# Expectation

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

$$Y=g(X_1, \ldots, X_n), \qquad g:\mathbb{R}^n \to \mathbb{R}^1,$$

is

$$\begin{aligned} E(Y) &\equiv \textstyle\sum_{y\in\mathcal{Y}} y\, p_Y(y) && (1)\\ &= \textstyle\sum_{\mathbf{x}=(x_1,\ldots,x_n)\in\mathcal{X}} g(x_1,\ldots,x_n)\, p_{\mathbf{X}}(x_1,\ldots,x_n) && (2)\\ &\equiv E[g(X_1,\ldots,X_n)] \end{aligned}$$

if $X_1, \ldots, X_n$ are discrete and the sum converges absolutely, or

$$\begin{aligned} E(Y) &\equiv \textstyle\int_{-\infty}^{\infty} y f_Y(y)\, dy && (3)\\ &= \textstyle\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} g(x_1,\ldots,x_n) f_{\mathbf{X}}(x_1,\ldots,x_n)\, dx_1\cdots dx_n && (4)\\ &\equiv E[g(X_1,\ldots,X_n)] \end{aligned}$$

if $Y$ and $X_1, \ldots, X_n$ are continuous and the integrals converges absolutely.





Proof. Like the univariate case.

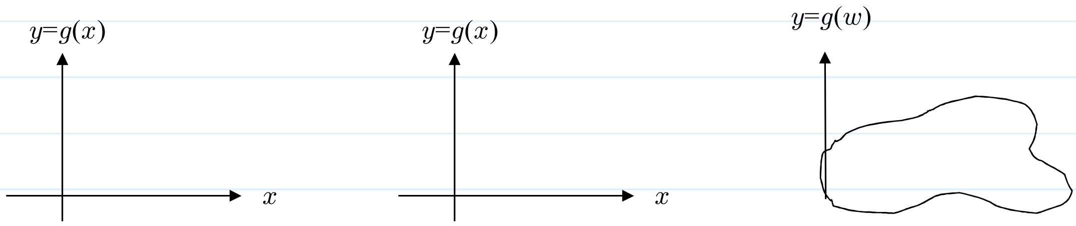
- **Q**: What if $Y$ is discrete and $X_1, \ldots, X_n$ are continuous?
- Notation.
  - Shorthand notation. Combine (1) and (3) by writing

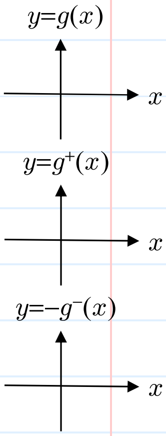
$$E(Y) = \int_{-\infty}^{\infty} y\, dF_Y(y) = \begin{cases} \sum_{y\in\mathcal{Y}} y\, p_Y(y), & \text{for discrete case,}\\ \int_{-\infty}^{\infty} y\, f_Y(y)\, dy, & \text{for continuous case,} \end{cases}$$

and combine (2) and (4) by writing

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

  - 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 $a=x_0<x_1<\cdots<x_n=b$ as $n\to\infty$ and $\max_{i=1,\ldots,n}(x_i - x_{i-1}) \to 0$.

[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.

- $g(X_1, \ldots, X_n)=X_i \Rightarrow E[g(X_1, \ldots, X_n)]=E(X_i) \equiv \mu_{X_i}$.
- $g(X_1, \ldots, X_n)=(X_i-\mu_{X_i})^2 \Rightarrow E[g(X_1, \ldots, X_n)]=Var(X_i) \equiv \sigma^2_{X_i}$.

➢ Example (Average distance between two points). Suppose that

$X, Y$ are i.i.d. ~ Uniform(0, 1).

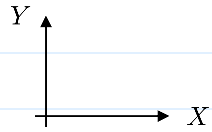
Let $D=|X-Y|$. Find $E(D)$.





- The joint pdf of $(X, Y)$ is

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




- $E(D) = \int_0^1\int_0^1 |x-y|\, dydx = \int_0^1 \left[\int_0^x (x-y)\, dy + \int_x^1 (y-x)\, dy\right] dx$

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

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

$$E(a_0+a_1X_1+\cdots+a_nX_n) = a_0+a_1\, E(X_1)+\cdots+a_n\, E(X_n).$$

Proof. $E(a_0 + a_1X_1 + \cdots + a_nX_n)$

$$\begin{aligned} &= \int_{\mathbb{R}^n} (a_0 + a_1x_1 + \cdots + a_nx_n)\, dF_{\mathbf{X}}(\mathbf{x}) \\ &= \int_{\mathbb{R}^n} a_0\, dF_{\mathbf{X}}(\mathbf{x}) + a_1 \int_{\mathbb{R}^n} x_1\, dF_{\mathbf{X}}(\mathbf{x}) \\ &\quad + \cdots + a_n \int_{\mathbb{R}^n} x_n\, dF_{\mathbf{X}}(\mathbf{x}) \\ &= a_0 + a_1E(X_1) + \cdots + a_nE(X_n). \end{aligned}$$

➢Corollary. Suppose that $\mu=E(X_1)=\cdots=E(X_n)$. Let

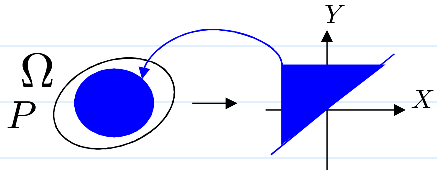
$$\overline{X}_n = \frac{X_1+\cdots+X_n}{n},$$

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

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

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

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




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

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

then $E(Z) = \int_0^\infty z\, dF_Z(z) \ge 0.$

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

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

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





- 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

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

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

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

Proof. We only prove it for the continuous case:

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

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

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

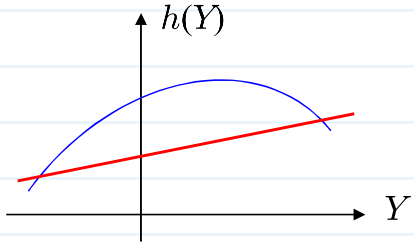
Proof. Let $g(X)=X$ and $h(Y)=Y$.

➢ **Q**: For independent r.v.'s $X$ and $Y$,

$E(X/Y)=E(X)/E(Y)$?

➢ Note. $E[h(Y)] \neq h(E(Y))$ in general, e.g.,

$E(1/Y) \neq 1/E(Y)$.




• Covariance and Correlation between 2 random variables

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

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

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

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





2. The *correlation* (*coefficient*) between $X$ and $Y$ is defined as

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

and denoted by $\rho_{XY}$.

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

▪ A special case of covariance:

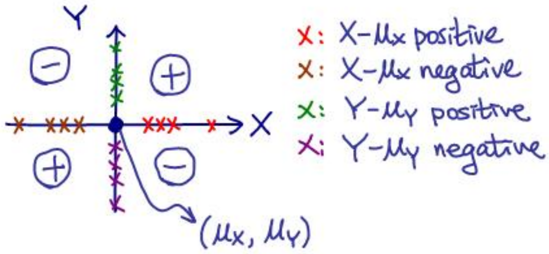
$Cov(X, X) = Var(X)$.

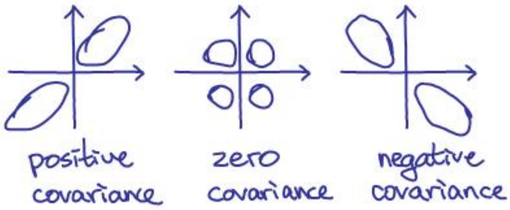
➢ Intuitive explanation of covariance and correlation

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

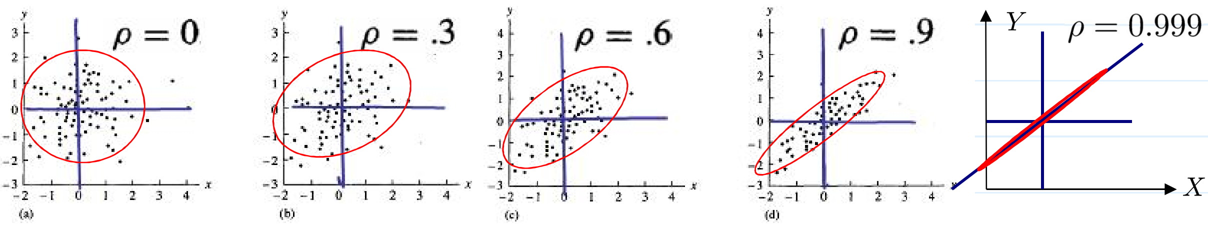
▪ Covariance is a measure of the *joint* variability of $X$ and $Y$, or their degree of association.

▪ Positive Covariance and Negative Covariance




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

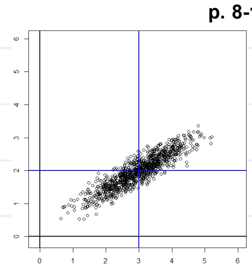



➢ Theorem. $Cov(X, Y) = E(XY) - \mu_X \mu_Y$.

Proof.
$$\begin{aligned} Cov(X,Y) &= E[(X-\mu_X)(Y-\mu_Y)] \\ &= E(XY - \mu_X Y - \mu_Y X + \mu_X \mu_Y) \\ &= E(XY) - \mu_X E(Y) - \mu_Y E(X) + \mu_X \mu_Y \\ &= E(XY) - \mu_X \mu_Y - \mu_Y \mu_X + \mu_X \mu_Y. \end{aligned}$$



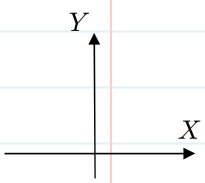




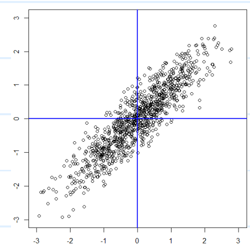
- Corollary. If $X$ and $Y$ are independent, then $Cov(X, Y)=0$, i.e., $X$ and $Y$ are uncorrelated.

  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 $X$~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 $(X_1, \ldots, X_m) \sim$ Multinomial$(n, m, p_1, \ldots, p_m)$, then

$$Cov(X_i, X_j) = -np_ip_j, \quad \text{for } 1 \le i \ne j \le m.$$

- Because $(X_1, X_2, X_3+\cdots+X_m) \sim$ Multinomial$(n, 3, p_1, p_2, p_3+\cdots+p_m)$, and

$$X_3+\cdots+X_m = n-X_1-X_2, \qquad p_3+\cdots+p_m = 1-p_1-p_2,$$

we have

$$\begin{aligned} E(X_1X_2) &= \sum x_1x_2 \binom{n}{x_1, x_2, n-x_1-x_2} p_1^{x_1} p_2^{x_2} (1-p_1-p_2)^{n-x_1-x_2} \\ &= \sum x_1x_2 \tfrac{n!}{x_1!x_2!(n-x_1-x_2)!} p_1^{x_1} p_2^{x_2} (1-p_1-p_2)^{n-x_1-x_2} \\ &= n(n-1)p_1p_2 \Big[\sum \tfrac{(n-2)!}{(x_1-1)!(x_2-1)!(n-x_1-x_2)!} \\ &\qquad \times (p_1)^{x_1-1} (p_2)^{x_2-1} (1-p_1-p_2)^{n-x_1-x_2}\Big] \\ &= n(n-1)p_1p_2. \end{aligned}$$

- WLOG, we can get $E(X_iX_j) = n(n-1)p_ip_j$, for $i \ne j$. Therefore,

$$\begin{aligned} Cov(X_i, X_j) &= E(X_iX_j) - E(X_i)E(X_j) \\ &= n(n-1)p_ip_j - (np_i)(np_j) = -np_ip_j. \end{aligned}$$





- And, for $i \ne j$,

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

- Cov & Cor for Sums of Random Variables

➢ Notation. In the following, let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ be r.v.'s and $-\infty < a_0, a_1, \ldots, a_n,$ $b_0, b_1, \ldots, b_m < \infty$ are constants.

➢ Recall. $E(a_0+a_1X_1+\cdots+a_nX_n) = a_0+a_1E(X_1)+\cdots+a_nE(X_n)$.

➢ Theorem (covariance of two sums).

$$\begin{aligned} &Cov(a_0 + a_1X_1 + \cdots + a_nX_n, b_0 + b_1Y_1 + \cdots + b_mY_m) \\ &= \sum_{i=1}^n \sum_{j=1}^m a_ib_jCov(X_i, Y_j). \end{aligned}$$

Proof. Let $S = a_0+a_1X_1+\cdots+a_nX_n$, and $T = b_0+b_1Y_1+\cdots+b_mY_m$, then

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

Therefore, $Cov(S,T) = E\{[S-E(S)][T-E(T)]\}$

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

$$= \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j Cov(X_i, Y_j).$$

➢ Theorem (variance of sum).

$$Var(a_0 + a_1X_1 + \cdots + a_nX_n)$$
$$= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j Cov(X_i, X_j)$$
$$= \sum_{i=1}^{n} a_i^2 Var(X_i)$$
$$+2\sum_{1\le i<j\le n} a_i a_j Cov(X_i, X_j).$$

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

- Corollary. If $X_1, \ldots, X_n$ are uncorrelated, then

$$Var(a_0 + a_1X_1 + \cdots + a_nX_n) = \sum_{i=1}^{n} a_i^2 Var(X_i).$$

- Corollary. If $X_1, \ldots, X_n$ are uncorrelated and

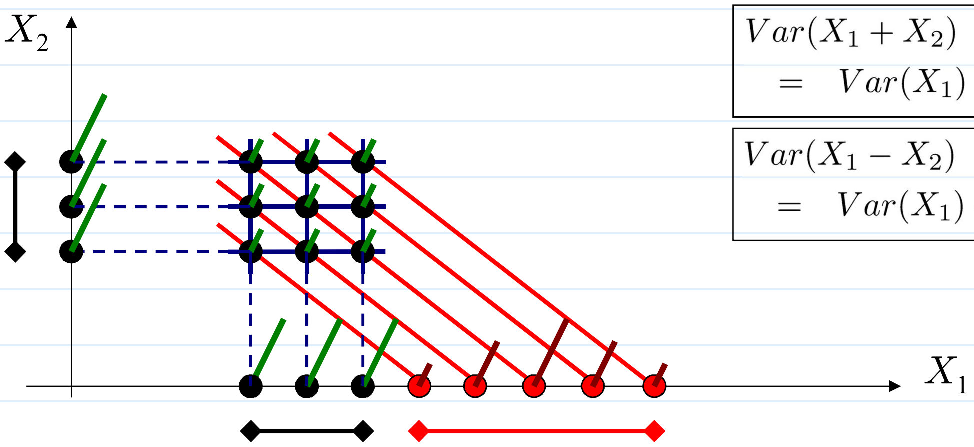
$$Var(X_1)=\cdots = Var(X_n) \equiv \sigma^2 < \infty,$$

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





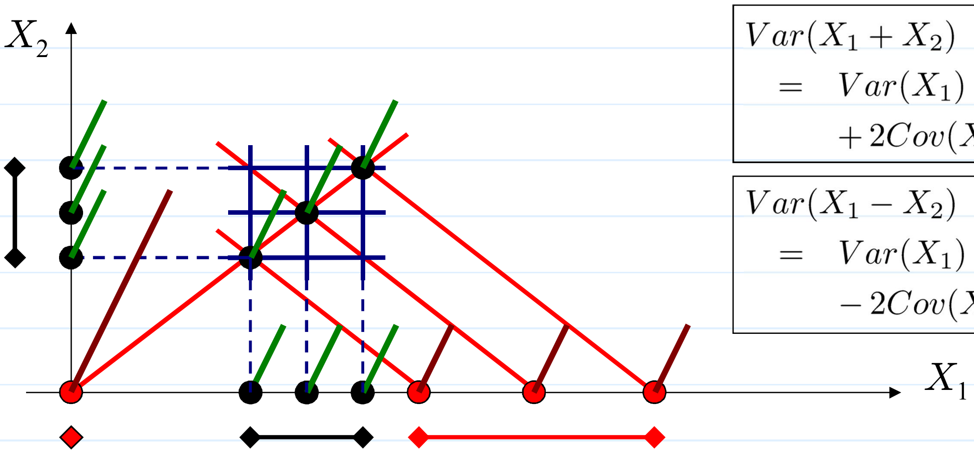
$Cov(X_1, X_2) = 0$




$Var(X_1 + X_2) = Var(X_1) + Var(X_2)$

$Var(X_1 - X_2) = Var(X_1) + Var(X_2)$

$Cov(X_1, X_2) > 0$




$Var(X_1 + X_2) = Var(X_1) + Var(X_2) + 2Cov(X_1, X_2)$

$Var(X_1 - X_2) = Var(X_1) + Var(X_2) - 2Cov(X_1, X_2)$

- Corollary. Suppose that $X_1, \ldots, X_n$ are uncorrelated and have same mean $\mu$ and variance $\sigma^2$. Let

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

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

Proof. $(n-1)S^2 = \sum_{i=1}^{n}(X_i - \overline{X}_n)^2$

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

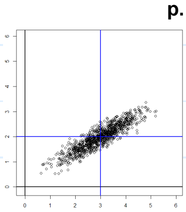
Therefore,

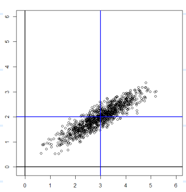
$$\begin{aligned}
(n-1)E(S^2) &= \left\{\sum_{i=1}^{n} E\left[(X_i-\mu)^2\right]\right\} - nE\left[(\overline{X}_n-\mu)^2\right] \\
&= n\sigma^2 - nVar(\overline{X}_n) = (n-1)\sigma^2.
\end{aligned}$$

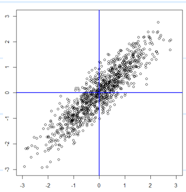




- Note. The previous three corollaries also hold if $X_1, \ldots, X_n$ are "uncorrelated" is replaced by "independent."







➢ Theorem ($\rho$ of linear transformation).

$$Cor(a_0+a_1X,\ b_0+b_1Y) = \text{sign}(a_1b_1)\times Cor(X, Y),$$

and

$$|Cor(a_0+a_1X,\ b_0+b_1Y)| = |Cor(X, Y)|,$$

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

Proof. Let $S=a_0+a_1X$ and $T=b_0+b_1Y$, then

$$Cov(S, T) = Cov(a_0+a_1X,\ b_0+b_1Y) = a_1b_1Cov(X, Y),$$

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

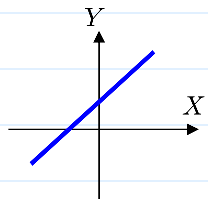
Therefore,

$$\rho_{ST} = \frac{Cov(S,T)}{\sigma_S\sigma_T} = \frac{a_1b_1Cov(X,Y)}{|a_1||b_1|\sigma_X\sigma_Y} = \frac{a_1b_1}{|a_1b_1|}\rho_{XY}.$$

- Theorem (some properties of $\rho$).

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

  (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).

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

  Similarly,

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

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





$$
\begin{aligned}
\rho_{XY} = 1 &\Leftrightarrow Var\left(\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y}\right) = 0, \\
&\Leftrightarrow P\left(\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y} = c\right) = 1,
\end{aligned}
$$

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 $X_1, \ldots, X_n$?
  - $g(X_1, \ldots, X_n) = X_i \Rightarrow E[g(\mathbf{X})] = \mu_{X_i}$: *mean* of $X_i$.
  - $g(X_1, \ldots, X_n) = (X_i - \mu_{X_i})^2 \Rightarrow E[g(\mathbf{X})] = \sigma^2_{X_i}$: *variance* of $X_i$.
  - $g(X_1, \ldots, X_n) = (X_i - \mu_{X_i})(X_j - \mu_{X_j})$ for $i \ne j$ $\Rightarrow E[g(\mathbf{X})] = \sigma_{X_i X_j}$: *covariance* of $X_i$ and $X_j$.
  - $g(X_1, \ldots, X_n) = [(X_i - \mu_{X_i})/\sigma_{X_i}][(X_j - \mu_{X_j})/\sigma_{X_j}]$ for $i \ne j$ $\Rightarrow E[g(\mathbf{X})] = \rho_{X_i X_j}$: *correlation coefficient* of $X_i$ and $X_j$.
  - Notes. $\mu_{X_i}, \sigma^2_{X_i}, \sigma_{X_i X_j}, \rho_{X_i X_j}$ are constants, not random

❖ **Reading**: textbook, Sec 7.1, 7.2, 7.4, 7.9

# Conditional Expectation

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

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

in the discrete case, or,

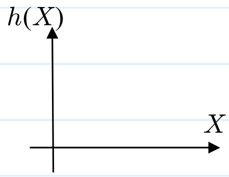
$$E_{\mathbf{Y}|\mathbf{X}}\left(h(\mathbf{Y})\middle| \mathbf{X}=\mathbf{x}\right) = \int_{\mathbb{R}^m} h(\mathbf{y})\ f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})\ d\mathbf{y},$$

in the continuous case,

provided that the sum or integral converges absolutely.

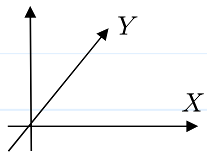
➢ Some Notes.

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




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

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








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

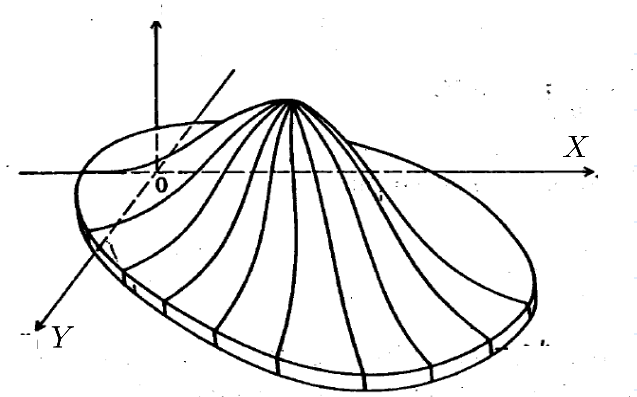
$$E_{\mathbf{Y}|\mathbf{X}}(h(\mathbf{Y})|\mathbf{X})$$

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

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

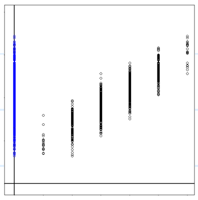
➢ $E_{Y|X}(Y|x^*)$: mean of $f_{Y|X}(y|x^*)$.

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

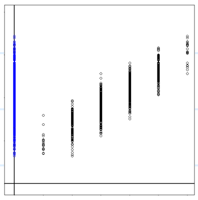
- Example. Sample a student from an elementary school. Let

$X$=age (unit: year), $Y$=height (unit: cm)

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

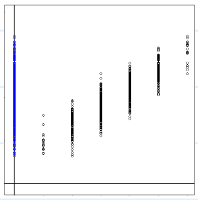


- $Y|X=x$: a random variable (unit: cm) that represents the height distribution of students with age=$x$.



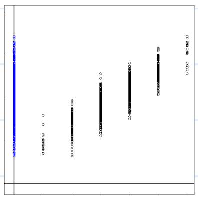
- $g(x)=E_{Y|X}(Y|X=x)$ or $E_{Y|X}(Y|x)$: a *function* maps from age (unit: year) to average height (unit: cm) of students with age=$x$.

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



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

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



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

  $Var_Y(Y)$: variation of height of ***all*** students.





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

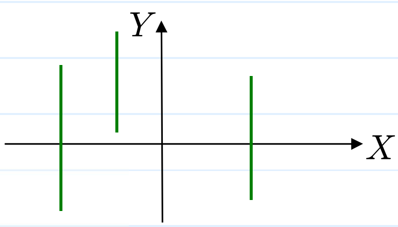
$$E_{\mathbf{X}}\{E_{\mathbf{Y}|\mathbf{X}}[h(\mathbf{Y})|\mathbf{X}]\}=E_{\mathbf{Y}}[h(\mathbf{Y})].$$

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

$$E_{\mathbf{X}}[E_{\mathbf{Y}|\mathbf{X}}(Y_i|\mathbf{X})]=E_{\mathbf{Y}}(Y_i).$$

Proof.

(only prove it for the continuous case)




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

➢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 $Y|X=x \sim$ hypergeometric$(n-x, N-R, W)$,

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

Because $X \sim$ hypergeometric$(n, N, R) \Rightarrow E_X(X) = n(R/N)$, and then

$$\begin{aligned} E_Y(Y) &= E_X[E_{Y|X}(Y|X)] = E_X[g(X)] \\ &= E_X\left[(n-X)\frac{W}{N-R}\right] = \frac{W}{N-R}[n - E_X(X)] \\ &= \frac{W}{N-R}\left(n - n\frac{R}{N}\right) = n\frac{W}{N}. \end{aligned}$$

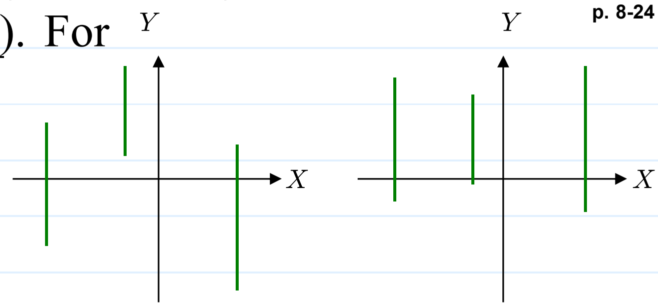
Note that $Y \sim$ hypergeometric$(n, N, W) \Rightarrow E_Y(Y) = n(W/N)$.





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

$$\begin{aligned} &Var_{\mathbf{Y}}(Y_i) \\ &\quad = Var_{\mathbf{X}}[E_{\mathbf{Y}|\mathbf{X}}(Y_i|\mathbf{X})] \\ &\quad + E_{\mathbf{X}}[Var_{\mathbf{Y}|\mathbf{X}}(Y_i|\mathbf{X})]. \end{aligned}$$




Proof. $$Var_{\mathbf{Y}|\mathbf{X}}(Y_i|\mathbf{x}) = E_{\mathbf{Y}|\mathbf{X}}(Y_i^2|\mathbf{x}) - [E_{\mathbf{Y}|\mathbf{X}}(Y_i|\mathbf{x})]^2,$$

and, $$\begin{aligned} &E_{\mathbf{X}}[Var_{\mathbf{Y}|\mathbf{X}}(Y_i|\mathbf{X})] \\ &= E_{\mathbf{X}}[E_{\mathbf{Y}|\mathbf{X}}(Y_i^2|\mathbf{X})] - E_{\mathbf{X}}\{[E_{\mathbf{Y}|\mathbf{X}}(Y_i|\mathbf{X})]^2\}. \end{aligned}$$

Also, $$\begin{aligned} &Var_{\mathbf{X}}[E_{\mathbf{Y}|\mathbf{X}}(Y_i|\mathbf{X})] \\ &= E_{\mathbf{X}}\{[E_{\mathbf{Y}|\mathbf{X}}(Y_i|\mathbf{X})]^2\} - \{E_{\mathbf{X}}[E_{\mathbf{Y}|\mathbf{X}}(Y_i|\mathbf{X})]\}^2. \end{aligned}$$

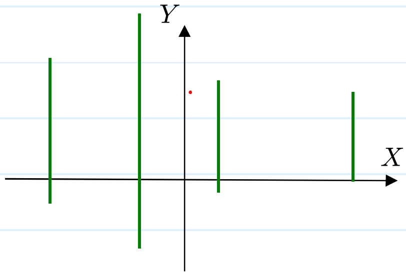
Now, $$\begin{aligned} Var_{\mathbf{Y}}(Y_i) &= E_{\mathbf{Y}}(Y_i^2) - [E_{\mathbf{Y}}(Y_i)]^2 \\ &= E_{\mathbf{X}}[E_{\mathbf{Y}|\mathbf{X}}(Y_i^2|\mathbf{X})] - \{E_{\mathbf{X}}[E_{\mathbf{Y}|\mathbf{X}}(Y_i|\mathbf{X})]\}^2 \\ &= E_{\mathbf{X}}[E_{\mathbf{Y}|\mathbf{X}}(Y_i^2|\mathbf{X})] - E_{\mathbf{X}}\{[E_{\mathbf{Y}|\mathbf{X}}(Y_i|\mathbf{X})]^2\} \\ &\quad + E_{\mathbf{X}}\{[E_{\mathbf{Y}|\mathbf{X}}(Y_i|\mathbf{X})]^2\} - \{E_{\mathbf{X}}[E_{\mathbf{Y}|\mathbf{X}}(Y_i|\mathbf{X})]\}^2 \\ &= E_{\mathbf{X}}[Var_{\mathbf{Y}|\mathbf{X}}(Y_i|\mathbf{X})] + Var_{\mathbf{X}}[E_{\mathbf{Y}|\mathbf{X}}(Y_i|\mathbf{X})]. \end{aligned}$$

➢ Corollary.

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

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

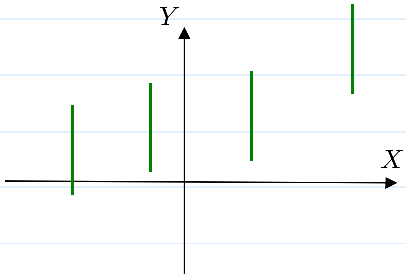
with probability one.




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

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

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?)





➢ Statistical modeling: assigning $(X, Y)$ a (known) joint distribution (cdf $F(x, y)$, pdf $f(x, y)$, or pmf $p(x, y)$)

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

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

➢ Predictor: considering the following three groups of $g$'s

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

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

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

Note. $G_1 \subset G_2 \subset G_3$.

➢ Question: Within each group, what is the "best" predictor?

➢ Criterion: minimizing mean square error

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

- Theorem (best *constant* predictor under MSE).

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

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

Proof.

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

- Theorem (best predictor under MSE).

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

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

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

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





where the last "=" comes from

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

Furthermore,

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

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

$$\begin{aligned}
&= Var_{X,Y}(Y-a-bX)+[E_{X,Y}(Y-a-bX)]^2\\
&= Var_{X,Y}(Y-bX)+(\mu_Y-a-b\mu_X)^2\\
&\geq Var_{X,Y}(Y-bX) \qquad (\Rightarrow \text{setting } a=\mu_Y-b\mu_X)\\
&= \sigma_Y^2+b^2\sigma_X^2-2b\sigma_{XY}\\
&= \sigma_X^2\left(b^2-2b\frac{\sigma_{XY}}{\sigma_X^2}+\frac{\sigma_{XY}^2}{\sigma_X^4}\right)+\sigma_Y^2-\frac{\sigma_{XY}^2}{\sigma_X^2}\\
&= \sigma_X^2\left(b-\frac{\sigma_{XY}}{\sigma_X^2}\right)^2+\sigma_Y^2(1-\rho_{XY}^2)\\
&\geq \sigma_Y^2(1-\rho_{XY}^2) \qquad \left(\Rightarrow \text{setting } b=\frac{\sigma_{XY}}{\sigma_X^2}=\frac{\sigma_{XY}}{\sigma_X\sigma_Y}\times\frac{\sigma_Y}{\sigma_X}=\rho_{XY}\frac{\sigma_Y}{\sigma_X}\right)
\end{aligned}$$





➢ 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 $\rho_{XY}$ is close to +1 or −1, and large if $\rho_{XY}$ 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

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

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

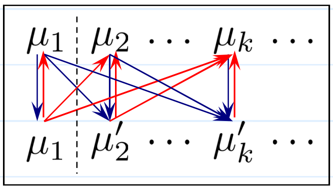
$$\mu'_k \equiv E[(X - \mu_X)^k] = \int_{-\infty}^{\infty} (x - \mu_X)^k \, dF_X(x), \quad k = 2, 3, \ldots,$$

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

➢ Some notes.

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

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

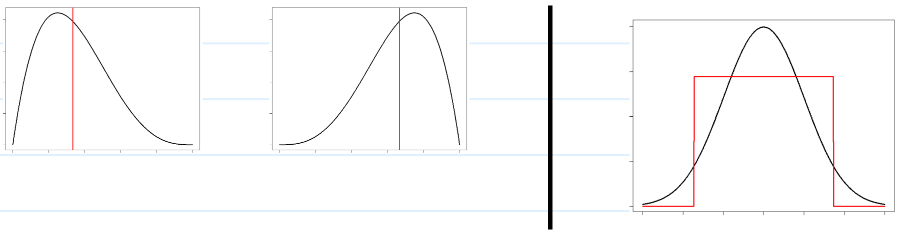


- In particular,
$$\begin{aligned} E(X) &= \mu_X = \mu_1, \quad \text{and}, \\ Var(X) &= \sigma_X^2 = \mu'_2 = \mu_2 - \mu_1^2. \end{aligned}$$





- 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): $\mu'_3/\sigma^3$.
  - ▫ Kurtosis (a measure of the "heavy tails"): $\mu'_4/\sigma^4$.



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

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

therefore,
$$\begin{aligned} \mu_X &= \mu_1 = 1/2, \quad \text{and}, \\ \sigma_X^2 &= \mu_2 - \mu_1^2 = 1/3 - (1/2)^2 = 1/12. \end{aligned}$$

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

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

- Recall. How to characterize a distribution?

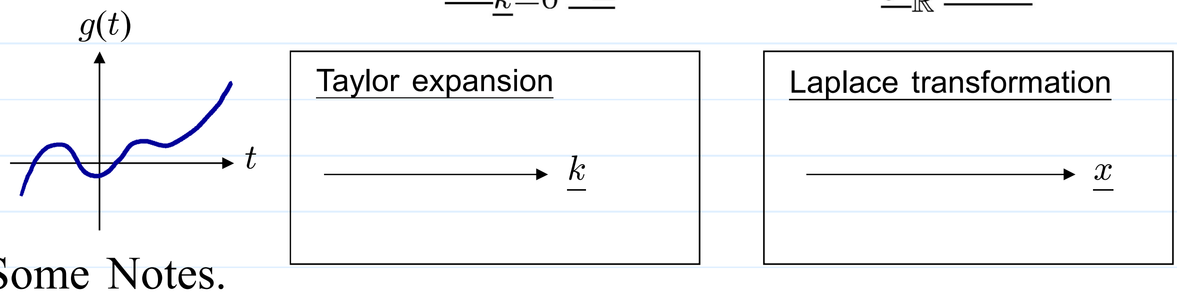
  (1) pdf/pmf, (2) cdf, (3) mgf

- 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 $t$.

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




➢ 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





➢ Example.

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

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

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

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

$$= \quad e^{-\lambda}\left(e^{\lambda e^t}\right) \sum_{x=0}^{\infty} \frac{e^{-(\lambda e^t)}(\lambda e^t)^x}{x!} = e^{-\lambda} e^{\lambda e^t} = e^{\lambda(e^t - 1)}.$$

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

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

$$= \quad \lambda\left(\frac{1}{\lambda - t}\right) \int_0^{\infty} (\lambda - t)\, e^{-(\lambda - t)x}\, dx = \frac{\lambda}{\lambda - t},$$

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

- A list of some mgfs (exercise)
  - If $X \sim$ binomial($n$, $p$),

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

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

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

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

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

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

• 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.

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 $X_1, \ldots, X_n$

➢ 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

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

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

$$M_X(0)=1,$$

and,

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

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

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

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

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

$$\cdots = \cdots$$

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

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

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

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

Because
$$M_X^{(k)}(t) = \frac{k!\,\lambda}{(\lambda-t)^{k+1}},$$
we get
$$\mu_k = M_X^{(k)}(0) = \frac{k!}{\lambda^k}.$$





- Theorem (MGF for linear transformation). For constants $a$ and $b$,
$$M_{a+bX}(t) = e^{at} M_X(bt).$$

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

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

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

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

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

➢Example. If $X_1, \ldots, X_n$ are i.i.d. $\sim$ geometric($p$), then
$S=X_1+\cdots+X_n \sim$ negative binomial($n$, $p$).

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

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

➢ 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 $S = a_0 + a_1 X_1 + \cdots + a_n X_n$, then

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

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

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

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

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

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:

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





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

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

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

$$= \left(p_1 e^{t_1} + \cdots + p_m e^{t_m}\right)^n.$$

- Some Properties of Joint mgf
  - ➢ $M_{X_1}(t) = M_{X_1, X_2, \ldots, X_n}(t, 0, \ldots, 0)$.
  - ➢ uniqueness theorem
  - ➢ $X_1, \ldots, X_n$ are independent if and only if

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

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

❖ **Reading**: textbook, Sec 7.7