MH2500 Probability and Introduction to Statistics

Handout 4 - Joint Distributions - II

Synopsis

We continue our discussion on density functions introduced in Handout 2. We focus on “joint” density functions with two or more continuous random variables.

- Continuous Random Variables,
- Independent Random Variables.
- Conditional Distribution.
- Functions of Jointly Distributed Random Variables
  - Sums and Quotients
  - ~~The general case~~ (Read it on your own, not in exam.)
- Ordered Statistics

Continuous Random variable

Suppose X and Y are continuous random variables with a joint cdf $F(x, y)$. Their **joint density function** is a piecewise continuous function of two variables, $f(x, y)$.

The density function $f(x, y)$ is nonnegative and $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dydx = 1$.

For any "reasonable" two-dimensional set A,

$$P((X, Y) \in A) = \iint_A f(u, v) dvdu$$

In particular, if $A = \{(X, Y) | X \leq x \text{ and } Y \leq y\}$,

$$F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(u, v) dvdu.$$

Continuous Random Variable

Remarks: We view double integrals as an iteration of two integrals.

For this course, the area A that we integrate over shall be easily understood, nothing sophisticated.

From the fundamental theorem of multivariable calculus, it follows that

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

Remarks: We assume this theorem is true without proof.

Although there are subtle differences between $\frac{\partial}{\partial x}$ and $\frac{d}{dx}$, you can just pretend that they are the same.

Continuous Random Variable

For small δ_x and δ_y, if f is continuous,

$$P(x \leq X \leq x + \delta_x, y \leq Y \leq y + \delta_y) = \int_x^{x+\delta_1} \int_y^{y+\delta_2} f(u, v) dvdu$$

$$\approx f(x, y)\delta_x \delta_y.$$

Thus the probability that (X, Y) is in a small neighbourhood of $(x, y)$is proportional to $f(x, y)$. Differential notation is sometimes useful.

$$P(x \leq X \leq x + dx, y \leq Y \leq y + dy) = f(x, y)dxdy.$$

Example A

Consider the bivariate density function

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

which is plotted in the figure below. Find $P(X > Y)$.

. . . .

[Solution:]
We begin with integrating f over the set $A := \{(x, y) | 0 \le y \le x \le 1\}$.

$$P(X > Y) =$$

$$= \quad .$$

Marginal cdf and marginal density

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

$$F_X(x) = P(X \le x) = \lim_{y \to \infty} F(x, y)$$
$$= \int_{-\infty}^{x} \int_{-\infty}^{\infty} f(u, y) dy du.$$

It follows that the density function of X alone, known as the **marginal density** of X, is,

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

In the discrete case, the marginal frequency function was found by summing the joint frequency function over the other variables. In the continuous case, It is found by integration.

Example B

Calculate the marginal density of X and the marginal density of Y in Example A.

. . . .

$$f_X(x) \;=\; \frac{12}{7}\int_0^1 (x^2 + xy)dy \;=\; \frac{12}{7}\left(x^2 + \frac{x}{2}\right).$$

Similarly,

$$\begin{aligned} f_Y(x) &= \\ &= \\ &= \qquad\qquad\qquad . \end{aligned}$$

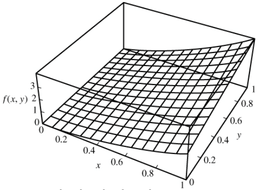

Figure 3.4 **The density function** $f(x,y) = \frac{12}{7}(x^2 + xy),\ 0 \le x \le 1,\ 0 \le y \le 1.$

More than two random variables

For a joint distribution of more than two continuous random variables, the formula is generalized in the obvious way.

For example, if X, Y, and Z are jointly continuous random variables with density function $f(x, y, z)$. Then the one-dimensional distribution of X is

$$f_X(x) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y, z) dydz.$$

and the two dimensional marginal distribution of X and Y is

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

Example

Let

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

Find the marginal density of X and of Y. Find the density function.

. . . .

$$F_X(x) = \lim_{y\to\infty} H(x,y) = \lim_{y\to 1} H(x,y) = 2x - x^2 - x + x^2 = x.$$

$$F_Y(y) = \lim_{x\to\infty} H(x,y) = \lim_{x\to 1} H(x,y) = 2y - y - y^2 + y^2 = y.$$

$$h(x,y) = \frac{\partial^2}{\partial x \partial y} H(x,y) =$$
$$=$$

Copula

A **copula** is a joint cdf of random variables that have uniform marginal distributions.

The example on the previous slide is a copula.

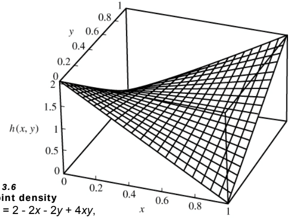

Figure 3.6
The joint density
h(x, y) = 2 - 2*x* - 2*y* + 4*xy*,
where 0 ≤ *x* ≤ 1 **and** 0 ≤ *y* ≤ 1,
which has uniform marginal densities.

Note that

$$P(U \leq u) = C(u, 1) = u \quad \text{and} \quad C(1, v) = v.$$

Copulas are used extensively in financial statistics.

Example D

Consider the joint density

$$f(x,y) = \begin{cases} \lambda^2 e^{-\lambda y}, & 0 \le x \le y,\ \lambda > 0\ ; \\ 0, & \text{elsewherse.} \end{cases}$$

Find the marginal distributions.

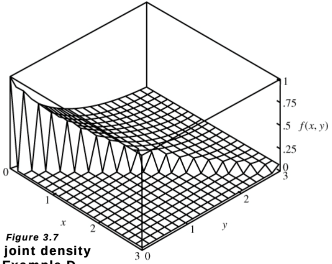

Figure 3.7
The joint density of Example D.

Plot region where $f(x, y)$ is nonzero.

Example D con't

First, the marginal density of X

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{XY}(x, y)dy \\ &= \int_{x}^{\infty} \lambda^2 e^{-\lambda y} dy \\ &= \end{aligned}$$

This is exponential distribution.

Next, for the marginal density of Y

$$\begin{aligned} f_Y(x) &= \int_{-\infty}^{\infty} f_{XY}(x, y)dx \\ &= \\ &= \end{aligned}$$

This is a gamma distribution.

Example F - Bivariate normal density

The bivariate normal density is given by

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}\right]\right)$$

The density depends on five parameters,

$$-\infty < \mu_X < \infty, \qquad -\infty < \mu_Y < \infty,$$
$$\sigma_X > 0, \qquad \sigma_Y > 0, \qquad -1 < \rho < 1.$$

Show that the marginal distributions of X and Y are $N(\mu_X, \sigma_X^2)$ and $N(\mu_Y, \sigma_Y^2)$, respectively.

Example F

The marginal density of X is given by

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y) dy.$$

Making the change of variables, $u = (x - \mu_X)/\sigma_X$ and $v = (y - \mu_Y)/\sigma_Y$, we find that

$$f_X(x) = \qquad\qquad\qquad\qquad\qquad\qquad\qquad .$$

By completing the square for part $v^2 - 2\rho uv$, we have

$$u^2 + v^2 - 2\rho uv = (v - \rho u)^2 + u^2(1 - \rho^2).$$

Example F

Substituting this into the integral, we find that

$$f_X(x) = \frac{1}{2\pi\sigma_X\sqrt{1-\rho^2}}\int_{-\infty}^{\infty}\exp\left(-\frac{1}{2(1-\rho^2)}\left[(v-\rho u)^2+u^2(1-\rho^2)\right]\right)dv$$
$$= \frac{1}{2\pi\sigma_X\sqrt{1-\rho^2}}e^{-u^2/2}\int_{-\infty}^{\infty}\exp\left(-\frac{1}{2(1-\rho^2)}(v-\rho u)^2\right).$$

The integral is recognized as that of a normal density with mean ρu and variance $(1-\rho^2)$, and so the integral is Hence

$$f_X(x) = \frac{\sqrt{1-\rho^2}\sqrt{2\pi}}{2\pi\sigma_X\sqrt{1-p^2}}e^{-u^2/2}$$
$$=$$

The marginal density of Y can be shown similarly.

Independent Random Variables

Definition

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

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

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

This definition holds for both continuous and discrete random variables. It is equivalent to their joint frequency/density function factors.
E.g. if the density function of a joint two continuous random variable factors, then the joint cdf ca be expressed as a product:

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

Independent Random Variables

The definition also implies that if X and Y are independent, then

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

Suppose X and Y are independent. If g and h are functions, then $Z = g(X)$ and $W = h(Y)$ are independent as well.

Sketch of argument:
Let $A(z)$ be the set of x such that $g(x) \leq z$, and let $B(w)$ be the set of y such that $h(y) \leq w$. Then

$$\begin{aligned} P(Z \leq z, W \leq w) &= P(X \in A(z), Y \in B(w)) \\ &= P(X \in A(z))P(Y \in B(w)) \\ &= P(Z \leq z)(P(W \leq w). \end{aligned}$$

Example A

Suppose that the point (X, Y) is uniformly distributed on the square $S = \{(x, y) | -1/2 \leq x \leq 1/2,\ -1/2 \leq y \leq 1/2\}$, i.e.,

$$f_{XY}(x, y) = \begin{cases} 1, & \text{for } (x, y) \text{ in } S; \\ 0, & \text{elsewhere.} \end{cases}$$

Find the marginal density functions. Are X and Y independent?

. . . .

By definition,

$$f_X(x) = \int_{-\frac{1}{2}}^{\frac{1}{2}} 1 dy = [y]_{-\frac{1}{2}}^{\frac{1}{2}} = 1, \qquad -\frac{1}{2} \leq x \leq \frac{1}{2}$$

Similarly, $f_Y(y) = 1$, and so X and Y are independent.

Example B

Rotate the square in Example A by 90 degrees to form a diamond. Sketch the diamond and show that the marginal density of X is nonnegative for $-1/2 \leq x \leq 1/2$ but it is not uniform.
Similar steps shows this is also true for Y. Show that $f_X(.5) > 0$, $f_Y(.5) > 0$ but $f_{XY}(0.5, 0.5) = 0$. Conclude that X and Y are not independent.

Example E

Suppose that a node in a communication network has the property that if two packets of information arrive within time τ of each other, they "collide" and then have to be retransmitted. If the times of arrival of the two packets are independent and uniform on $[0, T]$, what is the probability that they collide?

. . . .

The times of arrival of two packets, T_1 and T_2 are independent and uniform on $[0, T]$, so their joint density is the product of the marginals, or

$$f(t_1, t_2) = \frac{1}{T^2}$$

for t_1 and t_2 in the square with sides $[0, T]$. Therefore, (T_1, T_2) is uniformly distributed over the square.

Example E

The probability that the two packets collide is proportional to the area

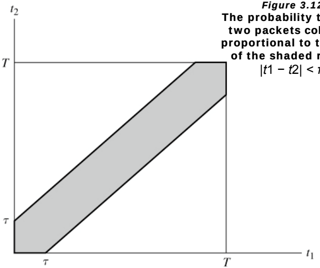

Figure 3.12
The probability that the two packets collide is proportional to the area of the shaded region
$|t1 - t2| < \tau$.

Area of shaded area = . Hence required probability is

$$\frac{T^2 - (T - \tau)^2}{T^2} = 1 - \left(1 - \frac{\tau}{T}\right)^2.$$

Conditional Distribution - Continuous

Definition

If X and Y are jointly continuous random variables, the **conditional density** of Y given X is defined to be

$$f_{Y|X}(y|x) = \begin{cases} \dfrac{f_{XY}(x,y)}{f_X(x)}, & \text{if } 0 < f_X(x) < \infty; \\ 0, & \text{otherwise.} \end{cases}.$$

Multiplying throughout by $f_X(x)$ gives

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

Integrating both sides over x allows the marginal density of Y to be expressed as

$$f_Y(y) = \int_{-\infty}^{\infty} f_{Y|X}(y|x)f_X(x)dx,$$

which is the law of total probability for the continuous case.

Example A

Recall example on Handout 3 slide 22.

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

Find the conditional densities, $f_{Y|X}$ and $f_{X|Y}$. What distribution are they?

. . . .

$$f_{Y|X}(y|x) = \qquad = \qquad , \qquad y \ge x.$$

This is exponential on the interval $[x, \infty)$.

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

This is the uniform distribution on the interval $[0, y]$.

Example D

The **rejection method** is commonly used to generate random variables from a density function, especially when the inverse of the cdf cannot be found in closed form and therefore the inverse cdf method, Proposition D in Section 2.3, cannot be used.

Suppose that f is a density function that is nonzero on an interval $[a, b]$ and zero outside the interval (a and b may be infinite). Let $M(x)$ be a function such that $M(x) \geq f(x)$ on $[a, b]$, and let

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

be a probability density function.

The idea is to choose M so that it is easy to generate random variables from m.

Example D con't

If $[a, b]$ is finite, m can be chosen to be the uniform distribution on $[a, b]$. The algorithm is as follow:

Step 1: Generate T with density m.

Step 2: Generate U, uniform on [0,1] and independent of T. If $M(T) \times U \leq f(T)$, then let $X = T$ (accept T). Otherwise, go to Step 1 (reject T).

Show that the density function of the random variable X obtained is f.

$$
\begin{aligned}
P(x \leq X \leq x + dx) &= P(x \leq T \leq x + dx | \text{accept}) \\
&= \frac{P(x \leq T \leq x + dx \text{ and accept})}{P(\text{accept})} \\
&= \frac{P(\text{accept } | x \leq T \leq x + dx) P(x \leq T \leq x + dx)}{P(\text{accept})}.
\end{aligned}
$$

Note that

$$P(\text{accept}|x \le T \le x + dx) = P\left(U \le \frac{f(x)}{M(x)}\right) = \frac{f(x)}{M(x)}$$

so that the numerator is

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

From the law of total probability, the denominator is

$$\begin{aligned} P(\text{accept}) &= P\left(U \le \frac{f(T)}{M(T)}\right) \\ &= \int_a^b \frac{f(t)}{M(t)} m(t)dt = \frac{1}{\int_a^b M(t)dt} \end{aligned}$$

where the last two steps follow from the definition of m and since f integrates to 1.

Finally we see that the numerator over the denominator is $f(x)dx$.

Functions of Jointly Distributed Random Variables

In Section 2.3, we introduced functions of a random variable. In this section, we extend our discussion to function of several random variables.

Suppose X and Y are discrete random variables taking values on the integers and having the joint frequency function $p(x, y)$. Let $Z = X + Y$.

To find the frequency function of Z, we note that $Z = z$ whenever $X = x$ and $Y = z - x$, where x is an integer. Summing over all x gives the $P(Z = z)$, i.e.,

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

If X and Y are independent so that $p(x, y) = p_X(x)p_Y(y)$, then

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

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

Sum of Continuous Random Variable

The continuous case is similar.

Suppose X and Y are continuous random variables and we first find the cdf of Z and differentiate to find the density. Since $Z \leq z$ whenever the point (X, Y) is in the region R_Z, we have

$$F_Z(z) = \iint_{R_Z} f(x, y) dxdy = \int_{-\infty}^{\infty} \int_{-\infty}^{z-x} f(x, y) dydx.$$

Making the change of variables $y = v - x$ in the inner integral gives

$$F_Z(z) = \int_{-\infty}^{\infty} \int_{-\infty}^{z} f(x, v - x) dvdx = \qquad\qquad\qquad\qquad dxdv.$$

Sum of Continuous Random Variable

Differentiating, and if $\int_{-\infty}^{\infty} f(x, z-x)dx$ is continuous at z, then

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

which is the analogue of the result for 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 A

Suppose the lifetime of a component is exponentially distributed and that an identical and independent backup component is available. The system operates as long as one of the components is functional; therefore, the distribution of the life of the system is that of the sum of two independent exponential random variables. Let T_1 and T_2 be independent exponentials with parameter λ, and let $S = T_1 + T_2$. Find $f_S(s)$.

This is given by the convolution of f_{T_1} and f_{T_2}

$$f_S(s) = \qquad\qquad\qquad\qquad .$$

Note the limits of integration. Beyond these limits, one of the two component densities is zero. Integrating, we have

$$f_S(s) = \qquad\qquad\qquad\qquad .$$

This is a gamma distribution with parameters 2 and λ.

Quotient

The procedure for finding the quotient of two continuous random variables is similar to that of the sum. First, find the cdf, then differentiate to find the density.

Suppose X and Y are continuous with joint density function f and $Z = Y/X$. The $F_Z(z) = P(Z \leq z)$ is the probability of the set of (x, y) such that $y/x \leq z$.

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

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

In particular, if X and Y are independent, then
$$f_Z(z) = \int_{-\infty}^{\infty} |x| f_X(x) f_Y(xz) dx.$$

Quotient, proof.

Proof:

If $X > 0$, this is the set $y \leq xz$; if $x < 0$, it is the set $y \geq xz$. Thus

$$F_Z(z) = \qquad\qquad\qquad\qquad\qquad .$$

Then by the change of variable $y = xv$ in the inner integral, we find that

$$\begin{aligned}
F_Z(z) &= \int_{-\infty}^{0}\int_{z}^{-\infty} xf(x,xv)dvdx + \int_{0}^{\infty}\int_{-\infty}^{z} xf(x,xv)dvdx \\
&= \int_{-\infty}^{0}\int_{-\infty}^{z} (-x)f(x,xv)dvdx + \int_{0}^{\infty}\int_{-\infty}^{z} xf(x,xv)dvdx \\
&= \int_{-\infty}^{z}\int_{-\infty}^{\infty} |x|f(x,xv)dxdv.
\end{aligned}$$

Example B

Suppose that$X$ and Y are independent standard normal random variables and that $Z = Y/X$. Prove that $f_Z(z) = \frac{1}{\pi(z^2+1)}$, for $= \infty < z < \infty$.

This density is called the **Cauchy density**. Like the standard normal density, the Cauchy density is symmetric about zero and bell-shaped. However, the tail of the Cauchy tend to zero very slowly compared to the tails of the normal. This can be interpreted as because a substantial probability that X in the quotient Y/X is near zero.

. . . .

Proof:
From previous slide and the density of standard normal,

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

Example cont

By symmetry of the integral about zero,

$$f_Z(z) = \int_0^{\infty} \frac{x}{\pi} e^{-x^2(1+z^2)/2} dx.$$

Making the change of variable $u = x^2$, we find that

$$f_Z(z) = \frac{1}{2\pi} \int_0^{\infty} e^{-u(1+z^2)/2} du.$$

Hence

$$\begin{aligned} f_Z(z) &= \left[-\frac{1}{\pi(z^2+1)} e^{-u(1+z^2)/2} \right]_0^{\infty} \\ &= -0 + \frac{1}{\pi(z^2+1)}. \end{aligned}$$

The General Case – Not in Exam

Suppose X and Y are jointly distributed continuous random variables, X and Y are mapped onto U and V by the transformation

$$u = g_1(x, y) \qquad v = g_2(x, y),$$

and the transformations can be inverted to obtain

$$x = h_1(u, v) \qquad y = h_2(u, v).$$

Suppose g_1 and g_2 have continuous partial derivatives and that for all x and y, the Jacobian

$$J(x, y) = \left(\frac{\partial g_1}{\partial x}\right)\left(\frac{\partial g_2}{\partial y}\right) - \left(\frac{\partial g_2}{\partial x}\right)\left(\frac{\partial g_1}{\partial y}\right) \neq 0.$$

Then the joint density of U and V is

$$f_{UV}(u, v) = f_{XY}(h_1(u, v)h_2(u, v))|J^{-1}(h_1(u, v), h_2(u, v))|$$

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

Ordered Statistics

Suppose $X_1, X_2, \ldots, X_n$ are independent random variables with a common cdf F and density f. Let U denote the maximum of the X_i's. Then the cdf of U is given by

$$\begin{aligned} F_U(u) &= \\ &= [F(u)]^n \end{aligned}$$

and so differentiating gives the density

$$f_U(u) = nf(u)[F(u)]^{n-1}.$$

Likewise, let V denote the minimum of the X_i's. Then

$$\begin{aligned} 1 - F_V(v) = P(V \geq v) &= P(X_1 \geq v)P(X_2 \geq v) \cdots P(X_n \geq v) \\ &= [1 - F(v)]^n. \end{aligned}$$

Hence

$$F_V(v) = 1 - [1 - F(v)]^n \qquad \text{and} \qquad f_V(v) = nf(v)[1 - F(v)]^{n-1}.$$

Example A

Suppose n system components are connected in series, i.e., the system fails if any one of them fails, and the lifetime of the components, $T_1, \ldots, T_n$, are independent random variables that are exponentially distributed with parameter λ: $F(t) = 1 - e^{-\lambda t}$.
Find the density function representing the length of time that the system operates.

. . . .

The length of time the system operates is the minimum, V, of the T_i's.

By the previous slide, the density is given by

$$\begin{aligned} f_V(v) &= n\lambda e^{-\lambda v}(e^{-\lambda v})^{n-1} \\ &= n\lambda e^{-n\lambda v}. \end{aligned}$$

Thus V is exponentially distributed with parameter $n\lambda$.

Example B

Suppose that a system has n components as described in Example A but are connected in parallel, which means that the system fails only when they all fail.
Find the density function representing the system's lifetime.

. . . .

Let U denote the system's lifetime. Then U is the maximum of n exponential random variables and has density

$$f_U(u) = \qquad\qquad\qquad\qquad .$$

We could expand the last term using the binomial theorem, and we will then obtain a weighted sum of exponential terms. Thus the density is not a simple exponential density, unlike in Example A.

Another approach - differential argument

Here is another approach to deriving the density of the maximum of independent random variables $X_1, X_2, \dots, X_n$ with a common cdf F.

Note that $f_U(u)$ consists of terms where one of the X_i's falls in the interval $(u, u + du)$ and the remaining $n - 1$ X_i's fall to the left of u.

Such a term has probability $[F(u)]^{n-1} f(u) du$, and since there are n such arrangements,

$$f_U(u) = n[F(u)]^{n-1} f(u).$$

This differential argument can be applied to more than just the maximum or minimum.

k-th order statistics

Definition

Suppose $X_1, X_2, \dots, X_n$ are independent continuous random variables with density $f(x)$. We sort the X_i's and denote by

$$X_{(1)} < X_{(2)} < X_{(3)} < \cdots < X_{(n)}.$$

Then $X_{(k)}$ is the **k-th order statistic**.

If $n = 2m + 1$ is odd, then $X_{(m+1)}$ is the **median** of the X_i's.

Note that X_1 is not necessarily equal to $X_{(1)}$, unless it also happens to be the minimum. (The probability that $X_1 = X_{(1)}$ is $\frac{1}{n}$.)

k-th order statistics

Theorem

The density of $X_{(k)}$, the k-th order statistics, is

$$f_k(x) = \frac{n!}{(k-1)!(n-k)!} f(x) F^{k-1}(x) [1 - F(x)]^{n-k}.$$

Proof: The event that $x \le X_{(k)} \le x + dx$ occurs if

- $k-1$ of the X_i's are less that x,
- one of the X_i's is in $[x, x+dx]$,
- $(n-k)$ of the X_i's are greater than $x + dx$.

Any one such arrangement has probability $F(x)^{k-1} f(x) [1 - F(x)]^{n-k} dx$. We multiply this by the total number of such arrangements, which is

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

and this completes the proof.

Example C

Suppose X_i's (for $1 \leq i \leq n$) are uniform on $[0, 1]$, Find the density of the k-th order statistic.

. . . .

By theorem on previous slide, the density of the k-th order statistics is

$$, \qquad (0 \leq x \leq 1).$$

This is the beta density. An interesting by product of this result is that since the density integrate to 1,

$$\int_0^1 x^{k-1}(1-x)^{n-k} dx = \frac{(k-1)!(n-k)!}{n!}.$$

Joint distribution of order statistics

Joint distributions of order statistics can also be worked out.

For example, to find the joint density of the minimum and maximum, we note that

$$x \leq X_{(1)} \leq x + dx \qquad \text{and} \qquad y \leq X_{(n)} \leq y + dy$$

if

- one of the X_i's ,
- $n - 2$ of the X_i's ,
- the remaining .

There are exactly $\frac{n!}{1!(n-2)!1!} = n(n-1)$ ways that this rearrangement occurs, and so

$$f(u, v) = n(n-1)f(v)f(u)[F(u) - F(v)]^{n-2}, \qquad u \geq v.$$

Joint distribution of order statistics

E.g. if the X_i's are uniform on $[0, 1]$, then

$$f(u, v) = n(n-1)[u-v]^{n-2}, \qquad 1 \geq u \geq v \geq 0.$$

We could define $R = X_{(n)} - X_{(1)}$ to study the range of the X_i's. The same kind of analysis then leads to

$$f_R(r) = \int_{-\infty}^{\infty} f(v+r, v) dv \qquad \text{(since we want } u - v \leq r\text{)}.$$

Example D

Find the distribution of the range, $U - V$, for the uniform $[0, 1]$ case.

. . . .

The integrand is for $0 \leq v \leq v + r \leq 1$, or equivalently, $0 \leq v \leq 1 - r$.

Then

$$f_R(r) = \qquad (0 \leq r \leq 1).$$

The corresponding cdf is

$$F_R(r) = \qquad , \qquad (0 \leq r \leq 1).$$