STA 103 Lecture 4: Joint Distributions

Instructor: Wookyeong Song

Department of Statistics, University of California, Davis

Aug 11th, 2025

Announcement

Midterm 1 Format Announcement

Hi all,

The midterm will be held next Wed from 12:10 pm - 1:00 pm (50 minutes). Below is the general information about the Midterm 1 format.

There will be **four Big problems** with each **3 sub problems.** We have a total of 12 subproblems, i.e., 1(a), 1(b), 1(c), 2(a), 2(b), 2(c), ... 4(a), 4(b), 4(c). The subproblems within the same main problem **may or may not** build on each other (e.g., (a) or (b) might or might not be used to solve (c)), but each main problem (1, 2, 3, 4) is **independent** from the others.

It means that the format (or style) will be similar to **Assignment 1**. However, the difficulty of Midterm 1 is adjusted so the exam can be completed in 50 minutes. (Assignment 1 was designed to take at least 3 hours, so you can expect each midterm problem to be more manageable.)

Details:

- Midterm 1 will cover Lecture 1 - 4.
- Each subproblems will be similar to one of (lecture note examples, assignment 1, and problem set 1).
- You will not encounter completely unfamiliar topics on the exam.
- Each question will have a clear, well-defined answer. No open-ended questions like ``give your thoughts," or ``discuss somethings."
- However, **showing your work is essential**: grader will check your derivations and may award partial credit even if your final answer is not correct.
- Likewise, you may not receive full credit if you just drop the answer without any derivation process.
- A printed formula sheet (with everything you need) will be provided at the start of the exam—**no need to print** your own.
- Please bring your **own calculator**.
- Please bring your **student ID**.

Tips for preparation:

I recommend mastering first both the material and the worked examples on the **lecture notes**. This will give you the solid foundation you need for the exams. But, to complete 12 subproblems in 50 minutes, you'll need not only a clear understanding of each concept but also enough practice to solve each problem in **about 4 minutes**. To build that speed, working through Assignment 1 and Problem Set 1 will be especially helpful.

Announcement

- To help you prepare for Midterm 1 next Wednesday, I will move my office hour next week only from Thursday (2:00–3:00 pm) to Tuesday (2:00–3:00 pm) in MSB 1143.

Dependence between two RVs X and Y

- **Motivation**: Suppose you roll two 6-sided dice, one red one blue. Let $X =$ the sum of the numbers showing for the red and blue die, $Y =$ the value of just the red die.

- **Question 1**: Suppose your betting that the red die is 1, i.e., $Y = 1$. What is the probability of winning?

- **Question 2**: Suppose you know the information that $X = 2$. What is the probability of $Y = 1$ given $X = 2$?

- **Question 3**: Suppose you know the information that $X = 7$. What is the probability of $Y = 1$ given $X = 7$?

Dependence between two RVs X and Y

- **Answer for Q2**: It turns out that if $X = 2$, then the adjusted PMF for Y is

$$p(y \mid X = 2) = \begin{cases} 1 & \text{if} \quad y = 1, \\ 0 & \text{if} \quad y = 0. \end{cases}$$

  (**Why?**)

- **Answer for Q3**: It turns out that if $X = 7$, then there is no adjustment of PMF for Y is

$$p(y \mid X = 7) = p(y) = 1/6, \quad \text{for } y = 1, 2, 3, 4, 5, 6.$$

  (**Why?**)

- **Message**: The fact that the conditional PMF of Y changes with different information on X indicates that X and Y are dependent random variables.

- More explicitly two random variables X, Y are dependent if knowing the value of one changes the PMF of the other.

Joint PMF for X and Y

- **Motivation**: We have seen that a PMF provides the blueprint for a single random variable X. However, when we have a pair of random variables (X, Y), it is no longer sufficient to use the PMF of each $p_X(x)$ (PMF of X), $p_Y(y)$ (PMF of Y) to characterize the randomness in the pair.

- **Definition 1**: The Joint Probability Mass Function (JPMF) $p(x, y)$ for a pair of random variables (X, Y), given in a table form is the following:

$X \backslash Y$	0	1	
0	$1/8$	0	
1	$2/8$	$1/8$	
2	0	$3/8$	
3	$1/8$	0	
			1

- **Example**: $p(3, 0) = P(X = 3, Y = 0) = \frac{1}{8}$.

Joint PMF for X and Y

- **Definition 2**: The Joint Probability Mass Function (JPMF) $p(x, y)$ for a pair of random variables (X, Y), given in a function form is the following:

$$p(x, y) = P(X = x \text{ and } Y = y).$$

- The input of a possible (X, Y) pair (x, y) and the output is the chance $P(X = x \text{ and } Y = y)$ that $X = x$ and $Y = y$.

Use the joint PMF to find any probability

- By summing the entries corresponding to the corresponding event. Box model is best illustrated with an example:

0	1	1	3	1	2	2	2
0	0	0	0	1	1	1	1

8 tickets

- Equivalently, we have a joint PMF in a table form:

$X \backslash Y$	0	1	
0	$1/8$	0	
1	$2/8$	$1/8$	
2	0	$3/8$	
3	$1/8$	0	
			1

Use the joint PMF to find any probability

- Suppose we have

$X \backslash Y$	0	1	
0	$1/8$	0	
1	$2/8$	$1/8$	
2	0	$3/8$	
3	$1/8$	0	
			1

- **Question 1**: What is $P(X = 2 \text{ and } Y = 1)$?
- **Question 2**: What is $P(X = 1 \text{ or } Y = 1)$?
- **Question 3**: What is $P(Y = 0)$?
- **Question 4**: What is $P(X = 1)$?

Use the joint PMF to find marginal PMF

- Notice that

$$P(X = x) = p_X(x) = \text{row sum } p(x, y) \text{ at row } x,$$
$$P(Y = y) = p_Y(y) = \text{column sum } p(x, y) \text{ at column } y.$$

- These PMFs: $p_X(x)$ and $p_Y(y)$ obtained from a joint PMF $p(x, y)$ are typically called the marginal PMFs for X and Y, respectively.

- **Example**:

$X \backslash Y$	0	1	$p_X(x)$
0	$1/8$	0	$1/8$
1	$2/8$	$1/8$	$3/8$
2	0	$3/8$	$3/8$
3	$1/8$	0	$1/8$
$p_Y(y)$	$4/8$	$4/8$	1

Use the joint PMF to find marginal PMF

- **Example**:

$X \backslash Y$	0	1	$p_X(x)$
0	$1/8$	0	$1/8$
1	$2/8$	$1/8$	$3/8$
2	0	$3/8$	$3/8$
3	$1/8$	0	$1/8$
$p_Y(y)$	$4/8$	$4/8$	1

- **Question**: What is $E(X)$, and what is $E(Y)$?

Marginal PMFs are **Not** Sufficient, We Need Joint PMF

- **Another example with same marginal PMFs**:

$X \backslash Y$	0	1	$p_X(x)$
0	$1/8$	0	1/8
1	$2/8$	$1/8$	3/8
2	$1/8$	$2/8$	3/8
3	0	$1/8$	1/8
$p_Y(y)$	4/8	4/8	1

- **Main message**: To provide a mathematical blueprint for characterizing the dependence between two RVs X and Y, we need joint PMF. The marginal PMFs alone is not sufficient.

Conditional Joint PMF

- Suppose X, Y have joint PMF:

$X \backslash Y$	0	1	$p_X(x)$
0	$1/8$	0	$1/8$
1	$2/8$	$1/8$	$3/8$
2	0	$3/8$	$3/8$
3	$1/8$	0	$1/8$
$p_Y(y)$	$4/8$	$4/8$	1

- For a given draw of (X, Y), suppose you told that $X \geq 1$ and $Y = 1$ but not the exact values of X and Y.

- **Question 1**: $P(X = 2 \mid X \geq 1 \text{ and } Y = 1)$?

Conditional Joint PMF

- Set the entries of $p(x, y)$ to zero that do not satisfy the given information, $X \geq 1$ and $Y = 1$.

$X \backslash Y$	0	1	
0	0	0	
1	0	$1/8$	
2	0	$3/8$	
3	0	0	
			$1/8 + 3/8 + 0 = 4/8$

- Re-normalize the entries to add to 1.

$X \backslash Y$	0	1	
0	0	0	
1	0	$(1/8)/(4/8)$	
2	0	$(3/8)/(4/8)$	
3	0	$0/(4/8)$	
			$(4/8)\ /\ (4/8)$

Conditional Joint PMF

The Joint PMF for $P\left(X=x, Y=y \mid X \geq 1 \text{ and } Y=1\right)$

$X \backslash Y$	0	1	$p_X\left(x \mid X \geq 1 \text{ and } Y=1\right)$
0	0	0	$0+0=0$
1	0	1/4	$0+1/4=1/4$
2	0	3/4	$0+3/4=3/4$
3	0	0	$0+0=0$
			1

- Getting back to the problem:

- **Answer for Question 1**:
$P(X=2 \mid X \geq 1 \text{ and } Y=1)=0+3/4=3/4.$

Conditional Joint PMF

- **Business example (customer behavior)**: a small online retailer tracks two variables across 100 recent customers.
  - X: Number of items purchased (0, 1, or 2)
  - Y: Whether the customer had a coupon ($0 =$ "No", $1 =$ "Yes").

- The joint distribution is summarized in the following table:

$X \backslash Y$	0 (No Coupon)	1 (Coupon)	
0	0.10	0.05	
1	0.20	0.25	
2	0.10	0.30	
			1

Conditional Joint PMF

- The joint distribution is summarized in the following table:

$X \backslash Y$	0 (No Coupon)	1 (Coupon)	
0	0.10	0.05	
1	0.20	0.25	
2	0.10	0.30	
			1

- Interpretation of marginal PMFs:
  - Marginal PMF of Y (Distribution of Coupon Usage): sum across rows:

$$p_Y(0) = 0.10 + 0.20 + 0.10 = 0.40$$
$$p_Y(1) = 0.05 + 0.25 + 0.30 = 0.60$$

  - Marginal PMF of X (Distribution of Item Purchased): sum across columns:

$$p_X(0) = 0.10 + 0.05 = 0.15$$
$$p_X(1) = 0.20 + 0.25 = 0.45$$
$$p_X(2) = 0.10 + 0.30 = 0.40$$

Computing $E\left(f(X,Y)\right)$

- Recall our shortcut for computing $E(f(X))$:

$$E(f(X)) = \sum_x f(x)p_X(x).$$

- A similar shortcut works with a joint PMF $p(x,y)$ for two RVs X and Y:

$$E(f(X,Y)) = \sum_{(x,y)} f(x,y)p(x,y),$$

$$E(f(X)) = \sum_{(x,y)} f(x)p_X(x),$$

$$E(f(Y)) = \sum_{(x,y)} f(y)p_Y(y).$$

- Let (X,Y) have the joint PMF:

$X \backslash Y$	0	1	
0	0.06	0.04	
1	0.54	0.36	
			1

- **Question**: Find $E(X+Y)$, $E(XY)$, and $\mathrm{Var}(XY)$.

Computing $E\left(f(X,Y)\right)$

Joint PMF

$X \backslash Y$	0	1	
0	0.06	0.04	
1	0.54	0.36	
			1

- Using our shortcut:

$$\begin{aligned} E(X+Y) &= \sum_{(x,y)} (x+y)p(x,y) \\ &= (0+0)\times 0.06 + (1+0)\times 0.54 \\ &+ (0+1)\times 0.04 + (1+1)\times 0.36. \end{aligned}$$

- Similarly,

$$\begin{aligned} E(XY) &= \sum_{(x,y)} (xy)p(x,y) \\ &= (0\cdot 0)\times 0.06 + (1\cdot 0)\times 0.54 \\ &+ (0\cdot 1)\times 0.04 + (1\cdot 1)\times 0.36. \end{aligned}$$

Computing $E\left(f(X,Y)\right)$

- We know that $\mathrm{Var}(XY) = E(X^2Y^2) - (E(XY))^2$. Then,

$$\begin{aligned} E(X^2Y^2) &= \sum_{(x,y)} (x^2y^2)p(x,y) \\ &= (0^2 \cdot 0^2) \times 0.06 + (1^2 \cdot 0^2) \times 0.54 \\ &+ (0^2 \cdot 1^2) \times 0.04 + (1^2 \cdot 1^2) \times 0.36. \end{aligned}$$

Then,

$$\mathrm{Var}(XY) = E(X^2Y^2) - (E(XY))^2.$$

Conditional Joint PMF

- Revisit the joint PMF for (X, Y):

$X \backslash Y$	0	1	$p_X(x)$
0	$1/8$	0	1/8
1	$2/8$	$1/8$	3/8
2	0	$3/8$	3/8
3	$1/8$	0	1/8
$p_Y(y)$	4/8	4/8	1

- For a given draw of (X, Y), suppose you told that $X \geq 1$ and $Y = 1$ but not the exact values of X and Y.

- **Question 2**: Your best prediction for X given $X \geq 1$ and $Y = 1$, i.e., $E(X \mid X \geq 1 \text{ and } Y = 1)$?

- **Question 3**: Your typical prediction error for X given $X \geq 1$ and $Y = 1$, i.e., $\mathrm{sd}(X \mid X \geq 1 \text{ and } Y = 1)$?

Conditional Joint PMF

The Joint PMF for $P\left(X=x, Y=y \mid X \geq 1 \text{ and } Y=1\right)$

$X \backslash Y$	0	1	$p_X\left(x \mid X \geq 1 \text{ and } Y=1\right)$
0	0	0	$0+0=0$
1	0	1/4	$0+1/4=1/4$
2	0	3/4	$0+3/4=3/4$
3	0	0	$0+0=0$
			1

- **Answer for Question 2**: Using the shortcut

$$
\begin{aligned}
E(X \mid X \geq 1 \text{ and } Y=1) &= \sum_x x p_X(x \mid X \geq 1 \text{ and } Y=1). \\
&= (0 \cdot 0)+(1 \cdot 1/4)+(2 \cdot 3/4)+(3 \cdot 0) \\
&= 7/4.
\end{aligned}
$$

Conditional Joint PMF

The Joint PMF for $P\left(X = x, Y = y \mid X \geq 1 \text{ and } Y = 1\right)$

$X \backslash Y$	0	1	$p_X\left(x \mid X \geq 1 \text{ and } Y = 1\right)$
0	0	0	0 + 0 =0
1	0	1/4	0 + 1/4 =1/4
2	0	3/4	0 + 3/4 = 3/4
3	0	0	0 + 0 = 0
			1

- **Answer for Question 3**: Using the shortcut

$$
\begin{aligned}
&\mathrm{Var}(X \mid X \geq 1 \text{ and } Y = 1) \\
&= E(X^2 \mid X \geq 1 \text{ and } Y = 1) - \left(E(X \mid X \geq 1 \text{ and } Y = 1)\right)^2 \\
&= \sum_x x^2 p_X(x \mid X \geq 1 \text{ and } Y = 1) - \left(\sum_x x p_X(x \mid X \geq 1 \text{ and } Y = 1)\right)^2. \\
&= (0^2 \cdot 0) + (1^2 \cdot 1/4) + (2^2 \cdot 3/4) + (3^2 \cdot 0) - (7/4)^2 \\
&= 13/4 - (7/4)^2 = 3/16.
\end{aligned}
$$

Conditional Joint PMF

The Joint PMF for $P\left(X=x, Y=y \mid X \geq 1 \text{ and } Y=1\right)$

$X \backslash Y$	0	1	$p_X\left(x \mid X \geq 1 \text{ and } Y=1\right)$
0	0	0	$0+0=0$
1	0	1/4	$0+1/4=1/4$
2	0	3/4	$0+3/4=3/4$
3	0	0	$0+0=0$
			1

- **Answer for Question 3 (Continue)**: Then we have

$$\begin{aligned}
\mathrm{sd}(X \mid X \geq 1 \text{ and } Y=1) &= \sqrt{\operatorname{Var}(X \mid X \geq 1 \text{ and } Y=1)} \\
&= \sqrt{\frac{3}{16}} = \frac{\sqrt{3}}{4}.
\end{aligned}$$

Independent Random Variables

The **formal** definition of independence for two random variables is that

- $p_X(x \mid Y = y) = p_X(x)$ for all possible x, y, or
- $p_Y(y \mid X = x) = p_Y(y)$ for all possible x, y.

- Here is an example of a pair X, Y such that knowing the value of Y does not change the PMF for X:

$X \backslash Y$	0	1	$p_X(x)$
0	0.06	0.04	0.1
1	0.54	0.36	0.9
			1

Independent Random Variables

Marginal PMF for X:

$X \backslash Y$	0	1	$p_X(x)$
0	0.06	0.04	0.1
1	0.54	0.36	0.9
			1

- Given $Y = 0$:

$X \backslash Y$	0	1	$p_X(x \mid Y = 0)$
0	0.06 / 0.6	0	0.1
1	0.54 / 0.6	0	0.9
			0.6 / 0.6

- Given $Y = 1$:

$X \backslash Y$	0	1	$p_X(x \mid Y = 1)$
0	0	0.04 / 0.4	0.1
1	0	0.36 / 0.4	0.9
			0.4 / 0.4

- Since $p_X(x \mid Y = y) = p_X(x)$ for all possible x, y, the randomness in X is unaffected by Y, then X and Y are independent.

Independent Random Variables

- But there is an easier way to check if X and Y are independent:

- **Fact**: Two RVs X and Y are independent whenever

$$p(x,y) = p_X(x)p_Y(y)$$

- Return to the example $p(x,y)$:

$X \backslash Y$	0	1	
0	0.06	0.04	
1	0.54	0.36	
			1

- The product of two PMFs $p_X(x)p_Y(y)$:

$X \backslash Y$	0	1	$p_X(x)$
0	0.6 $\times$ 0.1	0.4 $\times$ 0.1	0.1
1	0.6 $\times$0.9	0.4 $\times$ 0.9	0.9
$p_Y(y)$	0.6	0.4	1

- Note that two tables are **identical**.

Independent Random Variables

- **Summary**: Since the marginal PMFs are given by the row and column sums of the joint, this means X and Y are independent whenever each entry of the joint PMF equals the product of the corresponding row and column sum.

- If X and Y are **not** independent then they are said to be dependent.

Covariance

- **Motivation**: If two RVs X and Y are not independent (i.e., they are dependent), there are many ways to summarize how dependent they are. The most common is with covariance.

- **Definition**: If X and Y are two RVs then the covariance between X and Y is defined as

$$\mathrm{Cov}(X,Y) = E(XY) - E(X)E(Y) \quad \text{(easy to compute)},$$
$$\mathrm{Cov}(X,Y) = E\left((X - E(X))(Y - E(Y))\right) \quad \text{(easy to understand)}.$$

- **Interpretation**: It measures how much X and Y "co-vary together."

Properties of Covariance

- We will see that Cov is very closely related to Var. Indeed notice the similarity of the definitions:

$$\begin{aligned}\mathrm{Cov}(X,Y) &= E(XY) - E(X)E(Y)\\ \mathrm{Var}(X) &= E(X\cdot X) - E(X)E(X) = E(X^2) - (E(X))^2.\end{aligned}$$

- Notice therefore

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

- If X and Y are independent, then

$$\mathrm{Cov}(X,Y) = 0.$$

  However, $\mathrm{Cov}(X,Y) = 0$ does not imply the independence between X and Y.

Properties of Covariance

- **Formula**:

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y),$$
$$\mathrm{Var}(X-Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) - 2\mathrm{Cov}(X,Y).$$

- If X and Y are independent, then

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y),$$
$$\mathrm{Var}(X-Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

- Example:

$$\boxed{\begin{matrix}1\\2\end{matrix}} \quad \boxed{\begin{matrix}2\\1\end{matrix}} \quad \boxed{\begin{matrix}0\\1\end{matrix}}$$

Suppose $\boxed{\begin{matrix}X_1\\Y_1\end{matrix}}$ and $\boxed{\begin{matrix}X_2\\Y_2\end{matrix}}$ denote these two random tickets chosen with replacement from a box containing the following 3 tickets.

Properties of Covariance

- Selecting tickets with replacement implies that $\boxed{\begin{matrix} X_1 \\ Y_1 \end{matrix}}$ and $\boxed{\begin{matrix} X_2 \\ Y_2 \end{matrix}}$ are independent.

- **Question 1 (HW)**: Find $\mathrm{Var}(X_1 - Y_1)$.
  Hint: Let $W = X_1 - Y_1$, then generate PMF function of W in a table form:

w	$p(w)$
$\vdots$	$\vdots$

- **Question 2 (HW)**: Find $\mathrm{Var}(X_1 - Y_2)$.
  Hint: $\mathrm{Var}(X_1 - Y_2) = \mathrm{Var}(X_1) + \mathrm{Var}(Y_2) - 2\mathrm{Cov}(X_1, Y_2)$. Note that X_1 and Y_2 are independent, so that $\mathrm{Cov}(X_1, Y_2) = 0$. Then we have

$$\mathrm{Var}(X_1 - Y_2) = \mathrm{Var}(X_1) + \mathrm{Var}(Y_2).$$

  The problem boils down to find marginal PMFs of X_1 and Y_2.

Properties of Covariance

- Selecting tickets with replacement implies that $\boxed{\begin{matrix} X_1 \\ Y_1 \end{matrix}}$ and $\boxed{\begin{matrix} X_2 \\ Y_2 \end{matrix}}$ are independent.

- **Question 3 (HW)**: Are X_1 and Y_1 independent?

  Hint: To determine the independence, we need to create joint PMF of X and Y and check each entry of the joint PMF equals the product of the corresponding row and column sum (Lecture Note page $17-18$).