STAT 23400 Lecture 9: Expected Values, Covariance and Correlation

Review: Joint Continuous distributions

- **Joint pdf** for two continuous r.v.'s:

$$P[(X, Y) \in A] = \iint_A f(x, y)\, dx\, dy\,.$$

- **Marginal pdf**:

$$f_X(x) = \int_{\mathbb{R}} f(x, y) dy \qquad \text{for all } x \in \mathbb{R}$$

$$f_Y(y) = \int_{\mathbb{R}} f(x, y) dx \qquad \text{for all } y \in \mathbb{R}$$

- **Conditional probability distribution**:

$$f_{X|Y}(x \mid y) = \frac{f(x, y)}{f_Y(y)}\,, \quad f_{Y|X}(y \mid x) = \frac{f(x, y)}{f_X(x)}$$

- X and Y are **independent** if and only if

$$p(x, y) = p_X(x) p_Y(y) \qquad \text{if } X \text{ and } Y \text{ are discrete}\,,$$

$$f(x, y) = f_X(x) f_Y(y) \qquad \text{if } X \text{ and } Y \text{ are continuous}\,.$$

Outline for Today

- Expected Values, Covariance and Correlation (Section 5.2 in MMSA).
- Lab05: Central limit theorem (link posted on canvas).
- HW 5 is posted on canvas.

Expected Values

Expected Values of Functions of X, Y

For two r.v.'s X, Y with join pmf/pdf $f(x, y)$, the expected value of a function $g(X, Y)$ of X and Y is defined as

$$E[g(X,Y)] = \begin{cases} \sum_{x,y} g(x,y) f(x,y) & \text{in discrete case,} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f(x,y)\, dx\, dy & \text{in continuous case.} \end{cases}$$

Halloween Game Example Revisited

Recall that X is the smaller number of the two dice, and Y is the bigger number. The joint pmf is

	$f(x,y)$	y = 1	2	3	4	5	6
	1	1/36	2/36	2/36	2/36	2/36	2/36
	2	0	1/36	2/36	2/36	2/36	2/36
x	3	0	0	1/36	2/36	2/36	2/36
	4	0	0	0	1/36	2/36	2/36
	5	0	0	0	0	1/36	2/36
	6	0	0	0	0	0	1/36

Let us find $E(X)$ and $E(Y)$.

Recall the marginal distributions are

x	1	2	3	4	5	6
$p_X(x)$	11/36	9/36	7/36	5/36	3/36	1/36

y	1	2	3	4	5	6
$p_Y(y)$	1/36	3/36	5/36	7/36	9/36	11/36

What is $E(XY)$?

Expected Value of Linear combinations of X and Y

Thm. If $\mathbf{g(X,Y) = aX + bY}$ for any $a \in \mathbb{R}$ and any $b \in \mathbb{R}$, then $g(X,Y)$ is a **linear combination of X and Y** and

$$\mathbf{E[g(X,Y)] = aE(X) + bE(Y)}$$

Expected Value: Independent Case

Thm. If X and Y are **independent**, then for any functions g and h, we have

$$E[g(X)h(Y)] = E[g(X)]E[h(Y)].$$

Covariance

Measuring Dependence

For any dependent random variables X and Y,
How can we measure the relationship between X and Y?
How "strong" is this relationship?

- Positive Association, e.g., SAT score v.s. hours spent on studying.
- Negative Association: e.g., SAT score v.s. sleeping hours per night.

Covariance of two random variables

Covariance of two random variables X, Y with means μ_X, μ_Y. Take $g(X, Y) = (X - \mu_X) \cdot (Y - \mu_Y)$. The covariance of X and Y is defined as

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

$Cov(X, Y)$ is a measure of how the variables 'covary'.
$Cov(X, Y) > 0 \longrightarrow$ when X increases Y tends to increase.
$Cov(X, Y) < 0 \longrightarrow$ when X increases Y tends to decrease.

A **helpful formula:** $\boxed{Cov(X, Y) = E(XY) - \mu_X \mu_Y\,.}$

Properties of Covariance

The covariance has the following properties.

For any rv's X, Y, Z and constant numbers a, b, we have

- $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$
- Symmetry: $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$
- Homogeneity: $\mathrm{Cov}(aX, bY) = ab\mathrm{Cov}(X, Y)$
- The covariance of a r.v. and a constant number is 0, i.e., $\mathrm{Cov}(a, X) = 0$
- Right-linearity: $\mathrm{Cov}(X + Y, Z) = \mathrm{Cov}(X, Z) + \mathrm{Cov}(Y, Z)$
- Left-linearity: $\mathrm{Cov}(X, Y + Z) = \mathrm{Cov}(X, Y) + \mathrm{Cov}(X, Z)$

Proofs for Properties of Covariance

The proofs of the above are all straightforward from definition. Here we 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

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

Example 1: Application of the Properties of Covariance

Let X_1, X_2, X_3 be uncorrelated r.v.'s each with variances σ_1^2, σ_2^2, σ_3^2 respectively. Find $\mathrm{Cov}(X_1 - X_2, X_1 - X_3)$.

Example 2: Halloween Game Example Revisited

In the halloween game example, find the covariance of X and Y.

Correlation

Problems with Covariance

But, the "relationship" between X and Y should be the same no matter what units we choose to measure in.
Solution: Use **correlation coefficient** to measure linear relationship.

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

- Correlation is unitless
- We always have $-1 \leq \rho_{XY} \leq 1$
  - $\rho_{XY} = +1 \implies$ perfect, positive, linear relationship
  - $\rho_{XY} = 0 \implies$ no *linear* relationship (other relationship?)
  - $\rho_{XY} = -1 \implies$ perfect, negative, linear relationship

Halloween Game Example Revisited

Recall that, previously, in the Halloween Game Example, we have calculated that

$$Cov(X, Y) = \frac{1225}{1296} \approx 0.95.$$

Can we say the relationship between X and Y are almost perfectly linear?

Halloween Game Example Revisited

Recall that, previously, in the Halloween Game Example, we have calculated that

$$Cov(X, Y) = \frac{1225}{1296} \approx 0.95.$$

Can we say the relationship between X and Y are almost perfectly linear?

No, because covariance is not scale free. Need to consider correlation.

We can further obtain $\sigma_X = 2.88$, $\sigma_Y = 1.40$ and

$$\rho_{XY} = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} \approx 0.24\,,$$

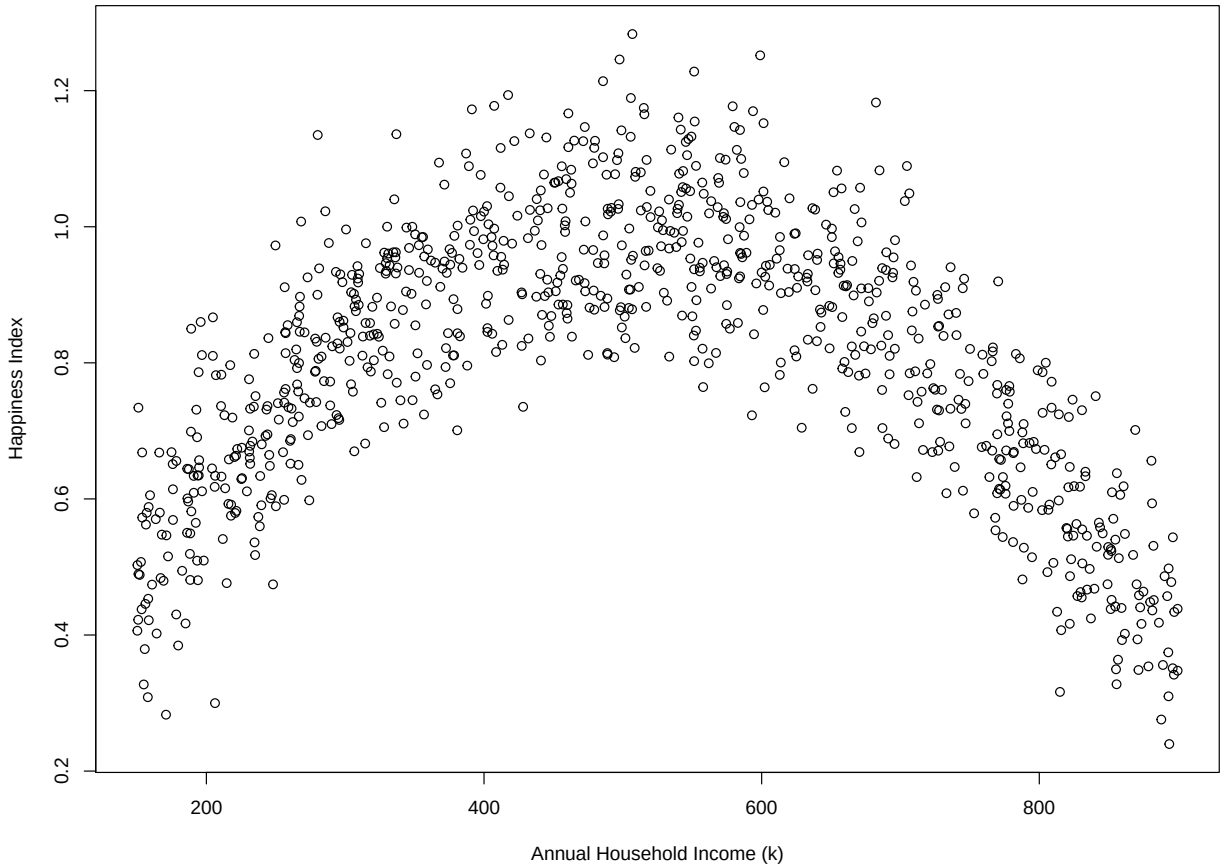
indicating a less strong linear relationship.

Correlation and Independence

What is the correlation if X and Y are independent?

Example: Zero Correlation Does NOT Imply Independence

Consider the relationship between the Annual Household Income (in thousand US dollars) and the Family Happiness Index.



Linear Combinations of Random Variables

Multivariate Random Variables: Joint Distribution

- We can define **joint probability mass function** for more than two r.v's, $\mathbf{X} = (X_1, X_2, \ldots, X_n)$, as

$$\begin{aligned} p(x_1, x_2, \ldots, x_n) =& P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) \\ =& P(\{X_1 = x_1\} \cap \{X_2 = x_2\} \cap \ldots \cap \{X_n = x_n\}). \end{aligned}$$

- Define the **joint probability density function** for continuous r.v.'s $\mathbf{X} = (X_1, \ldots X_n)$ such that

$$P[\mathbf{X} \in A] = \int_A \cdots \int f(x_1, \ldots x_n)\, dx_1\, dx_2 \ldots dx_n\,.$$

Multivariate Random Variables: Independence

- More generally, a sequence of random variables $X_1, X_2, \ldots, X_n$ are **(mutually) independent** if and only if

$$P(X_1 \in A_1, \ldots, X_n \in A_n) = P(X_1 \in A_1) \cdots P(X_n \in A_n).$$

  for all sequence of events $A_1, A_2, \ldots$

- Equivalently, the random variables $X_1, X_2, \ldots, X_n$ are **(mutually) independent** if and only if their joint distributions factors into the product of their marginal distributions.

$$f(x_1, x_2, \ldots, x_n) = f_1(x_1) f_2(x_2) \ldots f_n(x_n)$$

  for all $x_1, x_2, \ldots, x_n$.

Linear Combinations of Random Variables

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

$$a_1X_1 + a_2X_2 + \cdots + a_nX_n,$$

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 $X_1, X_2, \ldots, X_n$ with all $a_i = 1$.
- 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 $a_i = 1/n$.
- The difference $X - Y$ is a linear combination of X and Y with $a_1 = 1$, $a_2 = -1$.

Example: Linear Combinations

Suppose the bus fare is

\$2 for senior citizens, \$1 for children, and \$3 for all other people

Let

$$X = \text{the number of senior citizens on the bus,}$$
$$Y = \text{the number of children on the bus,}$$
$$Z = \text{the number of all other passengers on the bus}$$

The total amount of bus fares collected is then

$$2X + Y + 3Z$$

which is a linear combination of X, Y, Z.

Expected Values for Linear Combinations of RV's

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

Variance of a Linear Combination of RV's

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

Example: Expected Value and Variance of Binom(n,p)

Recall that, if $X \sim \text{Binom}(n, p)$, we have

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

We are now ready to prove these formulas using linear combinations.

Recall a Binomial random variable $X \sim \text{Binom}(n, p)$ is the total number of successes obtained in n independent Bernoulli trials. The expected value and variance of X are thus

$$\mathrm{E}(X) = \underbrace{\mathrm{E}(X_1)}_{=p} + \underbrace{\mathrm{E}(X_2)}_{=p} + \cdots + \underbrace{\mathrm{E}(X_n)}_{=p} = np$$

$$\mathrm{Var}(X) = \underbrace{\mathrm{Var}(X_1)}_{=p(1-p)} + \underbrace{\mathrm{Var}(X_2)}_{=p(1-p)} + \cdots + \underbrace{\mathrm{Var}(X_n)}_{=p(1-p)} = np(1-p)$$