STAT 24400 Lecture 9
4.3 Covariance & Correlation

Yibi Huang
Department of Statistics
University of Chicago

Expected Values for Functions of Several R.V.'s

Expected Values for Functions of Several R.V.'s

For **discrete** random variables $X_1, \ldots, X_n$ with joint PMF $p(x_1, \ldots, x_n)$, the expected value for $g(X_1, \ldots, X_n)$ is

$$\mathrm{E}(g(X_1, \ldots, X_n)) = \sum_{x_1, \ldots, x_n} g(x_1, \ldots, x_n) p(x_1, \ldots, x_n),$$

provided that $\sum_{x_1, \ldots, x_n} |g(x_1, \ldots, x_n)| p(x_1, \ldots, x_n) < \infty$.

For **continuous** random variables $X_1, \ldots, X_n$ with joint PDF $f(x_1, \ldots, x_n)$, the expected value for $g(X_1, \ldots, X_n)$ is

$$\mathrm{E}(g(X_1, \ldots, X_n)) = \int \cdots \int g(x_1, \ldots, x_n) f(x_1, \ldots, x_n) \mathrm{d}x_1 \ldots \mathrm{d}x_n,$$

provided that $\int \cdots \int |g(x_1, \ldots, x_n)| f(x_1, \ldots, x_n) \mathrm{d}x_1 \ldots \mathrm{d}x_n < \infty$.

$\mathrm{E}(aX + bY) = a\,\mathrm{E}(X) + b\,\mathrm{E}(Y)$

If $g(X, Y) = aX + bY$ for two random variables (discrete or continuous) X and Y and two constants a and b, we have

$$\mathrm{E}[g(X, Y)] = \mathrm{E}(aX + bY) = a\,\mathrm{E}(X) + b\,\mathrm{E}(Y).$$

Proof. We will prove it for continuous X and Y with joint PDF $f(x, y)$. The proof for the discrete case is similar.

$$\begin{aligned} \mathrm{E}(aX + bY) &= \iint (ax + by) f(x, y) \mathrm{d}x \mathrm{d}y, \quad \text{(by definition)} \\ &= \underbrace{\iint axf(x, y) \mathrm{d}x \mathrm{d}y}_{\text{Part I}} + \underbrace{\iint byf(x, y) \mathrm{d}x \mathrm{d}y}_{\text{Part II}} \end{aligned}$$

For Part I, we first integrate over y, and then over x.

$$\begin{aligned}\text{Part I} = \iint axf(x,y)\mathrm{d}x\mathrm{d}y &= a\int\left(\int xf(x,y)dy\right)\mathrm{d}x \\ &= a\int x\underbrace{\int f(x,y)dy}_{f_X(x)}\mathrm{d}x = a\underbrace{\int xf_X(x)\mathrm{d}x}_{\mathrm{E}(X)} = a\,\mathrm{E}(X)\end{aligned}$$

For Part II, we first integrate over x, and then over y.

$$\begin{aligned}\text{Part II} = \iint byf(x,y)\mathrm{d}x\mathrm{d}y &= b\int\left(\int yf(x,y)dx\right)\mathrm{d}y \\ &= b\int y\underbrace{\int f(x,y)dx}_{f_Y(y)}\mathrm{d}y = b\underbrace{\int yf_Y(y)\mathrm{d}y}_{\mathrm{E}(Y)} = b\,\mathrm{E}(Y)\end{aligned}$$

Putting Parts I & II together, we get

$$\mathrm{E}(aX+bY) = a\,\mathrm{E}(X) + b\,\mathrm{E}(Y).$$

Expected Value for Linear Combination of R.V.'s

The result $\mathrm{E}(aX + bY) = a\,\mathrm{E}(X) + b\,\mathrm{E}(Y)$ can be generalized to linear combinations of several random variables

$$\mathrm{E}(a_1X_1 + a_2X_2 + \cdots + a_nX_n) = a_1\,\mathrm{E}(X_1) + a_2\,\mathrm{E}(X_2) + \cdots + a_n\,\mathrm{E}(X_n),$$

no matter the rv's are discrete or continuous, independent or not.

$\mathrm{E}[g(X)h(Y)] = \mathrm{E}[g(X)]\,\mathrm{E}[h(Y)]$ if X, Y are independent

When X and Y are **independent**, for any functions g and h,

$$\mathrm{E}[g(X)h(Y)] = \mathrm{E}[g(X)]\,\mathrm{E}[h(Y)].$$

In particular, $\mathrm{E}(XY) = \mathrm{E}(X)\,\mathrm{E}(Y)$.

Proof. We prove the discrete case. The continuous case is similar. Using that $p(x,y) = p_X(x)p_Y(y)$ when X, Y are indep, one has

$$\begin{aligned}
\mathrm{E}[g(X)h(Y)] &= \sum\nolimits_{xy} g(x)h(y)\color{red}{p(x,y)} \\
&= \sum\nolimits_{x}\sum\nolimits_{y} g(x)h(y)\color{red}{p_X(x)p_Y(y)} \quad \text{(by independence)} \\
&= \underbrace{\sum\nolimits_{x} g(x)p_X(x)}_{\mathrm{E}[g(X)]}\underbrace{\sum\nolimits_{y} h(y)p_Y(y)}_{\mathrm{E}[h(Y)]} \\
&= \mathrm{E}[g(X)]\,\mathrm{E}[h(Y)]
\end{aligned}$$

Covariance

Covariance

The **covariance** of X and Y, denoted as $\text{Cov}(X, Y)$ or σ_{XY}, is defined as

$$\text{Cov}(X, Y) = \sigma_{XY} = \text{E}[(X - \mu_X)(Y - \mu_Y)],$$

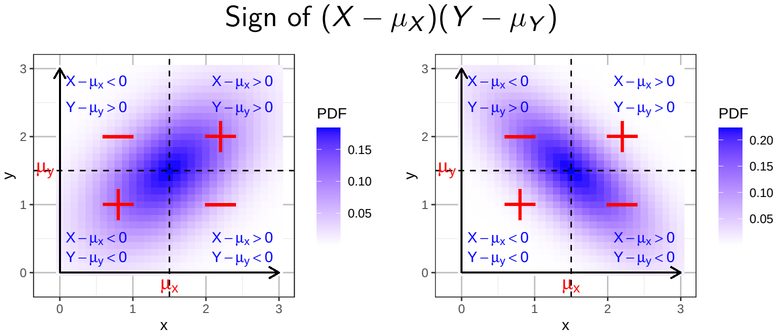
in which $\mu_X = \text{E}(X)$, $\mu_Y = \text{E}(Y)$

- Covariance is a generalization of variance as the variance of a random variable X is just the covariance of X with itself.

$$\text{Var}(X) = \text{Cov}(X, X) = \text{E}[(X - \mu_X)^2]$$

Sign of Covariance Reflects the Direction of (X, Y) Relation

- $\text{Cov}(X, Y) > 0$ means a *positive* relation between X, Y
  - When X increases, Y tends to increase
- $\text{Cov}(X, Y) < 0$ means a *negative* relation between X, Y
  - When X increases, Y tends to decrease



Shortcut Formula for Covariance

$$\boxed{\text{Cov}(X, Y) = \text{E}(XY) - \text{E}(X)\,\text{E}(Y)}$$

- Similar to the Shortcut Formula for Variance

$$\text{Var}(X) = \text{E}(X^2) - [\text{E}(X)]^2.$$

Shortcut Formula for Covariance

$$\boxed{\mathrm{Cov}(X,Y) = \mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y)}$$

- Similar to the Shortcut Formula for Variance

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2.$$

- If X, Y are independent,
  $\Rightarrow \mathrm{E}(XY) = \mathrm{E}(X)\,\mathrm{E}(Y) \Rightarrow \mathrm{Cov}(X,Y) = 0.$

Shortcut Formula for Covariance

$$\boxed{\mathrm{Cov}(X, Y) = \mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y)}$$

- Similar to the Shortcut Formula for Variance

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2.$$

- If X, Y are independent,
  $\Rightarrow \mathrm{E}(XY) = \mathrm{E}(X)\,\mathrm{E}(Y) \Rightarrow \mathrm{Cov}(X, Y) = 0.$

- However, $\mathrm{Cov}(X, Y) = 0$ does **not** imply the independence of X and Y. In this case, we say X and Y are *uncorrelated.*

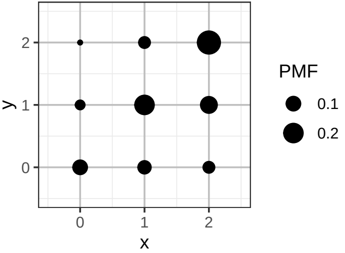
Proof of the Shortcut Formula for Covariance

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

Example (Gas Station) — E(XY)

Recall the joint PMF for the Gas Station Example in L05 is

		Y (full-service)		
	$p(x,y)$	0	1	2
X	0	0.10	0.04	0.02
self-	1	0.08	0.20	0.06
service	2	0.06	0.14	0.30

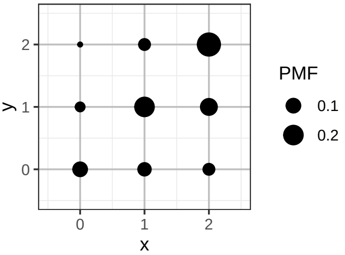


Guess Cov(X, Y) > 0 or < 0?

Example (Gas Station) — $\mathrm{E}(XY)$

Recall the joint PMF for the Gas Station Example in L05 is

	$p(x,y)$	Y (full-service) 0	1	2
X	0	0.10	0.04	0.02
self-	1	0.08	0.20	0.06
service	2	0.06	0.14	0.30



Guess $\mathrm{Cov}(X,Y) > 0$ or < 0?

$$
\begin{aligned}
\mathrm{E}(XY) &= \sum_{xy} xyp(x,y) \\
&= 0\cdot 0\cdot 0.10 + 0\cdot 1\cdot 0.04 + 0\cdot 2\cdot 0.02 \\
&\quad + 1\cdot 0\cdot 0.08 + 1\cdot 1\cdot 0.20 + 1\cdot 2\cdot 0.06 \\
&\quad + 2\cdot 0\cdot 0.06 + 2\cdot 1\cdot 0.14 + 2\cdot 2\cdot 0.30 \\
&= 1.8
\end{aligned}
$$

Example (Gas Station) — Covariance

Recall in L05, we obtained the marginal PMFs for X and for Y:

x	0	1	2
$p_X(x)$	0.16	0.34	0.50

$$E(X) = 0 \cdot 0.16 + 1 \cdot 0.34 + 2 \cdot 0.5 = 1.34$$

y	0	1	2
$p_Y(y)$	0.24	0.38	0.38

$$E(Y) = 0 \cdot 0.24 + 1 \cdot 0.38 + 2 \cdot 0.38 = 1.14$$

By the shortcut formula, the covariance is

$$\begin{aligned} \text{Cov}(X, Y) &= E(XY) - E(X)\,E(Y) \\ &= 1.8 - 1.34 \times 1.14 = 0.2724 > 0. \end{aligned}$$

When one service island is busy, the other also tends to be busy.

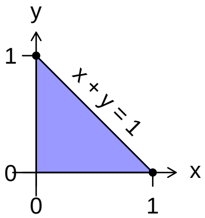
Example (Mixed Nuts)

Recall in L05, the joint PDF for

$$X = \text{ the weight of almonds, and}$$
$$Y = \text{ the weight of cashews}$$

in a can of mixed nuts is

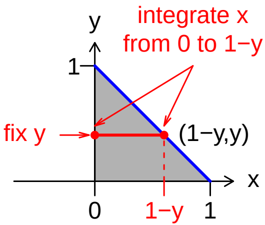
$$f(x, y) = \begin{cases} 24xy & \text{if } 0 \le x, y \le 1, x + y < 1 \\ 0 & \text{otherwise} \end{cases}$$



Before we calculate it, guess $\text{Cov}(X, Y) > 0$ or < 0?

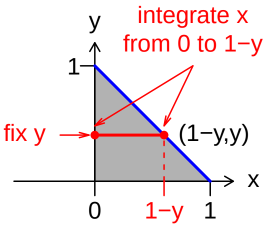
Example (Mixed Nuts) — $E(XY)$

$$E(XY) = \iint xyf(x,y)\mathrm{d}x\mathrm{d}y$$
$$= \int_0^1 \underbrace{\int_0^{1-y} 24x^2y^2\mathrm{d}x\,\mathrm{d}y}_{\text{see below}}$$



Example (Mixed Nuts) — E(XY)

$$\mathrm{E}(XY) = \iint xyf(x,y)\,\mathrm{d}x\mathrm{d}y = \int_0^1 \underbrace{\int_0^{1-y} 24x^2y^2\mathrm{d}x}_{\text{see below}}\,\mathrm{d}y$$



where

$$\int_0^{1-y} 24x^2y^2\,\mathrm{d}x = \left[8x^3y^2\right]_{x=0}^{x=1-y} = 8(1-y)^3y^2.$$

Putting it back to the double integral, we get

$$\mathrm{E}(XY) = \int_0^1 \int_0^{1-y} 24x^2y^2\,\mathrm{d}x\,\mathrm{d}y = \int_0^1 8(1-y)^3y^2\mathrm{d}y = \frac{2}{15}.$$

Example (Mixed Nuts) — Covariance

Recall in L05, we calculated the marginal PDF's for X and for Y:

$$f_X(x) = 12x(1-x)^2, \quad f_Y(y) = 12y(1-y)^2, \text{ for } 0 \le x, y \le 1.$$

using which we can calculate

$$\begin{aligned} \mathrm{E}(X) &= \int_0^1 x f_X(x)\mathrm{d}x = \int_0^1 12x^2(1-x)^2\mathrm{d}x \\ &= \int_0^1 12x^2 - 24x^3 + 12x^4\mathrm{d}x = \left[4x^3 - 6x^4 + \frac{12}{5}x^5\right]_0^1 = \frac{2}{5}. \end{aligned}$$

Likewise, $\mathrm{E}(Y) = 2/5$. The covariance by the shortcut formula is

$$\mathrm{Cov}(X, Y) = \mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y) = \frac{2}{15} - \frac{2}{5} \times \frac{2}{5} = -\frac{2}{75}$$

When a can of mixed nuts has more almond, it likely has less cashew, and vice versa.

More Properties of Covariance

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

In the following, a, b are constants. X, Y, Z are random variables

- Symmetry: $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$
- Scaling: $\mathrm{Cov}(aX, bY) = ab\,\mathrm{Cov}(X, Y)$
- Right-linearity: $\mathrm{Cov}(X + Y, Z) = \mathrm{Cov}(X, Z) + \mathrm{Cov}(Y, Z)$
- Left-linearity: $\mathrm{Cov}(X, Y + Z) = \mathrm{Cov}(X, Y) + \mathrm{Cov}(X, Z)$
- $\mathrm{Cov}(a, X) = 0$.

Proofs for Properties of Covariance

The proofs for these propertie are all straightforward from definition. We just prove the Right-linearity as an example.

$$\begin{aligned}
\mathrm{Cov}(X+Y, Z) &= \mathrm{E}((X+Y)Z) - \mathrm{E}(X+Y)\,\mathrm{E}(Z) \\
&= \mathrm{E}(XZ) + \mathrm{E}(YZ) - [\mathrm{E}(X) + \mathrm{E}(Y)]\,\mathrm{E}(Z) \\
&= \underbrace{\mathrm{E}(XZ) - \mathrm{E}(X)\,\mathrm{E}(Z)}_{\mathrm{Cov}(X,Z)} + \underbrace{\mathrm{E}(YZ) - \mathrm{E}(Y)\,\mathrm{E}(Z)}_{\mathrm{Cov}(Y,Z)} \\
&= \mathrm{Cov}(X, Z) + \mathrm{Cov}(Y, Z)
\end{aligned}$$

Note in the proof above, we used the property of expected value that

$$\begin{aligned}
\mathrm{E}(X+Y) &= \mathrm{E}(X) + \mathrm{E}(Y) \\
\mathrm{E}(XZ+YZ) &= \mathrm{E}(XZ) + \mathrm{E}(YZ)
\end{aligned}$$

Variance of Linear Combinations of Two Random Variables

Recall that expectation has the following linear property:

$$\mathrm{E}(aX + bY) = a\,\mathrm{E}(X) + b\,\mathrm{E}(Y).$$

We also have shown that $\mathrm{Var}(aX + b) = a^2\,\mathrm{Var}(X)$.

How about $\mathrm{Var}(aX + bY)$?

$$\boxed{\mathrm{Var}(aX + bY) = a^2\,\mathrm{Var}(X) + 2ab\,\mathrm{Cov}(X, Y) + b^2\,\mathrm{Var}(Y)}$$

- If X is independent of Y, $\mathrm{Var}(X \pm Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$

Proof of $\mathrm{Var}(aX + bY)$

$$
\begin{aligned}
\mathrm{Var}(aX + bY) &= \mathrm{Cov}(aX + bY, aX + bY) \\
&= \underbrace{\mathrm{Cov}(aX, aX + bY)}_{\downarrow} + \underbrace{\mathrm{Cov}(bY, aX + bY)}_{\downarrow} \quad \text{(right-linearity)} \\
&= \overbrace{\mathrm{Cov}(aX, aX) + \mathrm{Cov}(aX, bY)} + \overbrace{\mathrm{Cov}(bY, aX) + \mathrm{Cov}(bY, bY)} \quad \text{(left-linearity)} \\
&= \mathrm{Var}(aX) + 2\,\mathrm{Cov}(aX, bY) + \mathrm{Var}(bY) \quad \text{(symmetry)} \\
&= a^2\,\mathrm{Var}(X) + 2ab\,\mathrm{Cov}(X, Y) + b^2\,\mathrm{Var}(Y) \quad \text{(scaling)}
\end{aligned}
$$

Linear Combinations of Random Variables

For any random variables $X_1, X_2, \ldots, X_n$., a *linear combination* of $X_1, X_2, \ldots, X_n$ is

$$a_1 X_1 + a_2 X_2 + \cdots + a_n X_n,$$

where $a_1, a_2, \ldots, a_n$ are constant numbers. For example,

- The sum $X_1 + X_2 + \cdots + X_n$ is a linear combination of $X_1, \ldots, X_n$ with all a_i's $= 1$.
- The average
$$\frac{X_1 + X_2 + \cdots + X_n}{n}$$
is a linear combination of $X_1, X_2, \ldots, X_n$ with all a_i's $= 1/n$.
- The difference $X - Y$ is a linear combination of X and Y with $a_1 = 1$, $a_2 = -1$

Variance of a Linear Combination of RV's

$$\mathrm{Var}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 \,\mathrm{Var}(X_i) + 2\sum_{i<j} a_i a_j \,\mathrm{Cov}(X_i, X_j)$$

- There is a covariance term for every pair of X_i and X_j
- When $X_1, \ldots, X_n$ are independent, then

$$\mathrm{Var}(X_1 + X_2 + \ldots + X_n) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \ldots + \mathrm{Var}(X_n).$$

- When $\mathrm{Var}(X_i) = \sigma^2$ for $i = 1, \ldots, n$,
  and $\mathrm{Cov}(X_i, X_j) = \rho$ for $1 \leq i \neq j \leq n$, then

$$\mathrm{Var}(X_1 + \ldots + X_n) = n\sigma^2 + n(n-1)\rho.$$

Example: Variance of the Binomial Distribution

In L08, we computed the expected value for the Binomial distribution $\text{Bin}(n, p)$ is $E(X) = np$.

Today, we find its variance using linear combinations to be

$$\text{Var}(X) = np(1-p).$$

First for the special case $n = 1$, $X \sim \text{Bin}(n = 1, p)$, X only takes value 0 and 1 with the PMF below

x	0	1
$p(x)$	$1-p$	p

Hence

$$E(X) = \sum_{x=0,1} xp(x) = 0 \cdot (1-p) + 1 \cdot p = \boxed{p},$$

$$E(X^2) = \sum_{x=0,1} x^2 p(x) = 0^2 \cdot (1-p) + 1^2 \cdot p = p$$

$$\text{Var}(X) = E(X^2) - (E(X))^2 = p - p^2 = \boxed{p(1-p)}$$

For general n, recall a Binomial random variable $X \sim \text{Bin}(n, p)$ is the number of successes obtained in n independent Bernoulli trials. For each of the n trials, define

$$X_i = \begin{cases} 1 & \text{if success in the } i\text{th trial} \\ 0 & \text{if failure in the } i\text{th trial} \end{cases} \quad \Rightarrow X_i \sim \text{Bin}(n = 1, p).$$

$$\begin{aligned} \text{Then } X &= \text{the number of successes obtained in the } n \text{ trials} \\ &= X_1 + X_2 + \ldots + X_n, \end{aligned}$$

The expected value and variance of X are thus

$$\text{E}(X) = \underbrace{\text{E}(X_1)}_{=p} + \cdots + \underbrace{\text{E}(X_n)}_{=p} = np$$

$$\text{Var}(X) = \underbrace{\text{Var}(X_1)}_{=p(1-p)} + \cdots + \underbrace{\text{Var}(X_n)}_{=p(1-p)} = np(1-p)$$

since X_i's are indep. and each with mean p and variance $p(1-p)$ as $X_i \sim \text{Bin}(n = 1, p)$.

Example (Sample Mean)

Suppose $X_1, \ldots, X_n$ are *i.i.d.* rv's with mean μ and variance σ^2.

- *i.i.d.* = "independent and have an identical distribution"

Consider the *sample mean*

$$\overline{X} = \frac{1}{n}(X_1 + \cdots + X_n)$$

Then

$$\begin{aligned}
E(\overline{X}) &= \frac{1}{n}[E(X_1) + \ldots + E(X_n)] = \frac{1}{n}(\underbrace{\mu + \ldots + \mu}_{n \text{ copies}}) = \mu. \\
\text{Var}(\overline{X}) &= \frac{1}{n^2}\text{Var}(X_1 + X_2 + \ldots + X_n) \quad \text{since } \text{Var}(aX) = a^2 V(X) \\
&= \frac{1}{n^2}[\text{Var}(X_1) + \ldots + \text{Var}(X_n)] \quad \text{as all } X_i\text{'s are indep.} \\
&= \frac{1}{n^2}(\underbrace{\sigma^2 + \ldots + \sigma^2}_{n \text{ copies}}) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}
\end{aligned}$$

Example (Coupon Collector's Problem, p.127, textbook)

If each box of breakfast cereals contains a coupon,

- there are n different types of coupons,
- the coupon in any box is equally likely to be any of the n types,

let N be the number of boxes required to collect all n types of coupons. Find $\mathrm{E}(N)$ and $\mathrm{Var}(N)$.

Example (Coupon Collector's Problem, p.127, textbook)

If each box of breakfast cereals contains a coupon,

- there are n different types of coupons,
- the coupon in any box is equally likely to be any of the n types,

let N be the number of boxes required to collect all n types of coupons. Find $\mathrm{E}(N)$ and $\mathrm{Var}(N)$.

Sol. Let

- X_1 the number of boxes to get the first coupon (any type). Clearly $X_1 = 1$.
- $X_2 =$ the number of additional boxes required to collect a new type of coupons after collecting first type.
- $X_i =$ the number of additional boxes required to collect i types of coupons after collecting $i-1$ types, for $i = 1, 2, \ldots, n$.

Observe that $N = X_1 + X_2 + \cdots + X_n$

Coupon Collector's Problem — Expected Value

What's the distribution of X_i?

Coupon Collector's Problem — Expected Value

What's the distribution of X_i?

- For X_2, the chance to get a new type of coupon in the next box is $\frac{n-1}{n}$.

Coupon Collector's Problem — Expected Value

What's the distribution of X_i?

- For X_2, the chance to get a new type of coupon in the next box is $\frac{n-1}{n}$. So,

$$X_2 \sim \text{Geometric}(p = \frac{n-1}{n}).$$

- $X_3 \sim \text{Geometric}(p = \frac{n-2}{n})$

Coupon Collector's Problem — Expected Value

What's the distribution of X_i?

- For X_2, the chance to get a new type of coupon in the next box is $\dfrac{n-1}{n}$. So,

$$X_2 \sim \text{Geometric}(p = \frac{n-1}{n}).$$

- $X_3 \sim \text{Geometric}(p = \frac{n-2}{n})$
- In general, $X_i \sim \text{Geometric}(p = \frac{n-i+1}{n})$, $i = 1, 2, \ldots, n$.

Coupon Collector's Problem — Expected Value

What's the distribution of X_i?

- For X_2, the chance to get a new type of coupon in the next box is $\frac{n-1}{n}$. So,

$$X_2 \sim \text{Geometric}(p = \frac{n-1}{n}).$$

- $X_3 \sim \text{Geometric}(p = \frac{n-2}{n})$
- In general, $X_i \sim \text{Geometric}(p = \frac{n-i+1}{n})$, $i = 1, 2, \dots, n$.

Recall the expected value for Geometric(p) is $1/p$. We know

$$\begin{aligned} \mathrm{E}(N) = \sum_{i=1}^{n} \mathrm{E}(X_i) &= \sum_{i=1}^{n} \frac{n}{n-i+1} \\ &= n\left(\frac{1}{n} + \frac{1}{n-1} + \frac{1}{n-2} + \cdots + \frac{1}{1}\right) \\ &= n \sum_{r=1}^{n} \frac{1}{r} \approx n \log(n). \end{aligned}$$

Coupon Collector's Problem — Variance

Recall the variance for Geometric(p) is $\frac{1-p}{p^2}$. We get

$$\begin{aligned}
\mathrm{Var}(N) &= \sum_{i=1}^{n} \mathrm{Var}(X_i) \quad (\text{since } X_1, \cdots, X_n \text{ are indep.}) \\
&= \sum_{i=1}^{n} \frac{1 - \frac{n-i+1}{n}}{(n-i+1)^2/n^2} \\
&= n \sum_{i=1}^{n} \frac{i-1}{(n-i+1)^2} \\
&= n \left(0 + \frac{1}{(n-1)^2} + \frac{2}{(n-2)^2} + \frac{3}{(n-3)^2} + \cdots + \frac{n-1}{1^2} \right)
\end{aligned}$$

Correlation

How Large the Covariance Indicates a Strong Relation?

One can prove the Cauchy Inequality for covariance

$$[\mathrm{Cov}(X, Y)]^2 \leq \mathrm{Var}(X)\,\mathrm{Var}(Y)$$

How Large the Covariance Indicates a Strong Relation?

One can prove the Cauchy Inequality for covariance

$$[\mathrm{Cov}(X, Y)]^2 \leq \mathrm{Var}(X)\,\mathrm{Var}(Y)$$

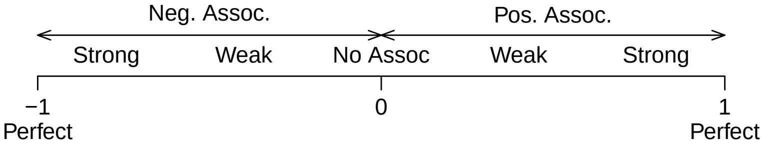
Moreover, the *covariance reaches its maximum possible magnitude if and only if X and Y has a perfect linear relation $Y = aX + b$, $a \neq 0$*.

Thus, one can assess the strength of linear relation between X, Y by comparing $\mathrm{Cov}(X, Y)$ with $\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}$.

Correlation

$$\text{Correlation } = \rho_{XY} = \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\,\text{Var}(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}.$$

- $-1 \leq \rho_{XY} \leq 1$ since $\text{Cov}(X, Y) \leq \sqrt{\text{Var}(X)\,\text{Var}(Y)}$
- The closer ρ_{XY} is to 1 or to -1, the stronger the linear relation between X and Y



- $\rho_{XY} = 1$ or -1 if and only if $Y = aX + b$ and $a \neq 0$, i.e., X and Y has an perfect linear relation

Covariance Is NOT Scale Invariant but Correlation Is!

Example. Let

- X = amount of time studying STAT 244 per week, and
- Y = final grade in STAT 244

If X is measured in minutes rather than in hours, $\text{Cov}(X, Y)$ would be 60 times as large.

The *strength* of XY relation should be the same no matter X is measured in minutes or in hours.

Correlation ρ_{XY} is *scale invariant* and has no unit.

$$\begin{aligned}
\text{Corr}(aX + c, bY + d) &= \frac{\text{Cov}(aX + c, bY + d)}{\sqrt{\text{Var}(aX + c)\,\text{Var}(bY + d)}} \\
&= \frac{ab\,\text{Cov}(X, Y)}{\sqrt{a^2\,\text{Var}(X) b^2\,\text{Var}(Y)}} \\
&= (\text{sign of } ab)\,\text{Corr}(X, Y)
\end{aligned}$$

Example (Gas Station) — Correlation

Recall in L05, we obtained the marginal PMFs for X and Y:

x	0	1	2
$p_X(x)$	0.16	0.34	0.50

,

y	0	1	2
$p_Y(y)$	0.24	0.38	0.38

$$\begin{aligned} \mathrm{E}(X^2) &= 0^2 \cdot 0.16 + 1^2 \cdot 0.34 + 2^2 \cdot 0.5 = 2.34 \\ \mathrm{Var}(X) &= \mathrm{E}(X^2) - (\mathrm{E}(X))^2 = 2.34 - 1.34^2 = 0.5444 \\ \mathrm{E}(Y^2) &= 0^2 \cdot 0.24 + 1^2 \cdot 0.38 + 2^2 \cdot 0.38 = 1.9 \\ \mathrm{Var}(Y) &= \mathrm{E}(Y^2) - (\mathrm{E}(Y))^2 = 1.9 - 1.14^2 = 0.6004 \end{aligned}$$

$$\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}} = \frac{0.2724}{\sqrt{0.5444 \times 0.6004}} \approx 0.476.$$

Example (Mixed Nuts) — Correlation

Recall in L05, we calculated the marginal pdf's for X and for Y:

$$f_X(x) = 12x(1-x)^2, \quad f_Y(y) = 12y(1-y)^2, \text{ for } 0 \le x, y \le 1.$$

using which we can calculate

$$\begin{aligned}
\mathrm{E}(X^2) &= \int_0^1 x^2 f_X(x)\mathrm{d}x = \int_0^1 12x^3(1-x)^2\mathrm{d}x \\
&= \int_0^1 12x^3 - 24x^4 + 12x^5\mathrm{d}x = 3x^4 - \frac{24x^5}{5} + 2x^6\Big|_0^1 = \frac{1}{5}
\end{aligned}$$

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2 = \frac{1}{5} - (\frac{2}{5})^2 = \frac{1}{25}$$

Similar, one can calculate $\mathrm{Var}(Y) = 1/25$

$$\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}} = \frac{-2/75}{\sqrt{(1/25)(1/25)}} = -\frac{2}{3} \approx -0.667.$$

Proof of $[\mathrm{Cov}(X, Y)]^2 \le \mathrm{Var}(X)\,\mathrm{Var}(Y)$

Since the variance of a random variable is always nonnegative,

$$\begin{aligned} 0 &\le \mathrm{Var}\left(\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}\right) \\ &= \mathrm{Var}\left(\frac{X}{\sigma_X}\right) + \mathrm{Var}\left(\frac{Y}{\sigma_Y}\right) + 2\,\mathrm{Cov}\left(\frac{X}{\sigma_X}, \frac{Y}{\sigma_Y}\right) \\ &= \frac{\mathrm{Var}(X)}{\sigma_X^2} + \frac{\mathrm{Var}(Y)}{\sigma_Y^2} + 2\underbrace{\frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}}_{=\rho} \\ &= \frac{\sigma_X^2}{\sigma_X^2} + \frac{\sigma_Y^2}{\sigma_Y^2} + 2\rho \\ &= 2(1+\rho), \quad \text{which implies } \rho \ge -1. \end{aligned}$$

Similarly, one can show that

$$0 \le \mathrm{Var}\left(\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y}\right) = 2(1-\rho), \quad \text{which implies } \rho \le 1.$$

This proves that $-1 \le \rho \le 1 \iff 1 \ge \rho^2 = \dfrac{[\mathrm{Cov}(X, Y)]^2}{\mathrm{Var}(X)\,\mathrm{Var}(Y)}$.

Proof that $\rho^2 = 1 \iff \mathrm{P}(Y = aX + b) = 1$

From

$$\mathrm{Var}\left(\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y}\right) = 2(1-\rho)$$

we see that $\rho = 1 \iff \mathrm{Var}\left(\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y}\right) = 0.$

The variance of a random variable W is 0 only if

$$\mathrm{P}(W = c) = 1, \quad \text{for some constant c.}$$

Thus $\rho = 1$ if and only if

$$\mathrm{P}\left(\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y} = c\right) = 1, \quad \text{for some constant } c.$$

Similarly, we can show $\rho = -1$ if and only if

$$\mathrm{P}\left(\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y} = c\right) = 1, \quad \text{for some constant } c.$$