Random Vectors

Ricardo Dahis

PUC-Rio, Department of Economics

Summer 2023

Contents

Contents

random vector

- **definition**: an n-dimensional random vector is a function from the sample space S into the n-dimensional Euclidean space $\mathbb{R}^n$

- **example**: consider the experiment of tossing two fair dice, and let X and Y denote the sum of the two dice and the absolute difference of the two dice, respectively

$$\mathbb{P}(X = 5, Y = 3) \quad = \quad \mathbb{P}(\{(1,4),(4,1)\}) \quad = \quad \frac{2}{36} \quad = \quad \frac{1}{18}$$

- **definition**: let (X, Y) denote a discrete bivariate random vector, then the joint pmf $f_{X,Y}(x, y)$ from $\mathbb{R}^2$ into $\mathbb{R}$ is given by $f(x, y) = \mathbb{P}(X = x, Y = y)$

- we can now discuss probabilities of events defined in terms of (X, Y).

joint pmf

- the joint pmf completely characterizes the probability distribution of a random vector (X, Y) just as in the univariate case

$$\mathbb{P}\big((X, Y) \in A\big) \quad = \quad \sum_{(x,y)\in A} f_{X,Y}(x, y)$$

- expectations are defined

$$\mathbb{E}[g(X, Y)] \quad = \quad \sum_{(x,y)\in\mathbb{R}^2} g(x, y) f_{X,Y}(x, y)$$

- fortunately, the expectation operator continues to have the same properties as before; in particular

$$\mathbb{E}[a\, g(X, Y) + b\, h(X, Y) + c] \quad = \quad a\, \mathbb{E}[g(X, Y)] + b\, \mathbb{E}[h(X, Y)] + c$$

properties of joint pdfs

joint pmf satisfies the usual properties (**verify**), namely

(i) $f_{X,Y}(x, y) \geq 0$ for any (x, y)

(ii) $\sum_{(x,y)\in\mathbb{R}^2} f_{X,Y}(x, y) = 1$

and thus it is a well-defined probability distribution.

marginal pmfs

- there may be events, probabilities, moments or expectations that involve only one of the random variables in the vector.

- **theorem** (CB 4.1.6): let (X, Y) denote a discrete bivariate random vector with joint pmf $f_{X,Y}(x, y)$, then the **marginal pmfs** of X and Y are respectively

$$\begin{aligned} f_X(x) &= \mathbb{P}(X = x) = \sum_{y \in \mathbb{R}} f_{X,Y}(x, y) \\ f_Y(y) &= \mathbb{P}(Y = y) = \sum_{x \in \mathbb{R}} f_{X,Y}(x, y) \end{aligned}$$

  we use the subscript X in $f_X(x)$ to emphasize the distinction from $f_{X,Y}(x, y)$.

same marginals, different joint pmfs

- **same marginal pmfs $\not\Rightarrow$ same joint pmfs.**
- counterexample: define

$$
\begin{aligned}
f_{X,Y}(0,0) &= f_{X,Y}(0,1) = \frac{1}{6} \\
f_{X,Y}(1,0) &= f_{X,Y}(1,1) = \frac{1}{3} \\
f_{X,Y}(x,y) &= 0 \text{ for any other } (x,y)
\end{aligned}
$$

the marginals are

$$
\begin{aligned}
f_X(0) &= \frac{1}{3}, \quad f_X(1) = \frac{2}{3} \\
f_Y(0) &= \frac{1}{2}, \quad f_Y(1) = \frac{1}{2}
\end{aligned}
$$

same marginals, different joint pmfs

- counterexample (cont'd): now define

$$
\begin{aligned}
g_{XY}(0,0) &= \frac{1}{12} \quad g_{XY}(0,1) = \frac{3}{12} \\
g_{XY}(1,0) &= \frac{5}{12} \quad g_{XY}(1,1) = \frac{3}{12} \\
g_{XY}(x,y) &= 0 \text{ for any other } (x,y)
\end{aligned}
$$

  the marginals are

$$
\begin{aligned}
g_X(0) &= \frac{1}{3}, \quad g_X(1) = \frac{2}{3} \\
g_Y(0) &= \frac{1}{2}, \quad g_Y(1) = \frac{1}{2}
\end{aligned}
$$

- $f_X(0) = g_X(0)$, $f_X(1) = g_X(1)$, $f_Y(0) = g_Y(0)$, $f_Y(1) = g_Y(1)$ but $f_{X,Y}(x,y) \neq g_{XY}(x,y)$.

- intuitive since marginals contain less information than joint pmfs.

joint and marginal pdfs

- **definition**: a function $f_{X,Y}(x, y)$ from $\mathbb{R}^2$ into $\mathbb{R}$ is the joint pdf of the continuous bivariate random vector (X, Y) if, for every $A \subset \mathbb{R}^2$,

$$\mathbb{P}\big((X, Y) \in A\big) = \iint_A f_{X,Y}(x, y)\,\mathrm{d}x\,\mathrm{d}y$$

  - the joint pdf is such that $f_{X,Y}(x, y) \geq 0$ for all $(x, y) \in \mathbb{R}^2$ and that $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y)\,\mathrm{d}x\,\mathrm{d}y = 1$
  - expectations are just like in the discrete case, but with integrals

$$\mathbb{E}[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{X,Y}(x, y)\,\mathrm{d}x\,\mathrm{d}y$$

- **definition**: the marginal pdfs are given by (you can also verify that this distribution is proper)

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{X,Y}(x, y)\,\mathrm{d}y, \quad -\infty < x < \infty \\ f_Y(y) &= \int_{-\infty}^{\infty} f_{X,Y}(x, y)\,\mathrm{d}x, \quad -\infty < y < \infty \end{aligned}$$

example

- example: let (X, Y) denote a continuous bivariate random vector with joint pdf $f_{X,Y}(x, y) = 6xy^2$ for (x, y) in the unit square and zero otherwise.

$$
\begin{aligned}
\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x, y)\,\mathrm{d}x\,\mathrm{d}y &= \int_0^1\int_0^1 6xy^2\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_0^1 \left(3x^2y^2\right)_0^1 \mathrm{d}y = \int_0^1 3y^2\,\mathrm{d}y = \left(y^3\right)_0^1 = 1 \\
f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\,\mathrm{d}y &= \int_0^1 6xy^2\,\mathrm{d}y = 6x\left(y^3/3\right)_0^1 = 2x \\
f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\,\mathrm{d}x &= \int_0^1 6xy^2\,\mathrm{d}x = 6y^2\left(x^2/2\right)_0^1 = 3y^2
\end{aligned}
$$

example

- example (cont'd): let (X, Y) denote a continuous bivariate random vector with joint pdf $f_{X,Y}(x, y) = 6xy^2$ for (x, y) in the unit square and zero otherwise.
  - Consider now calculating the probability that $X + Y \geq 1$.
  - The region over which we integrate is

$$\begin{aligned} A &= \{(x, y) : x + y \geq 1, 0 < x < 1, 0 < y < 1\} \\ &= \{(x, y) : x \geq 1 - y, 0 < x < 1, 0 < y < 1\} \\ &= \{(x, y) : 1 - y \leq x < 1, 0 < x < 1, 0 < y < 1\} \end{aligned}$$

  - So

$$\mathbb{P}(X + Y \geq 1) = \int_0^1 \int_{1-y}^1 6xy^2 \, \mathrm{d}x \, \mathrm{d}y = 0.9$$

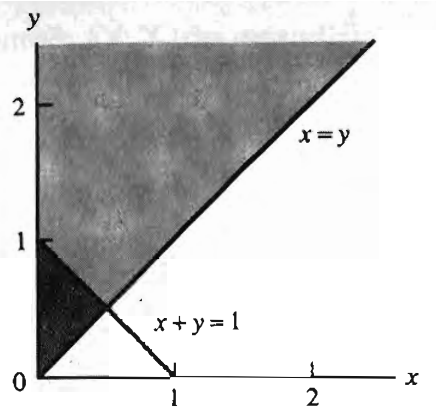
a more complicated example

- example 2: let (X, Y) denote a continuous bivariate random vector with joint pdf $f_{X,Y}(x, y) = e^{-y}$ for $0 < x < y < \infty$.

$$
\begin{aligned}
\mathbb{P}(X + Y \geq 1) &= 1 - \mathbb{P}(X + Y < 1) \\
&= 1 - \int_0^{1/2} \int_x^{1-x} e^{-y} \, \mathrm{d}y \, \mathrm{d}x \\
&= 1 - \int_0^{1/2} \left(e^{-x} - e^{-(1-x)} \right) \mathrm{d}x \\
&= 1 - \left(-e^{-\frac{1}{2}} + e^0 - e^{-\frac{1}{2}} + e^{-1} \right) \\
&= 2\, e^{-1/2} - e^{-1}
\end{aligned}
$$

given that $\Omega_{XY} = \{(x, y) : x + y \geq 1, 0 < x < y < \infty\}$ is the unbounded region with three sides given by $x = y$, $x + y = 1$, and $x = 0$

regions from the example



joint cdf

- the joint probability distribution of (X, Y) is also completely described with the joint cdf $F_{X,Y}(x, y) = \mathbb{P}(X \leq x, Y \leq y)$ for all $(x, y) \in \mathbb{R}^2$

- characterization: not very handy for discrete random vectors, but extremely useful for continuous random vectors given that

$$F_{X,Y}(x, y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f_{X,Y}(u, v)\,\mathrm{d}u\,\mathrm{d}v$$

  and hence, by the fundamental theorem of calculus,

$$\frac{\partial^2 F_{X,Y}(x, y)}{\partial x \partial y} = f_{X,Y}(x, y)$$

  at any continuity point of $f_{X,Y}(x, y)$

Contents

conditional probability

- **definition**: let (X, Y) denote a discrete bivariate random vector with joint pmf $f_{X,Y}(x, y)$ and marginals $f_X(x)$ and $f_Y(y)$, then the conditional pmf of Y given $X = x$ is

$$f_{Y|X}(y|x) \;=\; \mathbb{P}(Y = y|X = x) \;=\; \frac{f_{X,Y}(x, y)}{f_X(x)}$$

  for any x such that $f_X(x) = \mathbb{P}(X = x) > 0$

- just checking to be on the safe side...
  (i) $f_{Y|X}(y|x) \geq 0$ for every y given that $f_{X,Y}(x, y) \geq 0$ and $f_X(x) > 0$
  (ii) $\sum_y f_{Y|X}(y|x) = \frac{\sum_y f_{X,Y}(x,y)}{f_X(x)} = \frac{f_X(x)}{f_X(x)} = 1$

continuous random variables

- if X and Y are continuous random variables, then $\mathbb{P}(X = x) = 0$ for every value of x and hence we cannot divide the joint probability by the probability of the conditioning event

- however, we may still define the conditional probability of Y given $X = x$ analogously to the discrete case with pdfs replacing pmfs

- **definition**: let (X, Y) be a continuous bivariate random vector with joint pdf $f_{X,Y}(x, y)$ and marginals $f_X(x)$ and $f_Y(y)$, then the conditional pdf of Y given $X = x$ is

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

for any x such that $f_X(x) > 0$. Analogously,

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

for any y such that $f_Y(y) > 0$.

conditional expectation

- conditional pdfs/pmfs are useful not only to compute conditional probabilities, but also to calculate conditional expectations

$$\mathbb{E}[g(Y)|X=x] = \begin{cases} \sum_y g(y) f_{Y|X}(y|x) & \text{if discrete} \\ \int_{-\infty}^{\infty} g(y) f_{Y|X}(y|x)\,\mathrm{d}y & \text{if continuous} \end{cases}$$

- the conditional expectation satisfies all the properties of the usual expectation operator
- in particular, $\mathbb{E}(Y|X)$ provides the best guess at Y based on knowledge of X in a MSE sense (you can try to show this!)
- note that $f_{Y|X}(y|x)$ is function of x. So we really have a family of distributions, one for each x, possibly with different $\mathbb{E}(Y|X=x)$.
  - the notation $Y|X$ describes the entire family of distributions.

interesting example

- example: let (X, Y) have a joint pdf $f_{X,Y}(x, y) = e^{-y}$ for $0 < x < y < \infty$, then

$$
\begin{aligned}
f_X(x) &= \int_{-\infty}^{\infty} f_{X,Y}(x, y)\,\mathrm{d}y = \int_x^{\infty} e^{-y}\,\mathrm{d}y = e^{-x} \\
f_{Y|X}(y|x) &= \frac{f_{X,Y}(x, y)}{f_X(x)} = e^{-(y-x)} \text{ for } y > x
\end{aligned}
$$

and hence $X \sim \mathrm{Exp}(1)$ and $Y|X = x$ is also exponential with location parameter x

$$
\begin{aligned}
\mathbb{E}(Y|X = x) &= \int_x^{\infty} y f_{Y|X}(y|x)\,\mathrm{d}y = \int_x^{\infty} y e^{-(y-x)}\,\mathrm{d}y = 1 + x \\
\mathrm{var}(Y|X = x) &= \mathbb{E}(Y^2|X = x) - [\mathbb{E}(Y|X = x)]^2 \\
&= \int_x^{\infty} y^2 e^{-(y-x)}\,\mathrm{d}y - (1 + x)^2 = 1
\end{aligned}
$$

but why is it interesting?

- conditional variance does not depend on x, but does that mean it is equal to the unconditional variance?

$$f_Y(y) \;=\; \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}x \;=\; \int_0^y e^{-y}\,\mathrm{d}x \;=\; ye^{-y}$$

remember: the gamma distribution is given by

$$f(x|\alpha,\beta) \;=\; \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-\frac{x}{\beta}}, \quad \text{for}$$

for $0 < t < \infty$, $\alpha, \beta > 0$ and $\Gamma(\alpha) = (\alpha - 1)!$. Hence $Y \sim \mathrm{G}(\alpha, \beta)$, with $\alpha = 2$ and $\beta = 1$, implying that $\mathrm{var}(Y) = \alpha\beta^2 = 2$.

- even though the conditional variance does not depend on the value of x, knowledge of the latter considerably reduces the variability of Y

$$\mathrm{var}(Y|X = x) \;=\; c \;\not\Rightarrow\; \mathrm{var}(Y) \;=\; c$$

- we will come back to this point later

independence

- $\mathbb{E}[g(Y)|X]$ is a random variable whose values typically depend on the value of X, unless independent $(X \perp\!\!\!\perp Y)$

- definition: let (X, Y) denote a bivariate random vector with joint pdf/pmf $f_{X,Y}(x, y)$ and marginals $f_X(x)$ and $f_Y(y)$, then X and Y are independent if

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

- This immediately implies that $f_{Y|X}(y|x) = f_Y(y)$, since

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

and the knowledge of x does not inform the distribution of Y

a note of caution

- two pdfs that differ only a zero-measure set define the same probability distribution for (X, Y).
- so definition may fail to hold on sets with measure zero. But in this case X and Y are still independent.
- to see this, take $f_{X,Y}(x, y)$ and $f^*_{X,Y}(x, y)$ equal everywhere except on A for which $\int_A \int \mathrm{d}x\,\mathrm{d}y = 0$.
- let (X, Y) have pdf $f_{X,Y}(x, y)$, (X^*, Y^*) have pdf $f^*_{X,Y}(x, y)$, and $B \subset \mathbb{R}^2$. Then

$$\begin{aligned}
P((X, Y) \in B) &= \int_B \int f(x, y)\,\mathrm{d}x\,\mathrm{d}y = \int_{B\cap A^c} \int f(x, y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_{B\cap A^c} \int f^*(x, y)\,\mathrm{d}x\,\mathrm{d}y = \int_B \int f^*(x, y)\,\mathrm{d}x\,\mathrm{d}y \\
&= P((X^*, Y^*) \in B)
\end{aligned}$$

- for example, take $f_{X,Y}(x, y) = e^{-(x+y)}$ with $x, y > 0$, describing two independent exponential random variables.
- and take $f^*_{X,Y}(x, y) = f_{X,Y}(x, y)$ except that $f^*_{X,Y}(x, y) = 0$ if $x = y$ in $A = \{(x, x), x > 0\}$.

independence

- theorem (CB 4.2.7): let (X, Y) be a bivariate random vector with joint pdf $f(x, y)$. Then X and Y are independent if, and only if, there exist functions $g(x)$ and $h(y)$ such that, for every $x \in \mathbb{R}$ and $y \in \mathbb{R}$,

$$f(x, y) = g(x)h(y)$$

- proof ($\Rightarrow$): trivial setting $g(x) = f_X(x)$ and $h(y) = f_Y(y)$.
- proof ($\Leftarrow$): suppose that $f(x, y) = g(x)h(y)$ and define

$$\int_{-\infty}^{\infty} g(x)\,\mathrm{d}x = c \quad \text{and} \quad \int_{-\infty}^{\infty} h(y)\,\mathrm{d}y = d$$

so cd satisfies

$$\begin{aligned} cd &= \left(\int_{-\infty}^{\infty} g(x)\,\mathrm{d}x\right)\left(\int_{-\infty}^{\infty} h(y)\,\mathrm{d}y\right) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)\,\mathrm{d}x\,\mathrm{d}y = 1 \end{aligned}$$

- proof ($\Leftarrow$) (cont'd): the marginals are given by

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} g(x)h(y)\,\mathrm{d}y &= g(x)d \\ f_Y(y) &= \int_{-\infty}^{\infty} g(x)h(y)\,\mathrm{d}x &= h(y)c \\ &\Downarrow \\ f(x,y) &= g(x)h(y) = g(x)h(y)cd &= f_X(x)f_Y(y) \end{aligned}$$

establishing the desired result. ■

- example: Consider $f(x,y) = \frac{1}{384}x^2y^4e^{-y-\frac{x}{2}}$ with $x, y > 0$ and

$$g(x) = \begin{cases} x^2e^{-x/2} & x > 0 \\ 0 & x \le 0 \end{cases} \quad \text{and} \quad h(y) = \begin{cases} \frac{1}{384}y^4e^{-y} & y > 0 \\ 0 & y \le 0 \end{cases}$$

by theorem above, it follows immediately that X and Y are independent.

independence and support of the joint pdf

- the support set matters: independence can be ruled out in simple cases.
- denote the support of the marginals as $A = \{x : f_X(x) > 0\}$ and $B = \{y : f_Y(y) > 0\}$
- if X and Y independent, then $f(x, y) = f_X(x) f_Y(y) > 0$ on the set $\{(x, y) : x \in A, y \in B\}$
  - define $A \times B = \{(x, y) : x \in A, y \in B\}$, denoted cross-product set
  - if the set $\{(x, y) : f(x, y) > 0\}$ is not a cross-product, X and Y cannot be independent.
  - in one of the examples above, we have support set $0 < x < y < \infty$, so not only $0 < x, y < \infty$ but also $x < y$, so not independent.

independence and support of the joint pdf

- theorem (CB 4.2.10): let X and Y be independent variables
  (i) for any $A \subset \mathbb{R}$ and $B \subset \mathbb{R}$, $\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$. That is, the events $\{X \in A\}$ and $\{Y \in B\}$ are independent
  (ii) let $g(x)$ be a function of x and $h(y)$ be a function of y. Then

$$\mathbb{E}(g(X)h(Y)) = \mathbb{E}(g(X))\mathbb{E}(h(Y))$$

- proof (ii): Notice that

$$\begin{aligned}
\mathbb{E}(g(X)h(Y)) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)f(x,y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)f_X(x)f_Y(x)\,\mathrm{d}x\,\mathrm{d}y \\
&= \mathbb{E}(g(X))\mathbb{E}(h(Y))
\end{aligned}$$

independence and support of the joint pdf

- proof (i): Set $g(X) = 1(x \in A)$, $h(Y) = 1(y \in B)$. Notice that

$$
\begin{aligned}
\mathbb{P}(X \in A) &= \mathbb{E}\left[1(x \in A)\right] = \int_{-\infty}^{\infty} \mathcal{I}_A(x) f_X(x)\,\mathrm{d}x \\
\mathbb{P}(Y \in B) &= \mathbb{E}\left[1(y \in B)\right] = \int_{-\infty}^{\infty} \mathcal{I}_B(y) f_Y(y)\,\mathrm{d}y \\
\mathbb{P}(X \in A, Y \in B) &= \mathbb{E}\left[\mathcal{I}_{AB}(x, y)\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathcal{I}_{AB}(x, y) f_X(x) f_Y(y)\,\mathrm{d}x\,\mathrm{d}y
\end{aligned}
$$

and apply (ii). ■

independence and support of the joint pdf

- theorem (CB 4.2.12): let X and Y be independent random variables with moment generating functions $M_X(t)$ and $M_Y(t)$. Then the mgf of $Z = X + Y$ is

$$M_Z(t) = M_X(t)M_Y(t)$$

- proof:

$$M_Z(t) = \mathbb{E}\left(e^{tZ}\right) = \mathbb{E}\left(e^{t(X+Y)}\right) = \left(\mathbb{E}e^{tX}\right)\left(\mathbb{E}e^{tY}\right) = M_X(t)M_Y(t)$$

$\blacksquare$

- example/corollary (CB 4.2.14): let $X \sim N(\mu, \sigma^2)$ and $Y \sim N(\gamma, \tau^2)$, independent. Then $Z = X + Y \sim N(\mu + \gamma, \sigma^2 + \tau^2)$.

- proof: X and Y have mgf representations

$$\begin{aligned} M_X(t) &= e^{\mu t + \sigma^2 t^2/2} \\ M_Y(t) &= e^{\gamma t + \tau^2 t^2/2} \end{aligned}$$

then

$$M_Z(t) = e^{(\mu+\gamma)t + (\sigma^2+\tau^2)t^2/2}$$

which is the mgf of a normal random variable with mean $\mu + \gamma$ and variance $\sigma^2 + \tau^2$ ■

Contents

discrete random vectors

- let (X, Y) be a bivariate random vector with known probability distribution.
- Consider a new bivariate random vector (U, V) such that $U = g_1(X, Y)$ and $V = g_2(X, Y)$
  - $(U, V) \in B \Leftrightarrow (X, Y) \in A$, $A = \{(x, y) : \big(g_1(x, y), g_2(x, y)\big) \in B\}$
  - $\mathbb{P}\big((U, V) \in B\big) = \mathbb{P}\big((X, Y) \in A\big)$
  - keeping track of the support: from $\Omega_{X,Y} = \{(x, y) : f_{X,Y}(x, y) > 0\}$ to

$$\Omega_{U,V} = \{(u, v) : u = g_1(x, y),\ v = g_2(x, y) \text{ for some } (x, y) \in \Omega_{X,Y}\}$$

  - In the discrete case,

$$\begin{aligned} f_{UV}(u, v) &= \mathbb{P}(U = u, V = v) = \mathbb{P}\left((X, Y) \in \Omega^{(uv)}_{X,Y}\right) \\ &= \sum_{(x,y)\in\Omega^{(uv)}_{X,Y}} f_{X,Y}(x, y) \end{aligned}$$

where $\Omega^{uv}_{X,Y} = \{(x, y) \in \Omega_{X,Y} : g_1(x, y) = u, g_2(x, y) = v\}$.

sum of Poisson variables

example: let X and Y be independent Poisson random variables with joint pmf given by $f_{X,Y}(x,y) = \frac{e^{-\theta}\theta^x}{x!}\frac{e^{-\lambda}\lambda^y}{y!}$. also let $U = X + Y$ and $V = Y$.

- the support of the Poisson is $\Omega_{X,Y} = \{(x,y) : x \in \mathbb{N}, y \in \mathbb{N}\}$
- then $\Omega_{U,V} = \{(u,v) :\ v = 0, 1, 2, \dots \text{ and } u = v, v+1, v+2 \dots\}$
- $\Omega^{(uv)}_{X,Y}$ consists of only the single point $(u-v, v)$ and

$$f_{U,V}(u,v) = f_{X,Y}(u-v,v) = \frac{e^{-\theta}\theta^{u-v}}{(u-v)!}\frac{e^{-\lambda}\lambda^v}{v!}$$

- theorem/application: $X \sim \mathrm{P}(\theta)$, $Y \sim \mathrm{P}(\lambda)$ and $X \perp Y \Rightarrow X + Y \sim \mathrm{P}(\theta+\lambda)$

$$\begin{aligned} f_U(u) &= \sum_{v=0}^{u} \frac{e^{-\theta}\theta^{u-v}}{(u-v)!}\frac{e^{-\lambda}\lambda^v}{v!} = e^{-(\theta+\lambda)}\sum_{v=0}^{u}\frac{\theta^{u-v}\lambda^v}{v!(u-v)!} \\ &= \frac{e^{-(\theta+\lambda)}}{u!}\sum_{v=0}^{u}\binom{u}{v}\theta^{u-v}\lambda^v = \frac{e^{-(\theta+\lambda)}}{u!}(\theta+\lambda)^u \end{aligned}$$

binomial theorem is used in the last equality: $(x+y)^n = \sum_{k=0}^{n}\binom{n}{k}x^{n-k}y^n$.

continuous random vector

- let X and Y be continuous random variables with joint pdf $f_{X,Y}(x,y)$.
- as before, the support set $\Omega_{X,Y} = \{(x,y) : f_{X,Y}(x,y) > 0\}$ maps into

$$\Omega_{U,V} = \{(u,v) : u = g_1(x,y),\ v = g_2(x,y) \text{ for some } (x,y) \in \Omega_{X,Y}\}$$

- for now, assume that transformation $g : \Omega_{X,Y} \to \Omega_{U,V}$ is bijective: for each $(u,v) \in \Omega_{U,V}$ there is only one pair $(x,y) \in \Omega_{X,Y}$.
- we can solve the inverse transformation $x = h_1(u,v)$ and $y = h_2(u,v)$.

continuous random vector

- theorem: the pdf of (U, V) is given by

$$f_{U,V}(u,v) \quad = \quad f_{X,Y}\big(h_1(u,v), h_2(u,v)\big) \cdot |J|$$

  where J is the Jacobian of the transformation

$$J \quad = \quad \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} \quad = \quad \frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial x}{\partial v}\frac{\partial y}{\partial u}$$

  $x = h_1(u,v)$, $y = h_2(u,v)$ and $|\cdot|$ is the determinant.

- the term $|J|$ gives a "magnification factor" for area in going from u-v coordinates to x-y coordinates, just like in the univariate case.

- intuition for proof: draw rectangles in both coordinates and compute equivalent areas accounting for magnification.

product of betas

- example (CB 4.3.3): we want to find the distribution of the product of independent betas $X \sim \mathrm{B}(\alpha, \beta)$ and $Y \sim \mathrm{B}(\alpha + \beta, \gamma)$.

- each $B(\alpha, \beta)$ distribution is given by

$$f(x|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

  with $0 < x < 1$. So the joint distribution of X and Y is

$$f_{X,Y}(x, y) = \frac{\Gamma(\alpha+\beta+\gamma)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\gamma)} x^{\alpha-1}(1-x)^{\beta-1} y^{\alpha+\beta-1}(1-y)^{\gamma-1}$$

- we really don't care about V, but we choose one such that the mapping is bijective: let $U = XY$ and $V = X$, then $\Omega_{U,V} = \{(u, v) : 0 < u < v < 1\}$

- then we obtain the marginal for U to get the final answer.

product of betas

- so

$$
\begin{aligned}
f_{U,V}(u,v) &= f_{X,Y}(v,u/v)\left|\frac{\partial x}{\partial u}\frac{\partial y}{\partial v}-\frac{\partial y}{\partial u}\frac{\partial x}{\partial v}\right| \\
&= f_{X,Y}(v,u/v)\left|0(-u/v^2)-1(1/v)\right| \;=\; \frac{1}{v}\,f_{X,Y}(v,u/v) \\
&= \frac{\Gamma(\alpha+\beta+\gamma)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\gamma)}\,v^{\alpha-2}(1-v)^{\beta-1}\left(\frac{u}{v}\right)^{\alpha+\beta-1}\left(1-\frac{u}{v}\right)^{\gamma-1} \\
&= \frac{\Gamma(\alpha+\beta+\gamma)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\gamma)}\,u^{\alpha-1}\left(\frac{u}{v}-u\right)^{\beta-1}\left(1-\frac{u}{v}\right)^{\gamma-1}\frac{u}{v^2}
\end{aligned}
$$

marginal is also beta

- Taking the marginal for U,

$$
\begin{aligned}
f_U(u) &= \int_u^1 f_{U,V}(u,v)\,\mathrm{d}v \\
&= \frac{\Gamma(\alpha+\beta+\gamma)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\gamma)}\, u^{\alpha-1} \int_u^1 \left(\frac{u}{v} - u\right)^{\beta-1} \left(1 - \frac{u}{v}\right)^{\gamma-1} \frac{u}{v^2}\,\mathrm{d}v \\
&= \frac{\Gamma(\alpha+\beta+\gamma)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\gamma)}\, u^{\alpha-1}(1-u)^{\beta+\gamma-1} \\
&\qquad \times \int_u^1 \left(\frac{u/v - u}{1-u}\right)^{\beta-1} \left(\frac{1-u/v}{1-u}\right)^{\gamma-1} \frac{u}{v^2(1-u)}\,\mathrm{d}v \\
&= \frac{\Gamma(\alpha+\beta+\gamma)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\gamma)}\, u^{\alpha-1}(1-u)^{\beta+\gamma-1} \int_0^1 z^{\beta-1}(1-z)^{\gamma-1}\,\mathrm{d}z
\end{aligned}
$$

defining

$$
z = \frac{u/v - u}{1-u} \quad \Rightarrow \quad dz = -\frac{u}{v^2(1-u)} dv
$$

marginal is also beta

- so

$$
\begin{aligned}
f_U(u) &= \frac{\Gamma(\alpha+\beta+\gamma)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\gamma)}\, u^{\alpha-1}(1-u)^{\beta+\gamma-1} \underbrace{\int_0^1 z^{\beta-1}(1-z)^{\gamma-1}}_{=\frac{\Gamma(\beta)\Gamma(\gamma)}{\Gamma(\beta+\gamma)}} \,\mathrm{d}z \\
&= \frac{\Gamma(\alpha+\beta+\gamma)}{\Gamma(\alpha)\Gamma(\beta+\gamma)}\, u^{\alpha-1}(1-u)^{\beta+\gamma-1} \qquad\qquad U \sim \mathrm{B}(\alpha, \beta+\gamma)
\end{aligned}
$$

- where the last identity comes from recognizing the integrand as the kernel of a Beta pdf and using CB 3.3.17.

sum and difference of standard normals

- example (CB 4.3.4): let $X \sim N(0,1)$ and $Y \sim N(0,1)$ be independent, then $U = X + Y$ and $V = X - Y$ are also normal random variables

$$
\begin{aligned}
f_{U,V}(u,v) &= f_{X,Y}\left(\frac{u+v}{2}, \frac{u-v}{2}\right)\left|\frac{1}{2}\left(-\frac{1}{2}\right) - \frac{1}{2}\,\frac{1}{2}\right| = \frac{1}{2}\, f_{X,Y}\left(\frac{u+v}{2}, \frac{u-v}{2}\right) \\
&= \frac{1}{2}\,\frac{1}{2\pi}\, e^{-\frac{(u+v)^2}{8}}\, e^{-\frac{(u-v)^2}{8}} = \frac{1}{4\pi}\, e^{-\frac{u^2+2uv+v^2}{8} - \frac{u^2-2uv+v^2}{8}} \\
&= \frac{1}{4\pi}\, e^{-\frac{u^2}{4}}\, e^{-\frac{v^2}{4}} = \left(\frac{1}{\sqrt{2\pi}\,\sqrt{2}}\, e^{-\frac{u^2}{4}}\right)\left(\frac{1}{\sqrt{2\pi}\,\sqrt{2}}\, e^{-\frac{v^2}{4}}\right)
\end{aligned}
$$

- $U \sim N(0,2)$ and $V \sim N(0,2)$ are also independent (it actually holds as long as same variance)

independence

- there is a much simpler, but very important, situation in which the new variables U and V are independent

- theorem (CB 4.3.5): let X and Y be independent random variables, then $U = g(X)$ and $V = h(Y)$ are also independent.

- proof: consider the continuous case and define $\Omega_u = \{x : g(x) \leq u\}$ and $\Omega_v = \{y : h(y) \leq v\}$, then

$$\begin{aligned} F_{U,V}(u,v) &= \mathbb{P}(U \leq u, V \leq v) = \mathbb{P}(X \in \Omega_u, Y \in \Omega_v) \\ &= \mathbb{P}(X \in \Omega_u)\mathbb{P}(Y \in \Omega_v) \\ f_{U,V}(u,v) &= \frac{\partial^2}{\partial u \partial v} F_{U,V}(u,v) = \left(\frac{\mathrm{d}}{\mathrm{d}u}\mathbb{P}(X \in \Omega_u)\right)\left(\frac{\mathrm{d}}{\mathrm{d}v}\mathbb{P}(Y \in \Omega_v)\right) \end{aligned}$$

  the first term is a function only of u and the second term is a function only of v ■

find a partition if necessary

- in some situations of interest the transformation is not bijective...
- find a partition $A_0, A_1, \ldots, A_k$ of $\Omega_{X,Y}$, for which the set A_0 is such that $\mathbb{P}\big((X, Y) \in A_0\big) = 0$, whereas $(U, V) = \big(g_1(X, Y), g_2(X, Y)\big)$ is one-to-one from A_i to $\Omega_{U,V}$ for each $i = 1, \ldots, k$
- Then...

$$f_{U,V}(u, v) = \sum_{i=1}^{k} f_{X,Y}\big(h_{1i}(u, v), h_{2i}(u, v)\big)\,|J_i|$$

just like in the univariate case.

ratio of independent normal variables

example: let $U = X/Y$ and $V = |Y|$, with $X \sim N(0,1) \perp\!\!\!\perp Y \sim N(0,1)$

- $\Omega_{U,V} = \{(u,v) :\ v > 0\}$
- $A_0 = \{(x,y) :\ y = 0\},\ A_1 = \{(x,y) :\ y > 0\},\ A_2 = \{(x,y) :\ y < 0\}$
- $h_{11}(u,v) = uv,\ h_{21}(u,v) = v \Rightarrow |J_1| = |v \cdot 1 - u \cdot 0| = |v|$
- $h_{12}(u,v) = -uv,\ h_{21}(u,v) = -v \Rightarrow |J_2| = |(-v) \cdot (-1) + u \cdot 0| = |v|$
- Using the result above,

$$\begin{aligned} f_{U,V}(u,v) &= \frac{1}{2\pi}\, e^{-(uv)^2/2}\, e^{-v^2/2}\,|v| + \frac{1}{2\pi}\, e^{-(-uv)^2/2}\, e^{-(-v)^2/2}\,|v| \\ &= (v/\pi)\, e^{-(1+u^2)v^2/2} \qquad -\infty < u < \infty \qquad 0 < v < \infty \end{aligned}$$

ratio of independent normal variables

- the distribution of the ratio of independent normals is the marginal of U:

$$
\begin{aligned}
f_U(u) &= \int_0^\infty (v/\pi)\, e^{-(u^2+1)v^2/2}\, \mathrm{d}v \\
&= \frac{1}{2\pi} \int_0^\infty e^{-z(1+u^2)/2}\, \mathrm{d}z
\end{aligned}
$$

where we used $z = v^2 \Rightarrow \mathrm{d}z = 2v\,\mathrm{d}v$. By noticing that the integrand is kernel of exponential pdf with $\beta = \frac{2}{u^2+1}$, we get that

$$
\int_0^\infty e^{-z(1+u^2)/2}\, \mathrm{d}z = \frac{2}{1+u^2}
$$

and therefore

$$
f_U(u) = \frac{1}{2\pi} \frac{2}{1+u^2} = \frac{1}{\pi(1+u^2)} \qquad -\infty < u < \infty
$$

which is a Cauchy distribution. (intuitive, right?...)

Contents

hierarchy

- we have so far seen probability models in which a random variable has a single distribution, possibly depending on some fixed parameters

- **however...** it is sometimes useful to think about distributions with random parameters that follow themselves some known distribution

- **advantage** the main benefit is to handle intricate structures by means of a sequence of relatively simple models in a hierarchy

- classic example: how many eggs will survive on average if an insect lays a large number of eggs, each surviving with probability p?

binomial-Poisson hierarchy

- **let's make some assumptions...**
- large number of eggs is a random variable $N \sim \text{Poisson}(\lambda)$
- each egg's survival is independent and hence we may model their survival as Bernoulli trials $X|N \sim \text{Bin}(N, p)$

$$
\begin{aligned}
\mathbb{P}(X = x) &= \sum_{n=0}^{\infty} \mathbb{P}(X = x, N = n) = \sum_{n=0}^{\infty} \mathbb{P}(X = x | N = n)\, \mathbb{P}(N = n) \\
&= \sum_{n=x}^{\infty} \binom{n}{x} p^x (1-p)^{n-x} \frac{e^{-\lambda}\lambda^n}{n!} \\
&= e^{-\lambda}(\lambda p)^x \sum_{n=x}^{\infty} \frac{n!}{(n-x)!x!} (1-p)^{n-x} \frac{\lambda^{n-x}}{n!} \\
&= \frac{e^{-\lambda}(\lambda p)^x}{x!} \sum_{n=x}^{\infty} \frac{[(1-p)\lambda]^{n-x}}{(n-x)!} = \frac{e^{-\lambda}(\lambda p)^x}{x!} \sum_{t=0}^{\infty} \frac{[(1-p)\lambda]^t}{t!} \\
&= \frac{e^{-\lambda}(\lambda p)^x}{x!} e^{(1-p)\lambda} = \frac{(\lambda p)^x e^{-\lambda p}}{x!} \quad \Rightarrow \quad X \sim \text{Poisson}(\lambda p)
\end{aligned}
$$

law of iterated expectations (LIE)

- theorem: if X and Y are any two random variables, then

$$\mathbb{E}(X) = \mathbb{E}\big[\mathbb{E}(X|Y)\big]$$

  as long as the expectations exist.

- proof: $f_{X,Y}(x,y) = f_{X|Y}(x|y)f_Y(y)$ by definition, and hence

$$\begin{aligned}
\mathbb{E}(X) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} x f_{X|Y}(x|y)\,\mathrm{d}x\right] f_Y(y)\,\mathrm{d}y \\
&= \int_{-\infty}^{\infty} \mathbb{E}(X|y) f_Y(y)\,\mathrm{d}y = \mathbb{E}\big[\mathbb{E}(X|Y)\big] \qquad \blacksquare
\end{aligned}$$

- binomial-Poisson hierarchy: $\mathbb{E}(X) = \mathbb{E}\big[\mathbb{E}(X|N)\big] = \mathbb{E}(Np) = \lambda p$

mixture of distributions

- definition: a random variable X has a mixture distribution if the distribution of X depends on a quantity that also has a distribution

- any distribution arising from a hierarchy meets this definition

- example: Poisson(λp) is a mixture distribution as it results from the combination of a binomial distribution Bin(N, p) and $N \sim$ Poisson(λ)

- example: there is nothing to stop the hierarchy at two layers of structure there are now a large number of mother insects from which we draw one at random
  $X|N \sim \text{Bin}(N, p)$
  $N|\Lambda \sim \text{Poisson}(\lambda)$
  $\Lambda \sim \text{Exp}(\theta)$

- apart from aiding understanding, the hierarchical structure also helps with some moment calculations
- example: let X have a noncentral chi-squared distribution with p degrees of freedom and noncentrality parameter λ, then

$$f_X(x|\lambda, p) \;=\; \sum_{k=0}^{\infty} \frac{x^{p/2+k-1} e^{-x^2}}{\Gamma(p/2+k)2^{p/2+k}} \frac{\lambda^k e^{-\lambda}}{k!}$$

it is not so messy to compute $\mathbb{E}(X)$ if one realizes that $X|K \sim \chi^2_{p/2+K}$ and $K \sim \text{Poisson}(\lambda)$

$$\mathbb{E}(X) \;=\; \mathbb{E}\big[\mathbb{E}(X|K)\big] \;=\; \mathbb{E}(p+2K) \;=\; p+2\lambda$$

conditional variance identity

- theorem: $\text{var}(X) = \mathbb{E}\big[\text{var}(X|Y)\big] + \text{var}\big[\mathbb{E}(X|Y)\big]$
- proof:

$$
\begin{aligned}
\text{var}(X) &= \mathbb{E}[X - \mathbb{E}(X)]^2 \\
&= \mathbb{E}[X - \mathbb{E}(X|Y) + \mathbb{E}(X|Y) - \mathbb{E}(X)]^2 \\
&= \mathbb{E}[X - \mathbb{E}(X|Y)]^2 + \mathbb{E}[\mathbb{E}(X|Y) - \mathbb{E}(X)]^2 \\
&\qquad\qquad + 2\,\mathbb{E}\big\{[X - \mathbb{E}(X|Y)][\mathbb{E}(X|Y) - \mathbb{E}(X)]\big\} \\
&\overset{LIE}{=} \mathbb{E}\left(\mathbb{E}\big\{[X - \mathbb{E}(X|Y)]^2 | Y\big\}\right) + \text{var}\big[\mathbb{E}(X|Y)\big] \\
&\qquad\qquad + \ 2\,\mathbb{E}\left(\mathbb{E}\big\{[X - \mathbb{E}(X|Y)][\mathbb{E}(X|Y) - \mathbb{E}(X)] | Y\big\}\right) \\
&= \mathbb{E}\big[\text{var}(X|Y)\big] + \text{var}\big[\mathbb{E}(X|Y)\big] \\
&\qquad\qquad + 2\,\mathbb{E}\big\{[\mathbb{E}(X|Y) - \mathbb{E}(X|Y)][\mathbb{E}(X|Y) - \mathbb{E}(X)]\big\} \\
&= \mathbb{E}\big[\text{var}(X|Y)\big] + \text{var}\big[\mathbb{E}(X|Y)\big] \qquad \blacksquare
\end{aligned}
$$

Contents

how to gauge the strength of a relationship?

- let X and Y measure the weight and volume of a sample of water
  - if we gauge the pair (X, Y) in several samples and plot them
  - then data points should fall on a straight line in the absence of measurement errors

- let X and Y measure the body weight and height of a person
  - if we gauge the pair (X, Y) in several samples and plot them
  - then data points should also exhibit a upward trend, though not exactly a straight line

definitions

- the covariance between X and Y is

$$\mathrm{cov}(X, Y) \;=\; \mathbb{E}\big[(X - \mu_X)(Y - \mu_Y)\big] \;=\; \mathbb{E}(XY) - \mu_X \mu_Y,$$

  with $\mu_X = \mathbb{E}(X)$ and $\mu_Y = \mathbb{E}(Y)$, whereas the correlation is

$$\mathrm{corr}(X, Y) \;=\; \mathbb{E}\left(\frac{X - \mu_X}{\sigma_X}\,\frac{Y - \mu_Y}{\sigma_Y}\right) \;=\; \frac{1}{\sigma_X \sigma_Y}\,\mathrm{cov}(X, Y),$$

  with $\sigma_X = \sqrt{\mathrm{var}(X)}$ and $\sigma_Y = \sqrt{\mathrm{var}(Y)}$

- **independence** (CB 4.5.5): if X and Y are independent random variables then $\mathrm{cov}(X, Y) = \mathrm{corr}(X, Y) = 0$

counterexamples

Independence implies $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$, but not vice-versa.

1. let X be -1 or 1 with probability 0.5. Let Y be 0 if $X = -1$. If $X = 1$, Y is randomly -1 or 1 with probability 0.5. X and Y are not independent
   - however...

$$\begin{aligned}\mathbb{E}(XY) &= -1 \cdot 0 \cdot \mathbb{P}(X = -1) + 1 \cdot 1 \cdot \mathbb{P}(X = 1, Y = 1) \\ &\quad +1 \cdot -1 \cdot \mathbb{P}(X = 1, Y = -1) \\ &= 0\end{aligned}$$

   and $\mathbb{E}(X) = \mathbb{E}(Y) = 0$.

2. A standard normal distribution is such that $\mathbb{E}(X) = \mathbb{E}(X^3) = 0$. Take $Y = X^2$. Then

$$\mathrm{cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X) \cdot \mathbb{E}(Y) = 0$$

example of linear relationship

- example: let $X \sim \mathrm{U}(0,1) \perp\!\!\!\perp Z \sim \mathrm{U}(0,1/10)$ and $Y = X + Z$.

- then the joint pdf of (X, Y) is $f_{X,Y}(x,y) = 10$ for $0 < x < 1$ and $x < y < x + 1/10$, with

$$
\begin{aligned}
\mathbb{E}(Y) &= \mathbb{E}(X) + \mathbb{E}(Z) = 1/2 + 1/20 = 11/20 \\
\mathrm{cov}(X,Y) &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = \mathbb{E}\big[X(X+Z)\big] - \mathbb{E}(X)\mathbb{E}(X+Z) \\
&= \mathbb{E}(X^2) + \mathbb{E}(XZ) - \big[\mathbb{E}(X)\big]^2 - \mathbb{E}(X)\mathbb{E}(Z) \\
&= \mathrm{var}(X) = \frac{1}{12}(1-0)^2 = \frac{1}{12} \\
\mathrm{var}(Y) &= \mathrm{var}(X+Z) = \mathrm{var}(X) + \mathrm{var}(Z) = \frac{1}{12} + \frac{1}{1200} = \frac{101}{1200} \\
\mathrm{corr}(X,Y) &= \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X)\mathrm{var}(Y)}} = \frac{1/12}{\sqrt{1/12 \times 101/1200}} = \sqrt{\frac{100}{101}}
\end{aligned}
$$

example of nonlinear relationship

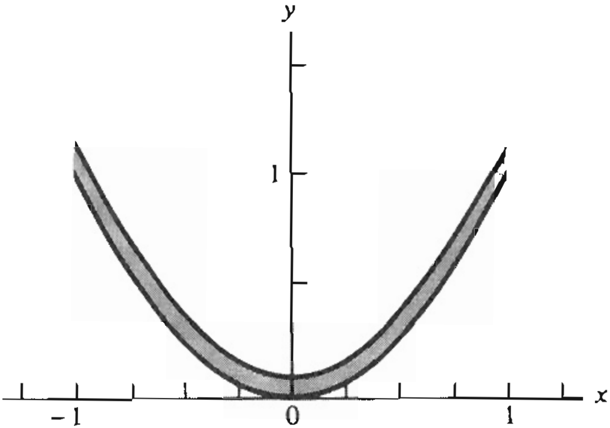
- example: let $X \sim \mathrm{U}(-1,1) \perp\!\!\!\perp Z \sim \mathrm{U}(0,1/10)$ and $Y = X^2 + Z$, then the joint pdf of (X, Y) is $f_{X,Y}(x,y) = 5$ for $-1 < x < 1$ and $x^2 < y < x^2 + 1/10$, with

$$\begin{aligned}
\mathrm{cov}(X,Y) &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) \\
&= \mathbb{E}\big[X(X^2+Z)\big] - \mathbb{E}(X)\mathbb{E}(X^2+Z) \\
&= \mathbb{E}(X^3) + \mathbb{E}(XZ) - \mathbb{E}(X)\mathbb{E}(X^2) - \mathbb{E}(X)\mathbb{E}(Z) \\
&= \mathbb{E}(X^3) + \mathbb{E}(X)\mathbb{E}(Z) - \mathbb{E}(X)\mathbb{E}(X^2) - \mathbb{E}(X)\mathbb{E}(Z) \\
&= 0
\end{aligned}$$

  given that $\mathbb{E}(X) = \mathbb{E}(X^3) = 0$ due to the symmetric nature of X

- there is a strong dependence between X and Y, but it is not linear...

how does it look like?



linear dependence

- theorem (CB 4.5.7): For any random variables X and Y,
  (i) $|\text{corr}(X, Y)| \leq 1$
  (ii) $|\text{corr}(X, Y)| = 1$ if and only if there exist numbers $a \neq 0$ and b such that $\mathbb{P}(Y = aX + b) = 1$, with $a > 0$ if $\text{corr}(X, Y) > 0$ and $a < 0$ if $\text{corr}(X, Y) < 0$

- proof of (i): define $h(t) = \mathbb{E}\big[(X - \mu_X)t + (Y - \mu_Y)\big]^2$, so $h(t) \geq 0$, $\forall t$

$$\begin{aligned} h(t) &= t^2\mathbb{E}(X - \mu_X)^2 + 2t\mathbb{E}(X - \mu_X)(Y - \mu_Y) + \mathbb{E}(Y - \mu_Y)^2 \\ &= t^2\sigma_X^2 + \sigma_Y^2 + 2t\,\text{cov}(X - \mu_X, Y - \mu_Y) \end{aligned}$$

and hence it can have at most one real root, implying a nonpositive discriminant,

$$\begin{aligned} \big[2\,\text{cov}(X, Y)\big]^2 - 4\sigma_X^2\sigma_Y^2 \leq 0 \quad &\Rightarrow \quad -\sigma_X\sigma_Y \leq \text{cov}(X, Y) \leq \sigma_X\sigma_Y \\ &\Rightarrow \quad |\text{corr}(X, Y)| \leq 1 \end{aligned}$$

linear dependence

- proof of (ii): now, $\text{corr}(X,Y) = 1 \Leftrightarrow \left[2t\,\text{cov}(X,Y)\right]^2 - 4t^2\sigma_X^2\sigma_Y^2 = 0$, i.e., $h(t)$ has a single root. Given that $\left[(X-\mu_X)t + (Y-\mu_Y)\right]^2 \geq 0$ for all t, $h(t) = 0$ if and only if

$$\begin{aligned} \mathbb{P}\left(\left[(X-\mu_X)t + (Y-\mu_Y)\right]^2 = 0\right) &= 1 \\ &\Downarrow \\ \mathbb{P}\left((X-\mu_X)t + (Y-\mu_Y) = 0\right) &= 1 \end{aligned}$$

which is equivalent to $\mathbb{P}(Y = aX + b) = 1$ with $a = -t = \frac{\text{cov}(X,Y)}{\sigma_X^2}$ and $b = \mu_X t + \mu_Y$ ■

- we will see that the Cauchy-Schwartz inequality considerably shortens the proof above

variance decomposition

- theorem (CB 4.5.6): if X and Y are any two random variables, and a and b are any two constants, then
$$\operatorname{var}(aX + bY) = a^2 \operatorname{var}(X) + b^2 \operatorname{var}(Y) + 2ab \operatorname{cov}(X, Y)$$

- proof: it follows from $\mathbb{E}(aX + bY) = a\mu_X + b\mu_Y$ that
$$\begin{aligned}
\operatorname{var}(aX + bY) &= \mathbb{E}\big[(aX + bY) - (a\mu_X + b\mu_Y)\big]^2 \\
&= \mathbb{E}\big[a(X - \mu_X) + b(Y - \mu_Y)\big]^2 \\
&= \mathbb{E}\big[a^2(X - \mu_X)^2 + b^2(Y - \mu_Y)^2 + 2\,a(X - \mu_X)\,b(Y - \mu_Y)\big] \\
&= a^2\big[\mathbb{E}(X - \mu_X)^2\big] + b^2\mathbb{E}\big[(Y - \mu_Y)^2\big] \\
&\quad +2\,ab\,\mathbb{E}\big[(X - \mu_X)(Y - \mu_Y)\big] \\
&= a^2 \operatorname{var}(X) + b^2 \operatorname{var}(Y) + 2ab \operatorname{cov}(X, Y) \qquad \blacksquare
\end{aligned}$$

- the variation in $X + Y$ is inferior to the sum of the variations in X and Y if $\operatorname{cov}(X, Y) < 0$ because large values of X are more likely to occur with small values of Y

bivariate normal

- definition: the bivariate normal distribution with parameters μ_X, μ_Y, $\sigma_X^2 > 0$, $\sigma_Y^2 > 0$ and $|\rho| < 1$

$$\begin{aligned} f_{X,Y}(x,y) &= \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{\rho}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right.\right. \\ &\qquad \left.\left. -2\rho\,\frac{x-\mu_X}{\sigma_X}\,\frac{y-\mu_Y}{\sigma_Y}\right]\right\} \end{aligned}$$

  for $-\infty < x < \infty$ and $-\infty < x < \infty$.

- the following properties hold (**proofs left as exercise**):
  - $\text{corr}(X, Y) = \rho$
  - $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$
  - $X|Y \sim N\left(\mu_X + \rho\,\frac{\sigma_X}{\sigma_Y}\,(Y - \mu_Y), \sigma_X^2(1-\rho^2)\right)$

Contents

joint, marginal and conditional probabilities

- **discrete**: the joint pmf of $\boldsymbol{X} = (X_1, \ldots, X_n) \subset \mathbb{R}^n$ is a function $f_{\boldsymbol{X}}(\boldsymbol{x})$ such that

$$\mathbb{P}(\boldsymbol{X} \in A) = \sum_{\boldsymbol{x} \in A} f_{\boldsymbol{X}}(\boldsymbol{x})$$

  for any $A \subset \mathbb{R}^n$

- **continuous**: the joint pdf of $\boldsymbol{X} = (X_1, \ldots, X_n) \subset \mathbb{R}^n$ is a function $f_{\boldsymbol{X}}(\boldsymbol{x})$ such that

$$\mathbb{P}(\boldsymbol{X} \in A) = \int \cdots \int_A f(x_1, \ldots, x_n)\, \mathrm{d}x_1 \cdots \mathrm{d}x_n$$

  for any $A \subset \mathbb{R}^n$

- **expectation**:

$$\mathbb{E}[g(\boldsymbol{x})] = \begin{cases} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(\boldsymbol{x})\, f_{\boldsymbol{X}}(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} & \text{if continuous} \\ \sum_{\boldsymbol{x} \in \mathbb{R}^n} g(\boldsymbol{x})\, f_{\boldsymbol{X}}(\boldsymbol{x}) & \text{if discrete} \end{cases}$$

joint, marginal and conditional probabilities

- marginals with respect to a subset of the variables can be obtained integrating with respect to the other variables

$$f(x_1, \ldots, x_k) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \ldots, x_n) \, \mathrm{d}x_{k+1} \cdots \mathrm{d}x_n$$

- similarly, the conditional pdf is

$$f(x_{k+1}, \ldots, x_n | x_1, \ldots, x_k) = \frac{f(x_1, \ldots, x_n)}{f(x_1, \ldots, x_k)}$$

- example: let

$$f(x_1, x_2, x_3, x_4) = \begin{cases} \frac{3}{4}(x_1^2 + x_2^2 + x_3^2 + x_4^2) & 0 < x_i < 1, i = 1, 2, 3, 4 \\ 0 & \text{o.w.} \end{cases}$$

- verify that:
  (i) $\int_0^1 \int_0^1 \int_0^1 \int_0^1 f(x_1, x_2, x_3, x_4)\, \mathrm{d}x_1\, \mathrm{d}x_2\, \mathrm{d}x_3\, \mathrm{d}x_4 = 1$
  (ii) $\mathbb{P}\left(X_1 < \frac{1}{2}, X_2 < \frac{3}{4}, X_4 > \frac{1}{2}\right) = \frac{3}{256}$
  (iii) $f(x_1, x_2) = \frac{3}{4}(x_1^2 + x_2^2) + \frac{1}{2}$
  (iv) $\mathbb{E}X_1 X_2 = \frac{5}{16}$
  (v) $f(x_3, x_4 | x_1, x_2) = \frac{x_1^2 + x_2^2 + x_3^2 + x_4^2}{x_1^2 + x_2^2 + \frac{2}{3}}$

multinomial distribution

- Bernoulli trials now have n distinct outcomes, with probabilities $p_1, \ldots, p_n$, common across trials. X_i represents the number of times that the ith outcome happened among m trials.

- example: toss a six-sided dice and let Z be the outcome. The dice is unbalanced and $\mathbb{P}(Z = z) = \frac{z}{21}$. Consider now tossing the dice ten times, and X_i counts the number of times i came up. Then $X = (X_1, X_2, \ldots, X_6)$ has a multinomial distribution with $m = 10$ trials, $n = 6$ possible outcomes, and

$$\begin{aligned} f(0,0,1,2,3,4) &= \frac{10!}{0!0!1!2!3!4!}\left(\frac{1}{21}\right)^0\left(\frac{2}{21}\right)^0\left(\frac{3}{21}\right)^1\left(\frac{4}{21}\right)^2\left(\frac{5}{21}\right)^3\left(\frac{6}{21}\right)^4 \\ &= 0.0059 \end{aligned}$$

multinomial distribution

- definition: let n and m denote positive integers, then the discrete random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ has a multinomial distribution with m trials and cell probabilities $0 \leq p_1, \ldots, p_n \leq 1$ such that $\sum_{i=1}^n p_i = 1$ if the joint pmf of $\boldsymbol{X}$ is given by

$$f_{\boldsymbol{X}}(x_1, \ldots, x_n) = \frac{m!}{x_1! \cdots x_n!} p_1^{x_1} \cdots p_n^{x_n} = m! \prod_{i=1}^n \frac{p_i^{x_i}}{x_i!}$$

for $\boldsymbol{x} = (x_1, \ldots, x_n)$ such that each integer $x_i \geq 0$ and $\sum_{i=1}^n x_i = m$

marginal and conditional pmfs of a multinomial

if the discrete random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ is multinomial with m trials and cell probabilities $0 \leq p_1, \ldots, p_n \leq 1$, (you may try to show these properties)

- the marginal of X_i is binomial $\mathrm{Bin}(m, p_i)$
- the conditional distribution of $(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$ given $X_i = x_i$ is multinomial with $m - x_i$ trials and cell probabilities $p_j/(1 - p_i)$ for $1 \leq j \neq i \leq n$
- there is some negative correlation given that $\sum_{i=1}^n X_i = m$ $\mathrm{corr}(X_i, X_j) = -mp_ip_j$ for $1 \leq i \neq j \leq n$

independence

- definition: let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ denote random vectors with joint pdf/pmf $f_{\boldsymbol{X}}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ and marginal pdf/pmf $f_{\boldsymbol{X}_i}(\boldsymbol{x_i})$, then they are mutually independent random vectors if, for every $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$,

$$f_{\boldsymbol{X}}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = f_{\boldsymbol{X_1}}(\boldsymbol{x_1}) \cdots f_{\boldsymbol{X}_n}(\boldsymbol{x_n}) = \prod_{i=1}^{n} f_{\boldsymbol{X}_i}(\boldsymbol{x_i})$$

- we now need to generalize the results we had for independent bivariate distributions

independence

if $X_1, \ldots, X_n$ are independent,

(1) let $g_1, \ldots, g_n$ be real-valued functions such that $g_i(x_i)$ is a function only of x_i.

$$\mathbb{E}\left[g_1(X_1)\cdots g_n(X_n)\right] = \prod_{i=1}^{n} \mathbb{E}\left[g_i(X_1)\right]$$

(2) let $M_{X_1}(t), \ldots, M_{X_N}(t)$ be the mgfs of $X_1, \ldots, X_N$ and $Z = \sum_{i=1}^{n} X_i$. Then

$$M_Z(t) = \prod_{i=1}^{n} M_{X_i}(t)$$

(3) let $a_1, \ldots, a_n, b_1, \ldots, b_n$ be fixed constants and $Z = \sum_{i=1}^{n} a_i X_i + b_i$. Then

$$M_Z(t) = \left(e^{t \sum b_i}\right) \prod_{i=1}^{n} M_{X_i}(a_i t)$$

independence

if $X_1, \ldots, X_n$ are independent,

(4) $X_1, \ldots, X_N$ are independent if, and only if, there exists functions $g_i(x_i)$ such that

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} g_i(x_i)$$

(5) $U_1 = g_1(X_1), \ldots, U_n = g_n(X_n)$ are also mutually independent

independence and normality

- example (CB 3.6.10): $X_i \sim N(\mu_i, \sigma_i^2)$, mutually independent. Let $a_1, \ldots, a_n, b_1, \ldots, b_n$ be fixed constants. Then

$$Z = \sum_{i=1}^{n} (a_i X_i + b_i) \sim N\left(\sum_{i=1}^{n}(a_i\mu_i + b_i), \sum_{i=1}^{n} a_i^2\sigma_i^2\right)$$

- proof: the mgf of a normal random variable is $M(t) = e^{\mu t + \sigma^2 t^2/2}$. Then

$$\begin{aligned} M_Z(t) &= \left(e^{t\sum b_i}\right)\prod_{i=1}^{n} e^{\mu_i a_i t + \sigma_i^2 a_i^2 t^2/2} \\ &= e^{t\sum(a_i\mu_i + b_i) + (\sum a_i^2\sigma_i^2)t^2/2} \end{aligned}$$

which is the mgf of a $N\left(\sum_{i=1}^{n}(a_i\mu_i + b_i), \sum_{i=1}^{n} a_i^2\sigma_i^2\right)$. ∎

multivariate normal

- the pdf of multivariate normal distributions is

$$f_X(x) \quad = \quad \frac{1}{(2\pi)^{n/2}} |\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)}$$

  for n-dimensional X. Denote $X \sim N(\mu, \Sigma)$.

- **lemma:** let $Z \sim N(0, I_n)$ and $X = \mu + \Sigma^{1/2} Z$. Then $X \sim N(\mu, \Sigma)$.

- proof: the distribution of Z is

$$f_Z(z) \quad = \quad \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2} z' z}$$

  and the transformation $x = \mu + \Sigma^{1/2} z$ has Jacobian $|\Sigma|^{-1/2}$.

- **lemma:** if $Y = AX + b$, then $Y \sim N(A\mu + b, A\Sigma A')$.

- proof: follows from previous slide.

multivariate normal

- take a partition $X = [X_1', X_2']'$, with $X \sim N(\mu, \Sigma)$ and let

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \text{ and } \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

- theorem: X_1 and X_2 are independent if and only if $\Sigma_{12} = 0$.
- proof ($\Rightarrow$): this is immediate (independent random variables imply zero correlation)
- proof ($\Leftarrow$): let $\Sigma_{12} = 0$ and write

$$\Sigma = \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{pmatrix}$$

then

$$\begin{aligned} f_X(x) &= \frac{1}{(2\pi)^{n/2}} |\Sigma|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu) \right\} \\ &= \frac{1}{(2\pi)^{n_1/2}} |\Sigma_{11}|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}(x_1-\mu_1)'\Sigma_{11}^{-1}(x_1-\mu_1) \right\} \\ &\quad \times \frac{1}{(2\pi)^{n_2/2}} |\Sigma_{22}|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}(x_2-\mu_2)'\Sigma_{22}^{-1}(x_2-\mu_2) \right\} \\ &= f_{X_1}(x_1) \cdot f_{X_2}(x_2) \end{aligned}$$

multivariate normal

- theorem: the conditional distribution of $X_1|X_2$ is $N(\mu_{1\cdot 2}, \Sigma_{11\cdot 2})$ with

$$\begin{aligned} \mu_{1\cdot 2} &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2) \\ \Sigma_{11\cdot 2} &= \Sigma_{11} + \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \end{aligned}$$

- proof: consider a random vector given by

$$\begin{bmatrix} X_1 - \Sigma_{12}\Sigma_{22}^{-1}X_2 \\ X_2 \end{bmatrix} = \begin{bmatrix} I & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$$

which is a linear transformation of a normal random vector X. The two subvectors $X_1 - \Sigma_{12}\Sigma_{22}^{-1}X_2$ and X_2 are uncorrelated,

$$\begin{aligned} Var\begin{bmatrix} X_1 - \Sigma_{12}\Sigma_{22}^{-1}X_2 \\ X_2 \end{bmatrix} &= \begin{bmatrix} I & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \begin{bmatrix} I & 0 \\ -\Sigma_{12}'\Sigma_{22}^{-1} & I \end{bmatrix} \\ &= \begin{bmatrix} \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} & 0 \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \begin{bmatrix} I & 0 \\ -\Sigma_{12}'\Sigma_{22}^{-1} & I \end{bmatrix} \\ &= \begin{bmatrix} \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} & 0 \\ 0 & \Sigma_{22} \end{bmatrix} \end{aligned}$$

therefore independent.

multivariate normal

- proof (cont'd): write

$$X_1 = \Sigma_{12}\Sigma_{22}^{-1}X_2 + (X_1 - \Sigma_{12}\Sigma_{22}^{-1}X_2)$$

  where the term in brackets is independent of X_2, so its conditional distribution given X_2 is consequently the same as its unconditional distribution, which is normal with mean $\mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2$ and variance $\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.

- then

$$\begin{aligned} E(X_1|X_2) &= E(\Sigma_{12}\Sigma_{22}^{-1}X_2|X_2) + E(X_1 - \Sigma_{12}\Sigma_{22}^{-1}X_2) \\ &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2) \end{aligned}$$

$$\begin{aligned} Var(X_1|X_2) &= Var(\Sigma_{12}\Sigma_{22}^{-1}X_2|X_2) + Var(X_1 - \Sigma_{12}\Sigma_{22}^{-1}X_2) \\ &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \end{aligned}$$

■

transformations of random vectors

- denote $\boldsymbol{U} = (U_1, \dots, U_n)$, with $U_i = g_i(X_1, \dots, X_n)$ for $i = 1, \dots, n$.
- let the support set be $\Omega_X = \{x : f_{\boldsymbol{X}}(\boldsymbol{x}) > 0\}$
- find partitions $A_0, A_1, A_2, \dots, A_k$ such that $\mathbb{P}(X \in A_0) = 0$ and g is a bijective transformation within each A_j, $j > 0$
- we then have inverse transformations $x_1 = h_{1j}(u_1, \dots, u_n), \dots, x_n = h_{nj}(u_1, \dots, u_n)$ for each $j > 0$
- the Jacobian term is given by

$$J_j = \begin{vmatrix} \frac{\partial x_1}{\partial u_1} & \frac{\partial x_1}{\partial u_2} & \cdots & \frac{\partial x_1}{\partial u_n} \\ \frac{\partial x_2}{\partial u_1} & \frac{\partial x_2}{\partial u_2} & \cdots & \frac{\partial x_2}{\partial u_n} \\ \vdots & \ddots & \cdots & \vdots \\ \frac{\partial x_n}{\partial u_1} & \frac{\partial x_n}{\partial u_2} & \cdots & \frac{\partial x_n}{\partial u_n} \end{vmatrix} = \begin{vmatrix} \frac{\partial h_{1j}(\boldsymbol{u})}{\partial u_1} & \frac{\partial h_{1j}(\boldsymbol{u})}{\partial u_2} & \cdots & \frac{\partial h_{1j}(\boldsymbol{u})}{\partial u_n} \\ \frac{\partial h_{2j}(\boldsymbol{u})}{\partial u_1} & \frac{\partial h_{2j}(\boldsymbol{u})}{\partial u_2} & \cdots & \frac{\partial h_{2j}(\boldsymbol{u})}{\partial u_n} \\ \vdots & \ddots & \cdots & \vdots \\ \frac{\partial h_{nj}(\boldsymbol{u})}{\partial u_1} & \frac{\partial h_{nj}(\boldsymbol{u})}{\partial u_2} & \cdots & \frac{\partial h_{nj}(\boldsymbol{u})}{\partial u_n} \end{vmatrix}$$

with $x_i = h_{ij}(\boldsymbol{u})$ for any $x_i \in A_j$ with $i = 1, \dots, n$ and $j = 1, \dots, k$

- then...

$$f_{\mathbf{U}}(u_1, \ldots, u_n) \;=\; \sum_{j=1}^{k} f_{\mathbf{X}}\left(h_{1j}(u_1, \ldots, u_n), \ldots, h_{nj}(u_1, \ldots, u_n)\right)|J_j|,$$

- example: joint pdf $f_{\mathbf{X}}(x_1, x_2, x_3, x_4) = 24e^{-x_1-x_2-x_3-x_4}$ with $0 < x_1 < x_2 < x_3 < x_4 < \infty$ and $U_1 = X_1$, $U_2 = X_2 - X_1$, $U_3 = X_3 - X_2$ and $U_4 = X_4 - X_3$
  - $X_1 = U_1$, $X_2 = U_1 + U_2$, $X_3 = U_1 + U_2 + U_3$, $X_4 = U_1 + U_2 + U_3 + U_4$
  - Jacobian

$$J \;=\; \begin{vmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{vmatrix} \;=\; 1$$

  - so $f_{\mathbf{U}}(u_1, \ldots, u_4) = 24e^{-4u_1-3u_2-2u_3-u_4}$

Contents

a lemma

- lemma: let $a, b > 0$ and $p, q > 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$, then $\frac{1}{p}a^p + \frac{1}{q}b^q \geq ab$ with equality if and only if $a^p = b^q$.

- sketch of proof: fix b and minimize

$$g(a) \quad = \quad \frac{1}{p}a^p + \frac{1}{q}b^q - ab$$

with respect to a. We get

$$\frac{dg(a)}{da} \quad = \quad 0 \quad \Rightarrow \quad a^{p-1} - b = 0 \quad \Rightarrow \quad b = a^{p-1}$$

The second derivative $\frac{\mathrm{d}^2 g(a)}{\mathrm{d}a^2} = (p-1)a^{p-1} > 0$, indeed a minimum. The value at the minimum is

$$\frac{1}{p}a^p + \frac{1}{q}a^{q(p-1)} - a^p \quad = \quad \frac{1}{p}a^p + \frac{1}{q}a^p - a^p \quad = \quad 0$$

since $\frac{1}{p} + \frac{1}{q} = 1 \Rightarrow q + p = pq \Rightarrow q(p-1) = p$. Equality holds if $b = a^{p-1} \Rightarrow a^p = b^q$. $\blacksquare$

Hölder's inequality

- theorem: let X and Y denote any two random variables and let p and q satisfy $\frac{1}{p}+\frac{1}{q}=1$, then

$$|\mathbb{E}(XY)| \;\leq\; \mathbb{E}|XY| \;\leq\; (\mathbb{E}|X|^p)^{1/p}(\mathbb{E}|Y|^q)^{1/q}$$

- proof: the first inequality follows from the fact that

$$-|XY| \leq XY \leq |XY| \;\Rightarrow\; -\mathbb{E}|XY| \leq \mathbb{E}(XY) \leq \mathbb{E}|XY|.$$

  to prove the second inequality, choose

$$a \;=\; \frac{|X|}{(\mathbb{E}|X|^p)^{1/p}} \quad \text{and} \quad b \;=\; \frac{|Y|}{(\mathbb{E}|Y|^q)^{1/q}}$$

  which, using the lemma, implies

$$\begin{aligned} \frac{1}{p}\frac{|X|^p}{(\mathbb{E}|X|^p)} + \frac{1}{q}\frac{|X|^q}{(\mathbb{E}|X|^q)} \quad &\geq \quad \frac{|X|}{(\mathbb{E}|X|^p)^{1/p}}\frac{|Y|}{(\mathbb{E}|Y|^q)^{1/q}} \\ &= \quad \frac{|XY|}{(\mathbb{E}|X|^p)^{1/p}(\mathbb{E}|Y|^q)^{1/q}} \end{aligned}$$

Hölder's inequality

- proof (cont'd): taking expectations on both sides,

$$
\begin{aligned}
\underbrace{\frac{1}{p}\frac{\mathbb{E}|X|^p}{(\mathbb{E}|X|^p)} + \frac{1}{q}\frac{\mathbb{E}|X|^q}{(\mathbb{E}|X|^q)}}_{=\frac{1}{p}+\frac{1}{q}=1} \quad &\geq \quad \frac{\mathbb{E}|XY|}{(\mathbb{E}|X|^p)^{1/p}(\mathbb{E}|Y|^q)^{1/q}} \\
&\Downarrow \\
\mathbb{E}|XY| \quad &\leq \quad (\mathbb{E}|X|^p)^{1/p}(\mathbb{E}|Y|^q)^{1/q}
\end{aligned}
$$

which completes the proof. ■

applications

$$\underline{\text{Hölder:}} \quad |\mathbb{E}(XY)| \;\leq\; \mathbb{E}|XY| \;\leq\; (\mathbb{E}|X|^p)^{1/p}(\mathbb{E}|Y|^q)^{1/q}$$

- selecting $p = q = 2$, we obtain the Cauchy-Schwarz inequality: for any random variables X and Y,

$$|\mathbb{E}(XY)| \;\leq\; \mathbb{E}|XY| \;\leq\; \sqrt{\mathbb{E}(X^2)}\sqrt{\mathbb{E}(Y^2)}$$

- covariance inequality: applying the Cauchy-Scwartz inequality to $X - \mu_X$ and $Y - \mu_Y$ yields

$$|\text{cov}(X, Y)| \;\leq\; \sigma_X \sigma_Y$$

  or, equivalently, that $|\text{corr}(X, Y)| \leq 1$.

- Lyapunov's inequality: set $Y = 1$, replace $|X|$ by $|X|^r$ for $1 < r < p$ and define $s = pr$ to obtain

$$(\mathbb{E}|X|^r)^{1/r} \leq (\mathbb{E}|X|^s)^{1/s}$$

  for $1 < r < s < \infty$

Minkowski's inequality

- theorem: let X and Y denote any two random variables, then

$$(\mathbb{E}|X+Y|^p)^{1/p} \leq (\mathbb{E}|X|^p)^{1/p} + (\mathbb{E}|Y|^p)^{1/p} \qquad 0 \leq p < \infty$$

- proof: triangular inequality $|X+Y| \leq |X| + |Y|$ ensures that

$$\begin{aligned}
\mathbb{E}|X+Y|^p &= \mathbb{E}\left(|X+Y||X+Y|^{p-1}\right) \\
&\leq \mathbb{E}\left(|X||X+Y|^{p-1}\right) + \mathbb{E}\left(|Y||X+Y|^{p-1}\right) \\
&\leq (\mathbb{E}|X|^p)^{1/p}\left(\mathbb{E}|X+Y|^{q(p-1)}\right)^{1/q} \\
&\quad + (\mathbb{E}|Y|^p)^{1/p}\left(\mathbb{E}|X+Y|^{q(p-1)}\right)^{1/q}
\end{aligned}$$

for $1/p + 1/q = 1$ where Hölder's inequality was applied twice.

Minkowski's inequality

- proof (cont'd): dividing by $\left(\mathbb{E}|X+Y|^{q(p-1)}\right)^{1/q}$,

$$\frac{\mathbb{E}|X+Y|^p}{(\mathbb{E}|X+Y|^{q(p-1)})^{1/q}} \leq (\mathbb{E}|X|^p)^{1/p} + (\mathbb{E}|Y|^p)^{1/p}$$

and since $\frac{1}{p} + \frac{1}{q} = 1 \Rightarrow p + q = pq \Rightarrow qp - q = p$,

$$\begin{aligned}\frac{\mathbb{E}|X+Y|^p}{(\mathbb{E}|X+Y|^{q(p-1)})^{1/q}} &= \frac{\mathbb{E}|X+Y|^p}{(\mathbb{E}|X+Y|^p)^{1/q}} \\ &= (\mathbb{E}|X+Y|^p)^{1-\frac{1}{q}} \\ &= (\mathbb{E}|X+Y|^p)^{\frac{1}{p}}\end{aligned}$$

which completes the proof. ■

Contents

Reference:

- Casella and Berger, Ch. 4

Exercises:

- 4.1, 4.4–4.7, 4.9, 4.10, 4.13, 4.15, 4.22, 4.24, 4.26, 4.30, 4.32, 4.37, 4.38, 4.41–4.43, 4.47, 4.58, 4.59.