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.

- Multivariate calculus is not a formal prerequisite for this course. We'll start by presenting a few details from multivariate calculus, as well as some results from analysis.

Partial Derivatives

- A partial derivative is an extension of a uni-variate derivative, where the function is a function of more than one variable.

Definition: Partial Derivative

Let $f(\mathbf{x})$ be a function $f : \mathbb{R}^d \to \mathbb{R}$. The partial derivative of f with respect to x_i, where $i \in \{1, 2, \ldots, d\}$ is

$$\frac{\partial}{\partial x_i} f(\mathbf{x}) = \frac{\partial f}{\partial x_i}$$

$$= \lim_{h \to 0} \frac{f(x_1, x_2, \ldots, x_i + h, \ldots, x_d) - f(x_1, x_2, \ldots, x_d)}{h}.$$

- We don't have to worry so much about the definition in this class. The thing to notice is that it's a directional derivative.

- Holding all other variables constant, it tells us the slope of the function in the x_i direction.

- Because it's definition is a limit, similar to the uni-variate case, the same derivative rules that you are familiar with from calculus also apply to partial derivatives, after treating the remaining variables as constants.

Example
Let $f(x_1, x_2, x_3) = x_1^2 x_2 - 10 x_2^2 x_3^3 + 43 x_1 - 7 \tan(4 x_2)$. Find $\frac{\partial}{\partial x_i} f(x_1, x_2, x_3)$ for all $i \in \{1, 2, 3\}$.

Schwarz's Theorem
Let $f : \mathbb{R}^d \to \mathbb{R}$ be twice-differentiable function. Higher order derivatives are written:

$$\frac{\partial^2}{\partial x_i \partial x_j} f = \frac{\partial}{\partial x_i} \left(\frac{\partial}{\partial x_j} f \right).$$

That is, we first take the derivative with respect to x_j, and then x_i. Shwarz's theorem states that the order can be swapped:

$$\frac{\partial^2}{\partial x_i \partial x_j} f = \frac{\partial^2}{\partial x_j \partial x_i} f.$$

Higher order integrals

- The next thing that we will review is higher-order integrals.

- This is a complex topic that we can't cover in detail in this class, but they do appear in this chapter and the next.

- For now, we'll just show some simple results that will be useful for our calculations.

- Recall one definition of integrals in the uni-case: calculating the area under the curve using an infinite Reimann approximation.

- That is, suppose we want to find the area under the function $f(x)$, on the interval $[a, b]$. We can use small, uniform sized rectangles to approximate the area.

- Let each rectangle have width Δx, with midpoint x_i^*, where the height of the ith rectangle is $f(x_i^*)$. Then:
$$A \approx f(x_1^*)\Delta x + f(x_2^*)\Delta x + \ldots + f(x_n^*)\Delta x.$$

- The exact area is equal to the limit as the boxes get smaller (and the number of boxes goes to infinity):
$$A = \int_a^b f(x)\, dx = \lim_{n \to \infty} \sum_{i=1}^n f(x_1^*)\, \Delta x.$$
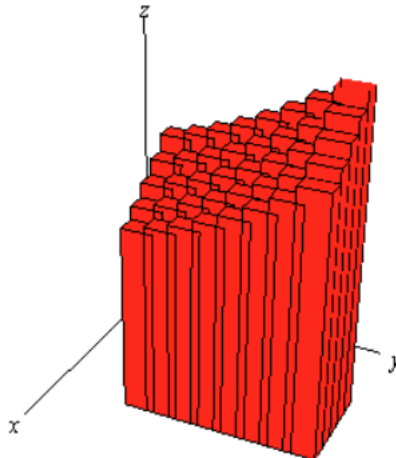
Figure 1: Credit: Paul's Online Notes: Section 15.1

- We now want to do the same thing, but in higher dimensions. Instead of using small rectangles to approximate the area, we now need to use small rectangular prisms.

- Let $z = f(x, y)$. Our goal is to find the volume between the function f and the x, y plane.

- We construct boxes such that the height of the ith box is given by $f(x_i^*, y_j^*)$.

- Then, the base of the rectangular prisms has area $\Delta A = \Delta x_i \times \Delta y_i$, and the height is $f(x_i^*, y_i^*)$.

- The total volume is approximately given by summing up the volume of each box:

$$V \approx \sum_{i=1}^{n} \sum_{j=1}^{m} f(x_i^*, y_j^*) \Delta A.$$

- We now can get the exact volume by taking the limit:

$$V = \iint_R f(x, y) \, dA = \lim_{n,m \to \infty} \sum_{i=1}^{n} \sum_{j=1}^{m} f(x_i^*, y_j^*) \Delta A$$

- *Fubini's theorem* gives us the standard approach to calculating these integrals

Fubini's Theorem
If we are evaluating a double integral on a rectangular region $R = X \times Y$, then the integral can be calculated as iterated uni-variate integrals:

$$\iint_R f(x, y) \, dA = \iint_{X \times Y} f(x, y) \, d(x, y)$$
$$= \int_X \left(\int_Y f(x, y) \, dy \right) dx$$
$$= \int_Y \left(\int_X f(x, y) \, dx \right) dy.$$

example

Let $R = [2, 4] \times [1, 2]$. Find
$$\iint_R 6xy^2 \, dA.$$

- One thing to note about higher-order integrals is that the area we are integrating (dA) doesn't have to be rectangular, and the limits of integration can be functions themselves.

- We'll see a few examples of this later, and I'll try to walk through these examples carefully. However, this isn't a class on multivariate calculus, so you won't be required to do calculations beyond what you can do using Fubini's theorem.

- There is a nice example in our textbook for polar-coordinate integration, which is perhaps the most common extension (Rice, 2007, Example A, Chapter 3.6.2).

Swapping order of limits and integrals

- An important result from measure theory (often covered in an analysis class) is known as the Dominated Convergence Theorem.

- We won't prove the theory here, nor state the complete theory, but we will state parts of it that will be useful to us.

- A proof of the parts of the theory that we are interested in can be found in our supplement text (Section 2.4 of Casella and Berger, 2024).

- A more complete treatment of this idea can be found in Resnick (Chapter 5, 2019).

- The next few theorems are all special cases of the Dominated Convergence Theorem.

Leibnitz's Rule
If $f(x, \theta)$ is differentiable with respect to θ, then
$$\frac{d}{d\theta} \int_a^b f(x, \theta) \, dx = \int_a^b \frac{\partial}{\partial \theta} f(x, \theta) \, dx.$$

- This is actually a special case of Leibnitz's Rule, which gives an expression if the limits of integration a, b, are actually differentiable functions of θ: $a(\theta), b(\theta)$.

- Thus, if the integral has finite range, we can change the order of the integral and derivative.

- Next, we consider changing the order of a limit and an integral.

Theorem: Limits and Integrals
Suppose the function $h(x, y)$ is continuous at y_0 for each x, and there exists a function $g(x)$ satisfying

(i) $|h(x, y)| \leq g(x)$ for all x and y.

(ii) $\int_{-\infty}^{\infty} g(x) \, dx < \infty$.

Then
$$\lim_{y \to y_0} \int_{-\infty}^{\infty} h(x, y) \, dx = \int_{-\infty}^{\infty} \lim_{y \to y_0} h(x, y) \, dx.$$

- Note that derivatives are just special types of limits:
$$\frac{\partial}{\partial \theta} f(x, \theta) = \lim_{\delta \to 0} \frac{f(x, \theta + \delta) - f(x, \theta)}{\delta}.$$

4

- Thus, theorems about interchanging integrals and integrals can be worked into a theorem about interchanging integrals and derivatives

Corollary: Derivatives and Integrals
Suppose $f(x, \theta)$ is differentiable in θ and there exists a function $g(x, \theta)$, and $\delta_0 > 0$ such that

$$\left| \frac{\partial}{\partial \theta} f(x, \theta) \right|_{\theta = \theta'} \leq g(x, \theta), \quad \text{for all } \theta' \text{ such that } |\theta' - \theta| \leq \delta_0,$$

and $\int_{-\infty}^{\infty} g(x, \theta) \, dx < \infty$. Then,

$$\frac{d}{d\theta} \int_{-\infty}^{\infty} f(x, \theta) \, d\theta = \int_{-\infty}^{\infty} \frac{\partial}{\partial \theta} f(x, \theta) dx.$$

- Similar results hold for interchanging summation, derivatives, and integrals.

- For finite sums, changing the order of sums and derivatives is a natural property of derivatives. It's not so simple with infinite sums.

Exchanging summation and derivatives
Suppose that $\sum_{x=0}^{\infty} h(\theta, x)$ converges for all θ in an interval (a, b), and (i) $\frac{\partial}{\partial \theta} h(\theta, x)$ is continuous in θ for each x, and (ii) $\sum_{x=0}^{\infty} \frac{\partial}{\partial \theta} h(\theta, x)$ converges uniformly on every closed bounded subinterval of (a, b). Then

$$\frac{d}{d\theta} \sum_{x=0}^{\infty} h(\theta, x) = \sum_{x=0}^{\infty} \frac{\partial}{\partial \theta} h(\theta, x).$$

- The necessary conditions stated above will hold for nearly all functions we'll consider in this class. You won't be asked to verify these conditions.

Exchaning sums and integrals
Suppose the series $\sum_{x=0}^{\infty} h(\theta, x)$ converges uniformly on $[a, b]$ and that, for each x, $h(\theta, x)$ is a continuous function of θ. Then

$$\int_a^b \sum_{x=0}^{\infty} h(\theta, x) \, d\theta = \sum_{x=0}^{\infty} \int_a^b h(\theta, x) \, d\theta.$$

- Like with differentiation, we can always exchange the order of summation and integration, if the sums are finite.

- You will not be asked to verify the conditions above, but I wanted to point out that these operations are theoretically supported when they inevitably come up in this class.

2 Joint Probabilities

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.

- Thus, any set $A \subset \mathbb{R}^2$, the joint cdf can give $P\big((X,Y) \in A\big)$. For continuous random variables:

$$P\big((X,Y) \in A\big) = \iint_A f(x,y)\, dy\, dx.$$

- 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...)

- $F(x_2, y_2)$ 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.

- For discrete random variables, the pmf just describes the joint frequencies: $P(X = x_i, Y = y_j)$ for all i, j.

- For continuous random variables, a joint pdf of (X,Y) is defined as functions $f : \mathbb{R}^2 \to \mathbb{R}$ that satisfy

$$F(X, y) = \int_{-\infty}^x \int_{-\infty}^y f(s,y)\, ds\, dt.$$

- The *bivariate*-Fundamental Theorem of Calculus implies:

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

- Similar ideas hold for higher-dimensions.

3 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.

x	y			
	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.

- 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:

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

$$= \binom{n}{n_i} p_i^{n_i} (1 - p_i)^{n - n_i}$$

4 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

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

Find $P(X > y)$.

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

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.

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

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

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 $-1 \leq \alpha \leq 1$:

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

- 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 $\alpha = 1/2$:

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

- 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

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

- For practice defining new random variables, 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 because of uniform probability, 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

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

- Transforming random variables from one coordinate system to another is an important idea, one we will talk more about later.

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

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

5   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).

x	y 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.

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

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

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

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

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

6 Conditional Distributions

6.1 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

$$
\begin{aligned}
P(X = x_i | Y = y_i) &= \frac{P(X = x_i, Y = y_i)}{P(Y = y_i)} \\
&= \frac{p_{XY}(x_i, y_i)}{p_Y(y_i)},
\end{aligned}
$$

- 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.

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.

6.2 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 $U \leq f(T)/M(T)$, 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 2.
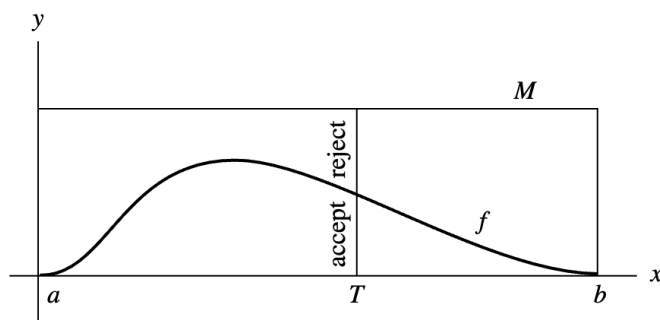

Figure 2: Illustration of the rejection method, copied from Rice (2007, Figure 3.15).

- 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: proof sketch

- With a little abuse of notation, we'll do a proof sketch that this works.

- Let I be the indicator event such that $I = 1$ if a sample is accepted, $I = 0$ otherwise.

- We want the distribution of accepted samples to follow $f(x)$, i.e., $f_{T|I=1}(x|I = 1) = f(x)$.

- For convenience, let's denote $\bar{M} = \int_a^b M(x)\,dx$, so that $m(x) = M(x)/\bar{M}$.

- First, using conditional probabilities, we have

$$f_X(x) = f_{T|I=1}(x|I=1) = \frac{f_{T,I=1}(x, I=1)}{P(I=1)} = \frac{P(I=1|T=x)f_T(x)}{P(I=1)}$$

- Note in the equation above, I switched from X to T; this matches our notation previously used, that we simulate a possible sample T, and accept if the particular condition is met, in which case, $X = T$.

- T was generated with the standardized function m. That is, the density $f_T(x) = m(x)$. Thus,

$$f_X(x) = f_{T|I=1}(x|I=1) = \frac{P(I=1|T=t)m(x)}{P(I=1)}.$$

- We have an expression for $m(x)$ already, now we want to find $P(I=1)$ and $P(I=1|T=x)$.

$$P(I=1) = \int P(I=1, T=t)\, dt$$
$$= \int P(I=1|T=t)m(t)\, dt.$$

- For fixed $T = t$, the probability of accepting the sample (i.e., $I = 1$) is determined by the uniform probability:

$$P(I=1|T=t) = P\big(M(t)U \le f(t)\big) = P\big(U \le f(t)/M(t)\big) = f(t)/M(t).$$

- Thus,

$$P(I=1) = \int P(I=1|T=t)m(t)\, dt$$
$$= \int \frac{m(t)f(t)}{M(t)}\, dt$$
$$= \int \frac{M(t)f(t)}{M(t)\bar{M}}\, dt$$
$$= \frac{1}{\bar{M}} \int f(t)\, dt$$
$$= \frac{1}{\bar{M}}.$$

- Therefore, the distribution of the accepted samples is:

$$f_{T|I=1}(x|I=1) = \frac{P(I=1|T=x)f_T(x)}{P(I=1)}$$
$$= \bar{M}P(I=1|T=x)m(x)$$
$$= \bar{M} \times \frac{f(x)}{M(x)} \times \frac{M(X)}{\bar{M}}$$
$$= f(x).$$

7   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:

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

- 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

Cauchy density
Let X and Y be independent, standard normal random variables. Find the pdf of $Z = X/Y$.

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

- 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 $m \leq n$. 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.

8 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 $P(X_{(n)} \leq x)$.

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

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

- 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 $F_{X_{(1)}}(x) = P(X_{(1)} \leq x)$.

- 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

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

- Therefore, by taking the derivative, we find

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

Theorem 3.1: Order statistic for continuous random variables

Let $X_1, \ldots, X_n$ be continuous random variables, with cdf and pdf F and f, respectively. If the ordered random variables are $X_{(1)}, \ldots, X_{(n)}$, then the cdf of $X_{(j)}$, $j \in \{1, 2, \ldots, n\}$ is

$$F_{X_{(j)}}(x) = \sum_{k=j}^{n} \binom{n}{k} \left[F(x)\right]^k \left[1 - F(x)\right]^{n-k}.$$

Proof.
- Let Y be a random variable that counts the number of X_i less than or equal to x.

- If we define the "success" as the event that $X_i \leq x$, then the probability of success is $p = F(x)$.

- Thus, Y is a binomial random variable with n trials, and probability $p = F(x)$.

- For $X_{(j)} \leq x$, that means that j of the X_i are less than or equal to x.

- Using our definition of Y, $X_{(j)} \leq x$ if and only if $Y \geq j$. Thus

$$F_{X_{(j)}}(x) = P(Y \geq j)$$
$$= \sum_{k=j}^{n} \binom{n}{k} \left[F(x)\right]^k \left[1 - F(x)\right]^{n-k}.$$

$\square$

Theorem 3.2: The pdf of ordered statistics

Consider the same setup as Theorem 3.1. That is, $X_{(1)}, \ldots, X_{(n)}$ represent the order statistics of a random sample. Then the pdf of $X_{(j)}$ is

$$f_{X_{(j)}}(x) = \frac{n!}{(j-1)!(n-j)!} f(x) \left[F(x)\right]^{j-1} \left[1 - F(x)\right]^{n-j}.$$

Proof. First, I want to note the following identity that will use

$$\binom{n}{k+1}(k+1) = \frac{n!}{(n-k-1)!(k+1)!}(k+1) = \frac{n!}{(n-k-1)!k!} = \frac{n!}{(n-k)!k!}(n-k) = \binom{n}{k}(n-k)$$

- Because we already have the cdf, we just need to differentiate to get the pdf.

$$f_{X_{(j)}}(x) = \frac{d}{dx} F_{X_{(j)}}(x)$$

$$= \frac{d}{dx} \sum_{k=j}^{n} \binom{n}{k} \left[F(x)\right]^k \left[1 - F(x)\right]^{n-k}$$

$$= \sum_{k=j}^{n} \binom{n}{k} \left[k f(x) \left(F(x)\right)^{k-1} \left(1 - F(x)\right)^{n-k} - (n-k) f(x) \left(F(x)\right)^k \left(1 - F(x)\right)^{n-k-1} \right]$$

$$= \sum_{k=j}^{n} \binom{n}{k} \left[k f(x) \left(F(x)\right)^{k-1} \left(1 - F(x)\right)^{n-k} \right] - \sum_{k=j}^{n} \binom{n}{k} \left[(n-k) f(x) \left(F(x)\right)^k \left(1 - F(x)\right)^{n-k-1} \right]$$

- There are many like terms in these sums, which becomes most evident with the identity we derived.

- In the first sum, let's "pull-out" the first term.

- In the second sum, the last term is zero because when $k = n$, we have a product $(n-k)$.

$$f_{X_{(j)}}(x) = \binom{n}{j} j \left(F(x)\right)^{j-1} \left(1 - F(x)\right)^{n-j} f(x)$$

$$+ \sum_{k=j+1}^{n} \binom{n}{k} \left[k f(x) \left(F(x)\right)^{k-1} \left(1 - F(x)\right)^{n-k} \right]$$

$$- \sum_{k=j}^{n-1} \binom{n}{k} \left[(n-k) f(x) \left(F(x)\right)^k \left(1 - F(x)\right)^{n-k-1} \right].$$

- Now I will expand the factorial in the first term, and do a change of variables for the first summation, shifting the indices down one so that the two sums match.

$$f_{X_{(j)}}(x) = \frac{n!}{(j-1)!(n-j)!} f(x) \left(F(x)\right)^{j-1} \left(1 - F(x)\right)^{n-j}$$

$$+ \sum_{k=j}^{n-1} \binom{n}{k+1} \left[k f(x) \left(F(x)\right)^k \left(1 - F(x)\right)^{n-k-1} \right]$$

$$- \sum_{k=j}^{n-1} \binom{n}{k} \left[(n-k) f(x) \left(F(x)\right)^k \left(1 - F(x)\right)^{n-k-1} \right].$$

- Finally, using the identity that we previously derived,

$$\binom{n}{k+1}(k+1) = \binom{n}{k}(n-k),$$

we see that the two sums are additive inverses of one another and cancel out, thus

$$f_{X_{(j)}}(x) = \frac{n!}{(j-1)!(n-j)!} f(x) \left[F(x)\right]^{j-1} \left[1 - F(x)\right]^{n-j}.$$

$\square$

Example: Uniform order statistic pdf
Let $X_1, \ldots, X_n$ be independent Uniform$(0,1)$ random variables, so that $f(x) = 1[0 < x < 1]$, and $F_X(x) = x$. Find the pdf of the jth order statistic.

- We can just use the formula we previously derived for order statistics:

$$f_{X_{(j)}}(x) = \frac{n!}{(j-1)!(n-j)!}(x)^{j-1}(1-x)^{n-j}, \quad x \in (0,1)$$

$$= \frac{\Gamma(n+1)}{\Gamma(j)\Gamma(n-j+1)}x^{j-1}(1-x)^{(n-j+1)-1}, \quad x \in (0,1),$$

- where the last equation comes from the definition of the gamma-function

- This final expression is the pdf of a beta$(j, n-j+1)$ random variable, which could be used to calculate the mean or variance.

- The joint distribution of two or more order statistics can also be derived, and is useful for deriving the density of some statistics of interest.

Theorem 3.3: joint pdf of order statistics
Consider the same setup as Theorem 3.1. That is, $X_{(1)}, \ldots, X_{(n)}$ represent the order statistics of a random sample. Then the joint pdf of $X_{(i)}, X_{(j)}, 1 \le i < j \le n$ is

$$f_{X_{(i)},X_{(j)}}(u,v) = \frac{n!}{(i-1)!(j-1-i)!(n-j)!}f(u)f(v)\big(F(u)\big)^{i-1}$$

$$\times \big(F(v)-F(u)\big)^{j-1-i}\big(1-F(v)\big)^{n-j},$$

for $-\infty < u < v < \infty$.

Example: Midrange and Range
Let $X_1, \ldots, X_n$ be independent uniform$(0, a)$ random variables, and denote the ith order statistic as $X_{(i)}$. The *range* of a sample is defined as the difference between the largest and smallest observations, $R = X_{(n)} - X_{(1)}$. Define the *midrange* as the midpoint between the minimum and maximum observations, $V = (X_{(1)} + X_{(n)})/2$. Find the pdf of both R and V.

- We first can apply Theorem 3.3 to get the joint density of $X_{(1)}$ and $X_{(n)}$:

$$f_{X_{(i)},X_{(j)}}(x_1, x_n) = \frac{n!}{(i-1)!(j-1-i)!(n-j)!}f(x_1)f(x_n)\big(F(x_1)\big)^{i-1}$$

$$\times \big(F(x_n)-F(x_1)\big)^{j-1-i}\big(1-F(x_n)\big)^{n-j}$$

$$= \frac{n!}{(1-1)!(n-1-1)!(n-n)!}f(x_1)f(x_n)\big(F(x_1)\big)^{1-1}$$

$$\times \big(F(x_n)-F(x_1)\big)^{n-1-1}\big(1-F(x_n)\big)^{n-n}$$

$$= \frac{n(n-1)}{a^2}\Big(\frac{x_n}{a}-\frac{x_1}{a}\Big)^{n-2}$$

$$= \frac{n(n-1)(x_n-x_1)^{n-2}}{a^n}, \quad 0 < x_1 < x_n < a.$$

- Now we can use the change of variables theorem. We find that $g_1(x_1, x_n) = x_n - x_1$, and $g_2(x_1, x_n) = (x_1 + x_n)/2$.

- Solving for R and V as functions of $X_{(1)}$ and $X_{(n)}$, we find the inverse transformations

$$h_1(r,v) = v - r/2, \quad h_2(r,v) = v + r/2.$$

- We can now calculate the Jacobian of the transformation:

$$J_h = \begin{vmatrix} \frac{\partial h_1}{\partial r} & \frac{\partial h_1}{\partial v} \\ \frac{\partial h_2}{\partial r} & \frac{\partial h_2}{\partial r} \end{vmatrix} = -\frac{1}{2} - \frac{1}{2} = -1.$$

- We need to calculate the support of (R, V). The transformation $g(x_1, x_n)$ maps the set $\{(x_1, x_n) : 0 < x_1 < x_n < a\}$ onto the set $\{(r, v) : 0 < r < a, r/2 < v < a - r/2\}$. To see this, note that $0 < x_2 - x_1 < a$, and the lower and upper bounds for v can be found by plugging in $x_1 = h_1(r, v)$ and $x_n = h_2(r, v)$ into the upper and lower inequalities of $0 < x_1 < x_n < a$, respectively.

- Thus, we find the joint density of (R, V) as

$$f_{R,V}(r, v) = \frac{n(n-1)\big(h_2(r, v) - h_1(r, v)\big)^{n-2}}{a^n}, \quad 0 < r < a, \ r/2 < v < a - r/2$$
$$= \frac{n(n-1)r^{n-2}}{a^n}, \quad 0 < r < a, \ r/2 < v < a - r/2.$$

- Finally, we can integrate to find the marginal densities:

$$f_R(r) = \int_{\mathbb{R}} f_{(R,V)}(r, v)\, dv$$
$$= \int_{r/2}^{a-r/2} \frac{n(n-1)r^{n-2}}{a^n}\, dv$$
$$= \frac{n(n-1)r^{n-2}(a-r)}{a^n}, \quad 0 < r < a.$$

- If $a = 1$, the density of R corresponds to that of a beta$(n-1, 2)$ distribution; alternatively, we can do a quick change of variable to see that R/a has a beta distribution, where a just scales the distribution.

- The bounds on the integral for integrating out r is a little bit less straightforward because r appears in both inequalities. In this case, it's helpful to plot the region https://www.desmos.com/calculator/tosungnfnu.

- We see we have a triangle that has a base on the v-axis, with height a. The triangle is defined by the equations $r = 2v$ and $r = 2(a - v)$, so the integral is calculated over two parts:

$$f_V(v) = \int_{\mathbb{R}} f_{(R,V)}(r, v)\, dr$$
$$= 1[0 < v \le a/2] \int_0^{2v} \frac{n(n-1)r^{n-2}}{a^n}\, dr + 1[a/2 < v \le a] \int_0^{2(a-b)} \frac{n(n-1)r^{n-2}}{a^n}\, dr$$
$$= \begin{cases} \frac{n(2v)^{n-1}}{a^n}, & 0 < v \le a/2 \\ \frac{n\big(2(a-v)\big)^{n-1}}{a^n}, & a/2 < v \le a \end{cases}$$

- The joint pdf of more than two order statistics can also be found.

- Perhaps the most useful is the joint pdf of all of the order statistics:

$$f_{X_{(1)},\dots,X_{(n)}}(x_1, \dots, x_n) = \begin{cases} n! f(x_1) \cdots f(x_n) & -\infty < x_1 < \dots < x_n < \infty \\ 0 & \text{otherwise} \end{cases}$$

- The $n!$ arises naturally because it is the number of ways to arrange n items.

- Conditional and marginal densities of order statistics can be found using formulas for conditional and marginal densities, as needed.

- There are a few more useful theorems / identities that are useful, and there derivation is similar.

- I'll just present these theorems without proof for now, but proofs can be found in Casella and Berger (2024, Chapter 5.4).

Theorem 3.4: Discrete random variables

Let $X_1, \ldots, X_n$ be a sample from a discrete distribution with pmf $p(x_i) = p_i$. If we define $P_0 = 0$, and for all $i \geq 1$ $P_i = \sum_{j=1}^{i} p_j$, then the cdf and pmf of the order statistic $X_{(j)}$ are given by

$$P(X_{(j)} \leq x_i) = \sum_{k=j}^{n} \binom{n}{k} P_i^k (1 - P_i)^{n-k},$$

and

$$P(X_{(j)} = x) = \sum_{k=j}^{n} \binom{n}{k} \left[P_i^k (1 - P_i)^{n-k} - P_{i-1}^k (1 - P_{i-1})^{n-k} \right].$$

Theorem 3.5: Joint distributions

If $X_{(1)}, \ldots, X_{(n)}$ denote the order statistics of a random sample from a continuous population with cdf $F(x)$ and pdf $f(x)$, then for any $1 \leq i < j \leq n$, the joint pdf of $X_{(i)}$ and $X_{(j)}$ is

$$f_{X_{(i)}, X_{(j)}}(u, v) = \frac{n!}{(i-1)!(j-1-i)!(n-j)!} f(u) f(v)$$
$$\times \left[F(u) \right]^{i-1} \left[F(v) - F(u) \right]^{j-1-i} \left[1 - F(v) \right]^{n-j}.$$

Acknowledgments

- Compiled on October 22, 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

Casella G, Berger R (2024). *Statistical inference.* Chapman and Hall/CRC. 4, 27

Resnick S (2019). *A probability path.* Springer. 4

Rice JA (2007). *Mathematical statistics and data analysis*, volume 371. 3 edition. Thomson/Brooks/Cole Belmont, CA. 1, 3, 15, 2

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. 11