Variance

X random variable, suppose $E[X^2] < \infty$.

Define the <u>variance</u> of X to be

$$\mu = E[X]$$

$$\begin{aligned}
\mathrm{var}(X) &\overset{\text{def}}{=} \boxed{E[X - EX]^2} \\
&= E[X^2 - 2\mu X + \mu^2] \\
&= E[X^2] - 2\mu EX + \mu^2 \\
&= E[X^2] - \mu^2 \\
&= \boxed{E[X^2] - E[X]^2}
\end{aligned}$$

<u>Ex:</u> $X \sim \mathrm{Unif}(a,b)$. Find $\mathrm{var}(X)$.

$$EX = \frac{a+b}{2}. \qquad f(x) = \frac{1}{b-a}$$

$$E[X^2] = \int_a^b x^2 f(x)\,dx = \frac{1}{b-a}\int_a^b x^2\,dx$$

$$= \frac{b^3 - a^3}{3(b-a)} = \frac{b^2 + ab + a^2}{3}$$

So $\mathrm{var}\,X = \dfrac{a^2 + ab + b^2}{3} - \left(\dfrac{a+b}{2}\right)^2 = \boxed{\dfrac{(a-b)^2}{12}}$

Some properties of variance:

Prop.: $\mathrm{var}(aX + b) = a^2\,\mathrm{var}(X)$

pf) $E[aX+b] = a\,EX + b$. Let $\mu = EX$.

$$\text{var}(aX+b) = E[aX+b-(a\mu+b)]^2$$
$$= E[\underbrace{a(X-\mu)}]^2$$
$$= a^2 \underbrace{E[X-\mu]^2}$$
$$= a^2\,\text{var}(X).$$

In particular, $\text{var}(b) = 0$. (Take $a=0$).

Conversely,

Prop: If $\text{var}\,X = 0$, then $X = EX$ with probability one, ie, $P(X \neq EX) = 0$.

pf) Let $c > 0$. Suffices to show

$$P(|X-EX| > c) = 0.$$

Why? Because $P(|X-EX| \leq 0) = 1$.

Since $|X-EX| \geq 0$, so $P(\underbrace{|X-EX| = 0}_{X=EX}) = 1$.

Let $\mathbb{1}_{\{|X-EX|>c\}}(x) = \begin{cases} 1 & \text{if } |x-EX| > c \\ 0 & \text{otherwise.} \end{cases}$

Then $|X-EX|^2 \geq c^2\,\mathbb{1}_{\{|X-EX|>c\}}$

$\{|X - EX| > c\}.$

because

$$|x - EX|^2 \begin{cases} > c^2 & \text{if } |x - EX| > c \\ \geq 0 & \text{if } |x - EX| \leq c. \end{cases}$$

while

$$1_{\{|X - EX| > c\}}(x) = \begin{cases} 1 & \text{if } |x - EX| > c \\ 0 & \text{if } |x - EX| \leq c \end{cases}$$

$$0 = E|X - EX|^2 \geq c^2 \underbrace{E\left[1_{\{|X - EX| > c\}}\right]}$$

discrete random variable

$$E\left[1_{\{|X - EX| > c\}}\right] = 1 \cdot P(|X - EX| > c)$$
$$+ 0 \cdot P(|X - EX| \leq c)$$
$$= P(|X - EX| > c).$$

So

$$0 \geq c^2 P(|X - EX| > c)$$

$$\Rightarrow P(|X - EX| > c) = 0.$$

Useful Fact: Suppose X has mean μ and variance σ^2 (where $\underline{\sigma = \sqrt{\operatorname{var} X}}$).

let $Y = \frac{X - \mu}{\sigma} = \frac{1}{\sigma}(X - \mu) = \frac{1}{\sigma}X - \frac{\mu}{\sigma}$

Then $EY = \frac{1}{\sigma}(EX - \mu) = 0$. standard deviation of X

and $\operatorname{var} Y = \frac{1}{\sigma^2} \operatorname{var} X = \frac{\sigma^2}{\sigma^2} = 1$.

Ex: Normal distribution. Suppose X has density function $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$.

$$EX = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} x e^{-x^2/2} dx = 0.$$

$$\operatorname{var} X = E[X^2] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-x^2/2} dx$$

Integration by parts.

$$u = \frac{1}{\sqrt{2\pi}} x, \quad dv = x e^{-x^2/2} dx$$

$$= \underbrace{-\frac{1}{\sqrt{2\pi}} x e^{-x^2/2} \Big|_{-\infty}^{\infty}}_{0} + \underbrace{\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2} dx}_{1}$$

$$= 1$$

$X \sim N(0,1)$

So if we define $Y = \sigma X + \mu$, then

$EY = \sigma EX + 0 + \mu = \mu$

and $\text{var}\, Y = \sigma^2 \text{var}\, X = \sigma^2$.

Covariance

The covariance of X and Y is

$$\boxed{\text{cov}(X,Y) = E[\underbrace{(X-EX)}_{u}\underbrace{(Y-EY)}_{v}]}$$

Note: $\text{var}(X) = E(X-EX)^2 = \text{cov}(X,X)$.

Can also write

$$\text{cov}(X,Y) = E[XY - X(EY) - (EX)Y + (EX)(EY)]$$

$$= E[XY] - EX \cdot EY - EX\,EY + EX\,EY$$

$$\boxed{\text{cov}(X,Y) = E[XY] - EX\,EY}$$

Thm: Spse $X_1, \dots, X_n$ rv's.

$$\text{var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \text{var}\, X_i + 2 \sum_{1 \le i < j \le n} \text{cov}(X_i, X_j)$$

pf) Let $\mu_i = EX_i$.

Then $E(\sum X_i) = \sum \mu_i$ $\quad EY = \sum \mu_i$

So $\text{var}(\underbrace{\sum X_i}_{Y}) = E(Y - EY)^2$

$= E(Y - \sum \mu_i)^2$

$$= E\left(\sum X_i - \sum \mu_i\right)^2$$

$$= E\left(\sum (X_i - \mu_i)\right)^2$$

Note:

$$\left(\sum_{i=1}^{n} a_i\right)^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j = \sum_{i=1}^{n} a_i^2 + 2\sum_{1\le i<j\le n} a_i a_j$$

Ex: $(a_1 + a_2 + a_3)^2$

$$\begin{aligned} = & \ a_1 a_1 + \boxed{a_1 a_2} + \boxed{a_1 a_3} && i=1 \\ & + a_2 a_1 + a_2 a_2 + \boxed{a_2 a_3} && i=2 \\ & + a_3 a_1 + a_3 a_2 + a_3 a_3 && i=3 \end{aligned}$$

$$= \sum_{i=1}^{3}\sum_{j=1}^{3} a_i a_j$$

$$\sum_{i=1}^{3} (a_i a_1 + a_i a_2 + a_i a_3)$$

$a_i = X_i - \mu_i$

$$\left[\sum_{i=1}^{n} (X_i - \mu_i)\right]^2 =$$

$$\sum_{i=1}^{n} (X_i - \mu_i)^2 + 2\sum_{1\le i<j\le n} (X_i - \mu_i)(X_j - \mu_j)$$

$$E\left[\sum_{i=1}^{n} (X_i - \mu_i)\right]^2$$

$$= \sum_{i=1} E(X_i - \mu_i)^2 = \sum \mathrm{var}(X_i)$$

$$+ 2 \sum_{i<j} \underbrace{E[(X_i - \mu_i)(X_j - \mu_j)]}_{\mathrm{cov}(X_i, X_j)}$$

$$= \sum_{i=1}^{n} \mathrm{var}\, X_i + 2 \sum_{1 \le i < j \le n} \mathrm{cov}(X_i, X_j)$$

We say that rv's $X_1, \dots, X_n$ are <u>uncorrelated</u> if, whenever $i \neq j$,

$$\mathrm{cov}(X_i, X_j) = 0.$$

In this case,

$$\mathrm{var}(X_1 + \dots + X_n) = \mathrm{var}(X_1) + \dots + \mathrm{var}(X_n)$$

<u>Note</u>: Independent rv's are uncorrelated.

$$\mathrm{cov}(X,Y) = E[XY] - [EX][EY]$$

$$0 \;\; \Rightarrow E[XY] = (EX)(EY)$$

$$E X_i X_j = E X_i \, E X_j$$

Independent $\Rightarrow E X_i X_j = E X_i \, E X_j$

$$\Rightarrow \mathrm{cov}(X_i, X_j) = 0.$$

Ex: X rv $\quad 0 < \mathrm{var}\, X < \infty$

$$X_1 = X, \quad X_2 = -X.$$

$$\text{var}(X_1 + X_2) = 0$$

but $\text{var}\, X_1 + \text{var}\, X_2 = 2\,\text{var}\, X$

Ex: $X_1 = X, \quad X_2 = X$

$$\text{var}(X_1 + X_2) = \text{var}(2X) = 4\,\text{var}\, X$$

$$\text{var}(X_1) + \text{var}\, X_2 = 2\,\text{var}\, X.$$

Ex: Bernoulli dist.

$$X = \begin{cases} 1 & \text{w/ prob } p \\ 0 & \text{w/ prob } 1-p. \end{cases}$$

$$EX = p(1) + (1-p)\,0 = p$$

$$EX^2 = p.$$

$$X^2 = \begin{cases} 1 & \text{if } X = 1 \\ 0 & \text{if } X = 0 \end{cases}$$

$$\text{var}\, X = EX^2 - (EX)^2$$

$$= p - p^2 = p(1-p).$$

Ex: Binomial dist. Sequence of indep trials w/ success prob p.

$$X_i = \begin{cases} 1 & \text{if } i\text{th trial is success} \\ 0 & \text{otherwise.} \end{cases}$$

$$S_n = X_1 + \dots + X_n$$

$\sim Bin(n, p).$

$$\mathrm{var}(S_n) = \sum \mathrm{var}(X_i)$$
$$= np(1-p).$$

Meaning of sign of covariance:

If $\mathrm{cov}(X_1, X_2) > 0$, then

if X_1 is large, then X_2 is more likely to be large.

Ex: $X_1 = $ ht, $X_2 = $ wt

$X_1 = $ score on 1st exam $\quad X_2 = $ score on 2nd exam

If $\mathrm{cov}(X_1, X_2) < 0$, then if X_1 large, then X_2 is less likely to be large.

$X_1 = $ # years of school completed

$X_2 = $ # dollars earned per year.

Correlation

X, Y rv's, sd $\sigma(X) > 0$, $\sigma(Y) > 0$

$$\mathrm{cor}(X,Y) = \underline{\mathrm{cov}(X,Y)} \quad \mathrm{cov}(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]$$

$\sigma(X)\,\sigma(Y)$

$$= E\Big[\underbrace{\frac{(X-\mu_X)}{\sigma(X)}}_{U} \cdot \underbrace{\frac{(Y-\mu_Y)}{\sigma_Y}}_{V}\Big]$$

mean 0, var 1.

<u>Prop:</u> Define $Y_1 = a_1 X_1 + b_1$,

$Y_2 = a_2 X_2 + b_2$.

$$\mathrm{cor}(Y_1, Y_2) = \frac{a_1}{|a_1|}\frac{a_2}{|a_2|}\,\mathrm{cor}(X_1, X_2)$$

± 1

Pf) $EY_i = a_i EX_i + b_i$

$$Y_i - EY_i = a_i X_i + b_i - (a_i EX_i + b_i)$$

$$= a_i X_i - a_i EX_i$$

$$= a_i (X_i - EX_i).$$

So $E[(Y_1 - EY_1)(Y_2 - EY_2)]$

$$= E[a_1(X_1 - EX_1)\cdot a_2(X_2 - EX_2)]$$

$$= a_1 a_2\, E[(X_1 - EX_1)(X_2 - EX_2)]$$

Recall: $\mathrm{var}(Y_i) = a_i^2\,\mathrm{var}(X_i)$

$$\Rightarrow \quad \sigma(Y_i) = |a_i|\,\sigma(X_i).$$

$$\Rightarrow \quad \sigma(Y_1)\,\sigma(Y_2) = |a_1|\,|a_2|\,\sigma(X_1)\,\sigma(X_2)$$

So $\quad \operatorname{corr}(Y_1, Y_2) = \dfrac{a_1}{|a_1|} \cdot \dfrac{a_2}{|a_2|} \operatorname{corr}(X_1, X_2).$

Ex: $-\operatorname{corr}(X, Y) = \operatorname{corr}(X, -Y)$

Fact: $\quad -1 \le \operatorname{corr}(X, Y) \le 1$

We'll prove this later...

Linear Prediction

X_1, X_2

Want: a, b to minimize $E\Big[\underbrace{(X_2 - (aX_1 + b))^2}_{\text{mean square error}}\Big]$

$$= E\left[X_2^2 - 2X_2(aX_1 + b) + (aX_1 + b)^2\right]$$

$$= E[X_2^2] - 2a\,E[X_1 X_2] - 2b\,E[X_2]$$

$f(a,b)$

$$+ a^2 E[X_1^2] + 2ab\,E[X_1] + b^2$$

Take partial derivatives wrt a and b.

(*) $\quad \dfrac{\partial f}{\partial a} = -2E[X_1 X_2] + 2a\,E[X_1^2] + 2b\,[EX_1] = 0$

$$(**)\quad \frac{\partial F}{\partial b} = -2E[X_2] + 2aE[X_1] + 2b = 0$$

Multiply $(**)$ by $-EX_1$ and add to $(*)$, get

$$-2E[X_1X_2] + 2E[X_1]E[X_2] + 2aE[X_1^2] - 2a(E[X_1])^2 = 0$$

$b = EX_2 - aEX_1$

$= \mu_2 - a\mu_1$

$$-E[X_1X_2] + E[X_1]E[X_2] + aE[X_1^2] - a(E[X_1])^2 = 0$$

$$a(E[X_1])^2 - aE[X_1^2] = E[X_1]E[X_2] - E[X_1X_2]$$

$$a = \frac{EX_1X_2 - EX_1EX_2}{EX_1^2 - [E(X_1)]^2} = \frac{cov(X_1, X_2)}{var\, X_1}$$

$$= \frac{corr(X_1, X_2)\sigma_1\sigma_2}{\sigma_1^2}$$

$$\boxed{a = corr(X_1, X_2)\frac{\sigma_2}{\sigma_1}}$$

$$\boxed{b = \mu_2 - a\mu_1}$$

The best straight line fit is

$$aX_1 + b = \mathrm{corr}(X_1, X_2)\frac{\sigma_2}{\sigma_1} X_1$$

$$+ \mu_2 - a\mu_1$$

$$= \boxed{\mu_2 + \mathrm{corr}(X_1, X_2)\frac{\sigma_2}{\sigma_1}(X_1 - \mu_1)}$$

when $X_1 = \mu_1$, get μ_2