

# Transformations of Random Variables

Justin Silverman

Penn State University

# Table of Contents

- 1 Measure Theory Made Ridiculously Simple
- 2 Transformation of Random Variables
- 3 Density Estimation Methods
- 4 Monte Carlo Integration
- 5 Example
- 6 Application - Transforming Dependent Variables in Regression
- 7 Want to learn more measure theory?

# Section 1

## Measure Theory Made Ridiculously Simple

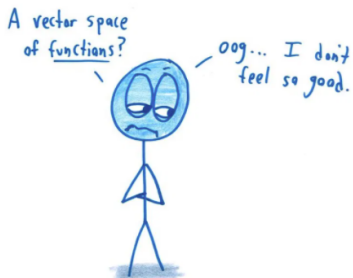
# Measure Theory Made Ridiculously Simple

These slides are based on a blog post I wrote a number of years ago.

<http://www.statsathome.com/2017/06/26/measure-theory-made-ridiculously-simple/>

# Emotions during this lecture

Abstractitude Sickness: the  
disorientation and nausea of  
ascending to a new level  
of abstraction



# Emotions during this lecture

The Flatlander's Glimpse: the wondrous sensation, however brief, of having successfully imagined an extra spatial dimension



# Why should I care?

Measure theory underlies the entire notion of random-variables, probability, statistics, and much of machine learning.

While this material is not essential to this course, it can give you a much deeper understanding of machine learning.

# Measures made simple

We talk about measuring probability.

Think of a measure  $\mu(x)$  as the probability that  $x$  is observed.

All of probabilistic modeling comes down to  $\mu(x)$ .



# Measures made simple

We talk about measuring probability.

Think of a measure  $\mu(x)$  as the probability that  $x$  is observed.

All of probabilistic modeling comes down to  $\mu(x)$ .

$x$  is a subset of possible observations from the larger set  $X$ .

$\mu(x)$  is a set function that has special constraints.

- $\mu$  must return results in the unit interval  $[0, 1]$ , returning 0 for the empty set and 1 for the entire space.
- $\mu$  must satisfy countable additivity. Simplified: The sum of  $\sum_i \mu(x_i) = \mu(\bigcup_i x_i)$ .

↑  
 $x_i$  disjoint

# Measures made simple

We talk about measuring probability.

Think of a measure  $\mu(x)$  as the probability that  $x$  is observed.

All of probabilistic modeling comes down to  $\mu(x)$ .


$x$  is a subset of possible observations from the larger set  $X$ .

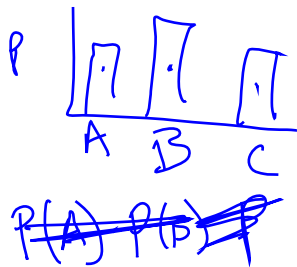
$\mu(x)$  is a set function that has special constraints.

- $\mu$  must return results in the unit interval  $[0, 1]$ , returning 0 for the empty set and 1 for the entire space.
- $\mu$  must satisfy *countable additivity*. Simplified: The sum of  $\sum_i \mu(x_i) = \mu(\bigcup_i x_i)$ .

**You can think of a probability measure as a generalization of a volume element.**

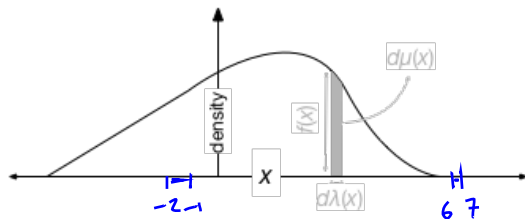
Why its hard to model with set functions directly.

$$X = \{A, B, C\}$$
$$\mu(\{A, B\})$$




# Measures made simple

You can think of a probability measure as a generalization of a volume element.



$$\begin{array}{c} \text{height} \cdot \text{width} = \text{probability} \\ \downarrow \qquad \downarrow \qquad \downarrow \\ \underline{f(x)} \cdot \underline{d\lambda(x)} = \underline{d\mu(x)} \end{array}$$

We often take  $\lambda(x)$  to be the *Lebesgue Measure*. Its like a uniform function over the sample space. The Lebesgue measure is probably what you would think to do before you even learned about measure theory.

# The Radon-Nykodym Derivative

$$f(x) = \frac{d\mu(x)}{d\lambda(x)}$$

(height = [infinitesimal] area / [infinitesimal] width)<sup>1</sup>

---

<sup>1</sup>Note, this is not strickly speaking the same type of derivative you learned in introductory calculus. This is a more general form of derivative.

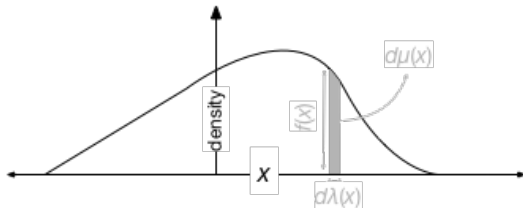
# The Radon-Nikodym Derivative

$$f(x) = \frac{d\mu(x)}{d\lambda(x)}$$

height = [infinitesimal] area / [infinitesimal] width

~~$P(A) =$~~   $\mu(A) = \int_{A \in X} f(x) d\lambda(x)$

This tells us what probability means  $\mu(A)$ . This tells us what a probability density means  $f(x)$ .



height · width = probability

$$\begin{array}{ccc} \updownarrow & \updownarrow & \updownarrow \\ f(x) & \cdot & d\lambda(x) = d\mu(x) \end{array}$$

## Where does Introductory Calculus Fit into all this?

if  $x \in \mathcal{R}^p$  and  $\lambda(x)$  is the Lebesgue Measure then you get introductory calculus:

$$f(x) = \frac{d\mu(x)}{dx}$$

$$\mu(A) = \int_{A \in X = \mathcal{R}^p} f(x) dx$$

# The Role of Sets and Set Theory in Measure Theory

- $\sigma$ -fields
- $\sigma$ -algebra
- Borel Sets
- etc. . .

This is where people get confused. But the underlying concepts are simple.

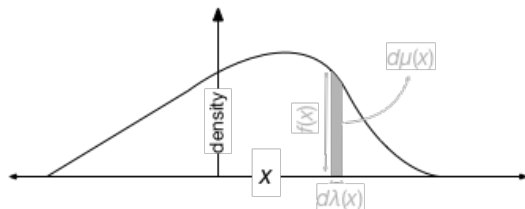


# The Role of Sets and Set Theory in Measure Theory

- $\sigma$ -fields
- $\sigma$ -algebra
- Borel Sets
- etc. . .

This is where people get confused. But the underlying concepts are simple.

**You can think of a probability measure as a generalization of a volume element.**



$$\begin{array}{ccccc} \text{height} \cdot \text{width} & = & \text{probability} \\ \updownarrow & & \updownarrow & & \updownarrow \\ f(x) \cdot d\lambda(x) & = & d\mu(x) \end{array}$$

## Section 2

# Transformation of Random Variables

# What do we mean by Transformation of Random Variables?

$$x \sim N(\mu, \sigma^2)$$

$$z = x^2$$

If I ask you, “what is the distribution of  $z$ ?”, what am I really asking?

## Transformation of Random Variables (Intuition)

If I draw  $(x_1, \dots, x_N)$  *iid* from  $x \sim N(\mu, \sigma^2)$ , and then transform each  $x_i$  ( $z_i = x_i^2$ ), then what is the distribution of the sample  $(z_1, \dots, z_N)$ ?

# The Mathematical Definition

Let  $X$  be a random variable over a set  $S$  such that we may write  $Pr(X \in A)$  where  $A \in S$ .<sup>2</sup>

---

<sup>2</sup> $Pr(X \in A)$  is the amount of probability \*measured\* in set  $A$ . (measure theory)

# The Mathematical Definition

Let  $X$  be a random variable over a set  $S$  such that we may write  $\Pr(X \in A)$  where  $A \in S$ .<sup>2</sup>

Let  $Y = r(X)$  for some function  $r: S \rightarrow T$ , that is  $Y$  exists in the set  $T$ .  
Let  $\underline{B} \in T$ .

---

<sup>2</sup> $\Pr(X \in A)$  is the amount of probability \*measured\* in set  $A$ . (measure theory)

# The Mathematical Definition

Let  $X$  be a random variable over a set  $S$  such that we may write  $Pr(X \in A)$  where  $A \in S$ .<sup>2</sup>

Let  $Y = r(X)$  for some function  $r : S \rightarrow T$ , that is  $Y$  exists in the set  $T$ . Let  $B \in T$ .

## The Definition of a Transformation of Random Variables

$$Pr(Y \in \underline{B}) = Pr(X \in r^{-1}(B))$$

**In words:** The probability that  $Y$  is in  $B$  is equal to the probability that  $X$  is within the set of values that get mapped into  $B$  under transformation.

---

<sup>2</sup> $Pr(X \in A)$  is the amount of probability \*measured\* in set  $A$ . (measure theory)

# Intuitive View of the Mathematical Definition

$$\Pr(Y \in B) = \Pr(X \in r^{-1}(B))$$

**In words:** The probability that  $Y$  is in  $B$  is equal to the probability that  $X$  is within the set of values that get mapped into  $B$  under transformation.

A function  $r : S \rightarrow T$ . How is a probability distribution on  $S$  transformed by  $r$  to a distribution on  $T$ ?

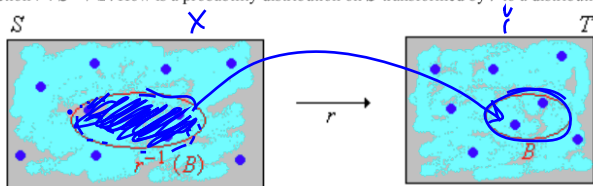


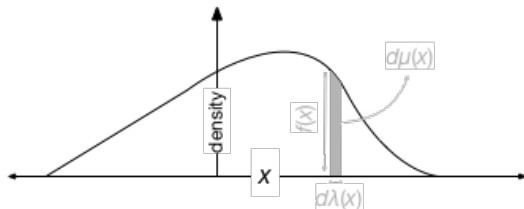
Figure 1: Image from:

<https://www.randomservices.org/random/dist/Transformations.html>

# Core Concept

Under a transformation of random variables, the ultimate probability measure stays the same (the same amount of probability, distributed over  $X$  in the same way as before the transformation).

What changes is our meter-stick – how we count probability over the new set over which  $Y$  is defined.



$$f_{\lambda}(x) = \frac{d\mu(x)}{d\lambda(x)}$$

The PDF of  $x \sim N(0, 1)$

$$\begin{array}{ccccc} \text{height} & \cdot & \text{width} & = & \text{probability} \\ \downarrow & & \downarrow & & \downarrow \\ f(x) & \cdot & d\lambda(x) & = & d\mu(x) \end{array}$$

$$f_{\omega}(x) = \frac{d\mu(x)}{d\lambda(x)} \frac{d\lambda(x)}{d\omega(x)}$$

The PDF of  $y = \sin(x)$



# Mathematical Tricks and Formulae

In special cases, e.g., when transforming discrete probability distributions or when working with continuous distributions and differentiable transforms, we can write formula that can be used to calculate the relationships between densities when transforming random variables.

## The Change of Variable Formula

Let  $x$  be a continuous random variable with density  $f(x)$ . Let  $r$  be a strictly increasing or strictly decreasing function<sup>a</sup> such that  $x = r^{-1}(y)$ . Then the probability density function  $g$  of  $y$  is given by

$$g(y) = f(x) \left| \frac{dx}{dy} \right|.$$

<sup>a</sup>on the set  $S$  over which  $x$  is defined

$$g(y) = r(f(x))$$

$$\begin{aligned} \rightarrow y &= r(x) \\ \rightarrow x &= r^{-1}(y) \end{aligned}$$

# An Example Problem

$$x \sim N(0, 1) \quad \checkmark$$

$$y = \sin(x)$$

Whats the density of  $y$ ?

$\sin(x)$  is not strictly increasing or decreasing. So, this becomes difficult... You can go to basic theory and spend a while trying to solve the problem explicitly, or you can use a computer to approximate just about any answer you might need.

# Computational Transformation of Random Variables

$$x \sim N(0, 1)$$

$$\underline{y} = \sin(x)$$

Whats the density of  $y$ ?

Lets go back to basics. This question is really asking: *If I draw  $(x_1, \dots, x_N)$  iid from  $x \sim N(\mu, \Sigma)$ , and then transform each  $x_i$  ( $y_i = \sin(x_i)$ ), then what is the distribution of the sample  $(y_1, \dots, y_N)$ ?*

So why don't we just sample  $(x_1, \dots, x_N)$  and transform them all? And then just plot the results?

## Section 3

### Density Estimation Methods

# Histograms vs. Kernel Density Estimates

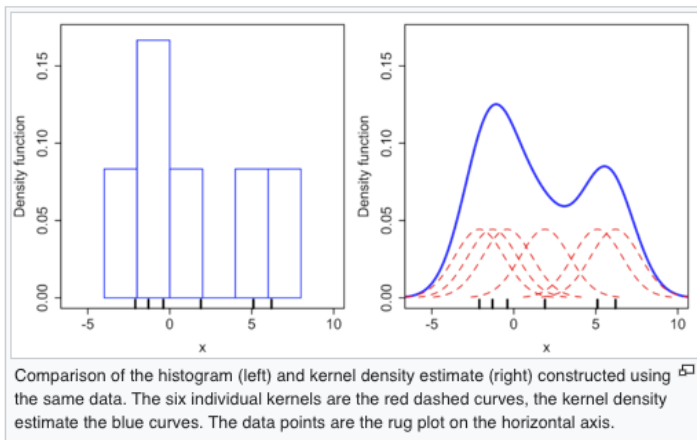


Image from: [https://en.wikipedia.org/wiki/Kernel\\_density\\_estimation](https://en.wikipedia.org/wiki/Kernel_density_estimation)

# Histograms

You should know how to make a histogram.

# Kernel Density Estimates (KDE)

Let  $(x_1, x_2, \dots, x_n)$  be *iid* samples drawn from some distribution with a univariate density  $f$  at a given point  $x$ . We are interested in estimating the shape of  $f$ .

It's Kernel Density Estimator is given by

$$\hat{f}_h(x) = \frac{1}{n} \sum_{i=1}^n K_h(x, x_i) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right)$$

## Section 4

# Monte Carlo Integration



# Monte Carlo Integration

There is a lot of theory we can go into here. I am going to cut through that and just give you what you need to know – the result is fairly intuitive if you think hard enough about it.

## Monte Carlo Approximation to Expectations.

For some random variable  $x$  that is described by a probability density  $p(x)$ , we wish to calculate the integral:

$$E[f(x)] = \int f(x) \underline{p(x)} dx.$$

We can approximate this integral using the following two steps:

- 1 Sample  $x_1, \dots, x_N \sim p(x)$
- 2 Calculate  $E[f(x)] \approx \frac{1}{N} \sum_i f(x_i)$  ←

As  $N \rightarrow \infty$  this approximation converges to the true expectation.

## Section 5

### Example

## Example Transformation

$$x \sim N(0, 1)$$

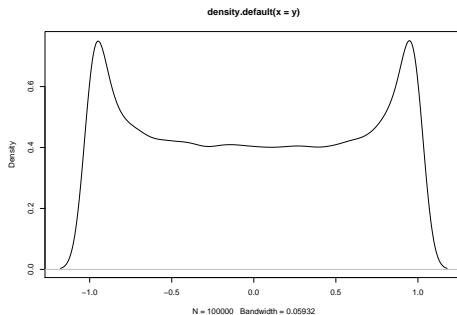
$$y = \sin(x)$$

# Plot the PDF of $y$

$$x \sim N(0, 1)$$

$$y = \sin(x)$$

```
# Three lines of code  
x <- rnorm(100000, mean = 0, sd = 1) # sample  
y <- sin(x)  
plot(density(y)) # Plot Kernel Density Estimate
```



## Calculate $E[y]$

$$x \sim N(0, 1)$$

$$y = \sin(x)$$

```
# Still, three lines of code  
x <- rnorm(100000, mean = 0, sd = 1) # sample  
y <- sin(x)  
mean(y) ←
```

```
## [1] 8.900217e-05 ←
```

This approximates the true answer of zero. The true answer is evident from the symmetry about zero of the density function on the prior slide.

# What is not correct

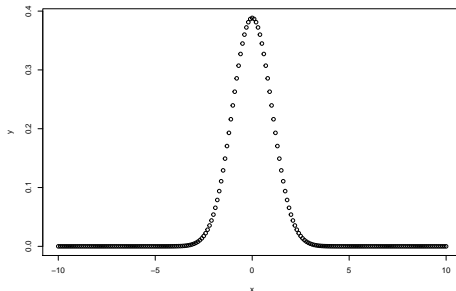
$$x \sim N(0, 1)$$

$$y = \sin(x)$$

$$y = x + 1$$

~~$g(y) = f(x)$~~   $\swarrow$

```
x <- seq(-10, 10, by=0.1)
p_x <- dnorm(x, mean = 0, sd = 1) # calculate the density function over a grid of x
y <- sin(p_x)
plot(x, y)
```



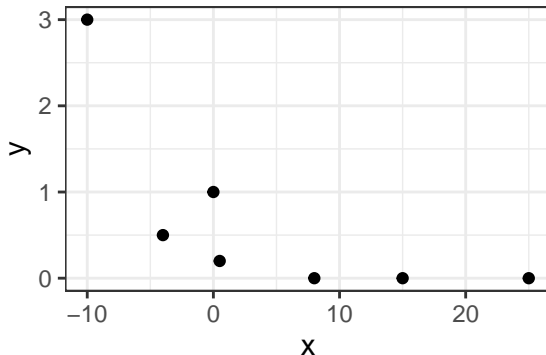
This has no meaning that I can think of. If the transform was  $y = x + 1$  then its easy to see that the resultant density would not even integrate to 1.

## Section 6

### Application - Transforming Dependent Variables in Regression

## A simple example: Regression with Positive Valued data.

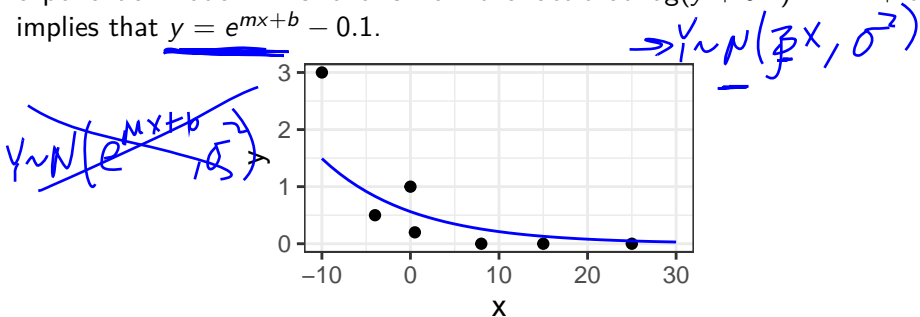
Joe is trying to model the amount of snow in his backyard as a function of the mid-day temperature. Let  $y_i$  denote a measurement of the amount of snow at one spot in his backyard (measured in inches) and let  $x_i$  denote the mid-day temperature on the day measurement (in celsius)  $y_i$  was taken. Here is the data Joe collects:





## A simple example: Regression with Positive Valued data.

Note the log-transformed model can instead be viewed as an equivalent exponential model. This follows from the fact that  $\log(y + 0.1) = mx + b$  implies that  $y = e^{mx+b} - 0.1$ .

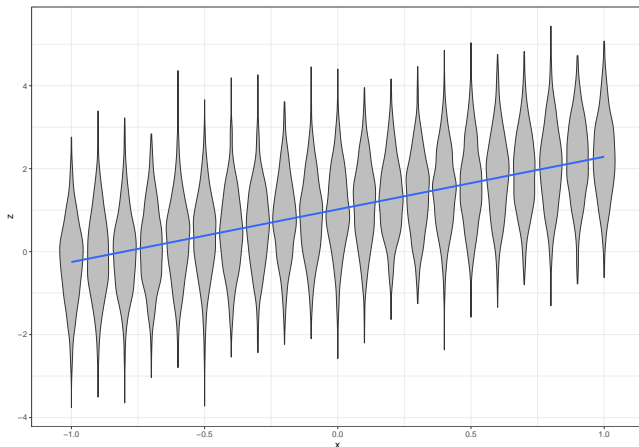


### Moral of the Story

The lesson here is that data has meaning. You need to consider that meaning when picking a representation / encoding. Also different encodings imply different models.

# What really happened here?

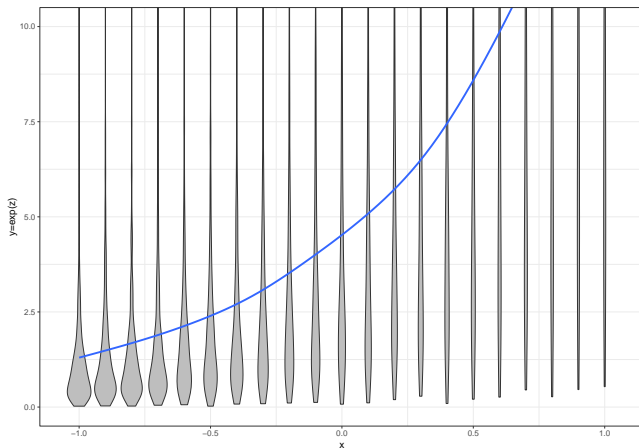
$$\log(y_i) = \underline{z_i \sim N(1 + 1.3x_i, 1)}$$



# What really happened here?

$$\log(y_i) = z_i \sim N(1 + 1.3x_i, 1)$$

$$y_i \sim \text{LogNormal}(1 + 1.3x_i, 1)$$



## Section 7

Want to learn more measure theory?

# Want to learn more measure theory?

- At a similar level to this lecture
- Fuller Picture (good book to learn from)
- A cool paper applied data analysis technique that could not happen without using measure theory explicitly

•