$G_1$ **Theorem** (best *constant* predictor under MSE). a constant function of $x$; minimum

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

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

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

$E_{X,Y}[h(Y)] = E_Y[h(Y)]$ (LNp.8-22)

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

$$E_Y(Y-c)^2 \overset{\text{Thm in LNp.5-19}}{=} Var_Y(Y) + (\mu_Y - c)^2 \geq Var_Y(Y)$$

LNp.8-24: minimum. cf. LNp.8-24

12/11

$G_3$ **Theorem** (best predictor under MSE).

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

minimum

The equality holds if and only if $g(x)=E_{Y|X}(Y|x)$. a function of $X$ only; cf. (LNp.8-21)

Proof. $E_{X,Y}[Y-g(X)]^2$

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

$$= E_{X,Y}[Y-E_{Y|X}(Y|X)]^2 + \underbrace{E_X}_{E_{X,Y}}[E_{Y|X}(Y|X)-g(X)]^2$$

a function of $X$ only

$$+ 2\cdot E_{X,Y}\{[Y-E_{Y|X}(Y|X)][E_{Y|X}(Y|X)-g(X)]\} \quad = 0$$

last "="

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

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

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



where the last "=" comes from

$$E_{X,Y}\{[Y-E_{Y|X}(Y|X)][E_{Y|X}(Y|X)-g(X)]\} \quad h(X,Y)$$

$$= E_X E_{Y|X}\left\{[Y-E_{Y|X}(Y|X)]\,[E_{Y|X}(Y|X)-g(X)] \,\middle|\, X\right\}$$

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

this is a constant when conditioned on $X$ ($x$)

important concept: mean is best predictor under MSE

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

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

Furthermore, (for (*) in LNp.8-27)

$$E_{X,Y}[Y-E_{Y|X}(Y|X)]^2$$

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

➢ Some notes for the best predictor in $G_3$

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

- $E_{Y|X}(Y|x)$ is the best predictor of $Y$ based on $X$, in the 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.)
  - $E_{Y|X}(Y|x)$ is called the regression function of $Y$ on $X$ 迴歸

$G_2$

**Theorem** (best *linear* predictor under MSE). $-1 \le \rho_{XY} \le 1$

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

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

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

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

$\because Var(Z) = E(Z^2) - [E(Z)]^2$

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

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

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

Thm in LNp.8-13

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

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

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

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



➢ 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 $Y$. $\xleftrightarrow{\text{cf.}}$ best in $G_3, G_1$ Which one require more information? more information better predictor
- $\sigma_Y^2(1-\rho_{XY}^2)$ is small if $\rho_{XY}$ is close to $+1$ or $-1$, and large if $\rho_{XY}$ is close to 0. intuition? check the correlation plots in LNp.8-9

MSE=0 if $\rho_{XY}=\pm1$ MSE=$\sigma_Y^2$ if $\rho_{XY}=0$

• A comparison of these minimum MSEs

$\because G_1 \subset G_2 \subset G_3$

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

➢ $\min_g E_{X,Y}[Y-g(X)]^2 \le \min_{a,b} E_{X,Y}[Y-(a+bX)]^2$ 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 $k^{\text{th}}$ *moments* of $X$ provided that the integral converges absolutely, and

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

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

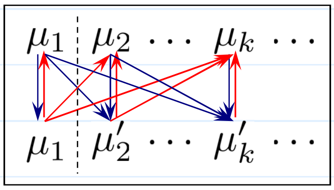
➢ Some notes.

- $\mu'_k = E[(X-\mu_X)^k] = E\left[\sum_{i=0}^{k} \binom{k}{i} (-\mu_X)^{k-i} X^i\right]$
  $= \sum_{i=0}^{k} \binom{k}{i} (-\mu_X)^{k-i} E(X^i) = \sum_{i=0}^{k} \binom{k}{i} (-\mu_X)^{k-i} \mu_i.$

  $\boxed{\mu_0 = E(X^0) = 1}$

- $\mu_k = E(X^k) = E\{[(X-\mu_X)+\mu_X]^k\}$
  $= \sum_{i=0}^{k} \binom{k}{i} (\mu_X)^{k-i} E[(X-\mu_X)^i]$
  $= \sum_{i=0}^{k} \binom{k}{i} (\mu_X)^{k-i} \mu'_i.$

  $\boxed{\mu'_0 = 1, \ \mu'_1 = 0}$



- In particular,

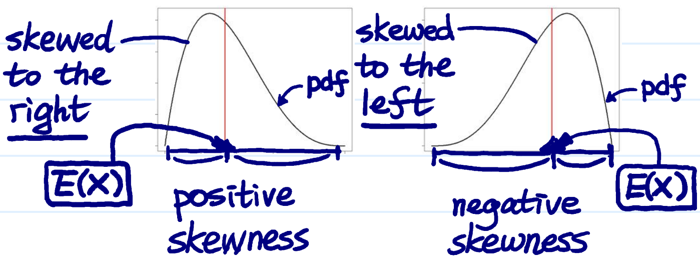
$$E(X) = \mu_X = \mu_1, \quad \text{and,}$$
$$Var(X) = \sigma_X^2 = \mu'_2 = \mu_2 - \mu_1^2. = E(X^2) - [E(X)]^2$$

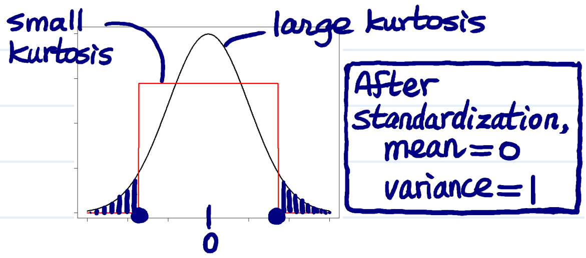


$\boxed{\text{Recall. mean, var, cov, cor}}$ $\boxed{\text{defined by expectation}}$

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

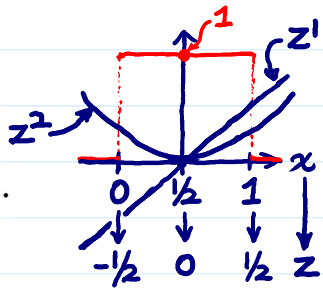
- ▫ Skewness (a measure of the asymmetry): $\mu'_3/\sigma^3 . = E\left(\frac{X-\mu}{\sigma}\right)^3$
- ▫ Kurtosis (a measure of the "heavy tails"): $\mu'_4/\sigma^4 . = E\left(\frac{X-\mu}{\sigma}\right)^4$




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

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

$\boxed{\text{skewness}=0, \ \text{kurtosis}=1.8}$

- 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

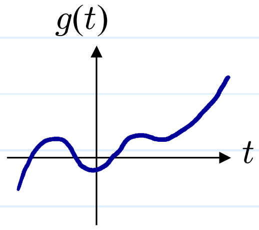
$$M_X(t) = E(e^{tX}) = \int_{-\infty}^{\infty} e^{tx}\, dF_X(x),$$

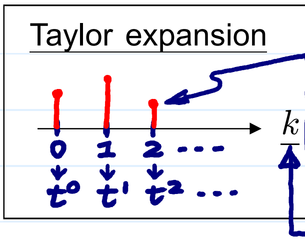
$= \int_{-\infty}^{\infty} e^{tx} f_X(x)dx$ for continuous case

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

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

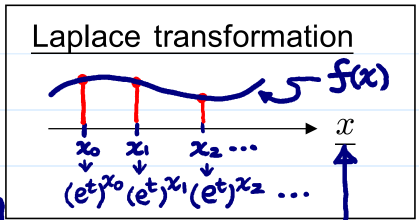
Q: how to express a function?




$a_k = \frac{g^{(k)}(0)}{k!}$

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($\lambda$), then for $-\infty<t<\infty$,

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

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

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

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

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

STO

This must be $>0$

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

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

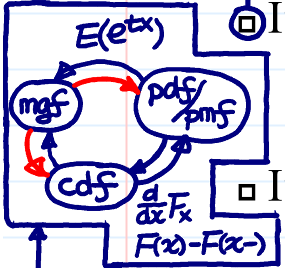
use negative binomial expansion (LNp.5-29)

$\alpha=1$, exponential distribution

- If $X \sim$ uniform$(\alpha, \beta)$, $M_X(\underline{t}) = \frac{e^{\beta \underline{t}} - e^{\alpha \underline{t}}}{\underline{t}(\beta-\alpha)}$.
- If $X \sim$ gamma$(\alpha, \lambda)$, $M_X(\underline{t}) = \left(\frac{\lambda}{\lambda-\underline{t}}\right)^{\alpha}$, for $t < \lambda$.

use STO

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




- If $X \sim$ beta$(\alpha, \beta)$, $M_X(\underline{t}) = 1 + \sum_{k=1}^{\infty} \left(\prod_{r=0}^{k-1} \frac{\alpha+r}{\alpha+\beta+r}\right) \frac{\underline{t}^k}{k!}$.
- If $X \sim$ normal$(\mu, \sigma^2)$, $M_X(\underline{t}) = e^{\mu \underline{t} + (\sigma^2/2)\underline{t}^2}$.

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

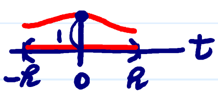
$$M_X(t) = M_Y(t),$$

for $|t|<h$, then

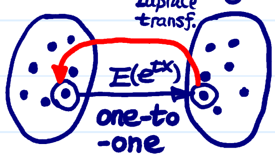
$$F_X(z) = F_Y(z)$$

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




distribution $\xrightarrow[\text{transf.}]{\text{Laplace}}$ mgf




X, Y have same dist.

does not mean X=Y

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 $M_X(t) = p_1 e^{a_1 t} + \cdots + p_k e^{a_k t}$, where $p_1+\cdots+p_k=1$, then $X$ is a discrete r.v. and its pmf is

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

by uniqueness Thm & mgf in LNp.8-34

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

$$M_X(0)=1,$$

can take derivative at $t=0$

and,

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

kth derivative

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

Know all moments ⇒ know dist.

This explains why it's called moment generating function.

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

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

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

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

$$\cdots = \cdots$$

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

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

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

➤ Example. If $X \sim$ exponential($\lambda$), 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, ...