G₁ ⊙ Theorem (best *constant* predictor under MSE).

a constant function of x — minimum

$$\boxed{G_1}\quad \underline{E_{X,Y}}\ \underline{(Y-\underline{c})^2} = \underline{E_Y}(Y-\underline{c})^2 \geq E_Y[Y-\underline{E_Y(Y)}]^2 = \underline{Var_Y(Y)}$$

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

Proof. $R(Y)=(Y-c)^2$: a function of Y

$E_{X,Y}[R(Y)]=E_Y[R(Y)]$

$\parallel$

$E_Y E_{X|Y}[R(Y)|Y]$ — $R(Y)$ — LNp.8-22

Thm in LNp.5-19

$$E_Y(Y-\underline{c})^2 \;\underline{=}\; Var_Y(Y)+(\mu_Y-\underline{c})^2 \;\geq\; Var_Y(Y)$$

cf. LNp.8-24

1/4

G₃ ⊙ Theorem (best predictor under MSE).

LNp.8-24 — minimum

$$\boxed{G_3}\quad E_{X,Y}[Y-\underline{g(X)}]^2 \geq E_{X,Y}[Y-\underline{E_{Y|X}(Y|X)}]^2 \;\underline{=}\; \underline{E_X}[\underline{Var_{Y|X}(Y|X)}]$$

(∗) cf. (LNp.8-21)

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

a function of X only

Proof.

$$\begin{aligned}
E_{X,Y}[Y-g(X)]^2 &= E_{X,Y}\{[\underline{Y}-E_{Y|X}(Y|\underline{X})]+[E_{Y|X}(Y|\underline{X})-g(\underline{X})]\}^2\\
&= E_{X,Y}[Y-E_{Y|X}(Y|X)]^2+\underset{E_{X,Y}}{\underline{E_X}}[E_{Y|X}(Y|X)-g(X)]^2\\
&\quad +2\cdot \underline{E_{X,Y}\{[Y-E_{Y|X}(Y|X)][E_{Y|X}(Y|X)-g(X)]\}} \;\leftarrow =0\\
&\boxed{=}\; E_{X,Y}[Y-E_{Y|X}(Y|X)]^2+\underline{E_X[E_{Y|X}(Y|X)-g(X)]^2}\\
&\boxed{\geq}\; E_{X,Y}[Y-E_{Y|X}(Y|X)]^2
\end{aligned}$$

a function of X only

last "="

=0 iff $g(x)=E_{Y|X}(Y|X)$

where the last "=" comes from

$$\underline{E_{X,Y}}\{[\underline{Y}-E_{Y|X}(Y|\underline{X})][E_{Y|X}(Y|\underline{X})-g(\underline{X})]\} \leftarrow R(X,Y)$$

$$\circleddash\ \underline{E_X}\,\underline{E_{Y|X}}\left\{[\underline{Y}-E_{Y|X}(Y|\overset{x}{X})]\,\underline{[E_{Y|X}(Y|\overset{x}{X})-g(\overset{x}{X})]}\,\middle|\,\underline{\overset{x}{X}}\right\}$$

By the law of total expectation (LNp.8-22) $\underline{E_{XY}}[R(X,Y)]=\underline{E_X}\,\underline{E_{Y|X}}[R(X,Y)|\underline{X}]$

this is a constant when conditioned on X

$$= E_X\{\underline{[E_{Y|X}(Y|X)-g(X)]}\ \underline{E_{Y|X}[Y-E_{Y|X}(Y|X)|\underline{X}]}\} = 0.$$

$= E_{Y|X}(Y|X) - \underline{E_{Y|X}(Y|X)}$

important concept: mean is best predictor under MSE

Furthermore, (for (∗) in LNp.8-27)

$$\underline{E_{X,Y}}[\underline{Y}-E_{Y|X}(Y|\underline{X})]^2 = \underline{E_X}\,\underline{E_{Y|X}}\{[Y-E_{Y|X}(Y|\overset{x}{X})]^2|\underline{\overset{x}{X}}\} = E_X[\underline{Var_{Y|X}(Y|X)}]$$

➢Some notes for the best predictor in G_3

- ⊙ $\underline{E_{Y|X}(Y|x)}$ is the best predictor of Y based on X, in the sense of mean square prediction error ← intuition ← check the graph in LNp.8-20
- ⊙ Its calculation requires to know the joint distribution of X and Y, or at least $E_{Y|X}(Y|x)$

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

- ▪ $\underline{E_{Y|X}(Y|x)}$ is called the regression function of Y on X — 迴歸

G_2 → ⊙ **Theorem** (best *linear* predictor under MSE). $-1 \leq \rho_{XY} \leq 1$

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

$$\text{minimum} \rightarrow \quad = \quad \sigma_Y^2(1-\rho_{XY}^2)$$

The equality holds if and only if $a=\mu_Y - b\mu_X$ and $b=\rho_{XY}\sigma_Y/\sigma_X$ ← unit=?

Proof. $E_{X,Y}(Y-a-bX)^2$ — $R(X,Y) \equiv Z$

∵ $Var(Z) = E(Z^2) - [E(Z)]^2$

$$= Var_{X,Y}(Y-a-bX) + [E_{X,Y}(Y-a-bX)]^2$$

$$= Var_{X,Y}(Y-bX) + (\mu_Y - a - b\mu_X)^2$$

$$\boxed{\geq} \quad Var_{X,Y}(Y-bX) \qquad (\Rightarrow \text{setting } a = \mu_Y - b\mu_X)$$

Thm in LNp.8-13

$$= \sigma_Y^2 + b^2\sigma_X^2 - 2b\sigma_{XY}$$

$$= \sigma_X^2\left(b^2 - 2b\frac{\sigma_{XY}}{\sigma_X^2} + \frac{\sigma_{XY}^2}{\sigma_X^4}\right) + \sigma_Y^2 - \frac{\sigma_{XY}^2}{\sigma_X^2}$$

$$= \sigma_X^2\left(b - \frac{\sigma_{XY}}{\sigma_X^2}\right)^2 + \sigma_Y^2(1-\rho_{XY}^2)$$

$$\boxed{\geq} \quad \sigma_Y^2(1-\rho_{XY}^2) \qquad \left(\Rightarrow \text{setting } b = \frac{\sigma_{XY}}{\sigma_X^2} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y} \times \frac{\sigma_Y}{\sigma_X} = \rho_{XY}\frac{\sigma_Y}{\sigma_X}\right)$$

➢ Some notes for the best *linear* predictor in G_2

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

best in G_3 — linear regression analysis — best in G_2

- Its calculation requires to know the means, variances, and covariance of X and Y. ← cf. → best in G_3, G_1 ← Which one require more information? → more information better predictor

MSE=0 if $\rho_{XY}=\pm 1$
MSE=σ_Y^2 if $\rho_{XY}=0$

⊙ $\sigma_Y^2(1-\rho_{XY}^2)$ is small if ρ_{XY} is close to +1 or −1, and large if ρ_{XY} is close to 0. ← intuition? ← check the correlation plots in LNp.8-9

- A comparison of these minimum MSEs

∵ $G_1 \cap G_2 \cap G_3$

➢ $\min_{a,b} E_{X,Y}[Y-(a+bX)]^2 \leq \min_c E_{X,Y}(Y-c)^2$ and the equality holds if and only if $\rho_{XY} = 0$.

➢ $\min_g E_{X,Y}[Y-g(X)]^2 \leq \min_{a,b} E_{X,Y}[Y-(a+bX)]^2$ and the equality holds if and only if $E_{Y|X}(Y|x) = \mu_Y + (\rho_{XY}\sigma_Y/\sigma_X)(x-\mu_X)$.

❖ **Reading**: textbook, Sec 7.6

Moment Generating Function

- Definition (Moment and Central Moment). If a random variable X has a cdf F_X, then

$$\mu_k \equiv E(X^k) = \int_{-\infty}^{\infty} x^k \, dF_X(x), \quad k = 1, 2, 3, \ldots,$$

are called the k^{th} *moments* of X provided that the integral converges absolutely, and

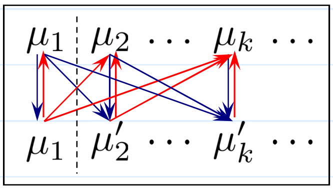
$$\mu'_k \equiv E[(X - \underset{\text{a constant}}{\mu_X})^k] = \int_{-\infty}^{\infty} (x - \mu_X)^k \, dF_X(x), \quad k = 2, 3, \ldots,$$

are called k^{th} *moment about the mean μ_X* or *central moment* of X provided that the integral converges absolutely.

➢Some notes.

- $\mu'_k = E[(X-\mu_X)^k] = E\left[\sum_{i=0}^{k} \binom{k}{i} (-\mu_X)^{k-i} X^i\right]$
  $= \sum_{i=0}^{k} \binom{k}{i} (-\mu_X)^{k-i} E(X^i) = \sum_{i=0}^{k} \binom{k}{i} (-\mu_X)^{k-i} \mu_i.$ $\boxed{\mu_0 = E(X^0) = 1}$
- $\mu_k = E(X^k) = E\{[(X - \mu_X) + \mu_X]^k\}$
  $= \sum_{i=0}^{k} \binom{k}{i} (\mu_X)^{k-i} E[(X-\mu_X)^i]$
  $= \sum_{i=0}^{k} \binom{k}{i} (\mu_X)^{k-i} \mu'_i.$ $\boxed{\begin{array}{l}\mu'_0 = 1 \\ \mu'_1 = 0\end{array}}$
- In particular,

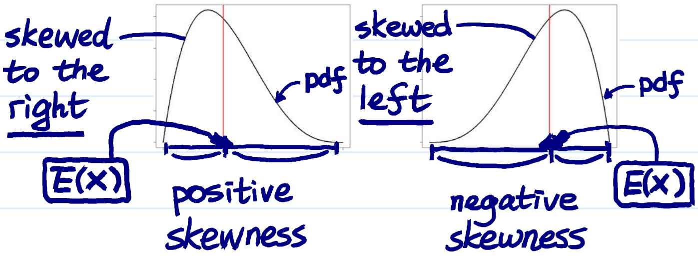
$$E(X) = \mu_X = \mu_1, \quad \text{and,}$$
$$Var(X) = \sigma_X^2 = \mu'_2 = \mu_2 - \mu_1^2. = E(x^2) - [E(x)]^2$$

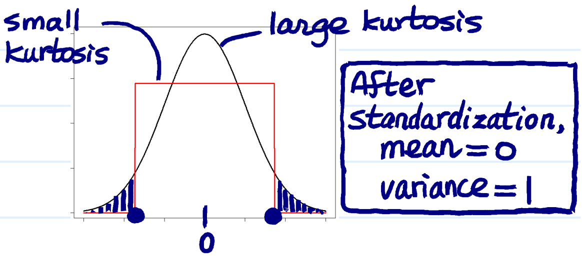


$\boxed{\text{Recall. mean, var, cov, cor}}$ $\boxed{\text{defined by expectation}}$

■ The (central) moments give a lot of useful information about the distribution in addition to mean and variance, e.g.,

- ☐ Skewness (a measure of the asymmetry): $\mu'_3/\sigma^3 = E\left(\frac{X-\mu}{\sigma}\right)^3$
- ☐ Kurtosis (a measure of the "heavy tails"): $\mu'_4/\sigma^4 = E\left(\frac{X-\mu}{\sigma}\right)^4$



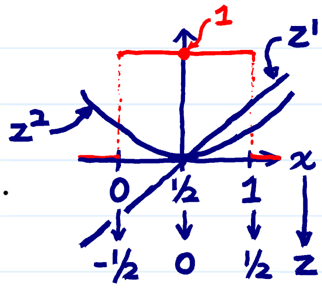


➢Example (Uniform). If $X \sim$ Uniform(0, 1), then

$$\mu_k = \int_0^1 x^k \, dx = \frac{1}{k+1},$$

therefore,

$$\mu_X = \mu_1 = 1/2, \quad \text{and,}$$
$$\mu'_2 \rightarrow \sigma_X^2 = \mu_2 - \mu_1^2 = 1/3 - (1/2)^2 = 1/12.$$



And, $\mu'_k = \int_0^1 (x-1/2)^k \, dx = \int_{-1/2}^{1/2} z^k \, dz$

$$= \frac{1}{k+1}\left[(1/2)^{k+1} - (-1/2)^{k+1}\right] = \begin{cases} 0, & k \text{ is odd,} \\ \frac{1}{(k+1)2^k}, & k \text{ is even.} \end{cases}$$

$\boxed{\begin{array}{l}\text{skewness} = 0 \\ \text{kurtosis} = 1.8\end{array}}$

- Recall. How to characterize a distribution?

  (1) pdf/pmf, (2) cdf, (3) mgf

- Definition (Moment Generating Function). If X is a random variable with the cdf F_X, then

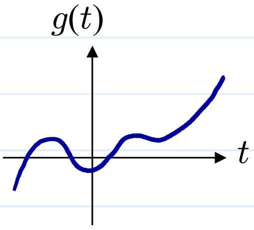
$$M_X(t) = E(e^{tX}) = \int_{-\infty}^{\infty} e^{tx}\, dF_X(x),$$

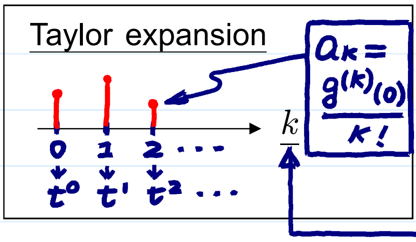
$= \int_{-\infty}^{\infty} e^{tx} f_X(x)dx$ for continuous case

is called the *moment generating function* (mgf) of X provided that the integral converges absolutely in some non-degenerate interval of t.

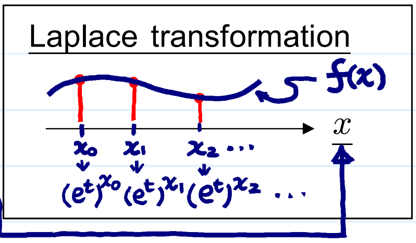
Q: how to express a function?

$$g(t) \boxed{=} \sum_{k=0}^{\infty} a_k t^k \qquad g(t) \boxed{=} \int_{\mathbb{R}} f(x)\, (e^t)^x\, dx$$





cf.



➢ Some Notes.

- The mgf is a function of the variable t.
- The mgf may only exist for some particular values of t. i.e., not all $t \in \mathbb{R}$
- $M_X(t)$ always exists at $t=0$ and $M_X(0)=1$ ← Thm in LNp.8-36

➢ Example.

- If X is a discrete r.v. taking on values x_i's with probability p_i's, $i=1, 2, 3, \ldots$, then

$$M_X(t) = E(e^{tX}) = \sum_{i=1}^{\infty} e^{tx_i} p_i.$$

- If $X \sim$ Poisson(λ), then for $-\infty<t<\infty$,

$$M_X(t) = E(e^{tX}) = \sum_{x=0}^{\infty} e^{tx} \times \frac{e^{-\lambda}\lambda^x}{x!}$$

$$= \quad e^{-\lambda}\left(e^{\lambda e^t}\right) \sum_{x=0}^{\infty} \underbrace{\frac{e^{-(\lambda e^t)}(\lambda e^t)^x}{x!}}_{\text{pmf of Poisson}(\lambda e^t)} = e^{-\lambda} e^{\lambda e^t} = e^{\lambda(e^t-1)}.$$

- If $X \sim$ exponential(λ), then for $t<\lambda$,

$$M_X(t) = E(e^{tX}) = \int_0^{\infty} e^{tx} \times \lambda\, e^{-\lambda x}\, dx$$

$$= \quad \lambda\left(\frac{1}{\lambda-t}\right) \int_0^{\infty} \underbrace{(\lambda-t)\, e^{-(\lambda-t)x}}_{\text{pdf of exponential}(\lambda-t)} dx = \frac{\lambda}{\lambda-t},$$

  and $M_X(t)$ does not exist for $t \geq \lambda$.

  STO

  This must be >0

- A list of some mgfs (exercise)
  - ▫ If $X \sim$ binomial(n, p), use binomial expansion (LNp.5-23)

$$M_X(t) = (1-p+pe^t)^n, \text{ for } t < -\log(1-p).$$

r = 1, geometric distribution

- If $X \sim$ negative binomial(r, p),

use negative binomial expansion (LNp.5-29)

$$M_X(t) = \left[\frac{pe^t}{1-(1-p)e^t}\right]^r, \text{ for } t < -\log(1-p).$$

α = 1, exponential distribution

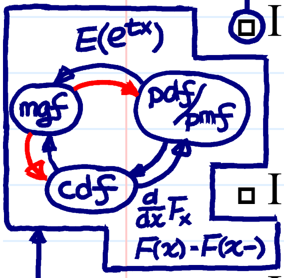
- If $X \sim$ uniform(α, β), $M_X(t) = \frac{e^{\beta t} - e^{\alpha t}}{t(\beta-\alpha)}$.

- If $X \sim$ gamma(α, λ),

use STO

$$M_X(t) = \left(\frac{\lambda}{\lambda-t}\right)^\alpha, \text{ for } t < \lambda.$$

use $e^u = \sum_{k=0}^{\infty} \frac{u^k}{k!}$ & STO



- If $X \sim$ beta(α, β), $M_X(t) = 1 + \sum_{k=1}^{\infty}\left(\prod_{r=0}^{k-1}\frac{\alpha+r}{\alpha+\beta+r}\right)\frac{t^k}{k!}$.

- If $X \sim$ normal(μ, σ^2), $M_X(t) = e^{\mu t+(\sigma^2/2)t^2}$. use STO

Theorem (Uniqueness Theorem). Suppose that the mgfs $M_X(t)$ and $M_Y(t)$ of random variables X and Y exist for all $|t|<h$ for some $h>0$. If

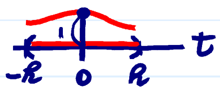
$$M_X(t) = M_Y(t),$$

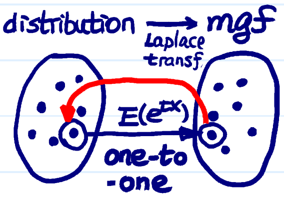
for $|t|<h$, then

$$\boxed{\text{does not mean } X=Y} \leftarrow F_X(z) = F_Y(z) \leftarrow \boxed{X, Y \text{ have same dist.}}$$

for all $z \in \mathbb{R}$, where F_X and F_Y are the cdfs of X and Y, respectively.

i.e. an open interval containing zero





Proof. Skipped (by the uniqueness theorem of Laplace transform.)

➢ Application of the uniqueness theorem

- When a mgf exists for all $|t|<h$ for some $h>0$, there is a *unique* distribution corresponding to that mgf.
- This allows us to use mgfs to find distributions of *transformed* random variables in some cases.
- This technique is most commonly used for *linear combinations* of independent random variables $X_1, \ldots, X_n$.

Find dist. of $X_1 + \cdots + X_n$ (Check the Thms in LNp.8-38)

➢ Example. If $M_X(t) = p_1e^{a_1t} + \cdots + p_ke^{a_kt}$, where $p_1+\cdots+p_k=1$, then X is a discrete r.v. and its pmf is

by uniqueness Thm & mgf in LNp.8-34

$$p_X(x) = \begin{cases} p_i, & \text{for } x = a_i,\ i = 1, \ldots, k, \\ 0, & \text{otherwise.} \end{cases}$$

- Theorem (Moments and MGF). If $M_X(t)$ exists for $|t|<h$ for some $h>0$, then

can take derivative at t = 0

$$M_X(0)=1,$$

related to the coefficients in the Taylor expansion of $M_X(t)$

know all moments ⇒ know dist.

and,

kth derivative

$$M_X^{(k)}(0) = \mu_k, \quad k = 1, 2, 3, \ldots$$

This explains why it's called moment generating function.

Proof. First, $M_X(0) = \int_{-\infty}^{\infty} e^{0\cdot x}\, dF_X(x) = \int_{-\infty}^{\infty} 1\, dF_X(x) = 1.$

$= F_X(x)\Big|_{-\infty}^{\infty} =$

$$M_X{}'(0) = \frac{d}{dt} M_X(t)\Big|_{t=0} = \left[\frac{d}{dt}\int_{-\infty}^{\infty} e^{tx}\, dF_X(x)\right]\Big|_{t=0}$$

$$= \int_{-\infty}^{\infty} \left(\frac{d}{dt} e^{tx}\Big|_{t=0}\right) dF_X(x) = \int_{-\infty}^{\infty} \left(x e^{tx}\big|_{t=0}\right) dF_X(x)$$

$$= \int_{-\infty}^{\infty} x \cdot 1\, dF_X(x) = E_X(X) = \mu_1.$$

$$\cdots = \cdots$$

$$M_X^{(k)}(0) = \frac{d^k}{dt^k} M_X(t)\Big|_{t=0} = \left[\frac{d^k}{dt^k}\int_{-\infty}^{\infty} e^{tx}\, dF_X(x)\right]\Big|_{t=0}$$

$$= \int_{-\infty}^{\infty} \left(\frac{d^k}{dt^k} e^{tx}\Big|_{t=0}\right) dF_X(x) = \int_{-\infty}^{\infty} \left(x^k e^{tx}\big|_{t=0}\right) dF_X(x)$$

$$= \int_{-\infty}^{\infty} x^k \cdot 1\, dF_X(x) = E_X(X^k) = \mu_k.$$

➢Example. If $X \sim$ exponential(λ), then $M_X(t) = \frac{\lambda}{\lambda - t}$. ← LNp.8-34

Because $$M_X^{(k)}(t) = \frac{k!\,\lambda}{(\lambda - t)^{k+1}},$$

we get $$\mu_k = M_X^{(k)}(0) = \frac{k!}{\lambda^k}.$$ → Then, can use kth moments to obtain mean, variance, skewness, kurtosis, ..., kth central moments, ...

- Theorem (MGF for linear transformation). For constants a and b,

$$M_{a+bX}(t) = e^{at} M_X(bt).$$ ← can be used to identify the dist. of a+bX from the dist. of X

Proof. $M_{a+bX}(t) = E_X[e^{t(a+bX)}] = e^{at} E_X[e^{(bt)X}] = e^{at} M_X(bt)$. Y

Recall. LNp.8-36 → Theorem (MGF for SUM of independent r.v.'s). If $X_1, \ldots, X_n$ are independent each with mgfs $M_{X_1}(t), \ldots, M_{X_n}(t)$, respectively, then the mgf of $S = X_1 + \cdots + X_n$ is

∵ independent

$S: \Omega \to \mathbb{R}^1$, $M_S: \mathbb{R}^1 \to \mathbb{R}^1$ → $$M_S(t) = M_{X_1}(t) \times \cdots \times M_{X_n}(t). \text{---} (*)$$

same t

Proof. $M_S(t) = E_S(e^{tS}) = E_{X_1,\ldots,X_n}[e^{t(X_1+\cdots+X_n)}]$ (Δ)

∵ $Y_1, \cdots, Y_n$ are indep.

$$= E_{X_1,\ldots,X_n}(e^{tX_1} \times \cdots \times e^{tX_n})$$ Y_1 Y_n

$$= E_{X_1}(e^{tX_1}) \times \cdots \times E_{X_n}(e^{tX_n}) = M_{X_1}(t) \times \cdots \times M_{X_n}(t).$$

Note. geometric is a special case of negative binomial with $n=1$

➢Example. If $X_1, \ldots, X_n$ are i.i.d. $\sim$ geometric(p), then

can use convolution approach (LNp.7-29) to prove it → $S = X_1 + \cdots + X_n \sim$ negative binomial(n, p).

By uniqueness Thm & mgf in LNp.8-35

Proof. $M_S(t) = M_{X_1}(t) \times \cdots \times M_{X_n}(t)$

$$= \frac{pe^t}{1-(1-p)e^t} \times \cdots \times \frac{pe^t}{1-(1-p)e^t} = \left[\frac{pe^t}{1-(1-p)e^t}\right]^n.$$

cf.

convolution approach (LNp.7-32,7-34)

➢ Example. If $X_1, \ldots, X_n$ are independent and

$$X_i \sim \text{normal}(\mu_i, \sigma_i^2), \text{ for } i=1, \ldots, n.$$

Let $S = a_0 + \underbrace{a_1X_1}_{Y_1} + \cdots + \underbrace{a_nX_n}_{Y_n}$, then → $\begin{cases} E(S) = \mu_* & (\text{LNp.8-4}) \\ Var(S) = \sigma_*^2 & (\text{LNp.8-13}) \end{cases}$

$Y_1 + \cdots + Y_n = Z$

$$S \sim \text{normal}\left(\underbrace{a_0 + a_1\mu_1 + \cdots + a_n\mu_n}_{=\mu_*}, \underbrace{a_1^2\,\sigma_1^2 + \cdots + a_n^2\,\sigma_n^2}_{=\sigma_*^2}\right).$$

Proof. $M_S(t) = e^{a_0 t} \times \prod_{i=1}^n e^{\mu_i(a_i t) + (\sigma_i^2/2)(a_i t)^2}$

$$= e^{(\underbrace{a_0 + a_1\mu_1 + \cdots + a_n\mu_n}_{\mu_*})t + \left[(\underbrace{a_1^2\sigma_1^2 + \cdots + a_n^2\sigma_n^2}_{\sigma_*^2})/2\right]t^2}.$$

By uniqueness Thm & mgf in LNp.8-35

Recall joint distribution

⊙ Definition (Joint Moment Generating Function). For random variables $X_1, \ldots, X_n$, their joint mgf is defined as

multi-dimensional Laplace transform

$(X_1, \cdots, X_n) : \Omega \to \mathbb{R}^n$
$M_{X_1, \cdots, X_n} : \mathbb{R}^n \to \mathbb{R}^1$

$$M_{X_1,\ldots,X_n}(t_1, \ldots, t_n) = E_{X_1,\ldots,X_n}\left(e^{t_1X_1 + \cdots + t_nX_n}\right)$$

different t_i's cf. (Δ) in LNp.8-38

provided that the expectation exists.

➢ Example. If $X_1, \ldots, X_m \sim$ multinomial$(n, m, p_1, \ldots, p_m)$, the joint pmf is:

$$\binom{n}{x_1, \ldots, x_m} p_1^{x_1} \cdots p_m^{x_m}$$

$0 \le x_i \le n$, $i = 1, \cdots, m$.
$x_1 + \cdots + x_m = n$

$$M_{X_1,\ldots,X_m}(t_1, \ldots, t_m)$$

$$= \sum_{\substack{0 \le x_i \le n,\ i=1,\ldots,m \\ x_1+\cdots+x_m=n}} e^{t_1x_1 + \cdots + t_mx_m} \binom{n}{x_1, \ldots, x_m} p_1^{x_1} \cdots p_m^{x_m}$$

$$= \sum_{\substack{0 \le x_i \le n,\ i=1,\ldots,m \\ x_1+\cdots+x_m=n}} (p_1 e^{t_1})^{x_1} \cdots (p_m e^{t_m})^{x_m}$$

By the multinomial expansion in LNp.7-16

$$= (p_1e^{t_1} + \cdots + p_me^{t_m})^n.$$

cf.
joint cdf → marginal cdf
joint pdf/pmf → marginal pdf/pmf
relationship between joint mgf & marginal mgf.

- Some Properties of Joint mgf

➢ $M_{X_1}(t) = M_{X_1, X_2, \ldots, X_n}(t, 0, \ldots, 0)$.

can be any X_i

➢ uniqueness theorem

Note. same property holds for pdf/pmf, cdf, mgf.

➢ $X_1, \ldots, X_n$ are independent if and only if

When setting $t_1 = \cdots = t_n = t$, it becomes the mgf of $X_1 + \cdots + X_n$ → (*) in LNp.8-38

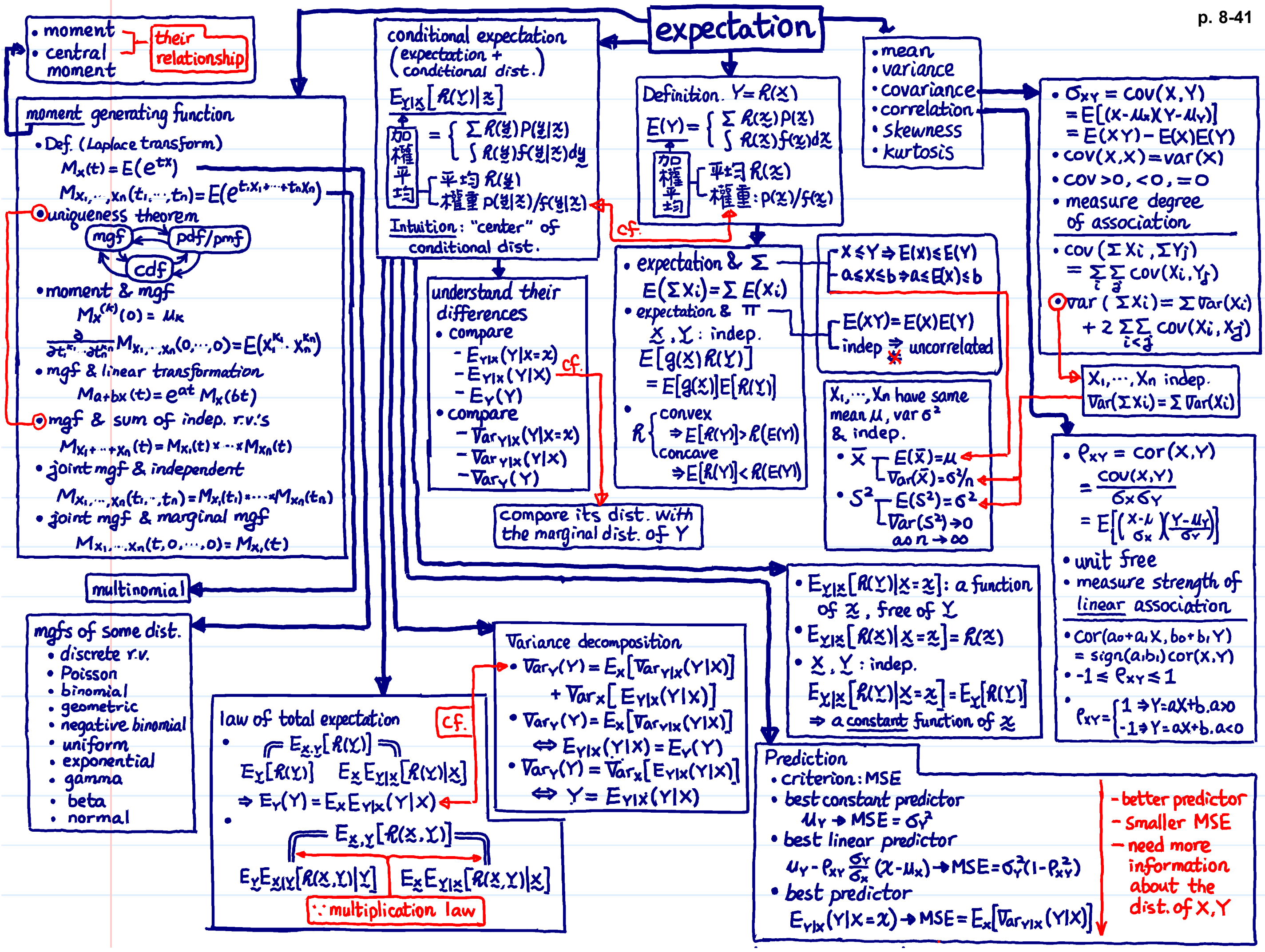
$$M_{X_1,\ldots,X_n}(t_1, \ldots, t_n) = M_{X_1}(t_1) \times \cdots \times M_{X_n}(t_n).$$

joint mgf = $\prod_{i=1}^n$ marginal mgf

different t_i's cf.

➢ $\dfrac{\partial^{k_1+\cdots+k_n}}{\partial t_1^{k_1} \cdots \partial t_n^{k_n}} M_{X_1,\ldots,X_n}(0, \ldots, 0) = E_{X_1,\ldots,X_n}(X_1^{k_1} \times \cdots \times X_n^{k_n})$.

❖ **Reading**: textbook, Sec 7.7



expectation

- moment
- central moment

their relationship

moment generating function

- Def. (Laplace transform)
  $M_X(t) = E(e^{tX})$
  $M_{X_1,\cdots,X_n}(t_1,\cdots,t_n) = E(e^{t_1X_1+\cdots+t_nX_n})$
- uniqueness theorem

- moment & mgf
  $M_X^{(k)}(0) = \mu_k$
  $\frac{\partial}{\partial t_1^{k_1}\cdots\partial t_n^{k_n}} M_{X_1,\cdots,X_n}(0,\cdots,0) = E(X_1^{k_1}\cdots X_n^{k_n})$
- mgf & linear transformation
  $M_{a+bX}(t) = e^{at} M_X(bt)$
- mgf & sum of indep. r.v.'s
  $M_{X_1+\cdots+X_n}(t) = M_{X_1}(t)\times\cdots\times M_{X_n}(t)$
- joint mgf & independent
  $M_{X_1,\cdots,X_n}(t_1,\cdots,t_n) = M_{X_1}(t_1)\times\cdots\times M_{X_n}(t_n)$
- joint mgf & marginal mgf
  $M_{X_1,\cdots,X_n}(t,0,\cdots,0) = M_{X_1}(t)$

multinomial

mgfs of some dist.

- discrete r.v.
- Poisson
- binomial
- geometric
- negative binomial
- uniform
- exponential
- gamma
- beta
- normal

conditional expectation (expectation + conditional dist.)

$E_{\underset{\sim}{Y}|\underset{\sim}{x}}[h(\underset{\sim}{Y})|\underset{\sim}{x}] = \begin{cases} \sum h(\underset{\sim}{y})p(\underset{\sim}{y}|\underset{\sim}{x}) \\ \int h(\underset{\sim}{y})f(\underset{\sim}{y}|\underset{\sim}{x})d\underset{\sim}{y} \end{cases}$

加權平均
- 平均 $h(\underset{\sim}{y})$
- 權重 $p(\underset{\sim}{y}|\underset{\sim}{x})/f(\underset{\sim}{y}|\underset{\sim}{x})$

Intuition: "center" of conditional dist.

cf.

Definition. $Y = h(\underset{\sim}{X})$

$E(Y) = \begin{cases} \sum h(\underset{\sim}{x})p(\underset{\sim}{x}) \\ \int h(\underset{\sim}{x})f(\underset{\sim}{x})d\underset{\sim}{x} \end{cases}$

加權平均
- 平均 $h(\underset{\sim}{x})$
- 權重: $p(\underset{\sim}{x})/f(\underset{\sim}{x})$

- mean
- variance
- covariance
- correlation
- skewness
- kurtosis

- $\sigma_{XY} = cov(X,Y)$
  $= E[(X-\mu_X)(Y-\mu_Y)]$
  $= E(XY) - E(X)E(Y)$
- $cov(X,X) = var(X)$
- $cov > 0, < 0, = 0$
- measure degree of association

- $cov(\sum X_i, \sum Y_j) = \sum_i\sum_j cov(X_i, Y_j)$
- $var(\sum X_i) = \sum Var(X_i) + 2\sum\sum_{i<j} cov(X_i, X_j)$

$X_1,\cdots,X_n$ indep.
$Var(\sum X_i) = \sum Var(X_i)$

- expectation & Σ
  $E(\sum X_i) = \sum E(X_i)$
- expectation & π
  $\underset{\sim}{X}, \underset{\sim}{Y}$: indep.
  $E[g(\underset{\sim}{X})h(\underset{\sim}{Y})] = E[g(\underset{\sim}{X})]E[h(\underset{\sim}{Y})]$
- $h\begin{cases} \text{convex} \Rightarrow E[h(Y)] > h(E(Y)) \\ \text{concave} \Rightarrow E[h(Y)] < h(E(Y)) \end{cases}$

- $X \le Y \Rightarrow E(X) \le E(Y)$
- $a \le X \le b \Rightarrow a \le E(X) \le b$

- $E(XY) = E(X)E(Y)$
- indep $\overset{\Rightarrow}{\cancel{\Leftarrow}}$ uncorrelated

$X_1,\cdots,X_n$ have same mean μ, var σ^2 & indep.
- $\bar{X}$
  - $E(\bar{X}) = \mu$
  - $Var(\bar{X}) = \sigma^2/n$
- S^2
  - $E(S^2) = \sigma^2$
  - $Var(S^2) \to 0$ as $n \to \infty$

understand their differences

- compare
  - $E_{Y|X}(Y|X=x)$
  - $E_{Y|X}(Y|X)$
  - $E_Y(Y)$
- compare
  - $Var_{Y|X}(Y|X=x)$
  - $Var_{Y|X}(Y|X)$
  - $Var_Y(Y)$

cf.

compare its dist. with the marginal dist. of Y

- $\rho_{XY} = cor(X,Y)$
  $= \frac{cov(X,Y)}{\sigma_X\sigma_Y}$
  $= E\left[\left(\frac{X-\mu}{\sigma_X}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right]$
- unit free
- measure strength of linear association

- $cor(a_0+a_1X, b_0+b_1Y) = sign(a_1b_1)cor(X,Y)$
- $-1 \le \rho_{XY} \le 1$
- $\rho_{XY} = \begin{cases} 1 \Rightarrow Y = aX+b, a>0 \\ -1 \Rightarrow Y = aX+b, a<0 \end{cases}$

- $E_{\underset{\sim}{Y}|\underset{\sim}{x}}[h(\underset{\sim}{Y})|\underset{\sim}{X}=\underset{\sim}{x}]$: a function of $\underset{\sim}{x}$, free of $\underset{\sim}{Y}$
- $E_{\underset{\sim}{Y}|\underset{\sim}{x}}[h(\underset{\sim}{X})|\underset{\sim}{X}=\underset{\sim}{x}] = h(\underset{\sim}{x})$
- $\underset{\sim}{X}, \underset{\sim}{Y}$: indep.
  $E_{\underset{\sim}{Y}|\underset{\sim}{x}}[h(\underset{\sim}{Y})|\underset{\sim}{X}=\underset{\sim}{x}] = E_{\underset{\sim}{Y}}[h(\underset{\sim}{Y})]$
  $\Rightarrow$ a constant function of $\underset{\sim}{x}$

Variance decomposition

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

law of total expectation

cf.

- $E_{\underset{\sim}{X},\underset{\sim}{Y}}[h(\underset{\sim}{Y})]$
  $E_{\underset{\sim}{Y}}[h(\underset{\sim}{Y})]$ $E_{\underset{\sim}{X}}E_{\underset{\sim}{Y}|\underset{\sim}{X}}[h(\underset{\sim}{Y})|\underset{\sim}{X}]$
  $\Rightarrow E_Y(Y) = E_XE_{Y|X}(Y|X)$
- $E_{\underset{\sim}{X},\underset{\sim}{Y}}[h(\underset{\sim}{X},\underset{\sim}{Y})]$
  $E_{\underset{\sim}{Y}}E_{\underset{\sim}{X}|\underset{\sim}{Y}}[h(\underset{\sim}{X},\underset{\sim}{Y})|\underset{\sim}{Y}]$ $E_{\underset{\sim}{X}}E_{\underset{\sim}{Y}|\underset{\sim}{X}}[h(\underset{\sim}{X},\underset{\sim}{Y})|\underset{\sim}{X}]$

∵ multiplication law

Prediction

- criterion: MSE
- best constant predictor
  $\mu_Y \to MSE = \sigma_Y^2$
- best linear predictor
  $\mu_Y - \rho_{XY}\frac{\sigma_Y}{\sigma_X}(x-\mu_X) \to MSE = \sigma_Y^2(1-\rho_{XY}^2)$
- best predictor
  $E_{Y|X}(Y|X=x) \to MSE = E_X[Var_{Y|X}(Y|X)]$

- better predictor
- smaller MSE
- need more information about the dist. of X, Y