Multivariate Distribution – Continuous

August 10, 2020
Questions like the following involve relationships between two variables in a population

“Is wage related to job experience, or age?”

“Does a woman whose husband earns more also have higher wages herself?”

- A (random) variable $X$ is **continuous** if its values fall within an interval $(a, b)$

- When $X$ and $Y$ are both continuous random variables, $(X, Y)$ is called a bivariate continuous random variable

To study these questions:

1. We collect data $(X_1, Y_1), ..., (X_n, Y_n)$ on $(X, Y)$
2. Using the data, we build a model to explain possible relationships between $X$ and $Y$ in the population
3. The model is a continuous bivariate probability distribution
Relationship of wages between spouses

**Women’s wage data**

University of Michigan Panel Study of Income Dynamics on 753 white married women in the US (1975-76):

<table>
<thead>
<tr>
<th>Woman</th>
<th>Workforce status (1=Yes, 0=No)</th>
<th>⋮</th>
<th>Hourly wage rate</th>
<th>Husband’s wage rate</th>
<th>⋮</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>⋮</td>
<td>3.3540</td>
<td>4.0288</td>
<td>⋮</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>⋮</td>
<td>1.3889</td>
<td>8.4416</td>
<td>⋮</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>⋮</td>
<td>4.5455</td>
<td>3.5807</td>
<td>⋮</td>
</tr>
<tr>
<td>⋮</td>
<td>⋮</td>
<td>⋮</td>
<td>⋮</td>
<td>⋮</td>
<td>⋮</td>
</tr>
<tr>
<td>428</td>
<td>1</td>
<td>⋮</td>
<td>4.0816</td>
<td>6.5844</td>
<td>⋮</td>
</tr>
<tr>
<td>429</td>
<td>0</td>
<td>⋮</td>
<td>0.0000</td>
<td>7.8529</td>
<td>⋮</td>
</tr>
<tr>
<td>⋮</td>
<td>⋮</td>
<td>⋮</td>
<td>⋮</td>
<td>⋮</td>
<td>⋮</td>
</tr>
<tr>
<td>753</td>
<td>0</td>
<td>⋮</td>
<td>0.0000</td>
<td>6.0897</td>
<td>⋮</td>
</tr>
</tbody>
</table>

- We use $X$ and $Y$, respectively, for a woman’s and her husband’s wage
- There are $n = 428$ pairs of observations of $(X, Y)$
We summarise the data using:

(a) 2 univariate histograms in the margins for $X$ and $Y$ individually
(b) 1 bivariate histogram for $(X, Y)$ – relationship between $X$ and $Y$
Continuous bivariate probability model

- We need a bivariate distribution for $(X, Y)$.
- The distribution helps to explain:
  1. The sample as summarised in the histogram.
  2. The population from which the sample is drawn.
- The distribution is given by a bivariate (joint) probability density function $f(x, y)$. 

![Graph](image.png)
(a) Histogram and (b) proposed pdf, $f(x)$, look similar

$$f(x) = \begin{cases} \frac{x}{4} e^{-x/2}, & 0 < x \\ 0, & \text{otherwise} \end{cases}$$
Univariate probability model (2)

\[ F(x) = \int_0^x f(x) \, dx = \int_0^x \frac{x}{4} e^{-x/2} \, dx. \]

Using integration by parts:

\[ \int u \frac{dv}{dx} \, dx = uv - \int v \frac{du}{dx} \, dx. \]
We let \( u = \frac{x}{2} \), so \( \frac{du}{dx} = \frac{1}{2} \); \( \frac{dv}{dx} = \frac{1}{2} e^{-x/2} \), so \( v = -e^{-x/2} \).

\[
\int_0^x \frac{x}{4} e^{-x/2} \, dx = \left[ -\frac{x}{2} e^{-x/2} \right]_0^x + \int_0^x \frac{1}{2} e^{-x/2} \, dx
= \left[ -\frac{x}{2} e^{-x/2} \right]_0^x + \left[ -e^{-x/2} \right]_0^x
= \left[ -(x/2 + 1)e^{-x/2} \right]_0^x
= -(x/2 + 1)e^{-x/2} - (-1)
= 1 - (x/2 + 1)e^{-x/2}, \quad 0 < x.
\]
A probability model for wages between a woman and her husband

- For (women’s) wages, we use \( f(x) = \frac{x}{4} e^{-x/2} \)
- Model for both wages: \( f(x, y) = \begin{cases} 
\left( \frac{x}{4} e^{-x/2} \right) \left( \frac{y}{4} e^{-y/2} \right), & 0 < x, y \\
0, & \text{otherwise}
\end{cases} \)

Probabilities are volumes under \( f(x, y) \), found by double integration
It saves time if we obtain the (joint) cdf \( F(x, y) = P(X \leq x, Y \leq y) \)

\[
F(x, y) = \int_0^x \int_0^y f(x, y) \, dy \, dx \\
= \int_0^x \int_0^y \left( \frac{x}{4} e^{-x/2} \right) \left( \frac{y}{4} e^{-y/2} \right) \, dy \, dx \\
= \left[ \frac{x}{4} e^{-x/2} \right]_0^x \left[ \frac{y}{4} e^{-y/2} \right]_0^y \\
= \left[ 1 - (x/2 + 1)e^{-x/2} \right] \left[ 1 - (y/2 + 1)e^{-y/2} \right], \quad 0 < x, y.
\]
Some calculations

\[ P(X \leq 3, Y \leq 4) = F(3, 4) \]
\[ = \left[ 1 - \left( \frac{3}{2} + 1 \right) e^{-\frac{3}{2}} \right] \left[ 1 - \left( \frac{4}{2} + 1 \right) e^{-\frac{4}{2}} \right] \]
\[ \approx 0.263 \]

\[ P(2 < X < 4, Y \leq 4) = P(X < 4, Y \leq 4) - P(X < 2, Y \leq 4) \]
\[ = F(4, 4) - F(2, 4) \]
\[ = \left[ 1 - \left( \frac{4}{2} + 1 \right) e^{-\frac{4}{2}} \right] \left[ 1 - \left( \frac{4}{2} + 1 \right) e^{-\frac{4}{2}} \right] - \left[ 1 - \left( \frac{2}{2} + 1 \right) e^{-\frac{2}{2}} \right] \left[ 1 - \left( \frac{4}{2} + 1 \right) e^{-\frac{4}{2}} \right] \]
\[ \approx 0.196 \]

\[ P(X > 4, Y \leq 4) = P(X < \infty, Y \leq 4) - P(X \leq 4, Y \leq 4) \]
\[ = F(\infty, 4) - F(4, 4) \]
\[ = \left[ 1 - \left( \frac{\infty}{2} + 1 \right) e^{-\frac{\infty}{2}} \right] \left[ 1 - \left( \frac{4}{2} + 1 \right) e^{-\frac{4}{2}} \right] - \left[ 1 - \left( \frac{4}{2} + 1 \right) e^{-\frac{4}{2}} \right] \left[ 1 - \left( \frac{4}{2} + 1 \right) e^{-\frac{4}{2}} \right] \]
\[ \approx 0.241 \]
Marginal probabilities

- We might wish to study one of the variables, say $Y$, using pdf $f(x, y)$.
- When we study $Y$, we are not interested in $X$ at all.

\[
P(Y \in (5, 10)) \equiv P\{X = \text{any value, } Y \in (5, 10)\} \\
\equiv P\{X \in (a, b), Y \in (5, 10)\} \\
= \int_{5}^{10} \int_{a}^{b} f(x, y)\,dx\,dy \\
= \int_{5}^{10} f(y)\,dy.
\]

- $f(y) = \int_{a}^{b} f(x, \text{fixed } y)\,dx$ is the marginal pdf of $Y$ at $y$. 

August 10, 2020 9 / 14
Another way of finding marginal probabilities

- Finding marginal probabilities using joint pdf can be tedious:

\[
P(Y \in (5, 10)) = \int_{5}^{10} \int_{0}^{\infty} f(y) \left( \frac{x}{4} e^{-x/2} \right) \left( \frac{y}{4} e^{-y/2} \right) dx \, dy
\]

- If we know the joint cdf, we can try

\[
F(y) \equiv P(Y \leq y) = P(\text{don’t care about } X, Y \leq y) \\
= P(X \leq \infty, Y \leq y) \\
\equiv F(\infty, y).
\]

\[
F(x, y) = \left[ 1 - \left( \frac{x}{2} + 1 \right)e^{-x/2} \right] \left[ 1 - \left( \frac{y}{2} + 1 \right)e^{-y/2} \right]
\]

\[
\Rightarrow F(y) = F(\infty, y) = \left[ 1 - \left( \infty/2 + 1 \right)e^{-\infty/2} \right] \left[ 1 - \left( y/2 + 1 \right)e^{-y/2} \right]
\]

\[
= \left[ 1 - \left( y/2 + 1 \right)e^{-y/2} \right]
\]

\[
\Rightarrow P(5 \leq Y \leq 10) = P(Y \leq 10) - P(Y \leq 5) \\
= F(10) - F(5) \\
= \left[ 1 - \left( 10/2 + 1 \right)e^{-10/2} \right] - \left[ 1 - (5/2 + 1)e^{-5/2} \right] \approx 0.247
\]
In our model, the pdf can be “factorised” as

\[ f(x, y) = \left( \frac{x}{4} e^{-x/2} \right) \left( \frac{y}{4} e^{-y/2} \right) = g(x)h(y), \quad g(x), h(y) \geq 0 \]

We call \( X, Y \) independent. Independence leads to simple analyses

\[ P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y) \]
\[ \Leftrightarrow F(x, y) = F(x)F(y) \Rightarrow G(x)H(y), \quad \forall G, H > 0 \]
\[ \Leftrightarrow f(x, y) = f(x)f(y) \]
\[ \Leftrightarrow \int \int f(x, y) \, dx \, dy = \int f(x) \, dx \int f(y) \, dy \]

Independence is a strong assumption that must be supported by evidence (theoretical or empirical) and not merely based on convenience.
Conditional probabilities can be used to answer questions such as
“The chance of a woman with low wage if her husband has low wage”

When $X$ and $Y$ are independent, calculations are simple because conditional probabilities become unconditional (marginal) probabilities, *eg.*

\[
P(\text{woman earns } \leq 5 \mid \text{husband earns } \leq 3) = P(X \leq 5 \mid Y \leq 3) \\
= P(X \leq 5) \\
= F(5) \\
= 1 - (5/2 + 1)e^{-5/2} \\
\approx 0.713
\]

\[
P(\text{woman earns } \leq 5 \mid \text{husband earns } = 3) = P(X \leq 5 \mid Y = 3) \\
= P(X \leq 5) \\
\approx 0.713
\]
When $X$ and $Y$ are not independent, we may be interested in

1. $X$ in an interval given $Y$ in an interval:

$$P(X \leq x|Y \leq y) = \frac{P(X \leq x, Y \leq y)}{P(Y \leq y)} = \frac{F(x, y)}{F(y)}$$

2. $X$ in an interval given $Y = y$. This is well defined because we are not calculating probability about $Y$. A heuristic (non-vigorous) approach ...

$$P(X \leq x|Y = y) = \frac{\int_{-\infty}^{x} f(x, y)\,dx}{\int_{-\infty}^{\infty} f(x, y)\,dx} = \left(\frac{\int_{-\infty}^{\infty} f(x, y)\,dx}{f(y)}\right) = \int_{-\infty}^{x} \frac{f(x, y)}{f(y)}\,dx \equiv F(x|y)$$

$$\frac{f(x, y)}{f(y)} \equiv f(x|y)$$

and $F(x|y)$ are, respectively, the conditional pdf and cdf of $X$ given $Y = y$. 
Find $P \left( Y \leq \frac{1}{4} \mid X = \frac{1}{3} \right)$ if $f(x, y) = \begin{cases} 6(x - y), & 0 \leq y \leq x \leq 1 \\ 0, & \text{otherwise} \end{cases}$.

\[
f(x) = \int_0^x 6(x - y) \, dy = 6 \left[ xy - \frac{y^2}{2} \right]_0^x = 6 \left[ x^2 - \frac{x^2}{2} \right] = 3x^2,
\]

\[
f(y) = \int_y^1 6(x - y) \, dx = 6 \left[ \frac{x^2}{2} - xy \right]_y^1 = 6 \left[ \frac{1}{2} - y \right] - 6 \left[ \frac{y^2}{2} - y^2 \right] = 3(1 - y)^2
\]

\[
f(y \mid x) = \frac{f(x, y)}{f(x)} = \frac{6(x - y)}{3x^2}, 0 \leq y \leq x \leq 1
\]

\[
P \left( Y \leq \frac{1}{4} \mid X = \frac{1}{3} \right) = \int_{-\infty}^{\frac{1}{4}} \frac{6\left(\frac{1}{3} - y\right)}{3\left(\frac{1}{3}\right)^2} \, dy
\]

\[
= \int_0^{\frac{1}{4}} \frac{6\left(\frac{1}{3} - y\right)}{3\left(\frac{1}{3}\right)^2} \, dy
\]

\[
= 18 \int_0^{\frac{1}{4}} \left(\frac{1}{3} - y\right) \, dy
\]

\[
= 18 \left[ \frac{1}{3}y - \frac{1}{2}y^2 \right]_{\frac{1}{4}}^0
\]

\[
= \frac{15}{16}
\]