Joint, conditional, and marginal distributions

David J. H. Shih

Intended learning outcomes

- Apply definitions and theorems regarding joint, conditional, and marginal distributions.
- Recognize and explain Simpson's paradox

Random vector

Definition

An n-dimensional **random vector** is a function from a sample space $\mathcal{S}$ to n-dimensional Euclidean space $\mathcal{R}$^N .

Joint probability mass function

Definition

Given a discrete bivariate random vector (X, Y), the joint probability mass function (pmf) is defined by

$$f_{X,Y}(x, y) \triangleq P_{X,Y}(X = x, Y = y).$$

Properties

A pmf $f_{X,Y}(x, y)$ satisfies

$$f_{X,Y}(x, y) \geq 0 \quad \forall\,(x, y) \in \mathcal{R}^2 \quad \text{and} \quad \sum_{(x,y)\in\mathcal{R}^2} f_{X,Y}(x, y) = 1.$$

Support

$f_{X,Y} : \mathcal{R} \times \mathcal{R} \to [0, 1]$ but we only defined $P_{X,Y}$ for $x \in \mathcal{X} \subseteq \mathcal{R}$ and $y \in \mathcal{Y}$.

Definition

The **support** of a distribution $f_X(x)$ is

$$\mathcal{X} = \{x : f_X(x) > 0\}.$$

Therefore, $f_{X,Y} = 0$ for $x \notin \mathcal{X}$ or $y \notin \mathcal{Y}$.

Marginal probability mass function

Definition

The marginal pmfs of random vector (X, Y) are defined by

$$f_X(x) \triangleq P_X(X = x) \quad f_Y(y) \triangleq P_Y(Y = y)$$

Theorem

Given a discrete random vector (X, Y) with joint pmf $f_{X,Y}(x, y)$, the marginal pmfs of $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ are given by

$$f_X(x) = \sum_{y \in \mathcal{Y}} f_{X,Y}(x, y) \quad f_Y(y) = \sum_{x \in \mathcal{X}} f_{X,Y}(x, y)$$

This theorem follows from the law of total probability.

Total law of probability

Given sample space $\mathcal{S}$, $\mathcal{A} \subseteq \mathcal{S}$,

$$P\left(\mathcal{A}\right) = \sum_i P\left(\mathcal{A} \cap \mathcal{B}_i\right)$$

where $\mathcal{B}_1, \mathcal{B}_2, \ldots \subseteq \mathcal{S}$ is a partition of $\mathcal{S}$, which is defined by

$$\mathcal{B}_i \cap \mathcal{B}_j = \emptyset \quad \forall i \neq j \quad \text{and} \quad \bigcup_i^{\infty} \mathcal{B}_i \quad = \mathcal{S}$$

Proof

It follows from set theory and Kolmogorov's probability axioms.

$$\mathcal{A} = \mathcal{A} \cap \mathcal{S} = \mathcal{A} \cap \left(\bigcup_i \mathcal{B}_i\right) = \bigcup_i \mathcal{A} \cap B_i$$

$$P\left(\mathcal{A}\right) = P\left(\bigcup_i \mathcal{A} \cap B_i\right) = \sum_i P\left(\mathcal{A} \cap \mathcal{B}_i\right) \quad \text{(additivity axiom)} \quad \blacksquare$$

Proof: Marginal pmf

Define $\mathcal{B}_y = \{s \in \mathcal{S} : Y(s) = y\}$.

Since Y is a map from $\mathcal{S}$ to $\mathcal{Y}$, there exists some y for every $s \in \mathcal{S}$. Then, $\bigcup_{y \in \mathcal{Y}} \mathcal{B}_y = \mathcal{S}$. Therefore, $\mathcal{B}_1, \mathcal{B}_2, \ldots$ is a partition of $\mathcal{S}$.

$$
\begin{aligned}
f_X(x) &= P_X(X = x) \\
&= P(\{s \in \mathcal{S} : X(s) = x\}) \\
&= \sum_{y \in \mathcal{Y}} P\left(\{s \in \mathcal{S} : X(s) = x\} \cap \mathcal{B}_y\right) \quad \text{(Law of total prob)} \\
&= \sum_{y \in \mathcal{Y}} P_{X,Y}(X = x, Y = y) = \sum_{y \in \mathcal{Y}} f_{X,Y}(x, y)
\end{aligned}
$$

The proof for $f_Y(y)$ follows similarly as above. $\blacksquare$

Marginalization as sweeping

Joint probability density function

Definition

Given a *continuous* random vector (X, Y), a joint probability density function (pdf) is a function $f_{X,Y}(x, y)$ such that, for every subset $\mathcal{A} \subset \mathcal{R}^2$,

$$P((X, Y) \in \mathcal{A}) = \int\int_{\mathcal{A}} f_{X,Y}(x, y)\, dx\, dy.$$

The notation $\int\int_{\mathcal{A}}$ means that the limits of integration are set so that the function is integrated over all $(x, y) \in \mathcal{A}$.

Properties

A pdf $f_{X,Y}(x, y)$ defined above also satisfies

$$f(x, y) \geq 0 \quad \forall\, (x, y) \in \mathcal{R}^2 \quad \text{and} \quad \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)\, dx\, dy = 1.$$

Marginal probability density function

Theorem

Given a *continuous* random vector (X, Y) with joint pdf $f_{X,Y}(x, y)$, the marginal pdf of X and Y are given by

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy \quad x \in \mathcal{X},$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\, dx \quad y \in \mathcal{Y}.$$

Conditional probability

Definition

Given events A and B, if $P(B) > 0$, then

$$P(A \mid B) \triangleq \frac{P(A \cap B)}{P(B)}.$$

Conditional probability distributions

Definition

Given a discrete (or continuous) random vector (X, Y) with joint pmf (or pdf) $f_{X,Y}(x, y)$ and marginal pmfs (or pdfs) $f_X(x)$ and $f_Y(y)$, for any x such that $f_X(x) > 0$, the conditional pmf (or pdf) of Y given that $X = x$ is defined by

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

Similarly, for any y such that $f_Y(y) > 0$,

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

Conditioning as slicing

Chain rule of probability

By re-arranging definition of conditional probability, we have

$$P\left(A_1 \cap A_2\right) = P\left(A_1 \mid A_2\right) P\left(A_2\right)$$

Of course, we can apply the definition of $P\left(A_2 \mid A_1\right)$ as well:

$$P\left(A_1 \cap A_2\right) = P\left(A_2 \mid A_1\right) P\left(A_1\right).$$

Chain rule of probability

If we have a sequence of J events, $A_1, A_2, \dots A_J$ in some *arbitrary order*, we can keep applying the definition of conditional probability.

$$
\begin{aligned}
&P\left(A_1 \cap A_2 \cap \dots A_J\right) \\
&= P\left(A_1 \mid A_2, A_3, \dots, A_J\right) P\left(A_2 \cap A_3 \cap \dots A_J\right) \\
&= P\left(A_1 \mid A_2, A_3, \dots, A_J\right) P\left(A_2 \mid A_3 \cap \dots A_J\right) P\left(A_3 \cap A_4 \cap \dots A_J\right) \\
&= \dots
\end{aligned}
$$

This then gives the **chain rule**:

$$P\left(\bigcap_{j=1}^{J} A_j\right) = \prod_{j=1}^{J} P\left(A_j \,\middle|\, \bigcap_{k=1}^{j-1} A_k\right).$$

This applies to probablity distributions as well.

Conditioning direction

Typically, we try to define the conditional probabilities based on our assumptions about the causal relationships If we believe A causes $B_1, B_2, \ldots B_J$, then in our model, it would be easier to define $P(B_1 \mid A), \ldots, P(B_J \mid A)$.

We can also choose to define $P(A \mid B_1, B_2, \ldots B_J)$ instead, but this can make the derivations more difficult if it is inconsistent with the underlying causal relationship.

Independence

Definition

Given random vector (X, Y) with joint pmf (or pdf) $f_{X,Y}(x, y)$ and marginal pmfs (or pdfs) $f_X(x)$ and $f_Y(x)$, X and Y are **independent** if and only if (iff), for every $x \in \mathcal{X}$ and $y \in \mathcal{Y}$,

$$f_{X,Y}(x, y) = f_X(x) f_Y(y).$$

If X are Y are independent, we write $X \perp Y$.

$X \perp Y \;\Rightarrow\; f_{Y|X}(y \mid x) = f_Y(y)$ for every $x \in \mathcal{X}$ and $y \in \mathcal{Y}$.

Lemma

Given random vector (X, Y), X and Y are independent if and only if (iff) there exists functions $g(x)$ and $h(y)$ such that, for every $x \in \mathcal{X}$ and $y \in \mathcal{Y}$,

$$f_{X,Y}(x, y) = g(x) h(y).$$

Simpson's paradox

$$P(Y = 1 \mid X = x_1) > P(Y = 1 \mid X = x_2).$$

However, conditioned on $Z = z$ for some z,

$$P(Y = 1 \mid X = x_1, Z = z) < P(Y = 1 \mid X = x_2, Z = z).$$

Example: Gender bias?

Y represents undergraduate admission. X represents gender.
Z represents department.

$$P\left(Y=1 \mid X=x_1\right) > P\left(Y=1 \mid X=x_2\right).$$

Table 1: Admission rates by gender

All	x_1	x_2
41%	**44%**	35%

$$P\left(Y=1 \mid x_1, z\right) < P\left(Y=1 \mid x_2, z\right), \ z \in \{1,2,4,6\}.$$
$$P\left(Y=1 \mid x_1, z\right) > P\left(Y=1 \mid x_2, z\right), \ z \in \{3,5\}.$$

Table 2: Admission rates by gender and department

Department	All	x_1	x_2
1	64%	62%	**82%**
2	63%	63%	**68%**
3	35%	**37%**	34%
4	34%	33%	**35%**
5	25%	**28%**	24%
6	6%	6%	**7%**
...			
All	41%	**44%**	35%

$$P\left(Y=1 \mid x_1, z\right) < P\left(Y=1 \mid x_2, z\right),\ z \in \{1,2,4,6\}.$$
$$P\left(Y=1 \mid x_1, z\right) > P\left(Y=1 \mid x_2, z\right),\ z \in \{3,5\}.$$

Table 3: Number of applicants by gender and department

Department	All	x_1	x_2
1	933	**825**	108
2	585	**560**	25
3	918	325	**593**
4	792	**417**	375
5	584	191	**393**
6	714	**373**	341
...			
All	12763	8442	4321

Causal effects?

$$P(Y = 1 \mid X = x_1) > P(Y = 1 \mid X = x_2) \quad \Rightarrow \quad X \not\perp Y.$$

Suppose that we know that Y does *not* affect X.

Does this mean that X affects Y?

Summary

Symbolic logic $\to$ Set theory $\to$ Probability theory
$\to$ Measure theory $\to$ Statistics $\to$ Data modelling

Casella & Berger 2002, sections 4.1-4.2.

Intended learning outcomes

- Apply definitions and theorems regarding joint, conditional, and marginal distributions.
- Recognize and explain Simpson's paradox