Multivariate distributions

So far we've looked at one and then two rvs at a time; easy to generalize to a finite number of rv $X_1, \dots, X_n$, n a positive finite integer.

Def. The joint CDF of n rvs $X_1, \dots, X_n$ is the function $F_{X_1, \dots, X_n}(x_1, \dots, x_n)$ specified by $F_{X_1, \dots, X_n}(x_1, \dots, x_n) = P(X_1 \le x_1, \dots, X_n \le x_n)$ and

More compact to use vector notation: $\underset{\sim}{X} = (X_1, \dots, X_n)$, $\underset{\sim}{x} = (x_1, \dots, x_n)$

$F_{\underset{\sim}{X}}(\underset{\sim}{x}) = P(X_1 \le x_1, \dots, X_n \le x_n)$ $\underset{\sim}{X}$ is said to be a random vector taking values in $\mathbb{R}^n$.

Def. n rv $(Y_1, \ldots, Y_n) = \underset{\sim}{Y}$ have a discrete joint distribution if the random vector $\underset{\sim}{Y}$ can only take on a finite or countably infinite # of possible values $(y_1, \ldots, y_n) \in \mathbb{R}^n$.

The joint PF (probability mass function) of $\underset{\sim}{Y}$ is

$$f_{Y_1 \ldots Y_n}(y_1, \ldots, y_n) = P(Y_1 = y_1, \ldots, Y_n = y_n)$$

or equivalently $f_{\underset{\sim}{Y}}(\underset{\sim}{y}) = P(\underset{\sim}{Y} = \underset{\sim}{y})$.

Example n patients in treatment group of a randomized clinical trial; $B_i = \begin{cases} 1 & \text{if patient } i \text{ has a good outcome} \\ 0 & \text{else} \end{cases}$.

If nothing else is known about the patients (e.g., age, disease burden at start of trial, ...) it would be reasonable to model the B_i as IID Bernoulli(θ) (same success probability).

$\underset{\sim}{B} = (B_1, \dots, B_n)$; $\underset{\sim}{b} = (b_1, \dots, b_n)$; $\underset{\sim}{B}$ has a discrete joint distribution $f_{\underset{\sim}{B}}(\underset{\sim}{b}) = p(B_1 = b_1, \dots, B_n = b_n)$.

PF

If θ were known you could use $f_{\underset{\sim}{B}}(\underset{\sim}{b} \mid \theta)$ to predict the dataset before it arrives: by the IID assumption $p(B_1 = b_1, \dots, B_n = b_n \mid \theta) = p(B_1 = b_1 \mid \theta) \cdots p(B_n = b_n \mid \theta)$

Recall that

$p(B_i = b_i \mid \theta) = \theta^{b_i}(1-\theta)^{1-b_i}$ for $b_i = 0, 1$ so

$$f_{\underset{\sim}{B}}(\underset{\sim}{b} \mid \theta) = \prod_{i=1}^{n} \theta^{b_i}(1-\theta)^{1-b_i} = \theta^{\sum_{i=1}^n b_i}(1-\theta)^{n-\sum_{i=1}^n b_i}$$

$$= \theta^s(1-\theta)^{n-s}$$

with $s = \sum_{i=1}^n b_i$.

Def. n rv $\Sigma_1, \dots, \Sigma_n$ have a continuous joint distribution if you can find a function $f_{\underset{\sim}{\Sigma}}$ on $\mathbb{R}^n$ such that for every (non-weird) subset $C \subset \mathbb{R}^n$

$$P[(Y_1,\dots,Y_n)\in C] = \int\cdots\int_C f_{Y_1\cdots Y_n}(y_1,\dots,y_n)\,dy_1\cdots dy_n$$

$f_{\underset{\sim}{Y}}(y)$ is the joint PDF (probability density function) of $\underset{\sim}{Y}$.

more compactly

$$P(\underset{\sim}{Y}\in C) = \int\cdots\int_C f_{\underset{\sim}{Y}}(y)\,dy.$$

Consequences of this def.

① If the joint dist. of $\underset{\sim}{Y}$ is continuous, then $f_{\underset{\sim}{Y}}(y) = \dfrac{\partial^n}{\partial y_1\cdots\partial y_n} F_{\underset{\sim}{Y}}(y)$.

mixed discrete/continuous with n rv

random vectors behave just as they do with 2 rv.

Example (trial continued)

more realistically, θ would be unknown, and you can think about the

joint dist. of $(\underset{\sim}{B}, \theta) = (B_1, \ldots, B_n, \theta)$, in which the B_i are discrete and $0 < \theta < 1$, is continuous.

Marginal distributions

If you know the joint PDF $f_{\underset{\sim}{Y}}$ of $\underset{\sim}{Y}$, you can work out the marginal distribution of any subset of $(Y_1, \ldots, Y_n)$ by integrating $f_{\underset{\sim}{Y}}(y)$ over the elements of $(Y_1, \ldots, Y_n)$ that are not in the subset.

Example

$\underset{\sim}{Y} = (Y_1, Y_2, Y_3, Y_4)$

$$f_{Y_1}(y_1) = \iiint f_{\underset{\sim}{Y}}(y)\, dy_2\, dy_3\, dy_4$$

$$f_{Y_2, Y_3}(y_2, y_3) = \iint f_{\underset{\sim}{Y}}(y)\, dy_1\, dy_4$$

and so on.

Similarly, you can work out a marginal CDF by sending the other components

to ∞: for example

$$F_{X_1}(x_1) = P(X_1 \le x_1) = P(X_1 \le x_1, X_2 < \infty, \dots, X_n < \infty)$$

$$= \lim_{x_2 \to +\infty, \dots, x_n \to +\infty} F_{\underset{\sim}{X}}(\underset{\sim}{x}).$$

Definition

n rvs $X_1, \dots, X_n$ are independent if

(non-weird)

for any n sets $A_1, \dots, A_n$ of real numbers

$$P(X_1 \in A_1, \dots, X_n \in A_n) = \prod_{i=1}^{n} P(X_i \in A_i).$$

Immediate consequences

① $X_1, \dots, X_n$ independent iff

$$F_{\underset{\sim}{X}}(\underset{\sim}{x}) = \prod_{i=1}^{n} F_{X_i}(x_i)$$

② $X_1, \dots, X_n$

independent iff $f_{\underset{\sim}{X}}(\underset{\sim}{x}) = \prod_{i=1}^{n} f_{X_i}(x_i)$

(2.10)

Def. Starting with a univariate PMF or PDF $f_{I_i}(y_i)$, n rvs $(I_1, \ldots, I_n)$ form a random sample of size n from f_{I_i} if the I_i are independent and all of them have marginal PMF or PDF f_{I_i} $\leftrightarrow$ i.e., if the I_i are an independent identically distributed (IID) sample from f_{I_i}.

Example

deer at UCSC: some have a disease (chronic wasting disease)

population: all deer living within UCSC boundary

sample: the observed deer

$N = ?$ (≈ 800)

	disease?
1	Is
2	a
:	
n	os

IID

	disease?
1	Is
2	a
:	
n	os

$n =$

y_1
:
y_n

1 = Y
0 = N

mean $\theta = ?$ (unknown)

mean $\bar{y} = \hat{\theta}$ ← estimate of

"y-bar"

↖ "theta-hat"

Shorthand for the diagram:

$(X_i \mid \theta) \overset{IID}{\sim} \text{Bernoulli}(\theta)$

$(i = 1, \dots, n)$

(end example forthwith)

Definition

Start with random vector

$\underset{\sim}{X} = (X_1, \dots, X_n)$; partition it into 2 subvectors $\underset{\sim}{X} = (\underset{\sim}{Y}, \underset{\sim}{Z})$, $\underset{\sim}{Y} = (Y_1, \dots, Y_k)$, $1 \le k \le n-1$

$\underset{\sim}{Z} = (Z_1, \dots, Z_{n-k})$

Then for every point $\underset{\sim}{z}$ for which $f_{\underset{\sim}{Z}}(\underset{\sim}{z}) > 0$, the conditional distribution of $\underset{\sim}{Y}$ given $\underset{\sim}{Z}$ is

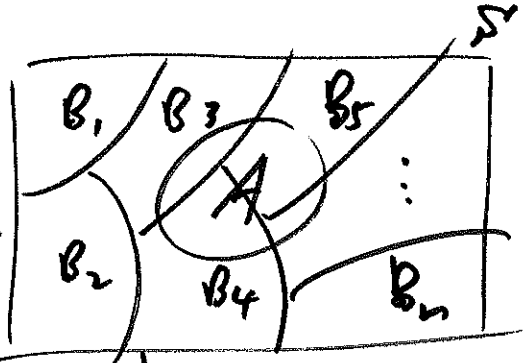
$$f_{\underset{\sim}{Y} \mid \underset{\sim}{Z}}(\underset{\sim}{y} \mid \underset{\sim}{z}) = \frac{f_{\underset{\sim}{Y}\underset{\sim}{Z}}(\underset{\sim}{y}, \underset{\sim}{z})}{f_{\underset{\sim}{Z}}(\underset{\sim}{z})}, \quad \underset{\sim}{y} \in \mathbb{R}^k$$

from which

$$f_{\underset{\sim}{Y}\underset{\sim}{Z}}(\underset{\sim}{y}, \underset{\sim}{z}) = f_{\underset{\sim}{Z}}(\underset{\sim}{z})\, f_{\underset{\sim}{Y} \mid \underset{\sim}{Z}}(\underset{\sim}{y} \mid \underset{\sim}{z}).$$

Multivariate Law of Total Probability

You'll recall that if A is an event & you're



trying to compute $P(A)$ & it's hard, one idea is to find another aspect of the world B upon which A depends, such that the events $B_1, \ldots, B_n$ form a partition;

then $P(A) = \sum_{i=1}^{n} P(A \cap B_i) = \sum_{i=1}^{n} P(B_i) P(A|B_i)$.

This has an analogue with continuous rvs:

---

using the notation in the definition of conditional distributions

$$f_{\underset{\sim}{X}}(\underset{\sim}{x}) = \int_{\mathbb{R}^{n-k}} \cdots \int f_{\underset{\sim}{Z}}(\underset{\sim}{z}) f_{\underset{\sim}{X}|\underset{\sim}{Z}}(\underset{\sim}{x}|\underset{\sim}{z}) \, d\underset{\sim}{z}.$$

$\underset{\sim}{x}$: like A; $\underset{\sim}{z}$: like B_i; $f_{\underset{\sim}{X}|\underset{\sim}{Z}}(\underset{\sim}{x}|\underset{\sim}{z})$: like $P(A|B_i)$

Multivariate Bayes's Theorem

using the same notation,

$$f_{\underset{\sim}{Z}|\underset{\sim}{X}}(\underset{\sim}{z}|\underset{\sim}{y}) = f_{\underset{\sim}{Z}}(\underset{\sim}{z})\, f_{\underset{\sim}{X}|\underset{\sim}{Z}}(\underset{\sim}{x}|\underset{\sim}{z})$$

(posterior info) (prior info) (likelihood info)

$\underset{\sim}{y}$ = unknown data

The usual application (normalizing constant) $\rightarrow f_{\underset{\sim}{X}}(\underset{\sim}{y})$.

of this in statistics is as follows.

Def. $\underset{\sim}{Z}$ a random vector with multivariate distribution $f_{\underset{\sim}{Z}}(\underset{\sim}{z})$; then random variables $X_1, \ldots, X_n$ are conditionally independent given $\underset{\sim}{Z}$ if for all $\underset{\sim}{z}$ with $f_{\underset{\sim}{Z}}(\underset{\sim}{z}) > 0$,

$$f_{\underset{\sim}{X}|\underset{\sim}{Z}}(\underset{\sim}{x}|\underset{\sim}{z}) = \prod_{i=1}^{n} f_{X_i|\underset{\sim}{Z}}(x_i|\underset{\sim}{z}).$$

Earlier example, revisited

Remember the machine with a θ dial that can make IID coin tossing with $P(\text{heads}) = \theta$?

We agreed earlier that, if θ is unknown to you, ① the results of the coin tossing $X_1, X_2, \ldots$ are dependent, because there is useful information in any subset of them for predicting any other subset, but ② the X_i become conditionally independent given θ, because once you know θ there's no longer any useful information in the X_i to predict other X_i.

This is why — in both the clinical trial example & the (nuts & bolts) example — we model the data values Y_i as

$(Y_i \mid \theta) \overset{\text{conditionally IID}}{\sim} \text{Bernoulli}(\theta)$.

Functions of a rv

Case 1: discrete

(univariate)

X discrete rv with pmf $f_X(x)$;

$Y = h(X)$ for some function h defined on {possible values of X}. Then

$$f_Y(y) = P(Y = y) = P[h(X) = y]$$

$$= \sum_{\{x : h(x) = y\}} f_X(x)$$

Example

Discrete

$X \sim \text{Uniform}\{1, 2, \ldots, 9\}$

The median of this distribution is 5;

$Y = |X - 5|$ ($= h(X)$) keeps track of how far X is from the median.

y	X such that $X+Y=y$	$P(Y=y)$
0	5	1/9
1	4 or 6	2/9
2	3 or 7	2/9
3	2 or 8	2/9
4	1 or 9	2/9
		1

Case 2: Continuous

X continuous rv with PDF $f_X(x)$;

$Y = h(X)$ as before.

The CDF $F_Y(y)$ can be worked out as follows:

$$F_Y(y) = P(Y \le y) = P[h(X) \le y]$$

$$= \int_{\{x : h(x) \le y\}} f_X(x)\,dx$$

and if Y is also continuous

$$f_Y(y) = \frac{d}{dy} F_Y(y)$$

(at every point y where F_Y is differentiable).

Example X = rate at which customers served in a queue at the bank

Natural to model X as continuous, with CDF F_X.
(also, $X > 0$)

Turns out that the average waiting time is $Y = \frac{1}{X} = h(X)$. You can get the PDF of Y in 2 steps:

① work out CDF of Y

② differentiate with respect to y

① (for $y > 0$)

$$F_Y(y) = P(Y \le y) = P[h(X) \le y]$$

$$= P\left(\frac{1}{X} \le y\right) = P\left(X \ge \frac{1}{y}\right)$$

since X is continuous

$$= 1 - P\left(X < \frac{1}{y}\right) = 1 - P\left(X \le \frac{1}{y}\right)$$

$$= 1 - F_X\left(\frac{1}{y}\right)$$ and now

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{d}{dy}\left(1 - F_X\left(\frac{1}{y}\right)\right)$$

chain rule

$$= -f_X\left(\frac{1}{y}\right)(-y^{-2}) = \frac{f_X\left(\frac{1}{y}\right)}{y^2}$$

(14 Aug 17)

Example: $X \sim \text{Uniform}[-1, +1]$ (continuous)

$Y = X^2$

find PDF of Y

First note that Y's possible values are $[0, 1)$.

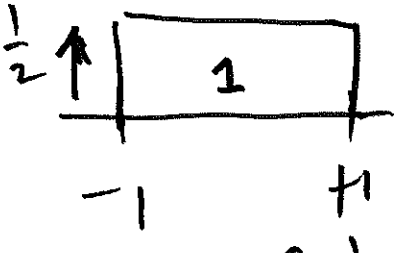
for $0 < y < 1$

① $F_Y(y) = P(Y \le y) = P(X^2 \le y)$ *

$= P(-\sqrt{y} \le X \le \sqrt{y}) = \int_{-\sqrt{y}}^{\sqrt{y}} f_X(x)\, dx$

$= \frac{1}{2} x \Big|_{-\sqrt{y}}^{\sqrt{y}}$

$= \sqrt{y}$

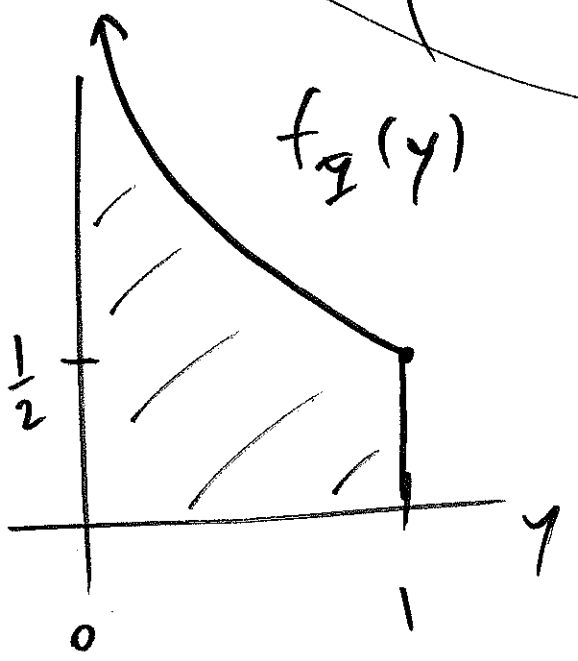


$$f_X(x) = \begin{cases} \frac{1}{2} & -1 \le x \le 1 \\ 0 & \text{else} \end{cases}$$

② thus

$$f_Y(y) = \frac{d}{dy} F_Y(y)$$

So $f_Y(y) = \begin{cases} \frac{d}{dy} y^{1/2} = \frac{1}{2\sqrt{y}} & \text{for } 0 < y < 1 \\ 0 & \text{else} \end{cases}$



This density is unbounded at 0 (!).

Easy theorem

X continuous rv with pdf $f_X(x)$,

$Y = aX + b$ $(a \neq 0)$ (linear transformation)

$$\rightarrow f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right).$$

Interesting and useful fact

X continuous with CDF $F_X(x)$;

what's the distribution of

$Y = F_X(X)$?

$$F_Y(y) = P(Y \le y) = P[F_X(X) \le y)$$

$$= P\left[X \le F_X^{-1}(y)\right] = F_X\left[F_X^{-1}(y)\right] = y$$

for $0 < y < 1$

But the dist. with $F_Y(y) = y$ for $0 < y < 1$ is the Uniform $(0,1)$ distribution (!)

---

probability Integral Transform

X continuous with CDF F_X, $Y = F_X(X)$

$\rightarrow Y \sim$ Uniform $(0,1)$ or $[\ \ ]$

---

why is this useful?

Converse is also true:

$Y \sim$ Uniform $[0,1]$, F_X a continuous CDF with quantile function

$F_X^{-1} \rightarrow X = F_X^{-1}(Y) \sim F_X$

This is the practical basis for the generation of many forms of pseudo-random numbers:

It turns out to be easy to generate pseudo-Uniform(0,1) values; therefore if you want to generate pseudo-random #s from a distribution with CDF F_X and F_X^{-1} is easy & fast to compute,

---

Algorithm (Quiz 6)

$U_1, \ldots, U_n \overset{IID}{\sim} \text{Uniform}(0,1)$

$\rightarrow F_X^{-1}(U_1), \ldots, F_X^{-1}(U_n) \overset{IID}{\sim} F_X$

Earlier Example revisited

If $X \sim \text{Exponential}(\lambda)$ $(\lambda > 0)$, its PDF is

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x > 0 \\ 0 & \text{else} \end{cases}$$ (11.91)

Earlier we saw that

$$F_X(x) = \begin{cases} 0 & x \le 0 \\ 1 - e^{-\lambda x} & x > 0 \end{cases}$$

and $F_{\Sigma}^{-1}(p) = -\frac{\log(1-p)}{\lambda}$ $(0<p<1)$

(R demo)

Now you can see immediately that if $U \sim \text{Uniform}(0,1)$ so is $(1-U)$, so to generate IID Exponential(λ) rv you just compute $-\frac{1}{\lambda}\log U_i$, $U_i \overset{IID}{\sim} \text{Uniform}(0,1)$

(rexp) R

Why do people want/need pseudo-random numbers?

Some stochastic (probabilistic) models of real-world phenomena are too complicated to fully characterize mathematically in closed form; one highly useful method in such situations is to conduct a simulation (computer-based) study driven by pseudo-random numbers.

Bedrock method in data science today.

The method used above for working out the distribution of $Y = \frac{1}{X}$ can be generalized, as follows. Some functions $h(X)$ are nice, in that they are both differentiable and one-to-one (invertible)

Calculus reminder (real-valued)

If $h(x)$ is differentiable and one-to-one (1-1) for x in the open interval (a, b), then h is either monotonically increasing or decreasing, and h is also continuous, so it transforms the interval (a, b) to another open interval $h[(a, b)] = (\alpha, \beta)$ called the image of (a, b) under h.

Since h is invertible, it makes sense

to talk about $y = h(x) \leftrightarrow x = h^{-1}(y)$.

Theorem: X continuous rv with PDF $f_X(x)$ and for which $P(a < X < b) = 1$ (a, b could be infinite); $Y = h(X)$, with h differentiable and 1-1 for $a < x < b$; (α, β) image of (a, b) under h; $h^{-1}(y)$ inverse function of $h(x)$ for $\alpha < y < \beta$ → PDF (chain rule),

of Y is

$$f_Y(y) = \begin{cases} f_X[h^{-1}(y)] \left| \frac{dh^{-1}(y)}{dy} \right| & \text{for } \alpha < y < \beta \\ 0 & \text{else} \end{cases}$$

Easy short-hand way to remember this: "Multiply" both sides by $|dy|$ to get $f_Y(y)|dy| = f_X(x)|dx|$

$y = h(x)$

$x = h^{-1}(y)$

Earlier example, revisited

$Y = h(X) = \frac{1}{X}$: average waiting time in the bank queue

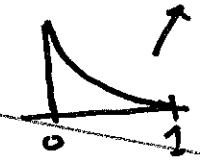
Here $y = h(x) = \frac{1}{x}$ so $x = h^{-1}(y) = \frac{1}{y}$

and $\frac{d}{dy}\frac{1}{y} = -\frac{1}{y^2}$; thus $f_Y(y) = \frac{f_X(\frac{1}{y})}{y^2}$ as before

---

Example At time 0, v organisms introduced into large tank of water with nutrients; X = rate of growth. Under one model that's realistic in some circumstances, at time t the predicted population size would be $Y = ve^{Xt}$ (exponential growth).

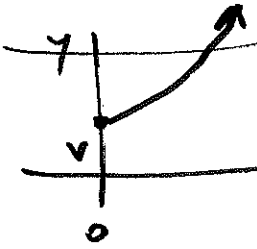
X unknown, modeled with $f_X(x) = \begin{cases} 3(1-x)^2 & 0 < x < 1 \\ 0 & \text{else} \end{cases}$



$y = h(x) = ve^{xt}$

$x = 0 \to y = v$

$x = 1 \to y = ve^{t}$ } image



strictly increasing

$$\frac{y}{v} = e^{xt} \rightarrow \log\left(\frac{y}{v}\right) = xt \rightarrow x = h^{-1}(y) = \frac{1}{t}\log\left(\frac{y}{v}\right)$$

$$\frac{d}{dy}\frac{1}{t}\log\left(\frac{y}{v}\right) = \frac{1}{t}\left(\frac{y}{v}\right)^{-1}\cdot\frac{1}{v} = \frac{1}{ty} \quad \text{RHS}$$

$$f_Y(y) = \begin{cases} \dfrac{3\left[1-\frac{1}{t}\log\left(\frac{y}{v}\right)\right]^2}{ty} & v < y < ve^t \\ 0 & \text{else} \end{cases}$$

(9 May 19)

Functions of 2 or more rvs

Case 1: discrete

n rvs $X_1, \ldots, X_n$ discrete joint dist. with joint PF $f_{\underset{\sim}{X}}(\underset{\sim}{x})$;

define $\begin{cases} Y_1 = h_1(X_1, \ldots, X_n) \\ \vdots \qquad\qquad \vdots \\ Y_m = h_m(X_1, \ldots, X_n) \end{cases}$ $(m \geq 1)$

real-valued $\quad (h_j : \mathbb{R}^n \rightarrow \mathbb{R})$