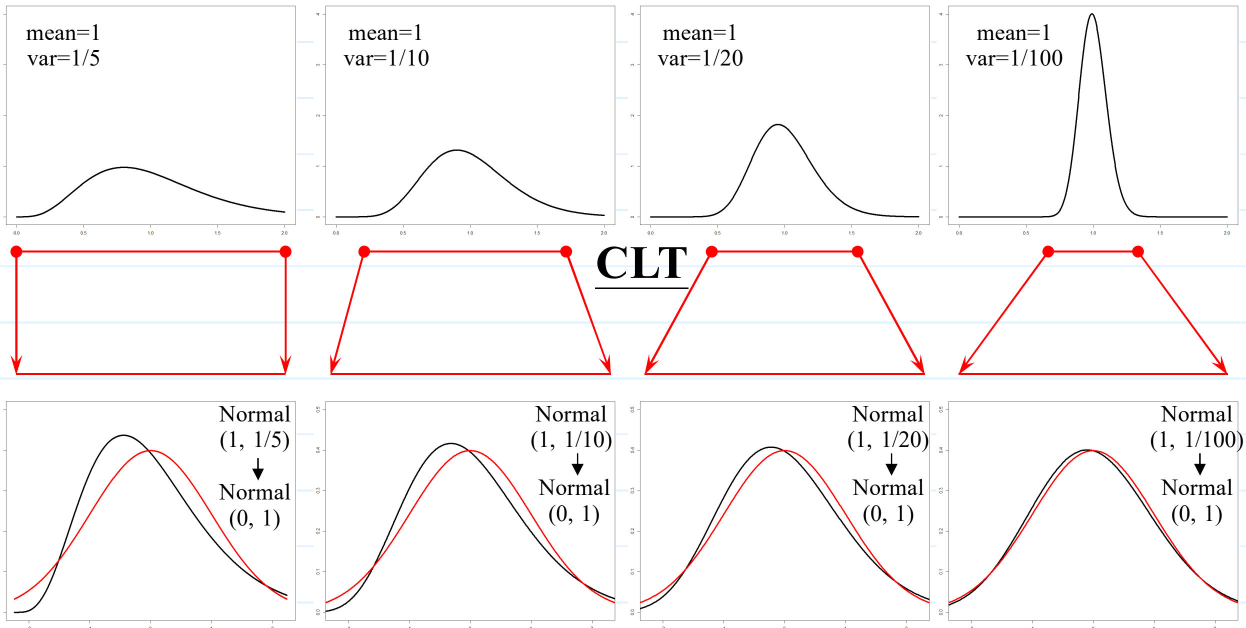
LLN

n=5 n=10 n=20 n=100

pdf of $\overline{X}_n$

$\dfrac{\overline{X}_n-\mu_X}{\sigma_X/\sqrt{n}}$

pdf of $\dfrac{\overline{X}_n-\mu_X}{\sigma_X/\sqrt{n}}$



Theorem 5.4 (Central Limit Theorem, TBp. 169)

Let $X_1, X_2, \ldots$ be i.i.d. with mean μ and variance σ^2. Let

$$\overline{X}_n \left(= \tfrac{1}{n}\sum_{i=1}^n X_i\right) \quad \text{and} \quad T_n \left(= n\overline{X}_n = \sum_{i=1}^n X_i\right)$$

be the average and the sum of data, respectively. Then,

$$\lim_{n\to\infty} P\left(\frac{T_n - n\mu}{\sigma\sqrt{n}} \le x\right) = \lim_{n\to\infty} P\left(\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \le x\right) = \Phi(x),$$

for $-\infty < x < \infty$, where $\Phi(x)$ is the cdf of $N(0,1)$.

Proof. Let $W_i = \frac{X_i-\mu}{\sigma}$, then $E(W_i)=0$ and $Var(W_i)=1$. Let

$$Z_n \equiv \tfrac{T_n-n\mu}{\sigma\sqrt{n}} = \tfrac{1}{\sqrt{n}}\sum_{i=1}^n W_i.$$

Let $M(t)$ be the mgf of W_i's and $M_{Z_n}(t)$ be the mgf of Z_n, then

$$M_{Z_n}(t) = \left[M\left(\tfrac{t}{\sqrt{n}}\right)\right]^n = \left[M(0) + M'(0)\tfrac{t}{\sqrt{n}} + \tfrac{1}{2}M''(0)\left(\tfrac{t}{\sqrt{n}}\right)^2 + o\left(\tfrac{1}{n}\right)\right]^n$$

$$= \left[1 + \tfrac{t^2}{2n} + o\left(\tfrac{1}{n}\right)\right]^n \longrightarrow e^{t^2/2} \quad \left[\text{because if } a_n \to a, \left(1+\tfrac{a_n}{n}\right)^n \to e^a\right]$$

Notes.

1. When mgf's do not exist, we can use chf's to prove it instead.
2. This is one of the simplest versions of CLT.

Example 5.6 (Normal approximation to Binomial distribution, TBp.187)

Let $X_1, X_2, \ldots, X_n$ be i.i.d. $\sim B(1,p)$, then $T_n \sim B(n,p)$. Note that $E(X_i) = p, Var(X_i) = p(1-p)$ and $E(T_n) = np, Var(T_n) = np(1-p)$. By CLT,

$$\frac{T_n - np}{\sqrt{np(1-p)}} \xrightarrow{d} N(0,1),$$

i.e., when n is large enough, we can approximate the distribution of $B(n,p)$ by $N(np, np(1-p))$.

Note: 1. how about those distributions that can be generated from a sum of some i.i.d. random variables? (example?)

2. **(cf.)** Poisson in Def. 4.7 (LNp.66) and Example 5.2 (LNp.92)

Example 5.7 (measurement error (or called sampling error), TBp. 186)

- Suppose that you want to know the average income of families living in Taipei.
- If you can ask every families their incomes, you will get the exact value of the average, denoted by μ.

- However, what if you only take a random sample of, say, 1000 families?
- The average income of the 1000 families, denoted by $\overline{X}_{1000}$, is a random variable. It has an error $\overline{X}_{1000} - \mu$, which is called measurement error or sampling error.
- By CLT, the error will be distributed normally, and we can approximate $P(|\overline{X}_{1000} - \mu| < c)$ using normal distribution no matter what the distribution of incomes is.

Example 5.8 (experimental error)

- It is usually true that an experimental error ϵ is a function of a number of component errors $\epsilon_1, \ldots, \epsilon_n$.
- for example, errors in the settings of experimental conditions, errors due to variation in raw materials, and so on.
- If each individual component error is fairly small, it is possible to approximate the overall error ϵ as a linear function of independently distributed component errors

$$\epsilon \approx a_1\epsilon_1 + \ldots + a_n\epsilon_n.$$

- By CLT, the distribution of ϵ will tend to normal as the number of component errors becomes large.
- This argument also offers a good justification for why in many statistical methods, such as in ANOVA or linear regression, the error part is assumed to be distributed normally.

Example 5.9 (cont. Ex.5.4 in LNp.94, Repeated Measurements)

Let $X_1, X_2, \ldots, X_n$ be i.i.d. with mean μ and variance σ^2.Then, by LLN, CLT, and Slutsky's theorem,

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{S_n} \xrightarrow{d} N(0,1),$$

because $\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \xrightarrow{d} N(0,1)$ and $S_n^2 \xrightarrow{P} \sigma^2 \left(\Rightarrow \frac{\sigma^2}{S_n^2} \xrightarrow{P} 1 \right)$.

❖ **Reading**: textbook, chapter 5

❖ **Further reading**: Roussas, chapter 8