Lecture 1: Review of Basic Probability

Ma Xuejun

School of Mathematical Sciences Soochow University https://xuejunma.github.io

2019.09.16

We would like to thank Professor Tian Maozai and Professor Wangie for sending slides and beneficial discussions.

Outline

- Syllabus
- Brief review of basic probability and statistics
 - Why is a random variable?
 - Transformations; independence; expectation
 - Important distributions
 - Some statistics

Terms

- Sample space; Measure; Random variable
- Transformation; Independence; Expectation; Conditional expectation; Variance & Standard deviation; Moment Generating Function; Characteristic function
- Common distributions
- Sample mean; Sample variance; Sample distribution
- Moment inequalities

Sample space and Measure

Outline

What do we mean by randomness?

- We construct an experiment, yet the result of the experiment has many possibilities.
 - Flip a coin, the result can be either head or tail
- Although we can not know the result beforehand, we do have some information about the result.
 - Approximately, there is equal chance for a head and a tail
- Randomness: the uncertainty of experiment results

Question: How to describe our information?

Sample space and Measure

Information 1. Possible outcomes

Definition 1.1.1: Sample space (Outcome space)

Let Ω be a sample space, which is a set containing all possible outcomes.

- Information 2. Probabilities for these possible outcomes
 - σ -field \mathcal{F} : a set of subsets of Ω which satisfies 3 rules.
 - Measurable space: (Ω, \mathcal{F})
 - ullet Event (measurable sets): element of ${\cal F}$
 - Probability measure P: for any element in the σ -field, assign it a probability, indicating the chance this event will happen
- (Ω, \mathcal{F}, P) (Probability space, measure space) is our information about the possible outcomes of this experiment. In short, we write it as the sample space Ω with probability P, or just Ω if there is no confusion.

σ -field

Let \mathcal{F} be a collection of subsets of a sample space. \mathcal{F} is called a σ -field (or σ -algebra) if and only if it has the following properties.

- **1** The empty set $\phi \in \mathcal{F}$
- ② If $A \in \mathcal{F}$, then the complement $A^c \in \mathcal{F}$.
- **3** If $A_i \in \mathcal{F}, i = 1, 2, ...$, then their union $\cup A_i \in \mathcal{F}$.
 - Measurable space: (Ω, \mathcal{F})
 - ullet Event (measurable sets): element of ${\cal F}$

 - Flip a coin, the result can be either head or tail $\Omega = \{H, T\}$, $\mathcal{F} = \{\dots\}$

Measure

Let Measurable space (Ω, \mathcal{F}) , A be a measurable space. A set function ν defined on \mathcal{F} is called a measure if and only if it has the following properties.

- $0 < \nu(A) < \infty$, for any $A \in \mathcal{F}$
- $\nu(\varnothing) = 0$
- \bullet If $A_i \in \mathcal{F}, i = 1, 2, ...$, and A_i 's are disjoint, i.e. $A_i \cap A_j = \emptyset$ for any $i \neq j$, then $\nu \left(\bigcup_{i=1}^{\infty} A_i \right) = \sum_{i=1}^{\infty} \nu(A_i)$
 - measure space: $(\Omega, \mathcal{F}, \nu)$
 - probability measure $\nu(\Omega)=1$. We usually denote it by Pinstead of ν , (Ω, \mathcal{F}, P) .
 - Flip a coin, the result can be either head or tail $\Omega = \{H, T\}$, $\mathcal{F} = \{\dots\}$
 - $\nu(A) = |A|$ the number of elements in $A \in (F)$.

$$P(A) = \frac{|A|}{|\Omega|} \stackrel{\triangleleft}{\triangleleft} \stackrel{\triangleleft}{\square} \stackrel{\triangleleft}{\wedge} \stackrel{\square}{\triangleleft} \stackrel{\square}{\wedge} \stackrel{\square}{\square} \stackrel{\square}{\square} \stackrel{\square}{\wedge} \stackrel{\square}{\square} \stackrel$$

Random Variables

What is of interest?

- Manufacturers Ω : all the combinations of good light bulbs and defective light bulbs. Need: proportion of defective light bulbs from a lot
- Market researchers Ω : survey results of all consumers for one product. Need: preference of all consumers about this product, with a scale 1-10.

Our interest:

- Not the details of Ω , but a special measurable characteristic of the outcomes!
- A random variable, is a mapping from Ω to R, which draws the measurable characteristic of interest

Example: an opinion poll. 50 people; 1: agree; 0 disagree:

- Ω has 2^{50} elements.
- interest: the number of people who agree out of 50. X =number of 1s recorded out of 50. $\mathcal{X} = \{0, 1, 2, \dots, 50\}$

Random Variable

Random Variable

- Let (Ω, \mathcal{F}) and (R, \mathcal{B}) (\mathcal{B} : Borel σ -field)be measurable spaces
- X is a function from Ω to R. The function X is called a random variable (r.v.; measurable function) if and only if

$$X^{-1}(B) = \{ \omega \in \Omega : X(\omega) \in B \} \subset \mathcal{F}$$

for any $B \subset \mathcal{B}$.

$$\Omega \longrightarrow X \qquad X_1 = X(\omega_1) = X(\omega_2)$$

$$\omega_2 \longrightarrow X \qquad X_2 = X(\omega_3)$$

$$X \longrightarrow X_2 = X(\omega_3)$$

Random Variable

Suppose we have a sample space

$$\Omega = \{\omega_1, \dots, \omega_n\}$$

with a probability function P.

- ullet We defined a random variable X with range $\mathcal{X} = \{x_1, \ldots, x_m\}.$
- We write

$$P_X(X = x_i) = P(\{\omega_j \in \Omega : X(\omega_j) = x_i\})$$

$$P_X(X \in B) = P(\{\omega \in \Omega : X(\omega) \in B\})$$

where P_X is an induced probability function \mathcal{X} .

- Notations:
 - Upper-case letters $X, Y, Z \dots$ to denote r.v.'s
 - Lower-case letters $x, y, z \dots$ to denote their possible values.



Example: Random variable

Example 1.4.3

- Consider the experiment of tossing a coin three times.
- H : Head; T : Tail.
- X: the number of heads obtained in the three tosses.

ω	HHH	HHT	HTH	THH	TTH	THT	HTT	TTT
$X(\omega)$	3	2	2	2	1	1	1	0

• $\mathcal{X} = \{0, 1, 2, 3\}$. The induced probability function on \mathcal{X} is given by

\overline{x}	0	1	2	3
$P_X(x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

$$P_X(X=1) = P(\{HTT, THT, TTH\}) = \frac{3}{8}.$$

Cumulative Density Function

Definition 1.5.1 Cumulative Density Function

The cumulative distribution function (CDF) of a random variable is defined by

$$F(x) = P(X \le x); -\infty < x < \infty$$

For all CDF's: there is

- F(x) is right-continuous. At each x, $\lim_{n\to\infty} F(y_n) = F(x)$ for any sequence $y_n \to x$ with $y_n > x$.
- \circ F(x) is non-decreasing.
- $\lim_{x\to-\infty} F(x) = 0, \lim_{x\to\infty} F(x) = 1.$

Any F(x) satisfying Properties 1-3 is a CDF for some random variable.

Example: Logistic distribution.

Example 1.5.5

$$F_X(x) = \frac{1}{1 + e^{-x}}$$

- $\bullet \lim_{x\to -\infty} F_X(x) = 0$
- $\lim_{x\to\infty} F_X(x) = 1$

•

$$f_X(x) = \frac{d}{dx} F_X(x) = \frac{e^{-x}}{(1 + e^{-x})^2} > 0$$

Discrete v.s. Continuous r.v.

• If X is discrete, then its probability mass function (pmf) is

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

• If X is continuous, then its probability density function (pdf) satisfies

$$P(X \in A) = \int_A f_X(x)dx = \int_A f(x)dx = \int_A dF(x)$$

and
$$f_X(x) = f(x) = F'(x)$$
.

• We say that X and Y have the same distribution(i.e. $X \stackrel{D}{=} Y$) if $P(X \in A) = P(Y \in A)$ for all A. $X \stackrel{D}{=} Y$) if only if $F_{\mathbf{Y}}(t) = F_{\mathbf{Y}}(t)$

 \boldsymbol{a}

Transformation

Given a r.v. X with density function $f_X(\cdot)$, it is often that we are interested in a transformation Y = g(X) which is defined as a known function g (either one-to-one or many-to-one) of X.

- Obviously, the composite function $g \circ X$ defines a new r.v. Y from Ω to R.
- Let Y = g(X).

$$P(Y \in A) = P(g(X) \in A)$$
$$= P(X \in g^{-1}(A))$$

where $g^{-1}(A) = \{x \in R, g(x) \in A\}$. In particular,

$$F_Y(y) = \Pr\{Y \in y\} = P(X \in g^{-1}(-\infty, y])$$

If X has pdf $f_X(x)$, then

$$F_Y(y) = \int_{g^{-1}(-\infty,y]} f_X(x)dx = \int_{\{x: g(x) \le y\}} f_X(x)dx$$

Example 2.1.2

Suppose X has a uniform distribution on the interval $(0,2\pi)$, that is

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

Consider $Y = \sin^2(X)$

$$P(Y \le y) = P(X \le x_1) + P(x_2 \le X \le x_3) + P(X \ge x_4)$$

= $2P(X \le x_1) + 2P(x_2 \le X \le \pi)$

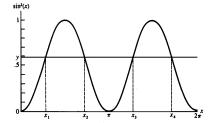


Figure 2.1.1. Graph of the transformation $y = \sin^2(x)$ of Example 2.1.2

• If q is increasing,

$$F_Y(y) = F_X(g^{-1}(y)).$$

If q is decreasing,

$$F_Y(y) = 1 - F_X(g^{-1}(y)).$$

Theorem 2.1.5

Let X have probability distribution function (pdf) $f_X(x)$ and Y = g(X), where g is a monotone function. Let

$$\mathcal{Y} = \{y : g^{-1}(y) \text{ is a possible value of } X\}.$$

Suppose $f_X(x)$ is continuous and that $g^{-1}(y)$ has a continuous derivative on \mathcal{Y} . Then the pdf of Y is given by

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right|, & y \in \mathcal{Y}, \\ 0, & \text{otherwise.} \end{cases}$$

Example 2.1.4

 $X \sim f_X(x) = 1I(0 < x < 1), F_X(x) = x. Y = q(x) = -\log x$ find its distribution.

Proof:

- $Y = q(x) = -\log x \Longrightarrow x = e^{-y}, q^{-1}(y) = e^{-y}$
- q is a decreasing function.

$$\frac{d}{dx}g(x) = \frac{d}{dx}(-\log x) = \frac{-1}{x} < 0, \ 0 < x < 1$$

•

$$F_Y(y) = P_Y(Y \le y) = P_X(g(X) \le y)$$

= $P_X(X \ge g^{-1}(y))$
= $1 - P_X(X \le g^{-1}(y)) = 1 - e^{-y}$

Example 2.1.6

Let

$$f_X(x) = \frac{1}{(n-1)!\beta^n} x^{n-1} e^{-x/\beta}, \quad 0 < x < \infty$$

be the Gamma pdf Y=1/X. Find the pdf of Y

Proof. $g^{-1}(y) = 1/y, \mathcal{Y} = (0, \infty), \left| \frac{d}{dy} g^{-1}(y) \right| = 1/y^2$. Therefore for all y > 0,

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

$$= \frac{1}{(n-1)!\beta^n} \left(\frac{1}{y} \right)^{n-1} e^{-1/(\beta y)} \frac{1}{y^2}$$

$$= \frac{1}{(n-1)!\beta^n} \left(\frac{1}{y} \right)^{n+1} e^{-1/(\beta y)}$$

 A special case of a pdf known as the inverted Gamma distribution. 4□ > 4回 > 4 = > 4 = > = 9 < 0</p>

Theorem 2.1.8

Let X have pdf $f_X(x)$, Let Y=g(X). Suppose there exists a partition A_0,A_1,\cdots,A_k such that $P(X\in A_0)=0$ and $f_X(x)$ is continuous on each A_i .

$$P(X \in \bigcup_{i=1}^{k} A_i) = 1.$$

Further, we have $g(\cdot)$ is monotone if restricted to A_i $i=1,2,\cdots,k$. Let

$$g_i^{-1}(y) = \{x \in A_i : g(x) = y\}$$

and assume $g_i^{-1}(y)$ has continuous derivative on $\mathcal Y$ for each i. Then

$$f_Y(y) = \begin{cases} \sum_{i=1}^k f_X(g_i^{-1}(y)) |\frac{d}{dy}g_i^{-1}(y)|, & y \in \mathcal{Y} \\ 0, & \text{otherwise} \end{cases}$$

• Remark Unfortunately, I found the above Theorem has very little practical use.

Example 2.1.9

Let $X \sim N(0,1), \ Y = X^2.$ we may use the above theorem to find the pdf of Y.

Proof:

- $g(x) = x^2$ is monotone on $(-\infty, 0)$ and on $(0, \infty)$.
- $\mathcal{Y} = (0, \infty)$.

$$A_0 = \{0\}$$

$$A_1 = (-\infty, 0), \ g_1(x) = x^2, \ g_1^{-1}(y) = -\sqrt{y}$$

$$A_2 = (0, \infty,), \ g_2(x) = x^2, \ g_1^{-1}(y) = \sqrt{y}$$

The pdf Y is

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \Phi(\sqrt{y}) \frac{1}{2} \frac{1}{\sqrt{y}} + \Phi(-\sqrt{y}) \frac{1}{2} \frac{1}{\sqrt{y}} = \frac{1}{\sqrt{y}} \Phi(\sqrt{y})$$

Probability integral transform

Theorem 2.1.10 Probability integral transform

Let X have continuous cdf $F_X(x)$ and define the random variable Y as $Y=F_X(X)$. Then Y is uniformly distributed on (0,1), that id

$$P(Y \le y) = y, \ 0 < y < 1.$$

- $F_X^{-1}(\tau) = \inf\{x : F_X(x) \ge \tau\}$
- Proof:

$$P_Y(Y \le y) = P_X(F_X(x) \le y)$$

$$= P_X(F_X^{-1}[F_X(x)] \le F_X^{-1}(y))$$

$$= P_X(X \le F_X^{-1}(y))$$

$$= F_X(F_X^{-1}(y))$$

$$= y$$

Independence

Theorem 4.2.10

Two r.v.'s X and Y are independent if and only if

$$P(X \in A; Y \in B) = P(X \in A)P(Y \in B)$$

for all A and B.

- F(x,y) = F(x)F(y) for any x and yf(x,y) = f(x)f(y) or p(x,y) = p(x)p(y)
- When X and Y are independent, h(X) and g(Y) are also independent, if h and g are well-defined functions.

Expectation

Definition:

$$\mathbf{E}(X) = \sum_{x} xp(x)$$

$$\mathbf{E}(X) = \int_{-\infty}^{\infty} xf(x)dx$$

- Properties:
 - $\mathbf{E}(g(X)) = \int_{-\infty}^{\infty} g(x) f(x) dx$
 - $\mathbf{E}(g(X,Y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f(x,y) dx dy$
 - If X_1, \ldots, X_n are independent, then

$$\mathbf{E}\Big(\prod_{i=1}^n X_i\Big) = \prod_{i=1}^n \mathbf{E}(X_i)$$

• Example 2.2.2

$$X \sim \exp(\lambda)$$
,

$$f_X(x) = \frac{1}{\lambda} e^{-x/\lambda} \quad x > 0.$$

Then $\mathbf{E}[X] = \lambda$.

• Example 2.2.3 $X \sim Binomial(n, p)$,

$$f_X(x) = \binom{n}{x} p^x (1-p)^{n-x} \ x = 0, 1, \cdots$$

Then $\mathbf{E}[X] = np$.

• Example 2.2.4 $X \sim \text{Cauchy}$,

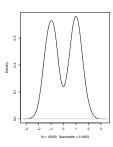
$$f_X(x) = \frac{1}{\pi} \cdot \frac{1}{1 + x^2} - \infty < x < \infty.$$

Then $\mathbf{E}[X]$ is not definded! (or do not exist).



26 / 55

$$X = 0.5N(-1, 0.5^2) + 0.5N(-1, 0.5^2)$$



```
1 rm(list=ls())
_{2} n <- 10000
3 \text{ comp} \leftarrow \text{sample}(c(0, 1), \text{ size} = n, \text{ prob} = c(0.5, 0.5),
```

Mixed normal distribution -

replace = T) $5 \times < - \text{rnorm}(n, \text{mean} = \text{ifelse}(\text{comp} == 0, -1, 1),$

sd = ifelse(comp == 0, 0.5, 0.5))6 7 plot (density(x), main="")

- Theorem 2.2.5 Let X be a r.v. and let a, b, and c be constants. Then for any functions $g_1(x)$ and $g_2(x)$ whose expectations exist.
 - (a) $\mathbf{E}[aq_1(X) + bq_2(X) + c] = a\mathbf{E}[q_1(X)] + b[q_2(X)] + c$
 - (b) If $q_1(x) > 0$ for all x, then $\mathbf{E}[q_1(X)] > 0$.
 - (c) If $g_1(x) \geq g_2(x)$ for all x, then $\mathbf{E}[g_1(X)] \geq \mathbf{E}[g_2(X)]$
 - (d) If $a < q_1(x) < b$ for all x, then $a < \mathbf{E}[q_1(x)] < b$
- Example 2.2.6

E(X) is the "center" of a distribution (or its r,v,) in the sense that

$$\min_{b} \mathbf{E}(X - b)^2 = \mathbf{E}[X - \mathbf{E}X]^2.$$

Homework

$$\min_{b} \mathbf{E} \rho_{\tau}(X-b)$$

Remark:
$$\rho_{\tau}(t) = \tau t I(t > 0) + (\tau - 1) t I(t < 0)$$
.

Variance & Standard Deviation

- Motivation: Describe the "spread" of r.v.
- Definition. $Var(x) = \mathbf{E}[(x-\mu)^2]$, where $\mu = \mathbf{E}(X)$, $sd(X) = \sqrt{Var(x)}$.
- Properties.
 - $Var(X) = \mathbf{E}(X^2) [\mathbf{E}(X)]^2$
 - If X_1, \ldots, X_n are independent, then

$$Var\left(\sum_{i=1}^{n}\right) = \sum_{i=1}^{n} Var(X_i)$$

The covariance is

$$\mathbf{Cov}(X,Y) = \mathbf{E}[(X - \mathbf{E}(X))(Y - \mathbf{E}(Y))] = \mathbf{E}(XY) - \mathbf{E}(X)\mathbf{E}(Y)$$

and the correlation coefficient is

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

For any two r.v.s with variance existed,

$$Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$$

Conditional Expectation

- Conditional Expectation of X when Y is given as y is that
 - $\mathbf{E}(X|Y=y) = \sum_{x} x p_{X|Y}(X|Y)$ for discrete r.v.
 - $\mathbf{E}(X|Y=y) = \int_{-\infty}^{\infty} x f_{X|Y}(x|y) dx$ for cont. r.v.
 - Interpretation: Note that X|Y=y is a new r.v., $\mathbf{E}(X|Y=y)$ is the expectation on this r.v.
- Law of Total Expectation

$$\mathbf{E}\Big[\mathbf{E}(X|Y)\Big] = \mathbf{E}(X)$$

I aw of Total Variance

$$Var(X) = Var\Big[\mathbf{E}(X|Y)\Big] + \mathbf{E}\Big[Var(X|Y)\Big]$$

Theorem 4.4.3

If X and Y are any two r.vs, then

$$\mathbf{E}(X) = \mathbf{E}\Big[\mathbf{E}(X|Y)\Big]$$

Proof:

$$\mathbf{E}X = \int \int x f(x, y) dx dy$$

$$= \int \left[\int x f(x|y) dx \right] f_Y(y) dy$$

$$= \int \mathbf{E}(X|y) f_Y(y) dy = \mathbf{E} \left[\mathbf{E}(X|Y) \right]$$

In general, the conditional expectation $\mathbf{E}[X|Y]$ can by defined as a r.v. g(Y) such that

$$\mathbf{E}[(X-g(Y))^2] = \inf_{\text{among all reasonable function } h} \mathbf{E}[(X-h(Y))^2]$$

or $\mathbf{E}[X|Y]$ is the function of Y which is "closest" to X in terms of mean square error.

Example 4.4.1 Hierarchical Model

 $Y \sim$ Number of eggs lay by a mother fish, and $X \sim$ Number of survivors (young fish). On the average, how many eggs will survive?

Then it is reasonable to assume

$$Y \sim Poisson(\lambda)$$

 $X|Y \sim Binomial(Y,p)$

So,

$$\mathbf{E}X = \mathbf{E} \Big[\mathbf{E}(X|Y) \Big]$$
$$= \mathbf{E}(pY)$$
$$= p\lambda$$

Example 4.4.5

$$X|Y \sim Binomial(Y,p)$$

 $Y|\Lambda \sim Poisson(\Lambda)$
 $\Lambda \sim exponential(\beta)$

Proof:

$$\begin{aligned} \mathbf{E}[X] &= \mathbf{E}[\mathbf{E}(X|Y)] \\ &= p\mathbf{E}[Y] \\ &= p\mathbf{E}[\mathbf{E}(Y|\Lambda)] \\ &= p\mathbf{E}[\Lambda] \\ &= p\beta. \end{aligned}$$

Theorem 4.4.7

For any two random variables X and Y,

$$Var(X) = E[Var(X|Y)] + Var[E(X|Y)]$$

provided that the expectation exist.

Proof:

$$Var(X) = \mathbf{E} \Big\{ [X - \mathbf{E}(X|Y) + \mathbf{E}(X|Y) - \mathbf{E}X]^2 \Big\}$$

$$= \mathbf{E} \Big\{ [X - \mathbf{E}(X|Y)]^2 + [\mathbf{E}(X|Y) - \mathbf{E}X]^2$$

$$+ 2[X - \mathbf{E}(X|Y)][\mathbf{E}(X|Y) - \mathbf{E}X] \Big\}$$

$$= \mathbf{E} \{ [X - \mathbf{E}(X|Y)]^2 \} + \mathbf{E} \{ [\mathbf{E}(X|Y) - \mathbf{E}X]^2 \}$$

$$= \mathbf{E} [Var(X|Y)] + Var[(\mathbf{E}X|Y)]$$

Moment Generating Function and Characteristic Function

- Moment Generating Function (MGF)
 - Definition: $M_X(t) = E(e^{tX})$: a function of t, not r.v.
 - If Y = aX + b, $M_Y(t) = e^{bt} M_X(at)$
 - If X and Y are independent, then $M_{X+Y}(t) = M_X(t)M_Y(t)$
- Characteristic Function
 - Definition: $\phi_X(t) = E[e^{itX}]$: a function of t; $i = \sqrt{-1}$.
 - Bounded: $\phi(t)$ | < 1
 - If X and Y are independent, then $\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t)$.

An example of two distribution functions but with the same moments.

Example 2.3.10

Consider the two pdfs given by

$$f_1(x) = \frac{1}{\sqrt{2\pi}x} e^{-(\log x)^2/2}, \quad 0 \le x < \infty,$$

$$f_2(x) = f_1(x)[1 + \sin(2\pi \log x)], \quad 0 \le x < \infty,$$

Then it can be shown if $X_1 \sim f_1(x)$,

$$\mathbf{E}[X_1^r] = e^{r^2/2}, \quad r = 0, 1, \cdots,$$

Now suppose that $X_2 \sim f_2(x)$, we have for $r = 0, 1, \cdots$

$$\mathbf{E}[X_2^r] = \int_0^\infty x^r f_1(x) [1 + \sin(2\pi \log x)] dx = \mathbf{E}[X_1^r] + \int_0^\infty x^r f_1(x) \sin(2\pi \log x) dx$$

$$\int_{0}^{\infty} x^{r} \frac{1}{\sqrt{2\pi}x} e^{-(\log x)^{2}/2} \sin(2\pi \log x) dx \quad y = \log x - r$$

$$= \int_{-\infty}^{\infty} e^{(y+r)r} \frac{1}{\sqrt{2\pi}} e^{-(y+r)^{2}/2} \sin(2\pi (y+r)) dy$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(y^{2}-r^{2})} \sin(2\pi y) dy \cdot \cos(2\pi r)$$

$$+ \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(y^{2}-r^{2})} \cos(2\pi y) dy \cdot \sin(2\pi r)$$

$$= 0 \quad r = 0, 1, \dots$$

since $e^{-\frac{1}{2}(y^2-r^2)}\sin(2\pi y)$ is an odd function.¹.

 $^{^{1}\}sin(A+B) = \sin A \cos B + \sin B \cos A$

However, we have the following theorem.

Theorem 2.3.11

Let $F_X(x)$ and $F_Y(y)$ be two cdfs all of whose moments exist.

- (a) If F_X and F_Y have bounded support , then $F_X(u)=F_Y(u)$ for all u iff $EX^r=EY^r$ for all $r=0,1,2,\cdots$
- (b) If the moment generating functions exist and $M_X(t)=M_Y(t)$ for all t in some neighborhood of 0 ,then $F_X(u)=F_Y(u)$ for all u.

Differentiating Under An Integral Sign

If a,b are finite and $f(x,\theta)$ is differentiable with respect to θ . Then we have

$$\frac{d}{d\theta} \int_{a}^{b} f(x,\theta) dx = \int_{a}^{b} \frac{\partial}{\partial \theta} f(x,\theta) dx.$$

But in statistics, we often need to evaluate $\frac{d}{d\theta} \int_{-\infty}^{\infty} f(x,\theta) dx$, which may or may not be $\int_{-\infty}^{\infty} \frac{\partial}{\partial \theta} f(x,\theta) dx$.

Theorem 2.4.2

Suppose the function h(x,y) is continuous at y_0 for each x, and there exists a function g(x) satisfying

- a) $|h(x,y)| \leq g(x)$, for all x and y;
- b) $\int_{-\infty}^{\infty} g(x)dx < \infty$.

Then

$$\lim_{y \to y_0} \int_{-\infty}^{\infty} h(x, y) dx = \int_{-\infty}^{\infty} \lim_{y \to y_0} h(x, y) dx.$$

Apply the above Theorem to the differentiation case, then we have

• Theorem 2.4.3 Suppose $f(x,\theta)$ is differentiable at $\theta=\theta_0$, and there exists a function $g(x,\theta_0)$ and a constant $\delta>0$ such that a) $\left|\frac{f(x,\theta_0+\triangle)-f(x,\theta_0)}{\triangle}\right|\leq g(x,\theta_0)$, for all x and $|\Delta|\leq \delta$;

b)
$$\int_{-\infty}^{\infty} g(x, \theta_0) dx < \infty$$
.

Then

$$\frac{d}{d\theta} \int_{-\infty}^{\infty} f(x,\theta) dx \mid_{\theta=\theta_0} = \int_{-\infty}^{\infty} \left[\frac{\partial}{\partial \theta} f(x,\theta) \mid_{\theta=\theta_0} \right] dx \quad (*)$$

• Corollary Suppose that there exists $\delta>0$ and function $g(x,\theta)$ such that $\left|\frac{\partial}{\partial \theta}f(x,\theta)\right|_{\theta=\theta'}\big|\leq g(x,\theta)$, for all θ' with $|\theta'-\theta|<\delta$, and $\int_{-\infty}^{\infty}g(x,\theta)dx<\infty$. Then (*) holds.

- Remark Finding bound $g(x,\theta)$ is cumbersome. We need to know that differentiating under the integral sign is not always automatic. In most situations, we just do it!!
- Example 2.4.6 $X \sim N(\mu, 1)$,

$$M_X(t) = E(e^{tX}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx} e^{-(x-\mu)^2/2} dx,$$

$$\frac{d}{dt}M_X(t) = \frac{d}{dt}E(e^{tX}) = E(\frac{\partial}{\partial t}e^{tX}) = E(Xe^{tX}).$$

For the exchange of operation of differentiation and summation, we have

- Theorem 2.4.8 Suppose that the series $\sum\limits_{x=0}^{\infty}h(\theta,x)$ converges for all θ in an interval (a,b) and
 - a) $\frac{\partial}{\partial \theta} h(\theta,x)$ is continuous in θ for each x;
 - b) $\sum\limits_{x=0}^{\infty} \frac{\partial}{\partial \theta} h(\theta,x)$ converges uniformly on every closed bounded subinterval of (a,b).

Then

$$\frac{d}{d\theta} \sum_{x=0}^{\infty} h(\theta, x) = \sum_{x=0}^{\infty} \frac{\partial}{\partial \theta} h(\theta, x).$$

• Theorem 2.4.10 Suppose that the series $\sum_{x=0}^{\infty} h(\theta, x)$ converges uniformly on [a,b] and that, for each x, $h(\theta,x)$ is a continuous function of θ . Then

$$\int_a^b \sum_{x=0}^\infty h(\theta,x) d\theta = \sum_{x=0}^\infty \int_a^b h(\theta,x) d\theta.$$

Important Distributions

- Discrete distributions:
 - Bernoulli r.v.: $X \sim Bernoulli(p), \ p(1) = p, \ p(0) = 1 p,$ p(x) = 0 if $x \neq 0$ and $x \neq 1$. It can also be written as $p^x(1-p)^{1-x}$ for x = 0, 1.
 - Binomial r.v.: $X \sim Binomial(n;p)$, $p(x) = \binom{n}{x} p^x q^{n-x}, x = 0,1,2,\ldots,n$. Summation of n Bernoulli random variables.
 - Poisson r.v.: $X \sim Pois(\lambda)$, $p(x) = \frac{\lambda^x}{x!}e^{-\lambda}$, $x = 0, 1, 2, \dots, n$.
- Continuous distributions:
 - Uniform r.v.: $X \sim Unif(a,b), f(x) = \frac{1}{b-a}, x \in (a,b)$
 - Exponential r.v.: $X \sim Exp(\lambda), f(x) = \lambda e^{-\lambda x}$
 - Normal r.v.: $X \sim N(\mu, \sigma^2), f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$

Multivariate Normal Distribution

• The d random vector $X \sim N(\mu, \Sigma,$

$$f(x) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^{\top} \Sigma^{-1}(x-\mu)\right)$$

- $AX + b \sim N(A\mu + b, A\Sigma^{-1}A^{\top})$
- Conditional distribution.

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \quad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

then

$$X_1|Y_2 \sim N\Big(\mu_1 + \Sigma_{11}\Sigma_{22}^{-1}(Y_2 - \mu_2), \Sigma_{11} - \sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Big)$$

Distribution

Statistics

Outline

- Sample mean $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$
- Sample variance: $S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i \bar{X}_n)^2$
- Sampling distribution of \bar{X}_n : $G_n(t) = P(\bar{X}_n \leq t)$

When it is normal:

• If $X \sim N(\mu, \Sigma^2)$, then \bar{X}_n and S_n^2 are independent,

$$\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\frac{(n-1)S_n^2}{\sigma^2} \sim \chi_{n-1}^2$$

Moment inequalities

Lemma 4.7.1

Let a and b be any two positive numbers, and let p and q be any positive numbers satisfying

$$\frac{1}{p} + \frac{1}{q} = 1.$$

Then

$$\frac{1}{p}a^p + \frac{1}{q}b^q \ge ab$$

with equality holds if and only if $a^p = b^q$.

• Proof: Consider for fixed b(or a),

$$g(a) = \frac{1}{p}a^p + \frac{1}{q}b^q - ab$$

with equality if and only if $a^p = b^q$.



Theorem 4.7.2 (H \ddot{o} lder's Inequality)

Let X and Y be any two random variables. Let p and q be any positive numbers satisfying

$$\frac{1}{p} + \frac{1}{q} = 1.$$

Then

$$|\mathbf{E}(XY)| \le \mathbf{E}|XY| \le (\mathbf{E}|X|^p)^{\frac{1}{p}} (\mathbf{E}|Y|^q)^{\frac{1}{q}}.$$

Proof: In the inequality (1), let

$$a = \frac{|X|}{(\mathbf{E}|X|^p)^{\frac{1}{p}}}, \quad b = \frac{|Y|}{(\mathbf{E}|Y|^q)^{\frac{1}{q}}}$$

After some simplification, and take expectation on the two sides of the inequality. The result can been obtained.

 Theorem 4.7.3 (Cauchy-Schwarz Inequality) For any two random variables X and Y,

$$|\mathbf{E}(XY)| \le \mathbf{E}|XY| \le (\mathbf{E}|X|^2)^{\frac{1}{2}} (\mathbf{E}|Y|^2)^{\frac{1}{2}}$$

• Example 4.7.4 (Covariance Inequality) If X and Y have means μ_X and μ_Y , and variances σ_X^2 and σ_Y^2 , respectively. We can apply the Cauchy-Schwarz Inequality to get

$$(\mathbf{Cov}(X,Y))^2 \le \sigma_X^2 \cdot \sigma_Y^2$$

Example

Let p > 1, then apply Hölder's Inequality. For any random variables X.

$$\mathbf{E}|X| \le \{\mathbf{E}|X|^p\}^{\frac{1}{p}} \tag{5.1}$$

If 1 < r < s, we have (Liapounov's Inequality)

$$\left(\mathbf{E}|X|^r\right)^{\frac{1}{r}} \le \left(\mathbf{E}|X|^p\right)^{\frac{1}{p}} \tag{5.2}$$

• Proof of (5.1) Let q be such that $\frac{1}{n} + \frac{1}{n} = 1$, then

$$\mathbf{E}|X| = \mathbf{E}|X| \cdot 1 \le (\mathbf{E}|X|^p)^{1/p} \cdot (\mathbf{E}1^q)^{1/q} = (\mathbf{E}|X|^p)^{1/p}.$$

• Proof of (5.2) Let s be such that s = pr, then s > 1.

$$\mathbf{E}(|X|^r) \le (\mathbf{E}(|X|^r)^p)^{1/p}.$$

Theorem 4.7.5 (Minkowski's Inequality)

Let X and Y be any two random variables. Then for 1

$$[\mathbf{E}|X+Y|^p]^{\frac{1}{p}} \le (\mathbf{E}|X|^p)^{\frac{1}{p}} + (\mathbf{E}|Y|^p)^{\frac{1}{p}}$$

Proof:

$$\mathbf{E}|X+Y|^{p} = \mathbf{E}(|X+Y||X+Y|^{p-1})$$

$$\leq \mathbf{E}(|X||X+Y|^{p-1}) + \mathbf{E}(|Y||X+Y|^{p-1})$$
 (5.3)

Using Hölder's Inequality,

$$\mathbf{E}(|X||X+Y|^{p-1}) \le (\mathbf{E}|X|^p)^{\frac{1}{p}} \left[\mathbf{E}|X+Y|^{q(p-1)} \right]^{\frac{1}{q}}$$
 (5.4)

where q is such that $\frac{1}{p}+\frac{1}{q}=1$ or $\frac{1}{q}=1-\frac{1}{p}$, i.e., $q=\frac{p}{p-1}$ or q(p-1)=p. Similarly,

$$\mathbf{E}(|Y||X+Y|^{p-1}) \le (\mathbf{E}|Y|^p)^{\frac{1}{p}} \left[\mathbf{E}|X+Y|^{q(p-1)} \right]^{\frac{1}{q}}$$
 (5.5)

So combine (5.4) and (5.5) with (5.3), divide through by $[\mathbf{E}(|X+Y|^{q(p-1)})]^{1/q}$, we have

$$\mathbf{E}|X+Y|^p \le (\mathbf{E}|X+Y|^p)^{\frac{p-1}{p}} \left[(\mathbf{E}|X|^p)^{\frac{1}{p}} + (\mathbf{E}|Y|^p)^{\frac{1}{p}} \right]$$

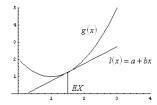
Theorem 4.7.7 (Jensen's Inequality)

For any random variable X, if g(x) is a convex function, then

$$\mathbf{E}g(X) \ge g(\mathbf{E}X)$$

- Equality holds if and only if, for any line a+bx that is tangent to g(x) at $x=\mathbf{E}X$, P(g(X)=a+bX)=1.
- If g(x) is linear, $g(\mathbf{E}X) = a + b\mathbf{E}X = \mathbf{E}g(X)$.

Remark For any twice differentiable function g(x), it is convex iff $g''(x) \ge 0$ for all x.



Example 4.7.8 (An inequality for means)

Let a_1, a_2, \dots, a_n be n non-negative numbers. Define

$$a_A = \frac{1}{n}(a_1 + a_2 + \dots + a_n)$$

$$a_G = [a_1 a_2 \dots a_n]^{1/n}$$

$$a_H = \frac{1}{\frac{1}{n}(\frac{1}{a_1} + \frac{1}{a_2} + \dots + \frac{1}{a_n})}$$

An inequality relating these means is

$$a_H \leq a_G \leq a_A$$
.

Remark The above inequality gives a reason for Maximum Likelihood Estimation(MLE).

Proof: Let X be a random variable with range a_1,\ldots,a_n , and $P(X=a_i)=1/n, n=1,\ldots,n$. Since $\log x$ is a concave function, $\mathbf{E}\log X \leq \log(\mathbf{E}X)$, hence

$$\log a_G = \frac{1}{n} \sum_{i=1}^n \log a_i = \mathbf{E} \log X \le \log(\mathbf{E}X)$$
$$= \log \left(\frac{1}{n} \sum_{i=1}^n a_i\right) = \log a_A$$

So, $a_G \leq a_A$. Furthermore,

$$\log \frac{1}{a_H} = \log \left(\frac{1}{n} \sum_{i=1}^n \frac{1}{a_i} \right) = \mathbf{E} \log \frac{1}{X} \ge \mathbf{E} \left(\log \frac{1}{X} \right) = -\log(\mathbf{E}X)$$
$$= -\log a_G = \log \left(\frac{1}{a_G} \right).$$

So, $a_G \geq a_H$.

Markov's Inequality

Markov's (Chebyshev's) Inequality

- If q is strictly increasing and positive on $(0, \infty)$, q(x) = q(-x).
- X is a r.v. such that $E[g(X)] < \infty$, then for each a > 0

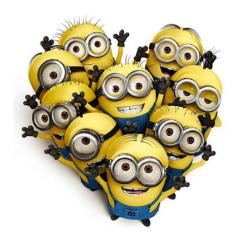
$$P(|X| \ge a) \le \frac{E[g(X)]}{g(a)}$$

Proof:

$$\begin{split} E[g(X)] &\geq E[g(X)I_{\{g(X)\geq g(a)\}}] \\ &\geq g(a)E[I_{\{g(X)\geq g(a)\}}] \\ &= g(a)E[I_{\{|X|\geq a\}}] \\ &= g(a)P(|X| \geq a) \end{split}$$

$$\begin{split} g(x) &= |x| \Longrightarrow P(|X| \ge a) \le \frac{E|X|}{a} \\ g(x) &= x^p \Longrightarrow P(|X| \ge a) \le \frac{E|g(X^p)|}{a^p} \\ g(x) &= x^2 \Longrightarrow P(|X - EX| \ge a) \le \frac{Var(X)}{a^2} \\ g(x) &= e^{t|x|} \Longrightarrow P(|X| \ge a) \le \frac{E\left[e^{t|X|}\right]}{e^{ta}} \end{split}$$

for some constant $t \geq 0$



Thank you