Mathematical Statistics I
Chapter 3: Joint Distributions

Jesse Wheeler

Contents

1 Introduction

Introduction

- This material is based on the textbook by Rice (2007, Chapter 3).

- Our goal is to better understand the joint probability structure of more than one random variable, defined on the same sample space.

- One reason that studying joint probabilities is an important topic is that it enables us to use what we know about one variable to study another.

Joint cdf

- Just like the univariate case, the joint behavior of two random variables, X and Y, is determined by the cumulative distribution function

$$F(x,y) = P(X \leq x, Y \leq y).$$

- This is true for both discrete and continuous random variables.

- The any set $A \subset \mathbb{R}^2$, the joint cdf can give $P\big((X,Y) \in A\big)$.

- For example, let A be the rectangle defined by $x_1 < X < x_2$, and $y_1 < Y < y_2$. (It helps to draw a picture...)

	y			
x	0	1	2	3
0	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	0
1	0	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$

Table 1: Frequency table for X and Y, flipping a fair coin three times.

- $F(x_2, y_2)$ gives $P(X < x_2, Y < y_2)$, an area that is too big, so we subtract off pieces

  - $F(x_2, y_1) = P(X < x_2, Y < y_1)$ (we already have the area $X < x_2$, but now subtract away the area $Y < y_1$).
  - $F(x_1, y_2) = P(X < x_1, Y < y_2)$ (Now subtracting the area $X < x_1$)
  - We have "double subtracted" the area $\{X < x_1, Y < y_1\}$, so we add it back.

$$P\big((X, Y) \in A\big) = F(x_2, y_2) - F(x_2, y_1) - F(x_1, y_2) + F(x_1, y_1).$$

- The definition also applies to more than two random variables.

- Let $X_1, \ldots, X_n$ be jointly distributed random variables defined on the same sample space. Then

$$F(x_1, x_2, \ldots, x_n) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n).$$

- Like the univariate case, we can also define the pmf and pdf of jointly distributed random variables as well.

2 Discrete Random Variables

Discrete Random Variables

Definition: Joint pmf
Let X and Y be discrete random variables define on the same sample space, and take on values $x_1, x_2, \ldots$ and $y_1, y_2, \ldots$, respectively. The *joint pmf* (or joint frequency function), is

$$p(x_i, y_j) = P(X = x_i, Y = y_j).$$

- For discrete RVs, it's often useful to describe the joint pmf as a frequency table.

- Suppose a fair coin is tossed 3 times. Let X denote the number of heads on the first toss, and Y the total number of heads.

- The sample space is
$$\Omega = \{hhh, hht, hth, thh, htt, tht, tth, ttt\}.$$

- The joint pmf can be expressed as the frequency table below (Table 1).

- Note that the probabilities in Table 1 sum to one.

- Using the probability laws we have already learned, we can calculate *marginal* probabilities.

$$
\begin{aligned}
p_Y(0) &= P(Y = 0) \\
&= P(Y = 0, X = 0) + P(Y = 0, X = 1) \\
&= \frac{1}{8} + 0 = \frac{1}{8} \\
p_Y(1) &= P(Y = 1) \\
&= P(Y = 1, X = 0) + P(Y = 1, X = 1) \\
&= \frac{2}{8} + \frac{1}{8} = \frac{3}{8}.
\end{aligned}
$$

- In general, to find the frequency function for Y and X, we just need to sum the appropriate columns or rows, respectively.

- $p_X(x) = \sum_i P(x, y_i)$ and $p_Y(y) = \sum_j P(x_j, y)$.

- The case with multiple random variables is similar:

$$
p_{X_i}(x_i) = \sum_{x_j : j \neq i} p(x_1, x_2, \ldots, x_n).
$$

- We can also get marginal frequencies for more than one variable:

$$
p_{X_i X_j}(x_i, x_j) = \sum_{x_k : k \notin \{i, j\}} p(x_1, x_2, \ldots, x_n).
$$

Example: Multinomial Distribution

- The *multinomial* distribution is a generalization of the binomial distribution.

- Suppose there are n independent trials, each with r possible outcomes, with probabilities $p_1, p_2, \ldots, p_r$, respectively.

- Let N_i be the total number of outcomes of type i in the n trials, with $i \in \{1, 2, \ldots, r\}$.

- The probability of any particular sequence $(N_1, N_2, \ldots, N_r) = (n_1, n_2, \ldots, n_r)$ is

$$
p_1^{n_1} p_2^{n_2} \cdots p_r^{n_r}
$$

- The total number of ways to do this was an identity from Chapter 1 (Proposition 1.3):

$$
\binom{n}{n_1 \; \cdots \; n_r}.
$$

- Combining this gives us the pmf of the multinomial distribution:

Multinomial Distribution
Let $N_1, N_2, \ldots, N_r$ be random variables that follow a multinomial distribution with parameters N and $(p_1, \ldots, p_r)$. The joint pmf is

$$
p(n_1, n_2, \ldots, n_r) = \binom{n}{n_1 \; \cdots \; n_r} p_1^{n_1} p_2^{n_2} \cdots p_r^{n_r}
$$

- The marginal distribution for any N_i can be found by summing the joint frequency function over the other n_j.

- While possible, this is a non-trivial algebraic exercise.

- The simple alternative is to reframe the problem: Let N_i be the number of successes in n trials, and $\tilde{N}_i = \sum_{j \neq i} N_j$ be the number of failures. The probability of success is still p_i, leaving the probability of failure to be $1 - p_i$.

- Thus, we see that the marginal distribution for N_i must follow a binomial distribution:

$$p_{N_i}(n_i) = \sum_{n_j : j \neq i} \binom{n}{n_1 \; \cdots \; n_r} p_1^{n_1} p_2^{n_2} \cdots p_r^{n_r}$$
$$= \binom{n}{n_i} p_i^{n_i} (1 - p_i)^{n - n_i}$$

3 Continuous Random Variables

Continuous Random Variables

- Let X, Y be continuous random variables with joint cdf $F(x, y)$.

- Their *joint density function* is a piecewise continuous function of two variables, $f(x, y)$.

- A few properties:

  - $f(x, y) \geq 0$ for all $(x, y) \in \mathbb{R}$ (or the support).
  - $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1$.
  - For any "measureable set" $A \subset \mathbb{R}^2$, $P\big((X, Y) \in A\big) = \int \int_A f(x, y) dx dy$
  - In particular, $F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(u, v) du dv$.

- From the fundamental theorem of multivariable calculus, it follows that

$$f(x, y) = \frac{\partial^2}{\partial x \partial y} F(x, y),$$

wherever the derivative is defined.

Finding joint probabilities
Let X, Y be jointly defined RVs with pdf

$$f(x, y) = \frac{12}{7}(x^2 + xy), \quad 0 \leq x \leq 1, \quad 0 \leq y \leq 1.$$

Find $P(X > y)$.

$$P(X > Y) = \frac{12}{7} \int_0^1 \int_0^x (x^2 + xy) dy dx$$
$$= \frac{9}{14}.$$

Marginal cdf

The *marginal cdf* of X, denoted F_X, is

$$
\begin{aligned}
F_X(x) &= P(X \le x) \\
&= P(X \le x \cap Y \in \mathbb{R}) = P(X \le x \cap Y < \infty) \\
&= \lim_{y \to \infty} F(x, y) \\
&= \int_{-\infty}^{x} \int_{-\infty}^{\infty} f(u, y) dy du.
\end{aligned}
$$

By taking the derivative of both sides of the equation, we get the *marginal density* of X:

$$
f_X(x) = F_X'(x) = \int_{-\infty}^{\infty} f(x, y) dy.
$$

Calculating Marginal Densities

Using the same joint distribution as the previous example, find the marginal density of X:

$$
\begin{aligned}
f_X(x) &= \int_Y f(x, y) dy \\
&= \frac{12}{7} \int_0^1 (x^2 + xy) dy \\
&= \frac{12}{7} \left(x^2 y + \frac{x}{2} y^2 \right) \Big|_0^1 \\
&= \frac{12}{7} \left(x^2 + \frac{x}{2} \right)
\end{aligned}
$$

More than two random variables

- For several jointly continuous random variables, we can make the obvious generalizations.

- That is, to find the *marginal* densities, we need to "marginalize-" or "integrate-" out the *nusaince* variables.

- This means integrating out any combination of variables that we want.

- Example: Let X, Y, and Z be jointly continuous RVs with pdf $f(x, y, z)$. Then the two-dimensional marginal distribution of X and Z is:

$$
f_{XZ}(x, z) = \int_{-\infty}^{\infty} f(x, y, z) dy.
$$

Example: constructing bivariate cdfs

- Suppose that $F(x)$ and $G(y)$ are cdfs for random variables X and Y, resp.

- It can be shown that the following function, $H(x, y)$, is always a bivariate cdf for all $-1 \le \alpha \le 1$:

$$
H(x, y) = F(x) G(y) \Big(1 + \alpha \big(1 - F(x) \big) \big(1 - G(y) \big) \Big).
$$

- Because $\lim_{x \to \infty} F(x) = \lim_{y \to \infty} G(x) = 1$, the marginal distributions are:

$$
\lim_{y \to \infty} H(x, y) = F(x)
$$

$$
\lim_{x \to \infty} H(x, y) = G(y)
$$

- Thus, we can use this approach to build an infinite number of biviariate distributions that have a particular marginal distribution.

- One important example is when the marginal distributions are uniformly distributed.

- Let $F(x) = x, 0 \leq x \leq 1$, and $G(y) = y, 0 \leq y \leq 1$.

- By selecting $\alpha = -1$, we have

$$H(x,y) = xy[1 - (1-x)(1-y)]$$
$$= x^2 y + y^2 x - x^2 y^2, \quad 0 \leq x, y \leq 1.$$

- The density is

$$h(x,y) = \frac{\partial^2}{\partial x \partial y} H(x,y)$$
$$= 2x + 2y - 4xy, \quad 0 \leq x, y \leq 1.$$

- Here is a link to a 3D rendering of this function.

- Now, let's select $\alpha = 1/2$:

$$H(x,y) = xy\left(1 + \frac{1}{2}\left(1 - F(x)\right)\left(1 - G(y)\right)\right)$$
$$= \frac{1}{2}x^2 y^2 - \frac{1}{2}x^2 y - \frac{1}{2}xy^2 + \frac{3}{2}xy.$$

- Taking the derivative, we get:

$$h(x,y) = \frac{\partial^2}{\partial x \partial y} H(x,y)$$
$$= 2xy - x - y + \frac{3}{2}, \quad 0 \leq x, y \leq 1.$$

- Here is a link to a 3D rendering of this function.

- The last two joint cdfs were examples of a *copula*.

  Definition: Copulas

  A copula is a joint cdf that has uniform marginal distributions.

- Let $C(u,v)$ be a copula. One immediate consequence of the definition is that if U and V are uniform random variables, then $P(U \leq u) = C(u,1) = u$, and $P(V \leq v) = C(1,v) = v$.

- Let $C(u,v)$ be a copula, we will restrict ourselves to the case where it is twice differentiable, such that $c(u,v) = \frac{\partial^2}{\partial u \partial v} C(u,v) \geq 0$.

- let F_X and F_Y be the cdfs of X and Y, resp.

- Now define $U = F_X(X)$, and $V = F_Y(Y)$. From Proposition 2.2, U and V are uniformly distributed.

- Now consider the function $H(x,y) = C(u,v) = C\big((F_X(x), F_Y(y)\big)$.

- Thus, by the property that $C(u, 1) = u$ and $C(1, v) = v$, we have

$$C\big(F_X(x), 1\big) = F_X(x)$$
$$C\big(1, F_Y(y)\big) = F_Y(y).$$

Therefore by definition, $F_{XY}(x, y) = H(x, y) = C\big((F_X(x), F_Y(y)\big)$.

- Using the chain rule, we can differentiate to obtain

$$f_{XY}(x, y) = c\big(F_X(x), F_Y(y)\big) f_X(x) f_Y(y).$$

- *Takeaway:* We took arbitrary marginal distributions F_X and F_Y, and created a family of joint density functions, defined by *any* copula. Thus: the marginal distributions do not determine the joint distribution.

- There is a Theorem known as Sklar's Theorem (Wikipedia contributors, 2025) that generalizes this statement: All joint distributions can be expressed using a copula and marginal distributions, *and* the representation is unique.

- That is, the copula can be thought of as a way to describe the dependence between the variables in any joint distribution.

Uniform on specific region

- So far when we have talked about *uniform distributions*, we think about being uniform over $[0, 1]$, or a higher dimensional box: $[a, b]^d$.

- It's often useful to have a uniform distribution for other regions of space.

- Let $R \subset \mathbb{R}^2$ be any region of interest. The two-dimensional uniform distribution over R is defined by the probability

$$P\big((X, Y) \in A\big) = \frac{|A|}{|R|},$$

where $|\,|$ denotes the measure of the area.

- Example: Suppose a point is chosen randomly in a disk of radius 1.

- The area of the disk is $\pi r^2 = \pi$, and therefore the joint pdf for the location (X, Y) is

$$f(x, y) = \begin{cases} \frac{1}{\pi} & x^2 + y^2 \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

- Now let R be the random variable denoting the distance of the point from the origin.

- Note that $R \leq r$ if and only if the point lies in a disk of radius r. This disk has area πr^2, and therefore the joint probability is

$$P(R \leq r) = \frac{\pi r^2}{\pi} = r^2, \quad 0 \leq r \leq 1.$$

- Taking a derivative, the corresponding density function is

$$f_R(r) = 2r, \quad 0 \leq r \leq 1.$$

- Now let us compute the marginal density of the x coordinate:

$$
\begin{aligned}
f_X(x) &= \int_{-\infty}^{\infty} f(x,y)dy \\
&= \int_{-\infty}^{\infty} \frac{1}{\pi} \times 1[x^2 + y^2 \leq 1]dy \\
&= \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \frac{1}{\pi}dy \\
&= \frac{2}{\pi}\sqrt{1-x^2}, \quad -1 \leq x \leq 1.
\end{aligned}
$$

4 Independent Random Variables

Independence

Definition: Independent Random Variables
Random variables $X_1, X_2, \ldots, X_n$ are said to be *independent* if their joint cdf factors into the product of their marginal cdf's:

$$
F(x_1, x_2, \ldots, x_n) = F_{X_1}(x_1)F_{X_2}(x_2)\cdots F_{X_n}(x_n)
$$

for all $x_1, x_2, \ldots, x_n$.

- This definition holds for both continuous and discrete random variables.

- For discrete RVs, it is equivalent to state that their joint pmf factors.

- For continous RVs, it is equivalent to state that their joint pdf factors.

- To see why this is true, consider the case of two RVs, X, Y.

- From the definition, if they are independent, then $F(x,y) = F_X(x)F_Y(y)$.

- Taking the second mixed partial derivative makes it immediately clear that the joint pdf $f(x,y)$ factors (assuming all densities exist).

- Conversely, suppose that the densities factor. Then by definition:

$$
\begin{aligned}
F(x,y) &= \int_{-\infty}^{x} \int_{-\infty}^{y} f(u,v)dvdu \\
&= \int_{-\infty}^{x} \int_{-\infty}^{y} f_X(u)f_Y(v)dvdu \\
&= \left(\int_{-\infty}^{x} f_X(u)du \right)\left(\int_{-\infty}^{y} f_Y(v)dv \right) \\
&= F_X(x)F_Y(y).
\end{aligned}
$$

- It can also be shown that the definition implies that if X and Y are independent, then

$$
P(X \in A, Y \in B) = P(X \in A)P(X \in B).
$$

- Furthermore, if g and h are functions on $\mathbb{R}$, then $Z = g(X)$ and $W = h(Y)$ are also independent.

	y			
x	0	1	2	3
0	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	0
1	0	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$

Table 2: Frequency table for X and Y, flipping a fair coin three times.

5 Conditional Distributions

5.1 Discrete random variables

Conditional distributions: discrete RVs

- If X and Y are jointly distributed discrete RVs, the *conditional probability* that $X = x_i$ given that $Y = y_i$ is

$$P(X = x_i | Y = y_i) = \frac{P(X = x_i, Y = y_i)}{P(Y = y_i)}$$
$$= \frac{p_{XY}(x_i, y_i)}{P_Y(y_i)},$$

- If $p_Y(y_i) = 0$, the probability above is defined to be zero.

- We denote this conditional probability as $p_{X|Y}$.

- It's important to note that the conditional pmf is a genuine pmf, as it is non-negative and sums to one.

- If X and Y are independent, $p_{Y|X}(y|x) = p_Y(y)$.

- Let's return to a previous joint pmf example (Table 2).

- The conditional frequency function of X given $Y = 1$ is:

$$p_{X|Y}(0|1) = \frac{2/8}{3/8} = 2/3$$
$$p_{X|Y}(1|1) = \frac{1/8}{3/8} = 1/3$$

- The definition of the conditional frequency can be reexpressed as

$$p_{XY}(x, y) = p_{X|Y}(x|y)p_Y(y).$$

- By summing up over all possible values of y, we have the following

$$p_X(x) = \sum_y p_{X|Y}(x|y)p_Y(y).$$

- Both of these identities resemble what we have already seen previously when talking about probabilities: The multiplication principle and the law of total probability.

9

Example: Counting particles

Suppose that a particle counter is imperfect; for each particle, it detects the particle with probability $0 < p < 1$. If the number of incoming particles in a unit of time is a Poisson distribution with parameter λ, what is the distribution of the number of counted particles?

Let N denote the true number of particles, and X the number of counted particles. Because the probability that a particle is counted is independent, then if $N = n$, we have n independent Bernoulli random variables. In other words, $X|N = n$ has a binomial distribution with parameters n and p. By the law of total probability,

$$
\begin{aligned}
P(X = k) &= \sum_n p_{X|N}(x|n) p_N(n) \\
&= \sum_{n=k}^{\infty} \binom{n}{k} p^k (1-p)^{n-k} \frac{\lambda^n e^{-\lambda}}{n!} \\
&= p^k e^{-\lambda} \sum_{n=k}^{\infty} \frac{n!}{(n-k)!k!} (1-p)^{n-k} \frac{\lambda^{(n-k)} \lambda^k}{n!} \\
&= \frac{(\lambda p)^k}{k!} e^{-\lambda} \sum_{n=k}^{\infty} \lambda^{n-k} \frac{(1-p)^{n-k}}{(n-k)!} \\
&= \frac{(\lambda p)^k}{k!} e^{-\lambda} \sum_{j=0}^{\infty} \lambda^j \frac{(1-p)^j}{j!} \\
&= \frac{(\lambda p)^k}{k!} e^{-\lambda} e^{\lambda(1-p)} \\
&= \frac{(\lambda p)^k}{k!} e^{-\lambda p}
\end{aligned}
$$

And therefore we see that the distribution of X is a Poisson distribution with parameter λp. This is a useful derivation for any situation where events occur following a Poisson process, and then with some probability p and additional, conditional event occurs. For instance, N might be the number of traffic accidents in a given time period, and each accident being fatal or non-fatal with probability p. Then X would be the number of fatal accidents.

5.2 Continuous Random Variables

The continuous case

- Although a formal argument is beyond the scope of this course, the definition for conditional density of $Y|X$ will be analogous to the discrete case.

Definition: Conditional density

Let X, Y be jointly continuous random variables with joint density $f_{XY}(x, y)$ and marginal densities $f_X(x)$ and $f_Y(y)$, respectively. Then the conditional density of Y given X is defined to be

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

if $0 < f_X(x) < \infty$, and 0 otherwise.

- A heuristic argument of why this definition makes sense is provided in Rice (2007, Section 3.5.2) using differentials.

- With our definition, we can express the joint density in terms of the marginal and conditional densities:

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

- We often use this expression to find marginal densities, using principles we have already discussed.

$$f_Y(y) = \int_{\mathbb{R}} f_{XY}(x, y) dx = \int_{\mathbb{R}} f_{Y|X}(y|x) f_X(x) dx.$$

- We can think of the above expression as the law of total probability for the continuous case.

Example: finding conditional densities

- Let X and Y be jointly distributed random variables with joint and marginal densities

$$f_{XY}(x, y) = \lambda^2 e^{-\lambda y}, \quad 0 \le x \le y$$
$$f_X(x) = \lambda e^{-\lambda x}, \quad x \ge 0$$
$$f_Y(y) = \lambda^2 y e^{-\lambda y}, \quad y \ge 0$$

- Note that if x is held constant, the joint density decays exponentially in y for $y \ge x$.

- If y is held constant, the joint density is constant for $0 \le x \le y$.

- Find the conditional densities for $Y|X$ and $X|Y$.

$$f_{Y|X}(y|x) = \frac{\lambda^2 e^{-\lambda y}}{\lambda e^{-\lambda x}} = \lambda e^{-\lambda(y-x)}, \quad y \ge x.$$

This density follows our intuition of what happens when x is fixed, namely that y appeared to decay exponentially. Now we see that $Y|X$ is exponentially distributed on the interval $[x, \infty)$. Now for $X|Y$,

$$f_{X|Y}(x|y) = \frac{\lambda^2 e^{-\lambda y}}{\lambda^2 y e^{-\lambda y}} = 1/y, \quad 0 \le x \le y.$$

Thus the conditional density of X given $Y = y$ is uniform on the interval $[0, y]$.

- Suppose we wanted to generate samples from the joint distribution (X, Y); how can this be done?

- Using what we have found about the conditional distributions, there are two simple ways for this to be done. Recall that the joint density is $f_{XY}(x, y) = f_{X|Y}(x|y) f_Y(y) = f_{Y|X}(y|x) f_X(x)$.

  1. We could generate X, which is an exponential random variable $(f_X(x))$. Then, we could generate Y conditioned on the simulated value of $X = x$, which follows an exponential distribution on the interval $[x, \infty)$.

  2. Similarly, we can note that Y has a gamma distribution, and therefore generate a y following a gamma distribution, and then generate a value from $X|Y = y$, which is uniform on $[0, y]$.

The rejection method

- We are often interested in generating random variables from a density function.

- If we have a closed form of the inverse cdf, we can use the "inverse cdf method" (Proposition 2.3).

- If a closed-form of the inverse cdf is not available, a commonly used approach is known as *rejection sampling*.

- Setup: let f be a density function we wish to simulate from, that is non-zero on an interval $[a, b]$.

- Pick a function $M(X)$ such that $M(x) \geq f(x)$ on $[a, b]$, and let

$$m(x) = \frac{M(x)}{\int_a^b M(x) dx}.$$

- Note that $m(x)$, as defined, is a probability density function. Then, to generate RV with density f, we can do the following:

  Step 1: Generate T with density m.

  Step 2: Generate $U \sim U[0, 1]$ independent of T. If $M(T) \times U \leq f(T)$, then we "accept" T as a sample $(X = T)$; *otherwise*, we "reject" and go back to Step 1.

Rejection Method Figure

- A geometric justification is randomly throwing a dart (uniformly) at Figure 1.
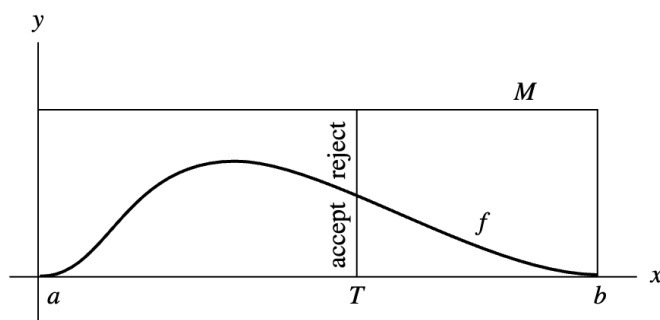

Figure 1: Illustration of the rejection method, copied from Rice (2007, Figure 3.15).

- If the dart lands below the curve f, record the x coordinate; otherwise, reject it. With enough throws, the distribution of x coordinates will be proportional to the height of the curve.

Rejection sampling

- A more formal argument using differentials is given in Rice (Example D 2007, Figure 3.15).

- In order for the rejection method to be worth-while (computationally efficient), it is important that the algorithm has high-acceptance (good choice of M), otherwise you may need a large number of samples because many are being rejected.

6 Functions of Jointly Distributed Random Variables

Convolutions

- Suppose that X and Y are discrete random variables that take values on the integers and joint pmf $p(x, y)$.

- Find the pmf of $Z = X + Y$.

- Note that $Z = z$ only when $X = x$ and $Y = z - x$, whenever x is an integer.

- Thus, using the law of total probability, we can write

$$p_Z(z) = \sum_{x=-\infty}^{\infty} p(x, z - x).$$

- If X and Y are independent, then $p(x, y) = p_X(x) P_Y(y)$, and

$$p_Z(z) = \sum_{x=-\infty}^{\infty} p_X(x) p_Y(z - x).$$

- This sum is called the *convolution* of the sequences p_X and p_Y.

- The continuous case is similar. Let X and Y be jointly continuous RVs, and $Z = X + Y$.

- If we want to find the cdf of Z, then:

$$
\begin{aligned}
F_Z(z) &= P(Z \le z) \\
&= P(X + Y \le z) \\
&= \int \int_{\{x+y \le z\}} f(x, y) \, dy \, dx \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{z-x} f(x, y) \, dy \, dx \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{z} f(x, v - x) \, dv \, dx \\
&= \int_{-\infty}^{z} \int_{-\infty}^{\infty} f(x, v - x) \, dx \, dv
\end{aligned}
$$

- Differentiating both sides, the fundamental theorem of calculus (with proper assumptions) gives

$$f_Z(z) = \int_{-\infty}^{\infty} f(x, z - x) dx$$

- Like in the discrete case, if X and Y are independent, then

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z - x) dx$$

- This integral is called the *convolution* of the functions f_X and f_Y.

Example: Sum of Exponential RVs
Suppose that the lifetime of an electrical component is exponentially distributed with rate λ, and that and independent and identical backup is available. If the system operates as long as one of the components is functional, and the components will not be replaced if they fail, what is the distribution of the life of the system?

Solution:

Let T_1 and T_2 denote the lifetimes of the two component, respectively. The lifetime of the system is $X = T_1 + T_2$. Thus, the pdf of X can be calculated as the convolution:

$$f_X(x) = \int_{-\infty}^{\infty} f_X(x) f_Y(z - x) dx$$

$$= \int_0^x \lambda e^{-\lambda t} \times \lambda e^{-\lambda(x-t)} dt$$

$$= \lambda^2 \int_0^x e^{-\lambda x} dt$$

$$= \lambda^2 x e^{-\lambda x}, \quad x \geq 0.$$

- In the previous example, note carefully the change in integration:

  - The exponential density is only positive when $t > 0$, and zero every where else.
  - Thus, from $(-\infty, 0)$, the integral is zero.
  - Similarly, we evaluate the density at $x - t$, and hence when $t > x$, the integral is also zero.

- You may notice that the density of $X = T_1 + T_2$ that we calculated is the same as a gamma distribution with parameters 2 and λ.

Quotients of random variables

- Let X and Y be jointly continuous random variables, and let $Z = Y/X$.

- Our derivation for the pdf of Z is similar as what we did with the sum: find the cdf, then take the derivative.

- $F_Z(z) = P(Z \leq z) = P(Y/X \leq z)$. Thus, we are interested in the probability of the set $\{x, y : y/x \leq z\}$.

- We have to be a little careful about what happens if $X = 0$, so we will split it into two parts:

  - If $x > 0$, then the set is $y \leq xz$.
  - If $x < 0$, then the set is $y \geq xz$.

  Thus,

  $$F_Z(z) = \int_{-\infty}^{0} \int_{xz}^{\infty} f(x, y) \, dy \, dx + \int_0^{\infty} \int_{-\infty}^{xz} f(x, y) \, dy \, dx.$$

- To remove dependence of the inner integrals on x, we make the change of variables $y = xv$:

  $$F_Z(z) = \int_{-\infty}^{0} \int_{z}^{-\infty} x f(x, xv) \, dv \, dx + \int_0^{\infty} \int_{-\infty}^{z} x f(x, xv) \, dv \, dx$$

  $$= \int_{-\infty}^{0} \int_{-\infty}^{z} (-x) f(x, xv) \, dv \, dx + \int_0^{\infty} \int_{-\infty}^{z} x f(x, xv) \, dv \, dx$$

  $$= \int_{-\infty}^{z} \int_{-\infty}^{\infty} |x| f(x, xv) \, dx \, dv$$

- And differentiating both sides, we obtain

  $$f_Z(z) = \int_{-\infty}^{\infty} |x| f(x, xz) \, dx.$$

- If X and Y are independent,

  $$f_Z(z) = \int_{-\infty}^{\infty} |x| f_X(x) f_Y(xz) dx.$$

Example: Cauchy density

- Let X and Y be independent, standard normal random variables.

- We wish to find the pdf of $Z = Y/X$.

- Using the expression we previously derived for the quotient of independent RVs, we have

$$f_Z(z) = \int_{-\infty}^{\infty} \frac{|x|}{2\pi} e^{-x^2/2} e^{-x^2 z^2/2} \, dx.$$

- Because the integrand is symmetric, we can re-express this as

$$\begin{aligned}
f_Z(z) &= 2 \int_0^{\infty} \frac{|x|}{2\pi} e^{-x^2/2} e^{-x^2 z^2/2} \, dx \\
&= \frac{1}{\pi} \int_0^{\infty} x e^{-x^2 \left((z^2+1)/2\right)} \, dx \\
&= \frac{1}{2\pi} \int_0^{\infty} e^{-u \left((z^2+1)/2\right)} \, du \\
&= \frac{1}{2\lambda\pi} \int_0^{\infty} \lambda e^{-u\lambda} \, du \\
&= \frac{1}{\pi(z^2+1)}, \quad -\infty < z < \infty.
\end{aligned}$$

- Here, I made the substitution $\lambda = (z^2 + 1)/2$, and the integral was calculated using the fact that the pdf of the exponential distribution integrates to one: $\int_0^{\infty} \lambda e^{-\lambda x} dx = 1$.

- This density is called the *Cauchy density*.

- Like the standard normal, the Cauchy density is symmetric about zero and bell-shaped, but the tails of the Cauchy tend to zero very slowly.

- Here is a link showing this comparison.

The General Case

- There is also a way to find the pdf of more general cases, though the derivation is outside the scope of this course.

- Let X and Y be jointly distributed, continuous RVs, and suppose we are interested in the joint pdf of $U = g_1(X, Y)$, $V = g_2(X, Y)$, where g_1 and g_2 are invertible functions with continuous partial derivatives.

- We will denote the inverse of g_1 and g_2 as $X = h_1(U, V)$ and $Y = h_2(U, V)$, respectively.

- The pdf of (U, V) can be calculated in two ways:

Proposition 3.1: Multivariate transformations
Under the assumptions above, the joint density of U and V is

$$\begin{aligned}
f_{UV}(u, v) &= f_{XY}\big(h_1(u, v), h_2(u, v)\big) \Big| J_h(u, v) \Big| \\
&= f_{XY}(x, y) \Big| J_g^{-1}(x, y) \Big|
\end{aligned}$$

for (u, v) such that $u = g_1(x, y)$ and $v = g_2(x, y)$ for some (x, y), and 0 otherwise.

- Above, we call J_f the *Jacobian determinant* (or just Jacobian) of f. It is equal to the matrix of partial derivatives:

$$J_h = \det \begin{bmatrix} \frac{\partial h_1}{\partial u} & \frac{\partial h_1}{\partial v} \\ \frac{\partial h_2}{\partial u} & \frac{\partial h_2}{\partial v} \end{bmatrix} = \left(\frac{\partial h_1}{\partial u} \right) \left(\frac{\partial h_2}{\partial v} \right) - \left(\frac{\partial h_2}{\partial u} \right) \left(\frac{\partial h_2}{\partial v} \right)$$

$$J_g = \det \begin{bmatrix} \frac{\partial g_1}{\partial x} & \frac{\partial g_1}{\partial y} \\ \frac{\partial g_2}{\partial x} & \frac{\partial g_2}{\partial y} \end{bmatrix} = \left(\frac{\partial g_1}{\partial x} \right) \left(\frac{\partial g_2}{\partial y} \right) - \left(\frac{\partial g_2}{\partial x} \right) \left(\frac{\partial g_2}{\partial y} \right)$$

- The reason these two expressions are equal is because $J_h = J_g^{-1}$, and we defined $x = h_1(u,v)$ and $y = h_2(u,v)$; you end up with the same result, but sometimes one might be an easier calculation than the other. Both versions are fairly common in textbooks and courses.

- I find the first line of the proposition more intuitive: it's a function of u and v, so let's start by calculating the Jacobian of the inverse transformation so that all variables are u and v.

- The textbook uses the second line approach. You might find this more intuitive, as it is a generalization of the single dimensional case.

Example: Polar Coordinates

- Suppose that X and Y are independent standard normal RVs. Their joint pdf is

$$f_{XY}(x,y) = \frac{1}{2\pi} e^{-(x^2/2) - (y^2/2)}.$$

- We wish the find the joint pdf of $R = \sqrt{X^2 + Y^2}$, and $\Theta = \arctan(y/x)$.

- Thus, we have

$$\begin{cases} g_1(x,y) = \sqrt{x^2 + y^2} = r \\ g_2(x,y) = \arctan(y/x) = \theta, \quad \text{if } x \neq 0, \text{ and } \theta = 0 \text{ o.w.} \end{cases}$$

- The inverse transformations are

$$\begin{cases} h_1(r,\theta) = r\cos\theta = x \\ h_2(r,\theta) = r\sin\theta = y. \end{cases}$$

- The Jacobian of the inverse transformation J_h is

$$\begin{aligned} J_h(r,\theta) &= \det \begin{bmatrix} \frac{\partial h_1}{\partial r} & \frac{\partial h_1}{\partial \theta} \\ \frac{\partial h_2}{\partial r} & \frac{\partial h_2}{\partial \theta} \end{bmatrix} \\ &= \det \begin{bmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{bmatrix} \\ &= r\cos^2\theta + r\sin^2\theta = r \end{aligned}$$

- Therefore, the joint distribution is

$$\begin{aligned} f_{R\Theta}(r,\theta) &= r f_{XY}(r\cos\theta, r\sin\theta) \\ &= \frac{r}{2\pi} e^{-r^2\cos^2\theta/2 - r^2\sin^2\theta/2} \\ &= \frac{r}{2\pi} e^{-r^2/2}. \end{aligned}$$

- As always, we can't forget the *support*, or values over which the density is positive. Here, because $(X,Y) \in \mathbb{R}^2$, the transformations imply that $\Theta \in [0, 2\pi]$, and $R \geq 0$.

Transformations of many variables

- Proposition 3.1 can be generalized to transformations of more than two random variables. If $X_1, \ldots, X_n$ have the joint density function $f_{X_1 \cdots X_n}$, and

$$Y_i = g_i(X_1, \ldots, X_n), \quad i = 1, \ldots, n$$
$$X_i = h_i(Y_1, \ldots, Y_n), \quad i = 1, \ldots, n$$

And if J_g is the determinant of the matrix with the ijth entry $\partial g_i / \partial x_j$, and J_h is the determinant of the matrix with entry $\partial h_i / \partial y_j$, then the joint density of $Y_1, \ldots, Y_n$ is

$$
\begin{aligned}
&f_{Y_1 \cdots Y_n}(y_1, \ldots, y_n) \\
&= f_{X_1 \cdots X_n}(x_1, \ldots, x_n) \big| J_g^{-1}(x_1, \ldots, x_n) \big| \\
&= f_{X_1 \cdots X_n}\big(h_1(y_1, \ldots, y_n), \ldots, h_n(y_1, \ldots, y_n)\big) \big| J_h(y_1, \ldots, y_n) \big|
\end{aligned}
$$

Final Comments

- In the transformation formulas, we always transform n variables to n variables. In practice, you might want to consider a transformation from $n \mapsto m$, with $m \le n$. In this case, there are two main approaches:

  - Start from scratch, just like we did for sums and quotients of random variables.
  - Create dummy variables to make $m = n$ (i.e., $Y_k = X_k$), calculate Jacobian, and then integrate out the dummy variables.

7 Order Statistics

Extrema and Order Statistics

- At times we are interested in properties of ordered collections of random variables.

- For instance, if we have n independent and identical random variables, what's the distribution of the *maximum* or *minimum* of this collection of random variables?

- More generally, we might want to find the distribution of the kth random variable, with $k \in \{1, 2, \ldots, n\}$.

Example: Maximum
Let $X_1, X_2, \ldots, X_n$ be independent and idenditcally distributed random variables, with distribution function $F(x)$ and density $f(x)$. Let $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ be the ordered random variables, such that $X_{(1)}$ is the minimum and $X_{(n)}$ is the maximum. Find the density function and pdf of $X_{(n)}$.

- To find the distribution function, we want to find $P(X_{(n)} \le x)$.

- If $X_{(n)}$ is the maximum of all X_i, then that means $X_{(n)} \le x$ if and only if $X_i \le x$ for all i. By the multiplication principle,

$$
\begin{aligned}
P(X_{(n)} \le x) &= P(X_1 \le x) P(X_2 \le x) \cdots P(X_n \le x) \\
&= F(x) F(x) \cdots F(x) \\
&= F(x)^n = F_{X_{(n)}}(x)
\end{aligned}
$$

- To find the density, we differentiate:

$$f_{X_{(n)}}(x) = n F(x)^{n-1} f(x)$$

Example: Maximum
Using the same setup as above, find the distribution and density functions for the minimum, $X_{(1)}$.

- As before, we are looking for the function $F_{X_{(1)}}(x) = P(X_{(1)} \le x)$.

- In this case, it's easier to find $P(X_{(1)} > x)$.

- The minimum is bigger than x if and only if all $X_i > x$. Thus

$$1 - F_{X_{(1)}}(x) = P(X_{(1)} > x) = \big(1 - F(x)\big)^n.$$

- Therefore, by taking the derivative, we find

$$F_{X_{(1)}}(x) = 1 - \big(1 - F(x)\big)^n$$
$$f_{X_{(1)}}(x) = nf(x)\big(1 - F(x)\big)^{n-1}$$

Acknowledgments

- Compiled on September 4, 2025 using R version 4.5.1.

- Licensed under the Creative Commons Attribution-NonCommercial license. Please share and remix non-commercially, mentioning its origin.

- We acknowledge students and instructors for previous versions of this course / slides.

References

Rice JA (2007). *Mathematical statistics and data analysis*, volume 371. 3 edition. Thomson/Brooks/Cole Belmont, CA.

Wikipedia contributors (2025). "Copula (statistics)#Sklar's Theorem: Wikipedia, The Free Encyclopedia." [Online; accessed 18-August-2025], URL https://en.wikipedia.org/w/index.php?title=Copula_(statistics)&oldid=1303484991.