

Expectation

- Recall. Expectation for univariate random variable.
- Theorem. For random variables $\underline{\mathbf{X}}=(X_1, \dots, X_n)$ with joint pmf $p_{\mathbf{X}}$ /pdf $f_{\mathbf{X}}$, the expectation of a univariate random variable \underline{Y} , where

$$\underline{Y}=g(\underline{X}_1, \dots, \underline{X}_n), \quad g:\mathbb{R}^n \rightarrow \mathbb{R}^1,$$

is
$$\underline{E}(Y) \equiv \sum_{y \in \mathcal{Y}} y p_Y(y) \quad (1)$$

$$= \sum_{\mathbf{x}=(x_1, \dots, x_n) \in \mathcal{X}} g(x_1, \dots, x_n) p_{\mathbf{X}}(x_1, \dots, x_n) \quad (2)$$

$$\equiv \underline{E}[g(X_1, \dots, X_n)]$$

if $\underline{X}_1, \dots, \underline{X}_n$ are discrete and the sum converges absolutely, or

$$\underline{E}(Y) \equiv \int_{-\infty}^{\infty} y f_Y(y) dy \quad (3)$$

$$= \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} g(x_1, \dots, x_n) f_{\mathbf{X}}(x_1, \dots, x_n) dx_1 \dots dx_n \quad (4)$$

$$\equiv \underline{E}[g(X_1, \dots, X_n)]$$

if \underline{Y} and $\underline{X}_1, \dots, \underline{X}_n$ are continuous and the integrals converges absolutely.

NTHU MATH 2810, 2023, Lecture Notes
made by S.-W. Cheng (NTHU, Taiwan)

Proof. Like the univariate case.

➤ **Q:** What if \underline{Y} is discrete and $\underline{X}_1, \dots, \underline{X}_n$ are continuous?

➤ Notation.

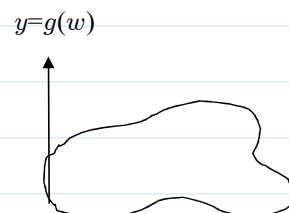
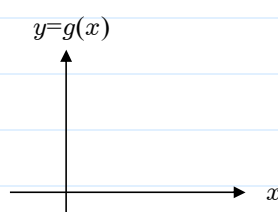
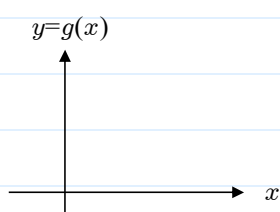
- Shorthand notation. Combine (1) and (3) by writing

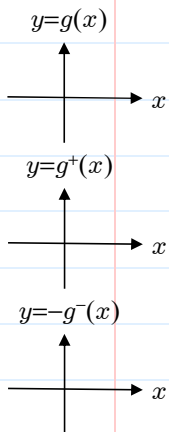
$$\underline{E}(Y) = \int_{-\infty}^{\infty} y dF_Y(y) = \begin{cases} \sum_{y \in \mathcal{Y}} y p_Y(y), & \text{for } \underline{\text{discrete}} \text{ case,} \\ \int_{-\infty}^{\infty} y f_Y(y) dy, & \text{for } \underline{\text{continuous}} \text{ case,} \end{cases}$$

and combine (2) and (4) by writing

$$\underline{E}[g(\underline{\mathbf{X}})] = \int_{\mathbb{R}^n} g(\mathbf{x}) dF_{\mathbf{X}}(\mathbf{x}) = \begin{cases} \sum_{\mathbf{x} \in \mathcal{X}} g(\mathbf{x}) p_{\mathbf{X}}(\mathbf{x}), & \text{for } \underline{\text{discrete}} \text{ case,} \\ \int_{\mathbb{R}^n} g(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}, & \text{for } \underline{\text{continuous}} \text{ case.} \end{cases}$$

- Riemann-Stieltjes Integral.





For example, for non-negative g , and non-decreasing, right-continuous F ,

$$\int_a^b g(x) dF(x) = \lim \sum_{i=1}^n g(x_i) [F(x_i) - F(x_{i-1})].$$

where the limit is taken over all $a=x_0 < x_1 < \dots < x_n = b$ as $n \rightarrow \infty$ and $\max_{i=1, \dots, n} (x_i - x_{i-1}) \rightarrow 0$.

[Recall. The integral of g over $(a, b]$ is defined as

$$\int_a^b g(x) dx = \lim \sum_{i=1}^n g(x_i) (x_i - x_{i-1}).]$$

➤ Note.

- $g(X_1, \dots, X_n) = X_i \Rightarrow E[g(X_1, \dots, X_n)] = E(X_i) \equiv \mu_{X_i}$.
- $g(X_1, \dots, X_n) = (X_i - \mu_{X_i})^2 \Rightarrow E[g(X_1, \dots, X_n)] = \text{Var}(X_i) \equiv \sigma_{X_i}^2$.

➤ Example (Average distance between two points). Suppose that X, Y are i.i.d. ~ Uniform(0, 1).

Let $D = |X - Y|$. Find $E(D)$.

NTHU MATH 2810, 2023, Lecture Notes
made by S.-W. Cheng (NTHU, Taiwan)

- The joint pdf of (X, Y) is

$$f(x, y) = \begin{cases} 1, & 0 \leq x \leq 1, 0 \leq y \leq 1, \\ 0, & \text{otherwise.} \end{cases}$$



- $E(D) = \int_0^1 \int_0^1 |x - y| dy dx = \int_0^1 \left[\int_0^x (x - y) dy + \int_x^1 (y - x) dy \right] dx$
 $= \int_0^1 \left[-\frac{1}{2}(y - x)^2 \Big|_{y=0}^x + \frac{1}{2}(y - x)^2 \Big|_{y=x}^1 \right] dx$
 $= \int_0^1 \frac{1}{2} [x^2 + (1 - x)^2] dx = \frac{1}{6} [x^3 - (1 - x)^3] \Big|_{x=0}^1 = \frac{1}{3}$.

- Theorem (Mean of Sum). For jointly distributed r.v.'s X_1, \dots, X_n and constants $-\infty < a_0, a_1, \dots, a_n < \infty$,

$$E(a_0 + a_1 X_1 + \dots + a_n X_n) = a_0 + a_1 E(X_1) + \dots + a_n E(X_n).$$

Proof. $E(a_0 + a_1 X_1 + \dots + a_n X_n)$

$$\begin{aligned} &= \int_{\mathbb{R}^n} (a_0 + a_1 x_1 + \dots + a_n x_n) dF_{\mathbf{X}}(\mathbf{x}) \\ &= \int_{\mathbb{R}^n} a_0 dF_{\mathbf{X}}(\mathbf{x}) + a_1 \int_{\mathbb{R}^n} x_1 dF_{\mathbf{X}}(\mathbf{x}) \\ &\quad + \dots + a_n \int_{\mathbb{R}^n} x_n dF_{\mathbf{X}}(\mathbf{x}) \\ &= a_0 + a_1 E(X_1) + \dots + a_n E(X_n). \end{aligned}$$

➤ Corollary. Suppose that $\mu = E(X_1) = \dots = E(X_n)$. Let

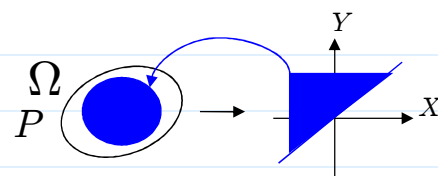
$$\bar{X}_n = \frac{X_1 + \dots + X_n}{n},$$

then, $E(\bar{X}_n) = \mu$.

➤ Corollary. If X and Y are r.v.'s with finite means and

$$P(X \leq Y) = 1,$$

then $E(X) \leq E(Y)$.



Proof. First, if Z is a random variable with finite mean and

$$P(Z \geq 0) = 1,$$

then $E(Z) = \int_0^{\infty} z dF_Z(z) \geq 0$.

For the general case, let $Z = Y - X$, then $Z \geq 0$ with probability one, and therefore, $0 \leq E(Z) = E(Y - X) = E(Y) - E(X)$.

➤ Corollary. If $P(a \leq X \leq b) = 1$ for some constants a, b , then

$$a \leq E(X) \leq b.$$

NTHU MATH 2810, 2023, Lecture Notes
made by S.-W. Cheng (NTHU, Taiwan)

• Theorem. If two random vectors $\mathbf{X} (\in \mathbb{R}^m)$ and $\mathbf{Y} (\in \mathbb{R}^n)$ are independent (i.e., $F_{\mathbf{X}, \mathbf{Y}}(\mathbf{x}, \mathbf{y}) = F_{\mathbf{X}}(\mathbf{x}) \times F_{\mathbf{Y}}(\mathbf{y})$, or

$$f_{\mathbf{X}, \mathbf{Y}}(\mathbf{x}, \mathbf{y}) = f_{\mathbf{X}}(\mathbf{x}) \times f_{\mathbf{Y}}(\mathbf{y}), \text{ or } p_{\mathbf{X}, \mathbf{Y}}(\mathbf{x}, \mathbf{y}) = p_{\mathbf{X}}(\mathbf{x}) \times p_{\mathbf{Y}}(\mathbf{y}),$$

then for $g: \mathbb{R}^m \rightarrow \mathbb{R}$ and $h: \mathbb{R}^n \rightarrow \mathbb{R}$,

$$E[g(\mathbf{X}) \times h(\mathbf{Y})] = E[g(\mathbf{X})] \times E[h(\mathbf{Y})].$$

Proof. We only prove it for the continuous case:

$$\begin{aligned} E[g(\mathbf{X})h(\mathbf{Y})] &= \int_{\mathbb{R}^m} \int_{\mathbb{R}^n} g(\mathbf{x})h(\mathbf{y}) f_{\mathbf{X}, \mathbf{Y}}(\mathbf{x}, \mathbf{y}) dy d\mathbf{x} \\ &= \int_{\mathbb{R}^m} \int_{\mathbb{R}^n} g(\mathbf{x})h(\mathbf{y}) f_{\mathbf{X}}(\mathbf{x}) f_{\mathbf{Y}}(\mathbf{y}) dy d\mathbf{x} \\ &= \int_{\mathbb{R}^m} g(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}) \left[\int_{\mathbb{R}^n} h(\mathbf{y}) f_{\mathbf{Y}}(\mathbf{y}) dy \right] d\mathbf{x} \\ &= \left[\int_{\mathbb{R}^m} g(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} \right] \left[\int_{\mathbb{R}^n} h(\mathbf{y}) f_{\mathbf{Y}}(\mathbf{y}) dy \right] \\ &= E[g(\mathbf{X})] E[h(\mathbf{Y})]. \end{aligned}$$

➤ Corollary. For 2 independent r.v.'s X and Y ,

$$E(XY) = E(X) \times E(Y).$$

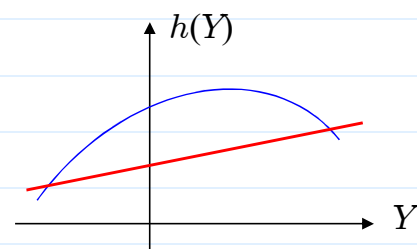
Proof. Let $g(X) = X$ and $h(Y) = Y$.

➤ Q: For independent r.v.'s X and Y,

$$\underline{E(X/Y)} = \underline{E(X)} / \underline{E(Y)}?$$

➤ Note. $\underline{E[h(Y)]} \neq h(\underline{E(Y)})$ in general, e.g.,

$$\underline{E(1/Y)} \neq 1/\underline{E(Y)}.$$



• Covariance and Correlation between 2 random variables

➤ Definition. Suppose that X and Y are two random variables with finite means $\underline{\mu_X}$, $\underline{\mu_Y}$ and variances $\underline{\sigma_X^2}$, $\underline{\sigma_Y^2}$, respectively.

1. Let $\underline{g(x, y)} = \underline{(x - \mu_X)(y - \mu_Y)}$, then

$$\begin{aligned} \underline{Cov(X, Y)} &\equiv \underline{E[g(X, Y)]} \\ &= \underline{E[(X - \mu_X)(Y - \mu_Y)]} \end{aligned}$$

is called the covariance between X and Y, denoted by $\underline{\sigma_{XY}}$.

NTHU MATH 2810, 2023, Lecture Notes
made by S.-W. Cheng (NTHU, Taiwan)

2. The correlation (coefficient) between X and Y is defined as ^{p. 8-8}

$$\underline{Cor(X, Y)} = \underline{\sigma_{XY}} / (\underline{\sigma_X} \underline{\sigma_Y})$$

and denoted by $\underline{\rho_{XY}}$.

3. X and Y are called uncorrelated if $\underline{\rho_{XY}} = 0$.

■ A special case of covariance:

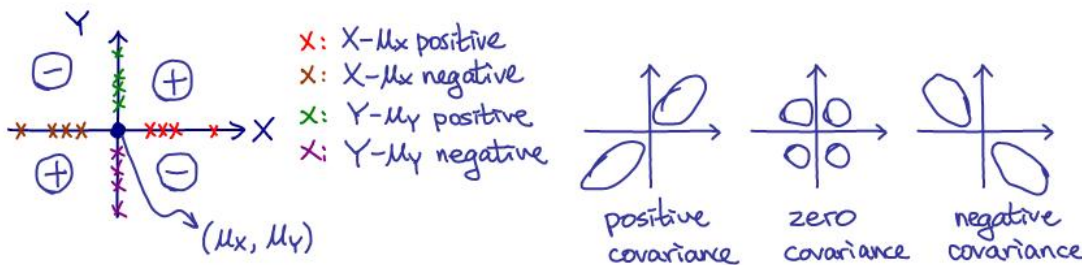
$$\underline{Cov(X, X)} = \underline{Var(X)}.$$

➤ Intuitive explanation of covariance and correlation

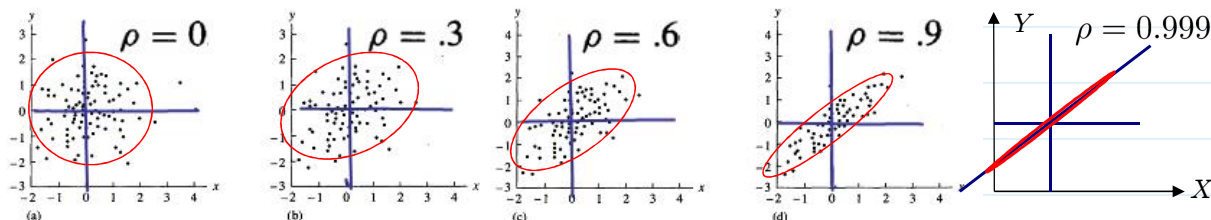
■ Covariance is the average value of the product of the deviation of X from its mean and the deviation of Y from its mean.

■ Covariance is a measure of the joint variability of X and Y, or their degree of association.

■ Positive Covariance and Negative Covariance



- Correlation Coefficient is unit free. (why?)
- Correlation coefficient measures the strength of the linear relationship between X and Y .



➤ Theorem. $Cov(X, Y) = E(XY) - \mu_X \mu_Y$.

Proof. $Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$

$$= E(XY - \mu_X Y - \mu_Y X + \mu_X \mu_Y)$$

$$= E(XY) - \mu_X E(Y) - \mu_Y E(X) + \mu_X \mu_Y$$

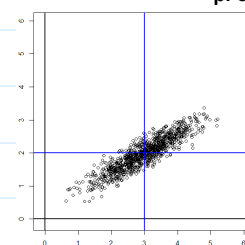
$$= E(XY) - \mu_X \mu_Y - \mu_Y \mu_X + \mu_X \mu_Y.$$

NTHU MATH 2810, 2023, Lecture Notes
made by S.-W. Cheng (NTHU, Taiwan)

- Corollary. If X and Y are independent, then $Cov(X, Y) = 0$, i.e., X and Y are uncorrelated.

Proof. When X, Y are independent,

$$E(XY) = E(X)E(Y) = \mu_X \mu_Y.$$

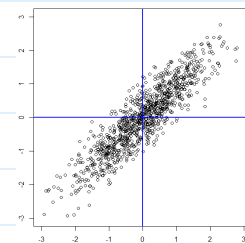


- However, the converse statement is not necessarily true.

(e.g., let $X \sim \text{Uniform}(-1, 1)$ and $Y = X^2$, then

$$Cov(X, Y) = 0,$$

but X and Y are not independent).



- Corollary.

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

Proof. By definition.

► Example. If $(X_1, \dots, X_m) \sim \text{Multinomial}(n, \underline{m}, p_1, \dots, p_m)$, then

$$\underline{Cov}(X_i, X_j) = \underline{-np_i p_j}, \quad \text{for } 1 \leq i \neq j \leq m.$$

■ Because $(X_1, X_2, \underline{X_3 + \dots + X_m}) \sim$

Multinomial $(n, \underline{3}, p_1, p_2, p_3 + \dots + p_m)$, and

$$\underline{X_3 + \dots + X_m} = n - X_1 - X_2, \quad \underline{p_3 + \dots + p_m} = 1 - p_1 - p_2,$$

we have

$$\begin{aligned} E(X_1 X_2) &= \sum \underline{x_1 x_2} \frac{\binom{n}{x_1, x_2, n-x_1-x_2} p_1^{x_1} p_2^{x_2} (1-p_1-p_2)^{n-x_1-x_2}}{x_1! x_2! (n-x_1-x_2)!} \\ &= \sum x_1 x_2 \frac{n!}{x_1! x_2! (n-x_1-x_2)!} p_1^{x_1} p_2^{x_2} (1-p_1-p_2)^{n-x_1-x_2} \\ &= n(n-1)p_1 p_2 \left[\sum \frac{(n-2)!}{(x_1-1)! (x_2-1)! (n-x_1-x_2)!} \right. \\ &\quad \left. \times (p_1)^{x_1-1} (p_2)^{x_2-1} (1-p_1-p_2)^{n-x_1-x_2} \right] \\ &= n(n-1)p_1 p_2. \end{aligned}$$

■ WLOG, we can get $\underline{E(X_i X_j)} = n(n-1)p_i p_j$, for $i \neq j$.

Therefore, $\underline{Cov}(X_i, X_j) = \underline{E(X_i X_j)} - \underline{E(X_i)} \underline{E(X_j)}$

$$= n(n-1)p_i p_j - (np_i)(np_j) = -np_i p_j.$$

NTHU MATH 2810, 2023, Lecture Notes
made by S.-W. Cheng (NTHU, Taiwan)

■ And, for $i \neq j$,

$$\underline{Cor}(X_i, X_j) = \frac{-np_i p_j}{\sqrt{np_i(1-p_i)} \sqrt{np_j(1-p_j)}} = -\sqrt{\frac{p_i p_j}{(1-p_i)(1-p_j)}}.$$

• Cov & Cor for Sums of Random Variables

► Notation. In the following, let $\underline{X_1, \dots, X_n}$ and

$\underline{Y_1, \dots, Y_m}$ be r.v.'s and $-\infty < \underline{a_0, a_1, \dots, a_n}$
 $\underline{b_0, b_1, \dots, b_m} < \infty$ are constants.

► Recall. $\underline{E(a_0 + a_1 X_1 + \dots + a_n X_n)} = \underline{a_0 + a_1 E(X_1) + \dots + a_n E(X_n)}$.

► Theorem (covariance of two sums).

$$\begin{aligned} \underline{Cov}(a_0 + a_1 X_1 + \dots + a_n X_n, b_0 + b_1 Y_1 + \dots + b_m Y_m) \\ = \sum_{i=1}^n \sum_{j=1}^m \underline{a_i b_j Cov}(X_i, Y_j). \end{aligned}$$

Proof. Let $S = a_0 + a_1 X_1 + \dots + a_n X_n$, and

$T = b_0 + b_1 Y_1 + \dots + b_m Y_m$, then

$$S - E(S) = \sum_{i=1}^n a_i (X_i - \mu_{X_i}),$$

$$T - E(T) = \sum_{j=1}^m b_j (Y_j - \mu_{Y_j}),$$

$$[S - E(S)][T - E(T)] = \sum_{i=1}^n \sum_{j=1}^m a_i b_j (X_i - \mu_{X_i})(Y_j - \mu_{Y_j}).$$

$$\begin{aligned} \text{Therefore, } Cov(S, T) &= E \{ [S - E(S)][T - E(T)] \} \\ &= \sum_{i=1}^n \sum_{j=1}^m a_i b_j E[(X_i - \mu_{X_i})(Y_j - \mu_{Y_j})] \\ &= \sum_{i=1}^n \sum_{j=1}^m a_i b_j Cov(X_i, Y_j). \end{aligned}$$

► Theorem (variance of sum).

$$\begin{aligned} &Var(a_0 + a_1 X_1 + \dots + a_n X_n) \\ &= \sum_{i=1}^n \sum_{j=1}^n a_i a_j Cov(X_i, X_j) \\ &= \sum_{i=1}^n a_i^2 Var(X_i) \\ &\quad + 2 \sum_{1 \leq i < j \leq n} a_i a_j Cov(X_i, X_j). \end{aligned}$$

Proof. $Cov(X_i, X_i) = Var(X_i)$ and $Cov(X_i, X_j) = Cov(X_j, X_i)$.

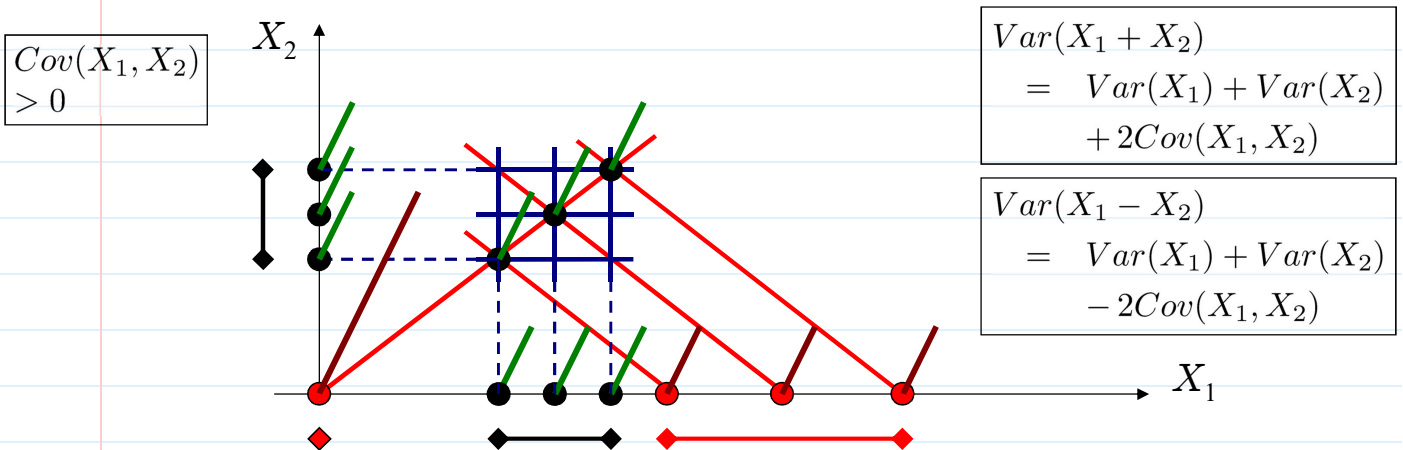
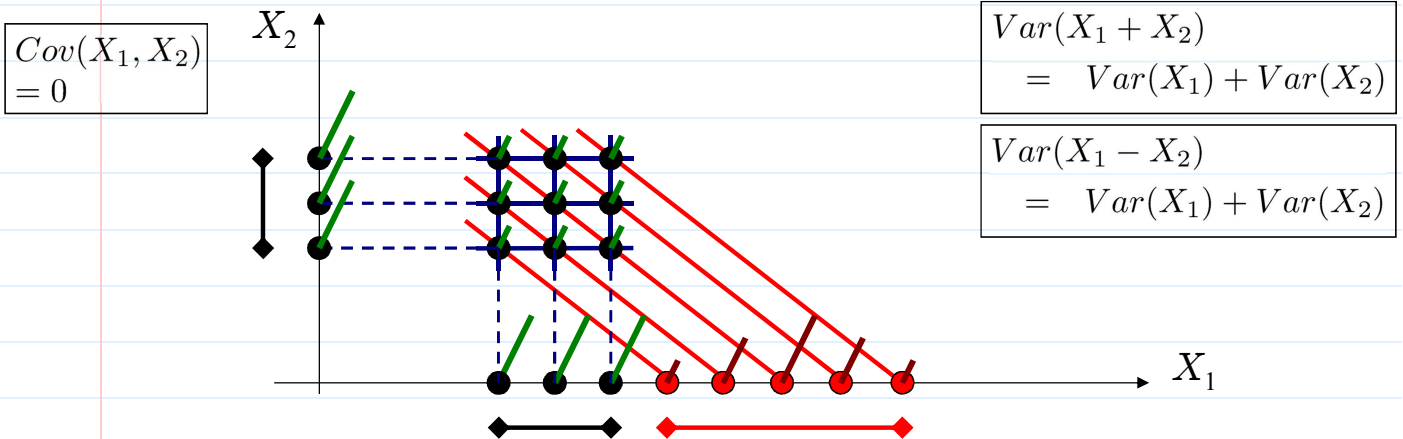
■ Corollary. If X_1, \dots, X_n are uncorrelated, then

$$Var(a_0 + a_1 X_1 + \dots + a_n X_n) = \sum_{i=1}^n a_i^2 Var(X_i).$$

■ Corollary. If X_1, \dots, X_n are uncorrelated and

$$Var(X_1) = \dots = Var(X_n) \equiv \sigma^2 < \infty,$$

then $Var(\bar{X}_n) = \sigma^2/n$.



■ Corollary. Suppose that X_1, \dots, X_n are uncorrelated and have same mean μ and variance σ^2 . Let

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{n - 1},$$

then $E(S^2) = \sigma^2$.

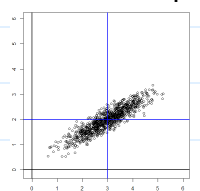
Proof.

$$\begin{aligned} (n - 1)S^2 &= \sum_{i=1}^n (X_i - \bar{X}_n)^2 \\ &= \sum_{i=1}^n [(X_i - \mu) - (\bar{X}_n - \mu)]^2 \\ &= [\sum_{i=1}^n (X_i - \mu)^2] + [\sum_{i=1}^n (\bar{X}_n - \mu)^2] \\ &\quad - 2(\bar{X}_n - \mu) [\sum_{i=1}^n (X_i - \mu)] \\ &= [\sum_{i=1}^n (X_i - \mu)^2] + n(\bar{X}_n - \mu)^2 - 2n(\bar{X}_n - \mu)^2 \\ &= [\sum_{i=1}^n (X_i - \mu)^2] - n(\bar{X}_n - \mu)^2. \end{aligned}$$

Therefore,

$$\begin{aligned} (n - 1)E(S^2) &= \left\{ \sum_{i=1}^n E[(X_i - \mu)^2] \right\} - nE[(\bar{X}_n - \mu)^2] \\ &= n\sigma^2 - nVar(\bar{X}_n) = (n - 1)\sigma^2. \end{aligned}$$

■ Note. The previous three corollaries also hold if X_1, \dots, X_n are “independent.”



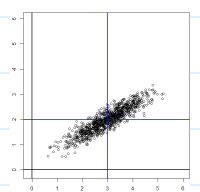
➤ Theorem (ρ of linear transformation).

$$\underline{Cor}(a_0 + a_1 X, b_0 + b_1 Y) = \text{sign}(a_1 b_1) \times \underline{Cor}(X, Y),$$

and

$$|\underline{Cor}(a_0 + a_1 X, b_0 + b_1 Y)| = |\underline{Cor}(X, Y)|,$$

i.e., $|\rho_{XY}|$ is invariant under location and scale changes.



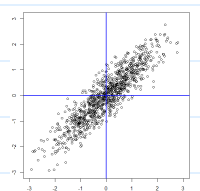
Proof. Let $S = a_0 + a_1 X$ and $T = b_0 + b_1 Y$, then

$$\underline{Cov}(S, T) = \underline{Cov}(a_0 + a_1 X, b_0 + b_1 Y) = a_1 b_1 \underline{Cov}(X, Y),$$

$$\underline{Var}(S) = a_1^2 \underline{Var}(X), \quad \text{and} \quad \underline{Var}(T) = b_1^2 \underline{Var}(Y).$$

Therefore,

$$\rho_{ST} = \frac{\underline{Cov}(S, T)}{\sigma_S \sigma_T} = \frac{a_1 b_1 \underline{Cov}(X, Y)}{|a_1| |b_1| \sigma_X \sigma_Y} = \frac{a_1 b_1}{|a_1 b_1|} \rho_{XY}.$$

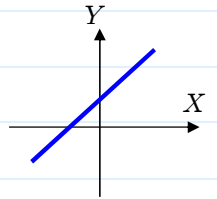


► Theorem (some properties of ρ).

(1) $-1 \leq \rho_{XY} \leq 1$. ($\Leftrightarrow |Cov(X, Y)| \leq \sigma_X \sigma_Y$)

(2) $\rho_{XY} = \pm 1$ if and only if there exist $a, b \in \mathbb{R}$

such that $P(Y = aX + b) = 1$.



(3) Furthermore, $\rho_{XY} = 1$, if $a > 0$ and $\rho_{XY} = -1$, if $a < 0$.

Proof of (1).

$$\begin{aligned} 0 &\leq Var\left(\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}\right) \\ &= Var\left(\frac{X}{\sigma_X}\right) + Var\left(\frac{Y}{\sigma_Y}\right) + 2Cov\left(\frac{X}{\sigma_X}, \frac{Y}{\sigma_Y}\right) \\ &= \frac{Var(X)}{\sigma_X^2} + \frac{Var(Y)}{\sigma_Y^2} + 2\frac{Cov(X, Y)}{\sigma_X \sigma_Y} \\ &= 1 + 1 + 2\rho_{XY} \Rightarrow \rho_{XY} \geq -1. \end{aligned}$$

Similarly,

$$0 \leq Var\left(\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y}\right) = 1 + 1 - 2\rho_{XY} \Rightarrow \rho_{XY} \leq 1.$$

Proof of (2) and (3). We see from the proof of (1),

NTHU MATH 2810, 2023, Lecture Notes
made by S.-W. Cheng (NTHU, Taiwan)

$$\rho_{XY} = 1 \Leftrightarrow Var\left(\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y}\right) = 0,$$

$$\Leftrightarrow P\left(\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y} = c\right) = 1,$$

where c is a constant.

$$\Leftrightarrow P\left(Y = \frac{\sigma_Y}{\sigma_X} X + c\sigma_Y\right) = 1.$$

Similarly, $\rho_{XY} = -1 \Leftrightarrow P\left(Y = -\frac{\sigma_Y}{\sigma_X} X + c\sigma_Y\right) = 1.$

- **Q:** How to use expectations to (roughly) characterize the distribution of random variables X_1, \dots, X_n ?

► $g(X_1, \dots, X_n) = X_i \Rightarrow E[g(\mathbf{X})] = \mu_{X_i}$: mean of X_i .

► $g(X_1, \dots, X_n) = (X_i - \mu_{X_i})^2 \Rightarrow E[g(\mathbf{X})] = \sigma_{X_i}^2$: variance of X_i .

► $g(X_1, \dots, X_n) = (X_i - \mu_{X_i})(X_j - \mu_{X_j})$ for $i \neq j$

$\Rightarrow E[g(\mathbf{X})] = \sigma_{X_i X_j}$: covariance of X_i and X_j .

► $g(X_1, \dots, X_n) = [(X_i - \mu_{X_i})/\sigma_{X_i}][(X_j - \mu_{X_j})/\sigma_{X_j}]$ for $i \neq j$

$\Rightarrow E[g(\mathbf{X})] = \rho_{X_i X_j}$: correlation coefficient of X_i and X_j .

► Notes. $\mu_{X_i}, \sigma_{X_i}^2, \sigma_{X_i X_j}, \rho_{X_i X_j}$ are constants, not random

Conditional Expectation

- Recall. $p_{Y|X}(y|x)$ or $f_{Y|X}(y|x)$ is a pmf/pdf for y (y : random, x : fixed).
- Definition. For random vectors \mathbf{X} and \mathbf{Y} , the conditional expectation of $Z=h(\mathbf{Y})$ given $\mathbf{X}=\mathbf{x}$, where $h: \mathbb{R}^m \rightarrow \mathbb{R}^1$, is

$$E_{Y|X} \left(h(\mathbf{Y}) \mid \mathbf{X} = \mathbf{x} \right) = \sum_{y \in \mathcal{Y}} h(y) p_{Y|X}(y|x),$$

in the discrete case, or,

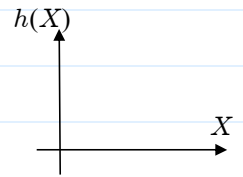
$$E_{Y|X} \left(h(\mathbf{Y}) \mid \mathbf{X} = \mathbf{x} \right) = \int_{\mathbb{R}^m} h(y) f_{Y|X}(y|x) dy,$$

in the continuous case,

provided that the sum or integral converges absolutely.

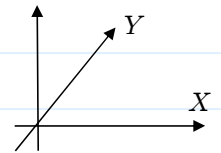
➤ Some Notes.

- $E_{Y|X}(h(\mathbf{Y}) \mid \mathbf{X}=\mathbf{x})$: a function of \mathbf{x} and free of \mathbf{Y} .
- $E_{Y|X}[h(\mathbf{X}) \mid \mathbf{X}=\mathbf{x}] = h(\mathbf{x})$.



- If \mathbf{X} and \mathbf{Y} are independent, then

$$E_{Y|X}(h(\mathbf{Y}) \mid \mathbf{X}=\mathbf{x}) = E_Y[h(\mathbf{Y})].$$



NTHU MATH 2810, 2023, Lecture Notes
made by S.-W. Cheng (NTHU, Taiwan)

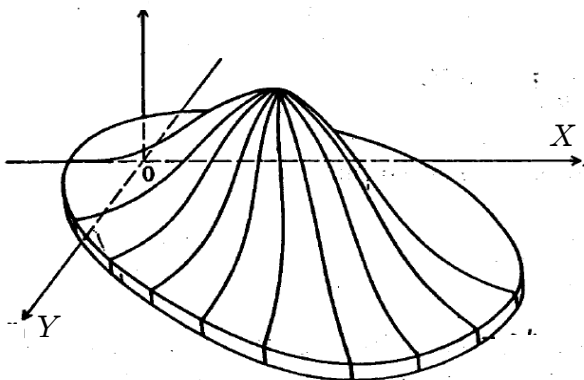
- Let $g(\mathbf{x}) = E_{Y|X}[h(\mathbf{Y}) \mid \mathbf{X}=\mathbf{x}]$, where $g: \mathbb{R}^n \rightarrow \mathbb{R}^1$, then we write

$$E_{Y|X}(h(\mathbf{Y}) \mid \mathbf{X})$$

when \mathbf{x} in g is replaced by \mathbf{X} (a fixed value replaced by a r.v.).

▢ Notice that $g(\mathbf{X})$ is a random variable.

$f(x, y)$: joint pdf



➤ $f(x, y)$: a joint pdf.

➤ Fix x^* , is $f(x^*, y)$ a pdf of y ? i.e.,

$$f_X(x^*) = \int_{-\infty}^{\infty} f(x^*, y) dy \stackrel{?}{=} 1.$$

➤ $f_{Y|X}(y|x^*) = f(x^*, y) / f_X(x^*)$ is a pdf of y since

$$\frac{\int_{-\infty}^{\infty} f(x^*, y) dy}{f_X(x^*)} = 1.$$

➤ $E_{Y|X}(Y|x^*)$: mean of $f_{Y|X}(y|x^*)$.

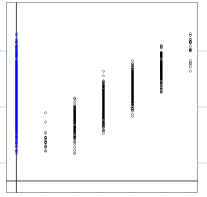
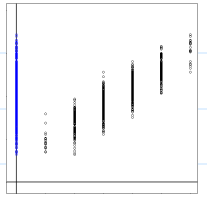
➤ Do it for any $x=x^*$, and get a function of x $\Rightarrow E_{Y|X}(Y|x)$

➤ Example. Sample a student from an elementary school. Let

$$\underline{X} = \text{age (unit: year)}, \quad \underline{Y} = \text{height (unit: cm)}$$

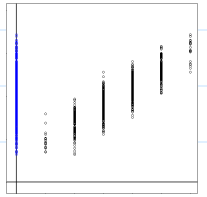
of the student. **Population:** all students of the school.

- $\underline{Y|X=x}$: a random variable (unit: cm) that represents the height distribution of students with age=x.
- $g(x) = \underline{E_{Y|X}}(Y|X=x)$ or $\underline{E_{Y|X}}(Y|x)$: a function maps from age (unit: year) to average height (unit: cm) of students with age=x.



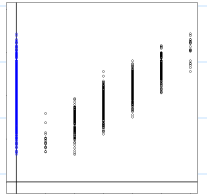
Note. $\underline{E_{Y|X}}(Y|x)$ is not a random variable.

- $g(X) = \underline{E_{Y|X}}(Y|X)$: a random variable because it is a function of age X , where X is a random variable.



Note. $g(X) = \underline{E_{Y|X}}(Y|X)$ is height, its unit is "cm".

- $\underline{Var_{Y|X}}(Y|X=x)$ & $\underline{Var_{Y|X}}(Y|X)$ defined similarly.
- $\underline{E_Y}(Y)$: average height of **all students**;
- $\underline{Var_Y}(Y)$: variation of height of **all students**.



• Theorem (Law of Total Expectation). For two random vectors $\underline{X} (\in \mathbb{R}^m)$ and $\underline{Y} (\in \mathbb{R}^n)$,

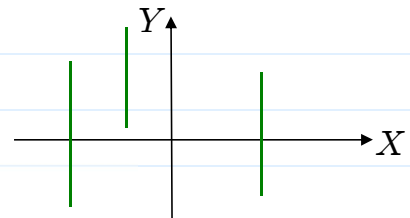
$$\underline{E_X}\{\underline{E_{Y|X}}[h(\underline{Y})|\underline{X}]\} = \underline{E_Y}[h(\underline{Y})].$$

In particular, let $h(\underline{Y}) = Y_i$, we have

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

Proof.

(only prove it for the continuous case)



$$\begin{aligned} & \underline{E_X}\{\underline{E_{Y|X}}[h(\underline{Y})|\underline{X}]\} \\ &= \int_{\mathbb{R}^m} \underline{E_{Y|X}}(h(\underline{Y})|\underline{x}) f_{\underline{X}}(\underline{x}) d\underline{x} \\ &= \int_{\mathbb{R}^m} \left[\int_{\mathbb{R}^n} h(\underline{y}) f_{\underline{Y|X}}(\underline{y}|\underline{x}) d\underline{y} \right] f_{\underline{X}}(\underline{x}) d\underline{x} \\ &= \int_{\mathbb{R}^n} \int_{\mathbb{R}^m} h(\underline{y}) \frac{f_{\underline{X}, \underline{Y}}(\underline{x}, \underline{y})}{f_{\underline{X}}(\underline{x})} f_{\underline{X}}(\underline{x}) d\underline{x} d\underline{y} \\ &= \int_{\mathbb{R}^n} h(\underline{y}) \left[\int_{\mathbb{R}^m} f_{\underline{X}, \underline{Y}}(\underline{x}, \underline{y}) d\underline{x} \right] d\underline{y} \\ &= \int_{\mathbb{R}^n} h(\underline{y}) f_{\underline{Y}}(\underline{y}) d\underline{y} \\ &= \underline{E_Y}[h(\underline{Y})]. \end{aligned}$$

➤ Example. If a sample of n balls is drawn without replacement from a box containing R red balls, W white balls, and $N-R-W$ blue balls. Let

X = # of red balls in the sample,

Y = # of white balls in the sample,

then, the joint pmf of (X, Y) is

$$p_{X,Y}(x, y) = \frac{\binom{R}{x} \binom{W}{y} \binom{N-R-W}{n-x-y}}{\binom{N}{n}},$$

Find $E_Y(Y)$.

Sol. Because $Y|X=x \sim \text{hypergeometric}(n-x, N-R, W)$,

$$g(x) \equiv \underline{E}_{Y|X}(Y|X=x) = (n-x)[W/(N-R)].$$

Because $X \sim \text{hypergeometric}(n, N, R) \Rightarrow \underline{E}_X(X) = n(R/N)$, and

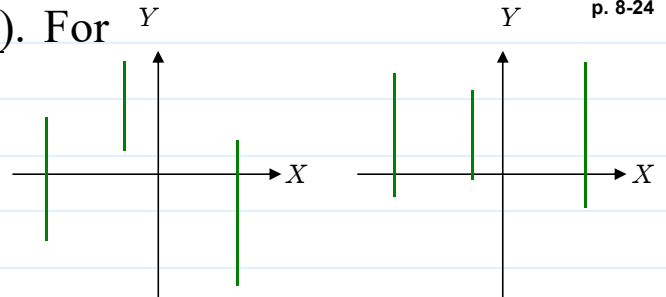
$$\begin{aligned} \text{then } \underline{E}_Y(Y) &= \underline{E}_X[\underline{E}_{Y|X}(Y|X)] = \underline{E}_X[g(X)] \\ &= \underline{E}_X\left[(n-X)\frac{W}{N-R}\right] = \frac{W}{N-R}[n - \underline{E}_X(X)] \\ &= \frac{W}{N-R}\left(n - n\frac{R}{N}\right) = n\frac{W}{N}. \end{aligned}$$

Note that $Y \sim \text{hypergeometric}(n, N, W) \Rightarrow \underline{E}_Y(Y) = n(W/N)$.

NTHU MATH 2810, 2023, Lecture Notes
made by S.-W. Cheng (NTHU, Taiwan)

• Theorem (Variance Decomposition). For two random vectors \mathbf{X} and \mathbf{Y} ,

$$\begin{aligned} \underline{Var}_Y(Y_i) &= \underline{Var}_X[\underline{E}_{Y|X}(Y_i|\mathbf{X})] \\ &\quad + \underline{E}_X[\underline{Var}_{Y|X}(Y_i|\mathbf{X})]. \end{aligned}$$



Proof. $\underline{Var}_{Y|\mathbf{X}}(Y_i|\underline{\mathbf{x}}) = \underline{E}_{Y|\mathbf{X}}(Y_i^2|\underline{\mathbf{x}}) - [\underline{E}_{Y|\mathbf{X}}(Y_i|\underline{\mathbf{x}})]^2,$

$$\begin{aligned} \text{and, } \underline{E}_X[\underline{Var}_{Y|\mathbf{X}}(Y_i|\mathbf{X})] &= \underline{E}_X[\underline{E}_{Y|\mathbf{X}}(Y_i^2|\mathbf{X})] - \underline{E}_X\{[\underline{E}_{Y|\mathbf{X}}(Y_i|\mathbf{X})]^2\}. \end{aligned}$$

$$\begin{aligned} \text{Also, } \underline{Var}_X[\underline{E}_{Y|\mathbf{X}}(Y_i|\mathbf{X})] &= \underline{E}_X\{[\underline{E}_{Y|\mathbf{X}}(Y_i|\mathbf{X})]^2\} - \{\underline{E}_X[\underline{E}_{Y|\mathbf{X}}(Y_i|\mathbf{X})]\}^2. \end{aligned}$$

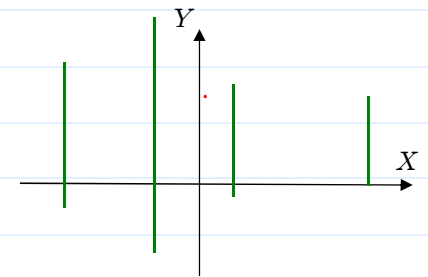
$$\begin{aligned} \text{Now, } \underline{Var}_Y(Y_i) &= \underline{E}_Y(Y_i^2) - [\underline{E}_Y(Y_i)]^2 \\ &= \underline{E}_X[\underline{E}_{Y|\mathbf{X}}(Y_i^2|\mathbf{X})] - \{\underline{E}_X[\underline{E}_{Y|\mathbf{X}}(Y_i|\mathbf{X})]\}^2 \\ &= \underline{E}_X[\underline{E}_{Y|\mathbf{X}}(Y_i^2|\mathbf{X})] - \underline{E}_X\{[\underline{E}_{Y|\mathbf{X}}(Y_i|\mathbf{X})]^2\} \\ &\quad + \underline{E}_X\{[\underline{E}_{Y|\mathbf{X}}(Y_i|\mathbf{X})]^2\} - \{\underline{E}_X[\underline{E}_{Y|\mathbf{X}}(Y_i|\mathbf{X})]\}^2 \\ &= \underline{E}_X[\underline{Var}_{Y|\mathbf{X}}(Y_i|\mathbf{X})] + \underline{Var}_X[\underline{E}_{Y|\mathbf{X}}(Y_i|\mathbf{X})]. \end{aligned}$$

➤ Corollary.

- $\underline{Var_Y(Y_i)} \geq \underline{E_X[Var_{Y|X}(Y_i|X)]}$ and the equality holds if and only if

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

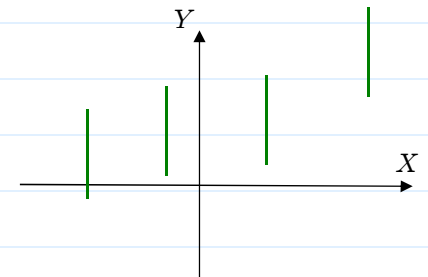
with probability one.



- $\underline{Var_Y(Y_i)} \geq \underline{Var_X[E_{Y|X}(Y_i|X)]}$ and the equality hold if and only if

$$\underline{Var_{Y|X}(Y_i|X)} = 0 \quad (\Rightarrow \quad \underline{Y_i} = \underline{E_{Y|X}(Y_i|X)})$$

with probability one.



❖ Reading: textbook, Sec 7.5

Conditional Expectation and Prediction

- Problem formulation: predicting the value of a r.v. Y on the basis of the observed value of a r.v. X

➤ Data: X and Y (example?)

NTHU MATH 2810, 2023, Lecture Notes
made by S.-W. Cheng (NTHU, Taiwan)

➤ Statistical modeling: assigning (X, Y) a (known) joint distribution (cdf $\underline{F(x, y)}$, pdf $\underline{f(x, y)}$, or pmf $\underline{p(x, y)}$)

➤ Objective: predicting Y by using a function of X, i.e.,

$$\underline{g(X)} \leftarrow \underline{\text{predictor}}$$

➤ Predictor: considering the following three groups of g's

(i) $\underline{G_1} = \{g(x) : \underline{g(x)} = \underline{c}, \text{ where } \underline{c} \in \mathbb{R}\}$

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

(iii) $\underline{G_3} = \{g(x) : \underline{g} \text{ is an } \underline{\text{arbitrary function}}\}$

Note. $\underline{G_1} \subset \underline{G_2} \subset \underline{G_3}$.

➤ Question: Within each group, what is the “best” predictor?

➤ Criterion: minimizing mean square error

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

- Theorem (best constant predictor under MSE).

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

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

Proof.

$$\begin{aligned} & \underline{E_Y}(Y - c)^2 \\ &= \underline{Var_Y}(Y) + (\mu_Y - c)^2 \\ &\geq \underline{Var_Y}(Y) \end{aligned}$$

- Theorem (best predictor under MSE).

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

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

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

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

NTHU MATH 2810, 2023, Lecture Notes
made by S.-W. Cheng (NTHU, Taiwan)

where the last “=” comes from

$$\begin{aligned} & \underline{E_{X,Y}} \{ \underline{[Y - E_{Y|X}(Y|X)][E_{Y|X}(Y|X) - g(X)]} \} \\ &= \underline{E_X} \underline{E_{Y|X}} \left\{ \underline{[Y - E_{Y|X}(Y|X)][E_{Y|X}(Y|X) - g(X)]} \mid \underline{X} \right\} \\ &= \underline{E_X} \{ \underline{[E_{Y|X}(Y|X) - g(X)] \underline{E_{Y|X}} \underline{[Y - E_{Y|X}(Y|X)]} \mid \underline{X}} \} = 0. \end{aligned}$$

Furthermore,

$$\begin{aligned} & \underline{E_{X,Y}} \underline{[Y - E_{Y|X}(Y|X)]^2} \\ &= \underline{E_X} \underline{E_{Y|X}} \{ \underline{[Y - E_{Y|X}(Y|X)]^2} \mid \underline{X} \} = \underline{E_X} \underline{[Var_{Y|X}(Y|X)]} \end{aligned}$$

➤ Some notes for the best predictor in G_3

- $E_{Y|X}(Y|x)$ is the best predictor of Y based on X , in the sense of mean square prediction error
- Its calculation requires 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

- Theorem (best linear predictor under MSE).

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

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

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

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

$$= \text{Var}_{X,Y}(Y - a - bX) + [E_{X,Y}(Y - a - bX)]^2$$

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

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

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

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

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

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

NTHU MATH 2810, 2023, Lecture Notes
made by S.-W. Cheng (NTHU, Taiwan)

➤ 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.
- Its calculation requires to know the means, variances, and covariance of X and Y .
- $\sigma_Y^2(1 - \rho_{XY}^2)$ is small if ρ_{XY} is close to $+1$ or -1 , and large if ρ_{XY} is close to 0 .

- A comparison of these minimum MSEs

➤ $\min_{a,b} E_{X,Y}[Y - (a + bX)]^2 \leq \min_c E_{X,Y}(Y - c)^2$ 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

$$\underline{\mu}_k \equiv E(\underline{X}^k) = \int_{-\infty}^{\infty} \underline{x}^k \underline{dF}_X(x), \quad k = 1, 2, 3, \dots,$$

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

$$\underline{\mu}'_k \equiv E[(\underline{X} - \underline{\mu}_X)^k] = \int_{-\infty}^{\infty} (x - \mu_X)^k dF_X(x), \quad k = 2, 3, \dots,$$

are called k^{th} moment about the mean $\underline{\mu}_X$ or central moment of \underline{X} provided that the integral converges absolutely.

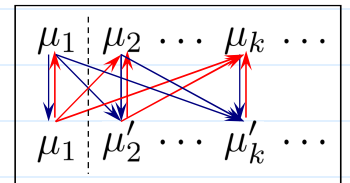
➤ Some notes.

$$\begin{aligned} \underline{\mu}'_k &= E[(\underline{X} - \underline{\mu}_X)^k] = E\left[\sum_{i=0}^k \binom{k}{i} (-\underline{\mu}_X)^{k-i} \underline{X}^i\right] \\ &= \sum_{i=0}^k \binom{k}{i} (-\underline{\mu}_X)^{k-i} E(\underline{X}^i) = \sum_{i=0}^k \binom{k}{i} (-\underline{\mu}_X)^{k-i} \underline{\mu}_i. \end{aligned}$$

$$\begin{aligned} \underline{\mu}_k &= E(\underline{X}^k) = E\{[(\underline{X} - \underline{\mu}_X) + \underline{\mu}_X]^k\} \\ &= \sum_{i=0}^k \binom{k}{i} (\underline{\mu}_X)^{k-i} E[(\underline{X} - \underline{\mu}_X)^i] \\ &= \sum_{i=0}^k \binom{k}{i} (\underline{\mu}_X)^{k-i} \underline{\mu}'_i. \end{aligned}$$

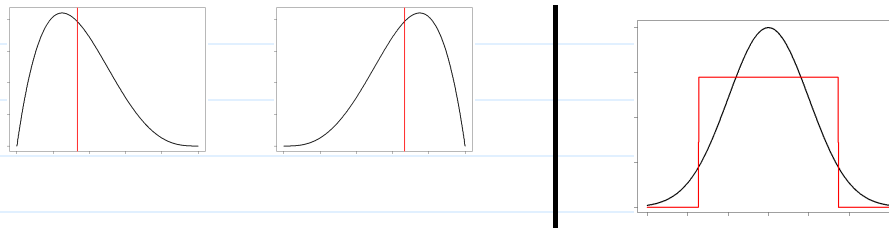
■ In particular,

$$\begin{aligned} \underline{E}(\underline{X}) &= \underline{\mu}_X = \underline{\mu}_1, \quad \text{and,} \\ \underline{Var}(\underline{X}) &= \underline{\sigma}_X^2 = \underline{\mu}'_2 = \underline{\mu}_2 - \underline{\mu}_1^2. \end{aligned}$$



NTHU MATH 2810, 2023, Lecture Notes
made by S.-W. Cheng (NTHU, Taiwan)

- 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): $\underline{\mu}'_3/\sigma^3$.
 - Kurtosis (a measure of the “heavy tails”): $\underline{\mu}'_4/\sigma^4$.



➤ Example (Uniform). If $\underline{X} \sim \text{Uniform}(0, 1)$, then

$$\underline{\mu}_k = \int_0^1 \underline{x}^k dx = \frac{1}{k+1},$$

therefore, $\underline{\mu}_X = \underline{\mu}_1 = 1/2$, and,

$$\underline{\sigma}_X^2 = \underline{\mu}_2 - \underline{\mu}_1^2 = 1/3 - (1/2)^2 = 1/12.$$

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

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

- Recall. How to characterize a distribution?

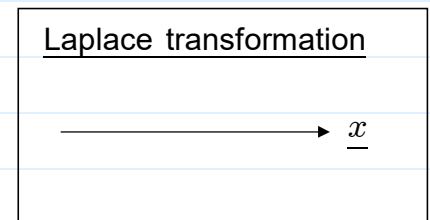
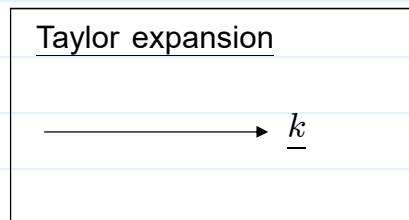
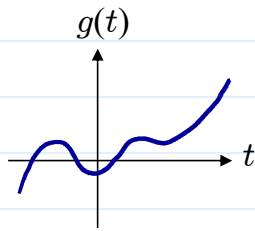
(1) pdf/pmf, (2) cdf, (3) mgf

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

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

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

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



➤ Some Notes.

- The mgf is a function of the variable t .
- The mgf may only exist for some particular values of t .
- $M_X(t)$ always exists at $t=0$ and $M_X(0)=1$

NTHU MATH 2810, 2023, Lecture Notes
made by S.-W. Cheng (NTHU, Taiwan)

➤ Example.

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

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

- If $X \sim \text{Poisson}(\lambda)$, then for $-\infty < t < \infty$,

$$\begin{aligned} M_X(t) &= E(e^{tX}) = \sum_{x=0}^{\infty} e^{tx} \times \frac{e^{-\lambda} \lambda^x}{x!} \\ &= e^{-\lambda} \left(e^{\lambda e^t} \right) \sum_{x=0}^{\infty} \frac{e^{-(\lambda e^t)} (\lambda e^t)^x}{x!} = e^{-\lambda} e^{\lambda e^t} = e^{\lambda(e^t - 1)}. \end{aligned}$$

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

$$\begin{aligned} M_X(t) &= E(e^{tX}) = \int_0^{\infty} e^{tx} \times \lambda e^{-\lambda x} dx \\ &= \lambda \left(\frac{1}{\lambda - t} \right) \int_0^{\infty} (\lambda - t) e^{-(\lambda - t)x} dx = \frac{\lambda}{\lambda - t}, \end{aligned}$$

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

- A list of some mgfs (**exercise**)

- If $X \sim \text{binomial}(n, p)$,

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

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

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

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

□ If $X \sim$ gamma(α, λ),

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

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

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

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

If

$$M_X(\underline{t}) = M_Y(\underline{t}),$$

for $|\underline{t}| < h$, then

$$F_X(\underline{z}) = F_Y(\underline{z})$$

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

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

NTHU MATH 2810, 2023, Lecture Notes
made by S.-W. Cheng (NTHU, Taiwan)

➤ Application of the uniqueness theorem

■ When a mgf exists for all $|\underline{t}| < h$ 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, \dots, X_n

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

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

• Theorem (Moments and MGF). If $M_X(\underline{t})$ exists for $|\underline{t}| < h$ for some $h > 0$, then

$$M_X(\underline{0}) = 1,$$

and,

$$M_X^{(k)}(\underline{0}) = \underline{\mu}_k, \quad k = 1, 2, 3, \dots$$

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

$$\begin{aligned} M_X'(\underline{0}) &= \left. \frac{d}{dt} M_X(t) \right|_{t=0} = \left[\frac{d}{dt} \int_{-\infty}^{\infty} e^{tx} dF_X(x) \right] \Big|_{t=0} \\ &= \int_{-\infty}^{\infty} \left(\left. \frac{d}{dt} e^{tx} \right|_{t=0} \right) dF_X(x) = \int_{-\infty}^{\infty} \left(\left. x e^{tx} \right|_{t=0} \right) dF_X(x) \\ &= \int_{-\infty}^{\infty} x \cdot \underline{1} dF_X(x) = E_X(X) = \underline{\mu}_1. \end{aligned}$$

... = ...

$$\begin{aligned} M_X^{(k)}(\underline{0}) &= \left. \frac{d^k}{dt^k} M_X(t) \right|_{t=0} = \left[\frac{d^k}{dt^k} \int_{-\infty}^{\infty} e^{tx} dF_X(x) \right] \Big|_{t=0} \\ &= \int_{-\infty}^{\infty} \left(\left. \frac{d^k}{dt^k} e^{tx} \right|_{t=0} \right) dF_X(x) = \int_{-\infty}^{\infty} \left(\left. x^k e^{tx} \right|_{t=0} \right) dF_X(x) \\ &= \int_{-\infty}^{\infty} x^k \cdot \underline{1} dF_X(x) = E_X(X^k) = \underline{\mu}_k. \end{aligned}$$

► Example. If $X \sim \text{exponential}(\lambda)$, then $M_X(t) = \frac{\lambda}{\lambda - t}$.

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

we get $\underline{\mu}_k = M_X^{(k)}(\underline{0}) = \frac{k!}{\lambda^k}.$

NTHU MATH 2810, 2023, Lecture Notes
made by S.-W. Cheng (NTHU, Taiwan)

• Theorem (MGF for linear transformation). For constants \underline{a} and \underline{b} , p. 8-38

$$M_{\underline{a} + \underline{b}X}(t) = e^{\underline{a}t} M_X(\underline{b}t).$$

Proof. $M_{\underline{a} + \underline{b}X}(t) = E_X[e^{t(\underline{a} + \underline{b}X)}] = e^{\underline{a}t} E_X[e^{(\underline{b}t)X}] = e^{\underline{a}t} M_X(\underline{b}t).$

• Theorem (MGF for SUM of independent r.v.'s). If X_1, \dots, X_n are independent each with mgfs $M_{X_1}(t), \dots, M_{X_n}(t)$, respectively, then the mgf of $\underline{S} = X_1 + \dots + X_n$ is

$$M_{\underline{S}}(t) = M_{X_1}(t) \times \dots \times M_{X_n}(t).$$

Proof. $M_{\underline{S}}(t) = E_{\underline{S}}(e^{t\underline{S}}) = E_{X_1, \dots, X_n}[e^{t(X_1 + \dots + X_n)}]$
 $= E_{X_1, \dots, X_n}(e^{tX_1} \times \dots \times e^{tX_n})$
 $= E_{X_1}(e^{tX_1}) \times \dots \times E_{X_n}(e^{tX_n}) = M_{X_1}(t) \times \dots \times M_{X_n}(t).$

► Example. If X_1, \dots, X_n are i.i.d. $\sim \text{geometric}(p)$, then

$\underline{S} = X_1 + \dots + X_n \sim \text{negative binomial}(n, p).$

Proof. $M_{\underline{S}}(t) = M_{X_1}(t) \times \dots \times M_{X_n}(t)$
 $= \frac{pe^t}{1 - (1-p)e^t} \times \dots \times \frac{pe^t}{1 - (1-p)e^t} = \left[\frac{pe^t}{1 - (1-p)e^t} \right]^n.$

➤ Example. If $\underline{X}_1, \dots, \underline{X}_n$ are independent and

$$\underline{X}_i \sim \underline{\text{normal}}(\underline{\mu}_i, \underline{\sigma}_i^2), \text{ for } i=1, \dots, n.$$

Let $\underline{S} = \underline{a}_0 + \underline{a}_1 \underline{X}_1 + \dots + \underline{a}_n \underline{X}_n$, then

$$\underline{S} \sim \underline{\text{normal}} \left(\underline{a}_0 + \underline{a}_1 \underline{\mu}_1 + \dots + \underline{a}_n \underline{\mu}_n, \underline{a}_1^2 \underline{\sigma}_1^2 + \dots + \underline{a}_n^2 \underline{\sigma}_n^2 \right).$$

Proof. $M_{\underline{S}}(t) = e^{\underline{a}_0 t} \times \prod_{i=1}^n e^{\underline{\mu}_i (\underline{a}_i t) + (\underline{\sigma}_i^2 / 2) (\underline{a}_i t)^2}$
 $= e^{(\underline{a}_0 + \underline{a}_1 \underline{\mu}_1 + \dots + \underline{a}_n \underline{\mu}_n) t + [(\underline{a}_1^2 \underline{\sigma}_1^2 + \dots + \underline{a}_n^2 \underline{\sigma}_n^2) / 2] t^2}.$

• Definition (Joint Moment Generating Function). For random variables $\underline{X}_1, \dots, \underline{X}_n$, their joint mgf is defined as

$$M_{\underline{X}_1, \dots, \underline{X}_n}(\underline{t}_1, \dots, \underline{t}_n) = E_{\underline{X}_1, \dots, \underline{X}_n} \left(e^{\underline{t}_1 \underline{X}_1 + \dots + \underline{t}_n \underline{X}_n} \right)$$

provided that the expectation exists.

➤ Example. If $\underline{X}_1, \dots, \underline{X}_m \sim \underline{\text{multinomial}}(n, m, \underline{p}_1, \dots, \underline{p}_m)$, the joint pmf is:

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

NTHU MATH 2810, 2023, Lecture Notes
made by S.-W. Cheng (NTHU, Taiwan)

$$M_{\underline{X}_1, \dots, \underline{X}_m}(\underline{t}_1, \dots, \underline{t}_m)$$

$$= \sum_{\substack{0 \leq x_i \leq n, i=1, \dots, m \\ x_1 + \dots + x_m = n}} e^{\underline{t}_1 x_1 + \dots + \underline{t}_m x_m} \binom{n}{x_1, \dots, x_m} \underline{p}_1^{x_1} \dots \underline{p}_m^{x_m}$$

$$= \sum_{\substack{0 \leq x_i \leq n, i=1, \dots, m \\ x_1 + \dots + x_m = n}} (\underline{p}_1 e^{\underline{t}_1})^{x_1} \dots (\underline{p}_m e^{\underline{t}_m})^{x_m}$$

$$= (\underline{p}_1 e^{\underline{t}_1} + \dots + \underline{p}_m e^{\underline{t}_m})^n.$$

• Some Properties of Joint mgf

➤ $M_{\underline{X}_1}(\underline{t}) = M_{\underline{X}_1, \underline{X}_2, \dots, \underline{X}_n}(\underline{t}, 0, \dots, 0).$

➤ uniqueness theorem

➤ $\underline{X}_1, \dots, \underline{X}_n$ are independent if and only if

$$M_{\underline{X}_1, \dots, \underline{X}_n}(\underline{t}_1, \dots, \underline{t}_n) = M_{\underline{X}_1}(\underline{t}_1) \times \dots \times M_{\underline{X}_n}(\underline{t}_n).$$

➤ $\frac{\partial^{k_1 + \dots + k_n}}{\partial t_1^{k_1} \dots \partial t_n^{k_n}} M_{\underline{X}_1, \dots, \underline{X}_n}(0, \dots, 0) = E_{\underline{X}_1, \dots, \underline{X}_n}(\underline{X}_1^{k_1} \times \dots \times \underline{X}_n^{k_n}).$