STAT 450: Statistical Theory

Distribution Theory

Reading in Casella and Berger: Ch 2 Sec 1, Ch 4 Sec 1, Ch 4 Sec 6.

Basic Problem: Start with assumptions about f or CDF of random vector $X = (X_1, \ldots, X_p)$. Define $Y = g(X_1, \ldots, X_p)$ to be some function of X (usually some statistic of interest like a sample mean or correlation coefficient or a test statistic). How can we compute the distribution or CDF or density of Y?

Univariate Techniques

In a *univariate* problem we have X a real-valued random variable and Y, a real valued transformation of X. Often we start with X and are told either the density or CDF of X and then asked to find the *distribution* of something like X^2 or $\sin(X)$ – a function of X which we call a *transformation*. I am going to show you the steps you should follow in an example problem.

Method 1: compute the CDF by integration and differentiate to find f_Y.

Example: Suppose $U \sim \text{Uniform}[0,1]$. What is the distribution of $-\log(X)$?.

Step 1: Make sure you give the transformed variable a name.

Let $Y = -\log U$.

Step 2: Write down the definition of the CDF of Y:

$$F_Y(y) = P(Y \le y).$$

Step 3: Substitute in the definition of Y in terms of the original variable – in this case, U:

$$F_Y(y) = P(Y \le y) = P(-\log(U) \le y).$$

Step 4: Solve the inequality to get, hopefully, one or more intervals for U. In this case

$$-\log(U) \le y$$

is the same as

$$\log(U) \ge -y$$

which is the same as

$$U \geq e^{-y}.$$

So

$$P(-\log(U) \leq Y) = P\left(U \geq e^{-y}\right).$$

Step 5: now use the assumptions to write the resulting probability in terms of F_U.

$$P\left(U \geq e^{-y}\right) = 1 - P\left(U < e^{-y}\right) = 1 - F_U\left(e^{-y}\right).$$

Step 6: Differentiate the resulting formula with respect to y. In this case the formula for $F_U(u)$ needs to be written carefully:

$$F_U(u) = \begin{cases} 1 & u > 1 \\ u & 0 < u < 1 \\ 0 & u \leq 0 \end{cases}$$

The quantity e^{-y} is never negative but if $y < 0$ $e^{-y} > 1$ so

$$1 - F_U(e^{-y}) = \begin{cases} 1 & y < 0 \\ 1 - e^{-y} & y \geq 0 \end{cases}$$

Now we can differentiate. For negative y the derivative is 0. For positive y the derivative is e^{-y}. At $y = 0$ the derivative does not exist and the density $f_Y(y)$ does not have a unique definition So our density is

$$f_Y(y) = \begin{cases} 0 & y < 0 \\ \text{undefined} & y = 0 \\ e^{-y} & y > 0 \end{cases}$$

This is the standard exponential density so we would finish the problem by saying that $-\log(U)$ has a standard exponential distribution.

I have done the individual steps in a bit of detail but here are the crucial pieces again in a chain of equalities:

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(-\log U \leq y) \\ &= P(\log U \geq -y) = P(U \geq e^{-y}) \\ &= \begin{cases} 1 - e^{-y} & y > 0 \\ 0 & y \leq 0\,. \end{cases} \end{aligned}$$

Differentiating we find

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \begin{cases} 0 & y < 0 \\ \text{undefined} & y = 0 \\ e^{-y} & y > 0 \end{cases}$$

so Y has standard exponential distribution.

Example: $Z \sim N(0,1)$, i.e.

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

and $Y = Z^2$. Then

$$\begin{aligned} F_Y(y) &= P(Z^2 \le y) \\ &= \begin{cases} 0 & y < 0 \\ P(-\sqrt{y} \le Z \le \sqrt{y}) & y \ge 0\,. \end{cases} \end{aligned}$$

Now differentiate

$$P(-\sqrt{y} \le Z \le \sqrt{y}) = F_Z(\sqrt{y}) - F_Z(-\sqrt{y})$$

to get

$$f_Y(y) = \begin{cases} 0 & y < 0 \\ \frac{d}{dy}\left[F_Z(\sqrt{y}) - F_Z(-\sqrt{y})\right] & y > 0 \\ \text{undefined} & y = 0\,. \end{cases}$$

Then

$$\begin{aligned} \frac{d}{dy} F_Z(\sqrt{y}) &= f_Z(\sqrt{y}) \frac{d}{dy}\sqrt{y} \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(-\left(\sqrt{y}\right)^2/2\right) \frac{1}{2} y^{-1/2} \\ &= \frac{1}{2\sqrt{2\pi y}} e^{-y/2}\,. \end{aligned}$$

(Similar formula for other derivative.) Thus

$$f_Y(y) = \begin{cases} \frac{1}{\sqrt{2\pi y}} e^{-y/2} & y > 0 \\ 0 & y < 0 \\ \text{undefined} & y = 0\,. \end{cases}$$

This is the density of a χ^2_1 random variable because the chi-squared distributing with ν degrees of freedom is defined to be the distribution of the sum of ν squared independent standard normal variables.

I am now going to rewrite this density using **indicator** notation which can be very useful:

$$1(y > 0) = \begin{cases} 1 & y > 0 \\ 0 & y \le 0 \end{cases}$$

which we use to write

$$f_Y(y) = \frac{1}{\sqrt{2\pi y}} e^{-y/2} 1(y > 0)$$

(changing definition unimportantly at $y = 0$).

Notice: I never evaluated F_Y before differentiating it. In fact F_Y and F_Z are integrals I can't do but I can differentiate them anyway. Remember the fundamental theorem of calculus:

$$\frac{d}{dx} \int_a^x f(y)\, dy = f(x)$$

at any x where f is continuous.

Summary: for $Y = g(X)$ with X and Y each real valued

$$\begin{aligned} P(Y \le y) &= P(g(X) \le y) \\ &= P(X \in g^{-1}(-\infty, y]) \, . \end{aligned}$$

Take d/dy to compute the density

$$f_Y(y) = \frac{d}{dy} \int_{\{x: g(x) \le y\}} f_X(x)\, dx \, .$$

Often can differentiate without doing integral.

Method 2: Change of variables.

Assume g is one to one. I do: g is increasing and differentiable. Interpretation of density (based on density $= F'$):

$$\begin{aligned} f_Y(y) &= \lim_{\delta y \to 0} \frac{P(y \le Y \le y + \delta y)}{\delta y} \\ &= \lim_{\delta y \to 0} \frac{F_Y(y + \delta y) - F_Y(y)}{\delta y} \end{aligned}$$

and

$$f_X(x) = \lim_{\delta x \to 0} \frac{P(x \le X \le x + \delta x)}{\delta x} .$$

Now assume $y = g(x)$. Define δy by $y + \delta y = g(x + \delta x)$. Then

$$P(y \le Y \le g(x + \delta x)) = P(x \le X \le x + \delta x) .$$

Get

$$\frac{P(y \le Y \le y + \delta y))}{\delta y} = \frac{P(x \le X \le x + \delta x)/\delta x}{\{g(x + \delta x) - y\}/\delta x} .$$

Take limit to get

$$f_Y(y) = f_X(x)/g'(x)$$

or

$$f_Y(g(x))g'(x) = f_X(x) .$$

Alternative view:

Each probability is integral of a density.

The first is the integral of the density of Y over the small interval from $y = g(x)$ to $y = g(x + \delta x)$. The interval is narrow so f_Y is nearly constant and

$$P(y \le Y \le g(x + \delta x)) \approx f_Y(y)(g(x + \delta x) - g(x)) .$$

Since g has a derivative the difference

$$g(x + \delta x) - g(x) \approx \delta x g'(x)$$

and we get

$$P(y \le Y \le g(x + \delta x)) \approx f_Y(y)g'(x)\delta x .$$

Same idea applied to $P(x \le X \le x + \delta x)$ gives

$$P(x \le X \le x + \delta x) \approx f_X(x)\delta x$$

so that

$$f_Y(y)g'(x)\delta x \approx f_X(x)\delta x$$

or, cancelling the δx in the limit

$$f_Y(y)g'(x) = f_X(x) .$$

If you remember $y = g(x)$ then you get

$$f_X(x) = f_Y(g(x))g'(x) .$$

Or solve $y = g(x)$ to get x in terms of y, that is, $x = g^{-1}(y)$ and then

$$f_Y(y) = f_X(g^{-1}(y))/g'(g^{-1}(y))\,.$$

This is just the change of variables formula for doing integrals.
Remark: For g decreasing $g' < 0$ but Then the interval $(g(x), g(x+\delta x))$ is really $(g(x+\delta x), g(x))$ so that $g(x) - g(x+\delta x) \approx -g'(x)\delta x$. In both cases this amounts to the formula

$$f_X(x) = f_Y(g(x))|g'(x)|\,.$$

Mnemonic:

$$f_Y(y)dy = f_X(x)dx\,.$$

Example: : $X \sim$ Weibull(shape α, scale β) or

$$f_X(x) = \frac{\alpha}{\beta}\left(\frac{x}{\beta}\right)^{\alpha-1} \exp\left\{-(x/\beta)^\alpha\right\} 1(x > 0)\,.$$

Let $Y = \log X$ or $g(x) = \log(x)$.
Solve $y = \log x$: $x = \exp(y)$ or $g^{-1}(y) = e^y$.
Then $g'(x) = 1/x$ and $1/g'(g^{-1}(y)) = 1/(1/e^y) = e^y$.
Hence

$$f_Y(y) = \frac{\alpha}{\beta}\left(\frac{e^y}{\beta}\right)^{\alpha-1} \exp\left\{-(e^y/\beta)^\alpha\right\} 1(e^y > 0)e^y\,.$$

For any y, $e^y > 0$ so indicator $= 1$. So

$$f_Y(y) = \frac{\alpha}{\beta^\alpha} \exp\left\{\alpha y - e^{\alpha y}/\beta^\alpha\right\}\,.$$

Define $\phi = \log\beta$ and $\theta = 1/\alpha$; then,

$$f_Y(y) = \frac{1}{\theta}\exp\left\{\frac{y-\phi}{\theta} - \exp\left\{\frac{y-\phi}{\theta}\right\}\right\}\,.$$

Extreme Value density with **location** parameter ϕ and **scale** parameter θ. (Note: several distributions are called Extreme Value.)

Marginalization

Simplest multivariate problem:

$$X = (X_1, \ldots, X_p), \qquad Y = X_1$$

(or in general Y is any X_j).

Theorem 1 *If X has density $f(x_1, \ldots, x_p)$ and $q < p$ then $Y = (X_1, \ldots, X_q)$ has density*

$$f_Y(x_1, \ldots, x_q) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \ldots, x_p)\, dx_{q+1} \ldots dx_p \,.$$

$f_{X_1,\ldots,X_q}$ is the **marginal** density of $X_1, \ldots, X_q$ and f_X the **joint** density of X but they are both just densities. "Marginal" just to distinguish from the joint density of X.

Example: : The function

$$f(x_1, x_2) = K x_1 x_2 1(x_1 > 0, x_2 > 0, x_1 + x_2 < 1)$$

is a density provided

$$P(X \in R^2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, x_2)\, dx_1\, dx_2 = 1 \,.$$

The integral is

$$\begin{aligned} K \int_0^1 \int_0^{1-x_1} & x_1 x_2 \, dx_1\, dx_2 \\ & = K \int_0^1 x_1 (1 - x_1)^2 \, dx_1/2 \\ & = K(1/2 - 2/3 + 1/4)/2 \\ & = K/24 \end{aligned}$$

so $K = 24$. The marginal density of x_1 is

$$\begin{aligned} f_{X_1}(x_1) = & \int_{-\infty}^{\infty} 24 x_1 x_2 \\ & \times 1(x_1 > 0, x_2 > 0, x_1 + x_2 < 1)\, dx_2 \\ =& 24 \int_0^{1-x_1} x_1 x_2 1(0 < x_1 < 1) dx_2 \\ =& 12 x_1 (1 - x_1)^2 1(0 < x_1 < 1) \,. \end{aligned}$$

This is a Beta$(2,3)$ density.
General problem has $Y = (Y_1, \ldots, Y_q)$ with $Y_i = g_i(X_1, \ldots, X_p)$.
Case 1: $q > p$. Y **won't** have density for "smooth" g. Y will have a **singular** or discrete distribution. Problem rarely of real interest. (But, e.g., residuals have singular distribution.)
Case 2: $q = p$. We use a change of variables formula which generalizes the one derived above for the case $p = q = 1$. (See below.)
Case 3: $q < p$. Pad out Y–add on $p - q$ more variables (carefully chosen) say $Y_{q+1}, \ldots, Y_p$. Find functions $g_{q+1}, \ldots, g_p$. Define for $q < i \le p$, $Y_i = g_i(X_1, \ldots, X_p)$ and $Z = (Y_1, \ldots, Y_p)$. Choose g_i so that we can use change of variables on $g = (g_1, \ldots, g_p)$ to compute f_Z. Find f_Y by integration:

$$f_Y(y_1, \ldots, y_q) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_Z(y_1, \ldots, y_q, z_{q+1}, \ldots, z_p) dz_{q+1} \ldots dz_p$$

Change of Variables

Suppose $Y = g(X) \in R^p$ with $X \in R^p$ having density f_X. **Assume g is a one to one ("injective") map,** i.e., $g(x_1) = g(x_2)$ if and only if $x_1 = x_2$. Find f_Y:
Step 1: Solve for x in terms of y: $x = g^{-1}(y)$.
Step 2: Use basic equation:

$$f_Y(y)dy = f_X(x)dx$$

and rewrite it in the form

$$f_Y(y) = f_X(g^{-1}(y)) \frac{dx}{dy}.$$

Interpretation of derivative $\frac{dx}{dy}$ when $p > 1$:

$$\frac{dx}{dy} = \left| \det \left(\frac{\partial x_i}{\partial y_j} \right) \right|$$

which is the so called **Jacobian**.

Equivalent formula inverts the matrix:

$$f_Y(y) = \frac{f_X(g^{-1}(y))}{\left|\frac{dy}{dx}\right|}$$

This notation means

$$\left|\frac{dy}{dx}\right| = \left|\det\begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_p} \\ & \vdots & & \\ \frac{\partial y_p}{\partial x_1} & \frac{\partial y_p}{\partial x_2} & \cdots & \frac{\partial y_p}{\partial x_p} \end{bmatrix}\right|$$

but with x replaced by the corresponding value of y, that is, replace x by $g^{-1}(y)$.

Example: : The density

$$f_X(x_1, x_2) = \frac{1}{2\pi}\exp\left\{-\frac{x_1^2 + x_2^2}{2}\right\}$$

is the **standard bivariate normal density**. Let $Y = (Y_1, Y_2)$ where $Y_1 = \sqrt{X_1^2 + X_2^2}$ and $0 \le Y_2 < 2\pi$ is angle from the positive x axis to the ray from the origin to the point (X_1, X_2). I.e., Y is X in polar co-ordinates.

Solve for x in terms of y:

$$\begin{aligned} X_1 &= Y_1\cos(Y_2) \\ X_2 &= Y_1\sin(Y_2) \end{aligned}$$

so that

$$\begin{aligned} g(x_1, x_2) &= (g_1(x_1, x_2), g_2(x_1, x_2)) \\ &= (\sqrt{x_1^2 + x_2^2}, \text{argument}(x_1, x_2)) \\ g^{-1}(y_1, y_2) &= (g_1^{-1}(y_1, y_2), g_2^{-1}(y_1, y_2)) \\ &= (y_1\cos(y_2), y_1\sin(y_2)) \\ \left|\frac{dx}{dy}\right| &= \left|\det\begin{pmatrix} \cos(y_2) & -y_1\sin(y_2) \\ \sin(y_2) & y_1\cos(y_2) \end{pmatrix}\right| \\ &= y_1\,. \end{aligned}$$

It follows that

$$f_Y(y_1, y_2) = \frac{1}{2\pi}\exp\left\{-\frac{y_1^2}{2}\right\} y_1 \times 1(0 \le y_1 < \infty)1(0 \le y_2 < 2\pi)\,.$$

Next: marginal densities of Y_1, Y_2?

Factor f_Y as $f_Y(y_1, y_2) = h_1(y_1)h_2(y_2)$ where

$$h_1(y_1) = y_1 e^{-y_1^2/2} 1(0 \le y_1 < \infty)$$

and

$$h_2(y_2) = 1(0 \le y_2 < 2\pi)/(2\pi)\,.$$

Then

$$\begin{aligned} f_{Y_1}(y_1) &= \int_{-\infty}^{\infty} h_1(y_1)h_2(y_2)\,dy_2 \\ &= h_1(y_1)\int_{-\infty}^{\infty} h_2(y_2)\,dy_2 \end{aligned}$$

so marginal density of Y_1 is a multiple of h_1. Multiplier makes $\int f_{Y_1} = 1$ but in this case

$$\int_{-\infty}^{\infty} h_2(y_2)\,dy_2 = \int_0^{2\pi} (2\pi)^{-1} dy_2 = 1$$

so that

$$f_{Y_1}(y_1) = y_1 e^{-y_1^2/2} 1(0 \le y_1 < \infty)\,.$$

(Special case of Weibull or Rayleigh distribution.) Similarly

$$f_{Y_2}(y_2) = 1(0 \le y_2 < 2\pi)/(2\pi)$$

which is the **Uniform**$((0, 2\pi)$ density.

Exercise: $W = Y_1^2/2$ has standard exponential distribution.

Recall: by definition $U = Y_1^2$ has a χ^2 distribution on 2 degrees of freedom.

Exercise: find χ_2^2 density.

Note: We show below factorization of density is equivalent to independence.