STAT 3503/8109 Lecture 5 Notes

Edoardo Airoldi

Scribe: Srikar Katta[*]

Fall 2020: September 28, 2020

1 Introduction

Today, we will be discussing the transformation theorem, an idea you can use to "simplify" your life and differentiates mathematics from statistics. For instance, in mathematics, $\mathbb{E}[X] = \int_X X f_X(x)dx$ - the expectation is just an integral - it is just an expression. However, in statistics, there are many ways of representing the expectation (discussed in lecture 3 notes), and it goes beyond just being an interval. If you have a random variable $Y = g(X)$, then we can find the expectation using the density of X and simplify our life. This comes from Casella & Berger Theorems 2.1.5 and 2.1.8 (PDF will be posted on Canvas).

2 Why do we need transformation theorem?

We are given $x_1 ... x_n \overset{\text{iid}}{\sim} exp(\theta)$. So, the density would be the following:

$$exp(\theta) = \theta e^{-\theta x_i}, \theta > 0, x_i \geq 0$$

We are also given that $y_1 ... y_m \overset{\text{iid}}{\sim} exp(\gamma) = \gamma e^{-\gamma y - i}$. Find the following:

$$\mathbb{E}[\sum x_i - \sum y_i]$$

[*] Please share any comments or suggestions with Srikar Katta at srikar@temple.edu

Answer: Notice,

$$\mathbb{E}[\sum x_i - \sum y_i] = \left(\int_{x_1} ... \int_{x_n} \sum_{i=1}^{n} \prod_{i=1}^{n} f_x(x_i) dx_n ... dx_1\right) - \left(\int_{y} 1 ... \int_{y} m \sum_{i=1}^{m} \prod_{i=1}^{m} f_y(y_i) dy_m ... dy_1\right)$$

. We just expanded the expectation into it's mathematical form with integrals. Here, a mathematician might "plug and chug" and find the answer. However, we want to be more creative and efficient with our resources. This is where the transformation theorem comes into play. It allows us to simplify these problems into something a lot more digestible.

Assume the following (the same statement with the introduction of different notation):

$$\begin{aligned} x_1 ... x_n &\overset{\text{iid}}{\sim} exp(\theta) \\ \tilde{x} &= \sum_{i=1}^{n} x_i \\ y_1 ... y_m &\overset{\text{iid}}{\sim} exp(\gamma) \\ \tilde{y} &= \sum_{i=1}^{m} y_i \end{aligned}$$

The transformation theorem allows us to find out that $\tilde{X} \sim gamma(\theta, n)$ and $\tilde{y} \sim gamma(\gamma, m)$ (we will see how later). Now, we can find $\mathbb{E}[\sum x_i - \sum y_i]$:

$$\begin{aligned} \mathbb{E}\left[\sum x_i - \sum y_i\right] &= \left[\int_{x_i > 0}^{\infty} \sum_{i=1}^{n} x_i f_{x_i}(x_i) dx\right] - \left[\int_{y_i > 0}^{\infty} \sum_{i=1}^{n} y_i f_{y_i}(x_i) dy\right] \\ &= \left[\int_{x_i > 0}^{\infty} \sum_{i=1}^{n} x_i \prod_{i=1}^{n} f_{x_i}(x_i) dx\right] - \left[\int_{y_i > 0}^{\infty} \sum_{i=1}^{n} y_i \prod_{i=1}^{n} f_{y_i}(y_i) dy\right] \\ &= \left[\int_{x_i > 0}^{\infty} \sum_{i=1}^{n} x_i \prod_{i=1}^{n} f_{gamma}(x_i | \theta, n) dx\right] - \left[\int_{y_i > 0}^{\infty} \sum_{i=1}^{n} y_i \prod_{i=1}^{n} f_{gamma}(x_i | \theta, n) dy\right] \\ &= \int_{\tilde{x}} \tilde{x} * gamma(\theta, n) d\tilde{x} - \int_{\tilde{y}} \tilde{y} * gamma(\gamma, m) d\tilde{y}. \end{aligned}$$

The transformation theorem will allow us to identify distributions for random variables dependent on our *known* random variables, making our computation much simpler since we now do not have to compute $n + m$ integrals, just two.

There are many situations in which we are looking for some aggregate random variable for many individual random variables (e.g., average price paid per purchase for many customers or the average grade on exam across students). We can simplify these problems from being multivariate ones to being univariate, simplifying our lives *drastically.* This is essentially the main use for the transformation theorem.

3 Facts we know about distributions

Statisticians know the following:

- The sum of normal random variables is a normal random variables
- The sum of exponential random variables is a gamma random variable
- The square of a normal random variable is actually χ^2 with some number of degrees of freedom

However, if you are not a data scientist/statistician who has proved these things, the transformation theorem would allow you to discover these without explicitly having known them beforehand. Now, we will give you a couple variants of the transformation theorem and work through a few examples.

4 Basic Transformation Theorem

First, let us recall some information from calculus. The derivative is a measure of the *instantaneous* rate of change for some function. In other words, it is a slope of a function at a certain point that we can climb up and down to find the maximum and minimum. Additionally, the integral measures the "area" under the curve.

Assume $X \sim f_x(x|\theta)$, a continuous random variable. Now, say we have Y, a deterministic function of X: $Y = g(X)$. We want to know the following: what is $f_x(y|\theta)$? Why do we want to know this? Well, $\mathbb{E}[x|\theta] = \int_x x f_x(x) dx$, by definition of the expectation operator. And $\mathbb{E}[y|\theta] = \mathbb{E}[g(x)|\theta] = \int_x g(x) f_x dx$, again by definition of the expectation operator. We also have another way of representing $\mathbb{E}[y|\theta]$ though. Again, by definition of the expectation operator, $\mathbb{E}[y|\theta] = \int_y y f_y(y) dy$. Now, if $g(x)f_x$ is complicated, that first representation of $\mathbb{E}[y|\theta]$ will be very difficult to compute. But in the second problem, we do not know what $f_y(y)$ is. The transformation theorem will help us find this.

Theorem 1. *If $y = g(x)$, then $x = g^{-1}(y)$ (with the caveat that this requires a unique inverse; we will discuss a solution if this is not the case). Then,*

$$f_y(y|\theta) = f_x(g^{-1}(y)|\theta) * \left|\frac{dg^{-1}(y)}{dy}\right|$$

Let us try to build an intuition for this with an example.

Example 4.1. *Assume $x \sim Normal(x|\mu, 1)$ and $y = e^x$. What is $f_y(y)$?*

If $x \in \mathbb{R}$, then $y \in \mathbb{R}, y > 0$. In words, because $y = e^x$, whose output is always greater than 0, y must be an element in the real numbers, greater than or equal to 0. So, the support[1] of x is the set of all real numbers while the support of y is the real numbers, greater than or equal to 0. This paragraph is saying the same thing - just using different terminology to get you more comfortable with the many ways of representing ideas in statistics.

Now, we must also find the inverse of g because the transformation theorem allows us to find f_y if we can find g^{-1}. From high school, we know that we can find the inverse by "swapping" x with y for the function in which we want to find the inverse of ($y = e^x$). So, $x = e^y \implies log(e) = y$, where log represents the natural log or log base e. We can check this is the true inverse by making sure that $y = g(g^{-1}(y))$ and $x = g^{-1}(g(x))$.

$$\begin{aligned}
g(g^{-1}(y)) &= g(log(y)) \\
&= e^{log(y)} \\
&= y \\
&AND \\
g^{-1}(g(x)) &= g^{-1}(e^x) \\
&= log(e^x) \\
&= x,
\end{aligned}$$

so $log(x)$ is in fact the inverse of g. Now, let us return to the computation of the transformation theorem itself:

$$\begin{aligned}
f_y(y|\mu) &= f_x(g^{-1}(y)|\theta) * \left|\frac{dg^{-1}(y)}{dy}\right| \\
&= f_x(log(y)|\theta) * \left|\frac{log(y)}{dy}\right|
\end{aligned}$$

[1] Recall, the support is just the set of all possible values that a random variable can adopt

Now, recall that since x follows a normal distribution, it has the following PDF:

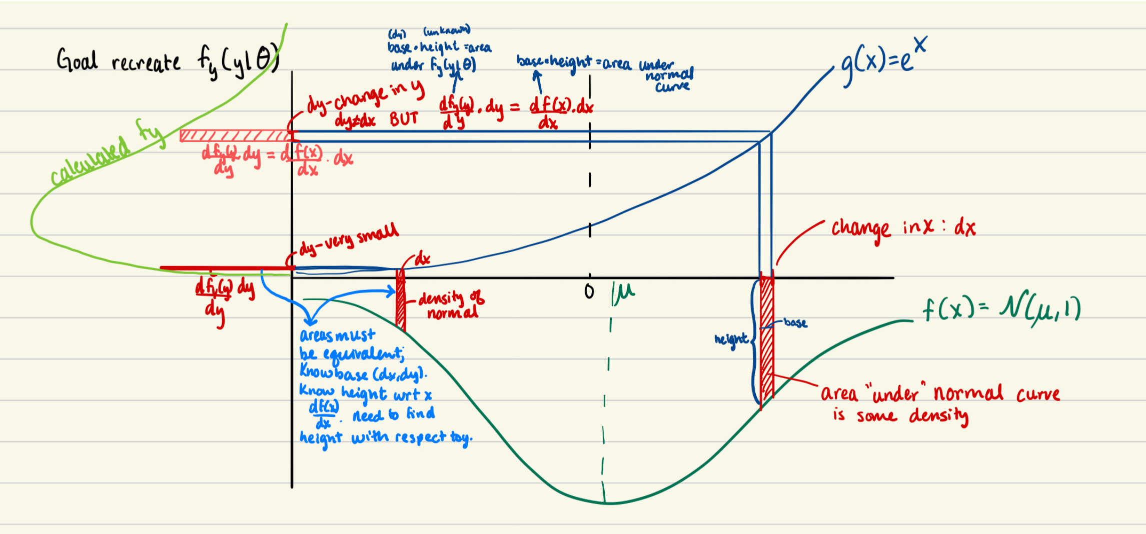
$$f_a(a|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(a-\mu)^2}{\sigma^2})}.$$

So, we can replace f_x in our transformation with this:

$$\begin{aligned}
f_y(y|\mu) &= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(log(y)-\mu)^2} \left| \frac{dlog(y)}{dy} \right| \\
&= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(logy-\mu)^2} * \frac{1}{y}, y \in \mathbb{R}, y > 0 \text{ because the derivative of } log(y) = \frac{1}{y} \\
&= \frac{1}{y\sqrt{2\pi}} e^{-\frac{1}{2}(logy-\mu)^2}, y \in \mathbb{R}^+,
\end{aligned}$$

which is the same thing as a log-normal distribution (see appendix in notes 4 for resources if this is unfamiliar).

Here is the picture. On the x-axis, we have our random variable $X \in \mathbb{R}$ (a number on the real number line). Above the origin is $y = g(x) = e^x$ and below the origin we have the normal distribution.



Goal recreate $f_y(y|\theta)$
(dy) (unknown)
base • height = area under $f_y(y|\theta)$
base • height = area under normal curve
$g(x)=e^x$
dy - change in y
dy≠dx BUT $\frac{df_y(y)}{dy} \cdot dy = \frac{df(x)}{dx} \cdot dx$
calculated fy
$\frac{df_y(y)}{dy} \cdot dy = \frac{df(x)}{dx} \cdot dx$
change in x : dx
dy - very small
dx
$\frac{df_y(y)}{dy}$ dy
density of normal
0
μ
$f(x) = \mathcal{N}(\mu,1)$
base
height
areas must be equivalent; Know base (dx, dy). Know height wrt x $\frac{df(x)}{dx}$. need to find height with respect to y.
area "under" normal curve is some density

Notice, this function, $g(x) = e^x$ is bounded below by 0. For every value on our x-axis, we will find a value at $g(x)$. So, now we will consider some very small change in x, the differential - dx. Over dx, we will find the area under the normal curve. Also, because g and the normal distribution share the same domain, we can find the area under the curve of $g(x)$ for the same differential. Now, this will lead us to some end points for $g(x)$ which we can then measure as a differential with respect to y, $dg(x)$. We know that we want the density from our normal distribution and the new distribution for the same differential in x to be the same. So, if the $dy > dx$, that's equivalent to saying dy has a larger "base." So, to have the same density, we must have a smaller height. And we are searching for this height to try and find the probability density function for y. Now, if $dx > dy$, then we could have a large area under the curve for the normal distribution but a small "base" from the differential with respect to y. So, the "height" of the probability distribution function of y will be greater to keep the density the same. So from our formula, $\left|\frac{dg^{-1}(y)}{dx}\right|$, just represents some correction factor to expand and contract the change from x to y.

Let us review the transformation theorem so far. We have some random variable X with a probability density function $f_x(x|\theta)$ and some random variable y that is a function of x, so $y = g(x)$. So, we can find the probability density function of y using the probability density function with x:

$$f_y(y|\theta) = f_x(g^{-1}(y)|\theta)\left|\frac{dg^{-1}(y)}{dx}\right|$$

Let us look at a slightly more complicated problem.

Example 4.2. *Let X be a random variable such that $X \sim f_x(x|n,\beta) = \frac{1}{(n-1)!\beta^n}x^{n-1}e^{\frac{-x}{\beta}}$. This is the density of the Gamma distribution, but this fact is secondary. All that matters is that f_x is some expression. Now, define $y = g(x) = \frac{1}{x}$.*

If we were to draw this, we would see that x takes on only positive values so y takes on only positive values. Now, we can find $g^{-1}(y) = \frac{1}{y}$. How do we quantify the relationship between dx and dy? Well, it is the same thing as saying $\frac{dx}{dy} = \frac{dg^{-1}(y)}{dy} = \frac{-1}{y^2}$, from our calculus derivative rules. This expression quantifies how much x changes for every little bit of dy. So, if we want to find $f_y(y|n,\beta)$, we just need to follow our transformation theorem formula:

$$\begin{aligned}
f_y(y|n,\beta) &= f_x(x|n,\beta)\left|\frac{dx}{dy}\right| \\
&= f_x(g^{-1}(y)|n,\beta)\left|\frac{dx}{dy}\right| \\
&= \frac{1}{(n-1)!\beta^n}(\frac{1}{y})^{n-1}e^{\frac{-1}{\beta y}} * \left|-\frac{1}{y^2}\right| \\
&= \frac{1}{(n-1)!\beta^n} * (\frac{1}{y})^{n-1}e^{\frac{-1}{\beta y}}\frac{1}{y^2}
\end{aligned}$$

So, if you know what the inverse is in terms of y, then we can easily find the probability density function of y.

5 Variant of The Transformation Theorem

So far we have considered a function of X, our random variable, such that there exists an inverse of $y = g(x)$, so that $x = g^{-1}(y)$. Suppose we still have a random variable X with density $f_x(x|\theta)$ and suppose that the support of X is some set A. Now, suppose $y = g(x)$ and assume that $x = g^{-1}(y)$ does not exist on the whole set A. For example, suppose $X \in \mathbb{R}$ but $y = x^2$. Then, there is no function that maps y to all the values in the $\mathbb{R}$ since $y = x^2 \geq 0$ always. However, we can do the following:

$$\begin{aligned} x &= g_1^{-1}(y) \text{ exists on } A_1 \subset A \\ x &= g_2^{-1}(y) \text{ exists on } A_2 \subset A, A_1 \cap A_2 = \emptyset. \end{aligned}$$

This means that we can find all the values of x again by considering two different "inverses" of g in which we can map y to every value in A and therefore map y to x somehow. Return to our $y = x^2$ example. With definitions of functions, x^2 as a function of x is okay because there is only one output for every input. However, $x = \sqrt{y}$ as a function of $y \in \mathbb{R}^+$ is not okay because one y can return $\pm\sqrt{x}$. So, we can split it up so that $x = \sqrt{y}$ for $x \in \mathbb{R}^+$ and $x = \sqrt{y}$ for $x \in \mathbb{R}^-$. So, we have one output for each function and one returns all positive $\sqrt{y}$ and another returns all negative $\sqrt{y}$. So, to solve such problems, we split the transformation theorem on the positive and negative axes and compose them together. Here is what the theorem itself says:

If $x \sim f_x(x|\theta), y = g(x)$, then

$$f_y(y|\theta) = \sum_{i=1}^{n} f_x(g_i^{-1}(y)|\theta) \left| \frac{dg_i^{-1}(y)}{dy} \right|, x \in A_i$$

So, we break down the $x - axis$ into pieces and now we sum them up together.

Example 5.1. *Let $X \sim N(0, 1)$ and $y = x^2$. What is $f_y(y)$?*

First, we must find $g^{-1}(y)$:

$$g_{-1}(y) = \begin{cases} g_1^{-1}(y) = \sqrt{y} & \text{if } x \geq 0 \\ g_2^{-1}(y) = -\sqrt{y} & \text{if } x < 0. \end{cases}$$

So,

$$\begin{aligned} f_y(y) &= \frac{1}{\sqrt{2\pi}} e^{-(\sqrt{y})^2} * \left| \frac{d\sqrt{y}}{dy} \right| + \frac{1}{\sqrt{2\pi}} e^{-(-\sqrt{y})^2} * \left| \frac{d-\sqrt{y}}{dy} \right| \\ &= \frac{1}{\sqrt{2\pi}} e^{-(\sqrt{y})^2} * \left| \frac{d\sqrt{y}}{dy} \right| + \frac{1}{\sqrt{2\pi}} e^{-(-\sqrt{y})^2} * \left| \frac{d-\sqrt{y}}{dy} \right| \\ &= \frac{1}{\sqrt{2\pi}} e^{-(\sqrt{y})^2} * \left| \frac{1}{2\sqrt{y}} \right| + \frac{1}{\sqrt{2\pi}} e^{-(-\sqrt{y})^2} * \left| \frac{-1}{2\sqrt{y}} \right| \\ &= \frac{1}{\sqrt{2\pi}} e^{-(\sqrt{y})^2} * \frac{1}{2\sqrt{y}} + \frac{1}{\sqrt{2\pi}} e^{-(-\sqrt{y})^2} * \frac{-1}{2\sqrt{y}} \\ &= \left(\frac{1}{2\sqrt{2\pi y}} + \frac{1}{2\sqrt{2\pi y}} \right) e^{-y}, y \geq 0 \\ &= \frac{1}{\sqrt{2\pi y}} e^{-y}, y \geq 0, \end{aligned}$$

which is a $\chi^2_{(2)}$, a chi squared distribution with two degrees of freedom. The key takeaways: because a single inverse did not exist, we had to break it up and built it back.

6 Summary

So so far, we have talked about a model: this involves a data generating process and a problem statement. The DGP will lay out what is variable and what is constant and will allow you to simulate some "fake" data. The problem statement will allow you to identify what is observed and what is not. So, the DGP will lead to the probability(random variables, latent variables | constants). DGP will give you the probabilities but the problem statement differentiates between random and latent quantities. And this then allows us to compute the likelihood. Today, what we learned will allow us to represent the likelihood as the probability of some function g of the random variables.