5/9/19 Lecture 12

Multivariate distributions

So far we've looked at one and then two rvs at a time; easy to generalize to a finite # of rv $Y_1, \dots, Y_n$ when n is a finite positive integer

<u>Def:</u> The joint CDF of n rvs $Y_1, \dots, Y_n$ is the function $F_{Y_1 \cdots Y_n}(y_1, \dots, y_n)$ specified by

$$F_{Y_1, \dots, Y_n}(y_1, \dots, y_n) = P(\underbrace{Y_1 \le y_1 \, , \, \dots, Y_n \le y_n}_{\text{, = and}})$$

More compact to use vector notation:

$$\underset{\sim}{Y} = (Y_1, \dots, Y_n), \quad \underset{\sim}{y} = (y_1, \dots, y_n)$$

$$F_{\underset{\sim}{Y}}(\underset{\sim}{y}) = P(Y_1 \le y_1, \dots, Y_n \le y_n)$$

$\underset{\sim}{Y}$ is said to be a random vector taking values in $\mathbb{R}^n$

<u>Def:</u> n rv $(Y_1, \dots, Y_n) = \underset{\sim}{Y}$ have a discrete joint distribution if the random vector $\underset{\sim}{Y}$ can only take on a finite or countably infinite # of possible values $(y_1, \dots, y_n) \in \mathbb{R}^n$. The joint pmf (probability mass function) of $\underset{\sim}{Y}$ is $f_{Y_1 \cdots Y_n}(y_1, \dots, y_n) = P(Y_1 = y_1, \dots, Y_n = y_n)$

or equivalently $f_{\underset{\sim}{Y}}(\underset{\sim}{y}) = P(\underset{\sim}{Y} = \underset{\sim}{y})$

<u>Example:</u> n patients in treatment group of a randomized clinical trial; $B_i = \begin{cases} 1 & \text{if patient } i \text{ has a good outcome} \\ 0 & \text{else} \end{cases}$

If nothing else is known about the patients (e.g. age, disease burden at start of trial, ...) it would be reasonable to model the B_i as IID Bernoulli(θ) same success probability

Good outcome $\leftrightarrow$ remission

$(Y_i \mid \theta_i) \sim \text{Bernoulli}(\theta_i)$: Ideal model

$(i = 1, \ldots, n)$

but can't be fit because there are too many unknowns

We have n unknown $(\theta_1, \ldots, \theta_n)$

n 1/0 indicators $(B_1, \ldots, B_n)$

$\underset{\sim}{B} = (B_1, \ldots, B_n)$; $\underset{\sim}{b} = (b_1, \ldots, b_n)$; $\underset{\sim}{B}$ has a discrete joint distribution $f_{\underset{\sim}{B}}(\underset{\sim}{b}) = P(B_1 = b_1, \ldots, B_n = b_n)$

1 or 0

If θ were known you could use $f_{\underset{\sim}{B}}(\underset{\sim}{b})$ to predict the dataset before it arrives!

by the IID assumption $P(B_1 = b_1, \ldots, B_n = b_n \mid \theta)$

$$= P(B_1 = b_1 \mid \theta) \cdots P(B_n = b_n \mid \theta)$$

Recall that

$P(B_i = b_i \mid \theta) = \theta^{b_i}(1-\theta)^{1-b_i}$ for $b_i = 0, 1$

so $f_{\underset{\sim}{B}}(\underset{\sim}{b} \mid \theta) = \prod_{i=1}^{n} \theta^{b_i}(1-\theta)^{1-b_i} = \theta^{\sum_{i=1}^{n} b_i}(1-\theta)^{n - \sum_{i=1}^{n} b_i}$

$= \theta^{s}(1-\theta)^{n-s}$ with $\sum_{i=1}^{n} b_i$

Def: n rv $Y_1, \ldots, Y_n$ have a continuous joint distribution if you can find a function $f_{\underset{\sim}{Y}}$ on $\mathbb{R}^n$ s.t. for every (non-weird) subset $C \subset \mathbb{R}^n$

$$P[(Y_1, \ldots, Y_n) \in C] = \int_C \cdots \int f_{Y_1 \cdots Y_n}(y_1, \ldots, y_n)\, dy_1 \cdots dy_n$$

$f_{\underset{\sim}{Y}}(\underset{\sim}{y})$ is the joint CDF (probability density function) of $\underset{\sim}{Y}$

$$\boxed{\text{More compactly } P(\underset{\sim}{Y} \in C) = \int_C \cdots \int f_{\underset{\sim}{Y}}(\underset{\sim}{y})\, dy}$$

Consequences of this def:

1) If the joint dist. of $\underset{\sim}{Y}$ is continuous then

$$f_{\underset{\sim}{Y}}(\underset{\sim}{y}) = \frac{\partial^n}{\partial y_1 \cdots \partial y_n} \cdot F_{\underset{\sim}{Y}}(\underset{\sim}{y})$$

Mixed discrete/continuous random vectors behave with n rv just as they do with 2 rv

Ex: Clinical trial continued

More realistically, θ would be unknown, and you can think about the joint distribution of $(\underset{\sim}{B}, \theta) = (B_1, \ldots, B_n, \theta)$, in which the B_i are discrete and $0 < \theta < 1$ is continuous

Marginal distributions

If you know the joint PDF $f_{\underset{\sim}{Y}}$ of $\underset{\sim}{Y}$, you can work out the marginal distribution of any subset of $(Y_1, \ldots, Y_n)$ by integrating $f_{\underset{\sim}{Y}}(\underset{\sim}{y})$ over the elements of $(Y_1, \ldots, Y_n)$ that are not in the subset

Ex: $\underset{\sim}{Y} = (Y_1, Y_2, Y_3, Y_4)$

$$f_{Y_1}(y_1) = \iiint f_{\underset{\sim}{Y}}(\underset{\sim}{y})\, dy_2\, dy_3\, dy_4$$

$$f_{Y_2, Y_3}(y_2, y_3) = \iint f_{\underset{\sim}{Y}}(\underset{\sim}{y})\, dy_1\, dy_4$$ and so on

Similarly, you can work out a marginal CDF by sending the othe components to ∞

Ex: $F_{Y_1}(y_1) = P(Y_1 \le y_1) = P(Y_1 \le y_1, Y_2 < \infty, \dots, Y_n < \infty)$

$$= \lim_{y_2 \to \infty, \dots y_n \to \infty} F_{\underset{\sim}{Y}}(\underset{\sim}{y})$$

Def: n rvs $Y_1, \dots, Y_n$ are independent if for any non-weird sets $A_1, \dots, A_n$ of real #s

$$P(Y_1 \in A_1, \dots, Y_n \in A_n) = \prod_{i=1}^{n} P(Y_i \in A_i)$$

Immediate Consequences:

1) $Y_1, \dots, Y_n$ independent iff $F_{\underset{\sim}{Y}}(\underset{\sim}{y}) = \prod_{i=1}^{n} F_{Y_i}(y_i)$

2) $Y_1, \dots Y_n$ independent iff $f_{\underset{\sim}{Y}}(\underset{\sim}{y}) = \prod_{i=1}^{n} f_{Y_i}(y_i)$

Def: Starting with a univariate pmf or pdf $f_{Y_i}(y_i)$, n rvs $(Y_1, \dots, Y_n)$ form a random sample of size n from f_{Y_i} if the Y_i are independent and all of them have marginal pmf or pdf f_{Y_i} $\to$ ie if the Y_i are an independent identically distributed (IID) sample from f_{Y_i}.

Ex: Deer at UCSC: Some have a disease (chronic wasting disease)

Population	Sample
all deer living within UCSC boundaries on May 9 2019	The observed deer

disease? $\quad\quad\quad\quad\quad\quad$ disease? $\quad\quad\quad\quad\quad$ $1 = Y \quad 0 = N$

$$N \approx 800 \updownarrow \begin{bmatrix} 1\text{s} \\ \text{and} \\ 0\text{s} \end{bmatrix} \xrightarrow{\text{IID}} \begin{array}{c} 1 \\ 2 \\ \vdots \\ n \end{array} \begin{bmatrix} 1\text{s} \\ \text{and} \\ 0\text{s} \end{bmatrix} \updownarrow n = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

mean $= \theta = ?$ (unknown) $\quad\quad$ mean $\bar{y} = \hat{\theta}$ — estimate of

Shorthand for the diagram: $(Y_i \mid \theta) \stackrel{\text{IID}}{\sim} \text{Bernoulli}(\theta)$
$(i = 1, \ldots, n)$

Def: Start with random vector $\underset{\sim}{X} = (X_1, \ldots, X_n)$; partition it into 2 subvectors $\underset{\sim}{X} = (\underset{\sim}{Y}, \underset{\sim}{Z})$, $\underset{\sim}{Y} = (Y_1, \ldots, Y_k)$
$\underset{\sim}{Z} = (Z_1, \ldots, Z_{n-k})$ $\quad\quad 1 \le k \le n-1$

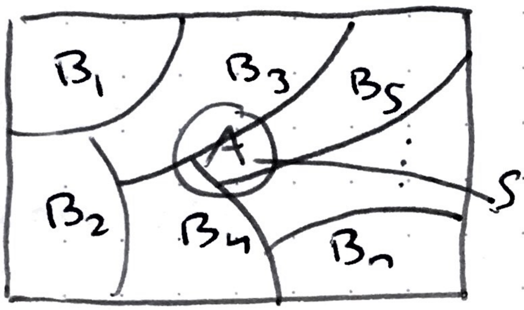
Then for every point $\underset{\sim}{z}$ for which $f_{\underset{\sim}{Z}}(\underset{\sim}{z}) > 0$, the conditional distribution of $\underset{\sim}{Y}$ given $\underset{\sim}{Z}$ is

$$f_{\underset{\sim}{Y} \mid \underset{\sim}{Z}}(\underset{\sim}{y} \mid \underset{\sim}{z}) = \frac{f_{\underset{\sim}{Y}\underset{\sim}{Z}}(\underset{\sim}{y}, \underset{\sim}{z})}{f_{\underset{\sim}{Z}}(\underset{\sim}{z})}, \quad \underset{\sim}{y} \in \mathbb{R}^k$$

from which $f_{\underset{\sim}{Y}\underset{\sim}{Z}}(\underset{\sim}{y}, \underset{\sim}{z}) = f_{\underset{\sim}{Z}}(\underset{\sim}{z})\, f_{\underset{\sim}{Y} \mid \underset{\sim}{Z}}(\underset{\sim}{y} \mid \underset{\sim}{z})$

Multivariate Law of Total Probability

If A is an event and you're trying to compute $P(A)$ and it's hard, one idea is to find another aspect of the world B upon which A depends, such that the events $B_1, \ldots, B_n$ form a partition; then



$$P(A) = \sum_{i=1}^{n} P(A \cap B_i)$$

$$= \sum_{i=1}^{n} P(B_i) P(A \mid B_i)$$

This has an analogue with continuous rvs (using the notation in the definition of conditional distributions):

$$\underbrace{f_{\mathbf{Y}}(\mathbf{y})}_{\text{like } A} = \int \cdots \int_{\mathbb{R}^{n-k}} \underbrace{f_{\mathbf{Z}}(\mathbf{z})}_{\text{like } B_i} \underbrace{f_{\mathbf{Y}|\mathbf{Z}}(\mathbf{y} \mid \mathbf{z})}_{\text{like } P(A \mid B_i)}$$

Multivariate Bayes' Theorem

Using the same notation, $\left(\begin{matrix}\text{posterior}\\ \text{info}\end{matrix}\right) = \dfrac{\left(\begin{matrix}\text{prior}\\ \text{info}\end{matrix}\right)\left(\begin{matrix}\text{likelihood}\\ \text{info}\end{matrix}\right)}{(\text{normalizing constant})}$

$$f_{\mathbf{Z}|\mathbf{Y}}(\underbrace{\mathbf{z}}_{\text{unknown}} \mid \underbrace{\mathbf{y}}_{\text{data}}) = \frac{f_{\mathbf{Z}}(\mathbf{z})\, f_{\mathbf{Y}|\mathbf{Z}}(\mathbf{y} \mid \mathbf{z})}{f_{\mathbf{Y}}(\mathbf{y})}$$

The usual application of this in statistics is as follows.

<u>Def:</u> Z_1 a random vector with multivariate distribution $f_{Z_1}(\underset{\sim}{z})$; then random variables

$X_1, \ldots, X_n$ are conditionally independent given Z_1 if for all $\underset{\sim}{z}$ with $f_{Z_1}(\underset{\sim}{z}) > 0$,

$$f_{X \mid Z_1}(\underset{\sim}{x} \mid \underset{\sim}{z}) = \prod_{i=1}^{n} f_{X_i \mid Z_1}(x_i \mid \underset{\sim}{z})$$

<u>Revisit ex:</u> Remember the machine with a θ dial that can make IID coin tosses with $P(\text{heads}) = \theta$?

We agreed that if θ is unknown to you

1) the results of the coin tosses $Y_1, Y_2, \ldots$ are dependent, because there is useful information in any subset of them for predicting any other subset, but...

2) the Y_i become conditionally independent given θ, because there's no longer any useful information in the Y_i to predict other Y_i

This is why in both the clinical trial example and the nuts & bolts example we model the data values Y_i as $(Y_i \mid \theta) \underset{\sim}{\text{conditionally IID}}$ Bernoulli(θ)

<u>Functions of a rv</u>

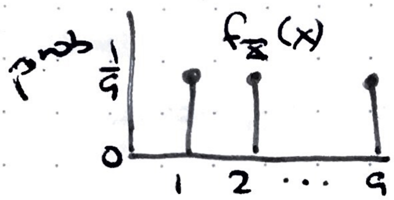
Case 1: discrete – X discrete rv with pmf $= f_X(x)$ ← univariate

$Y = h(x)$ for some function h defined on {possible values of X}

Then $f_Y(y) = P(Y=y) = P[h(X)=y] = \sum_{\{x : h(x)=y\}} f_X(x)$

PMF spike plot with support of integers 1–9

$\boxed{\text{Ex: } X \sim \overset{\text{Discrete}}{\text{Uniform}} \{1, 2, \ldots, 9\}}$



At the same height because it's uniform

Median = 5

$Y = |X - 5| = h(X)$: keeps track of how far X is from the median

y	X such that $Y=y$	$P(Y=y)$
0	5	1/9
1	4 or 6	2/9
2	3 or 7	2/9
3	2 or 8	2/9
4	1 or 5	2/9
		1

Case 2: Continous: X continous rv with PDF $f_X(x)$

$Y = h(X)$ as before

The CDF can be worked out as follows:

$$F_Y(y) = P(Y \leq y) = P[h(X) \leq y] = \int_{\{x : h(x) \leq y\}} f_X(x)\,dx$$

and if Y is also continous $f_Y(y) = \frac{d}{dy} F_Y(y)$
(at every point y where F_Y is differentiable)

Example: X = rate at which customers served in a queue at the bank

Natural to model X as continuous with CDF F_X and rates $X > 0$

Turns out that the average waiting time is $Y = \frac{1}{X} = h(x)$

You can get the PDF of Y in 2 steps:

1) Work out CDF of Y
2) Differentiate with respect to y

1) (for $y > 0$)

$$F_Y(y) = P(Y \leq y) = P(h(X) \leq y)$$

$$= P\left(\frac{1}{X} \leq y\right) = P\left(X \geq \frac{1}{y}\right)$$

since X is continuous

$$= 1 - P\left(X < \frac{1}{y}\right) = 1 - P\left(X \leq \frac{1}{y}\right)$$

$$= 1 - F_X\left(\frac{1}{y}\right)$$

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{d}{dy}\left[1 - F_X\left(\frac{1}{y}\right)\right]$$

chain rule

$$= -f_X\left(\frac{1}{y}\right)(-y^{-2}) = \frac{f_X\left(\frac{1}{y}\right)}{y^2}$$

Ex: $X \sim \text{Uniform}[-1, 1]$ (continuous)

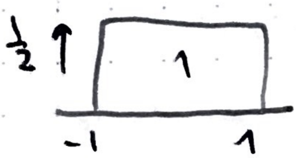
$Y = X^2$ Find the PDF of Y

Take the support for X and see how it transforms through the function $h(X)$ into the support for Y

Note that Y's possible values are $[0, 1]$

1) for $0 < y < 1$ $\quad F_Y(y) = P(Y \le y) = P(X^2 \le y)$

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



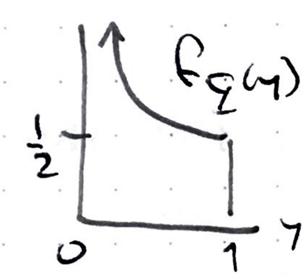
$$= \int_{-\sqrt{y}}^{\sqrt{y}} f_X(x)\,dx$$

$$f_X(x) = \begin{cases} \frac{1}{2} & \text{for } -1 \le x \le 1 \\ 0 & \text{else} \end{cases}$$

$$= \frac{1}{2} \times \Big|_{-\sqrt{y}}^{\sqrt{y}} = \sqrt{y}$$

2) Thus $f_Y(y) = \frac{d}{dy} F_Y(y)$

$$\text{so } f_Y(y) = \begin{cases} \frac{d}{dy} y^{1/2} = \frac{1}{2\sqrt{y}} & \text{for } 0 < y < 1 \\ 0 & \text{else} \end{cases}$$



This density is unbounded at 0

Easy theorem

X continuous rv with pdf $f_X(x)$,

$Y = aX + b$ $(a \neq 0)$ linear transformation

$$\Rightarrow f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right)$$

<u>Interesting & useful fact</u>: X continuous with CDF $F_X(x)$. What's the distribution of $Y = F_X(X)$?

$$F_Y(y) = P(Y \le y) = P(F_X(X) \le y) = P(X \le F_X^{-1}(y))$$

$$= F_X(F_X^{-1}(y)) = y \quad \text{for } 0 < y < 1$$

But the dist. with $F_Y(y) = y$ for $0 < y < 1$ is the Uniform $(0,1)$ distribution

<u>Probability Integral Transform</u>

X continuous with CDF F_X, $Y = F_X(X)$

$\rightarrow Y \sim \text{Uniform}(0,1)$

<u>Why is this useful?</u>

Converse is also true: $Y \sim \text{Uniform}(0,1)$

F_X continuous CDF with quantile function

$F_X^{-1} \rightarrow X = F_X^{-1}(Y) \sim F_X$

This is the practical basis for the generation of many forms of pseudo-random numbers:

It turns out to be easy to generate pseudo-Uniform(0,1) values; therefore if you want to generate pseudo-random #s from a distribution with CDF F_X and F_X^{-1} is easy and fast to compute

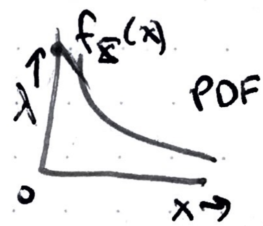
Algorithm $u_1, \dots, u_n \overset{IID}{\sim}$ uniform $(0,1)$

$F_{\underline{X}}^{-1}(u_1), \dots, F_{\underline{X}}^{-1}(u_n) \overset{IID}{\sim} F_{\underline{X}}$ (Quiz 6)

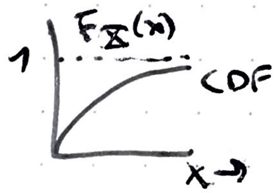
Earlier example revisted

If $\underline{X} \sim$ Exponential $(\lambda \mid \lambda > 0)$, its PDF is

$$f_{\underline{X}}(x) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x > 0 \\ 0 & \text{else} \end{cases}$$



$$F_{\underline{X}}(x) = \begin{cases} 0 & x \le 0 \\ 1 - e^{-\lambda x} & x > 0 \end{cases}$$



$$F_{\underline{X}}^{-1}(p) = \frac{-\log(1-p)}{\lambda} \quad (0 < p < 1$$

Now you can see immediately that if $U \sim$ Uniform$(0,1)$ so is $(1-U)$ so to generate IID Exponential λ rv you just compute $-\frac{1}{\lambda} \log u_i$, $u_i \overset{IID}{\sim}$ Uniform$(0,1)$

Why do people want/need pseudorandom #s?

Some stochastic (probabilistically) models of real world phenomena are too complicated to fully characterize mathematically in closed form; one highly useful method in such situations is to conduct a computer-based simulation study driven by pseudo-random #s.

The method used above for working out the distribution of $\underline{Y} = \frac{1}{\underline{X}}$ can be generalized

Some functions $h(x)$ are nice, in that they are both differentiable and one-to-one (invertible)

Calculus Reminder

If real-valued $h(x)$ is differentiable and one-to-one (1-1) for x in the open interval (a,b) then h is either monotonically increasing or decreasing, and h is also continuous, so it transforms the interval (a,b) to another open interval $h[(a,b)] = (\alpha, \beta)$ called the image of (a,b) under h. Since h is invertible, it makes sense to talk about $y = h(x) \leftrightarrow x = h^{-1}(y)$

Theorem: X continuous rv with PDF $f_X(x)$ and for which $P(a < X < b) = 1$; $Y = h(X)$, with h differentiable and 1-1 for $a < x < b$

could be infinite

(α, β) image of (a,b) under h; $h^{-1}(y)$ inverse function of $h(x)$ for $\alpha < y < \beta$ $\rightarrow$ PDF of Y is

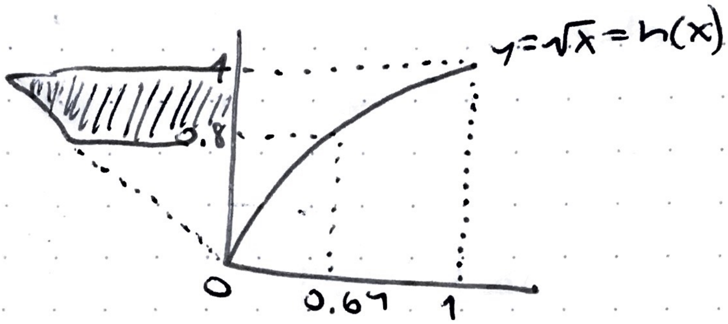
$$f_Y(y) = \begin{cases} f_X[h^{-1}(y)] \left| \dfrac{dh^{-1}(y)}{dy} \right| & \text{for } \alpha < y < \beta \\ 0 & \text{else} \end{cases}$$

$$f_Y(y) = f_X[h^{-1}(y)] \cdot \left| \frac{d}{dy} h^{-1}(y) \right| = f_X(x) \left| \frac{d}{dy} x \right|$$

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

$\mathbb{X} \sim \text{Uniform}(0,1)$ (continuous)

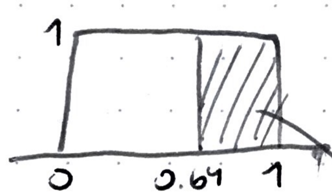
$\mathbb{Y} \sim \mathbb{X} \quad y = h(x) = \sqrt{x}$



$x = h^{-1}(y) = y^2$

$\frac{d}{dy} h^{-1}(y) = |2y|$

$= 2y \uparrow$



$f(x) = \begin{cases} 1 & \text{for } 0 < x < 1 \\ 0 & \text{else} \end{cases}$

this has to be compressed in the region between 0.8 and 1 in the first graph

$f_{\mathbb{Y}}(y) = \begin{cases} \overset{1}{\cancel{f_{\mathbb{X}}(x)}} \cdot 2y & \text{for } 0 < y < 1 \\ 0 & \text{else} \end{cases}$

$\boxed{\begin{array}{c} y = h(x) \\ x = h^{-1}(y) \end{array}}$

Easy short-hand: "Multiply both sides by $|dy|$ to get $f_{\mathbb{Y}}(y)\,|dy| = f_{\mathbb{X}}(x)\,|dx|$

<u>Earlier ex. revisited</u>

$\mathbb{Y} = h(\mathbb{X}) = \frac{1}{\mathbb{X}}$: average waiting time in the bank queue

Here $y = h(x) = \frac{1}{x}$ so $x = h^{-1}(y) = \frac{1}{y}$ and

$\frac{d}{dy}\frac{1}{y} = -\frac{1}{y^2}$; thus $f_{\mathbb{Y}}(y) = \dfrac{f_{\mathbb{X}}\left(\frac{1}{y}\right)}{y^2}$ as before

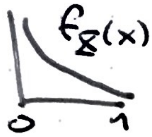
Example

At time 0, population of ν organisms introduced into large tank of water w/ nutrients

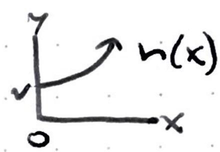
$\mathbb{X}$ = rate of growth

Under one model that's realistic in some circumstances, at time t the predicted population size would be $\mathbb{Y} = \nu e^{\mathbb{X}t}$ (exponential growth)

$\mathbb{X}$ unknown, modeled with $f_{\mathbb{X}}(x) = \begin{cases} 3(1-x)^2 & 0<x<1 \\ 0 & \text{else} \end{cases}$



$y = h(x) = \nu e^{xt}$



strictly increasing

$$\left.\begin{array}{l} x=0 \to y=\nu \\ x=1 \to y=\nu e^{t} \end{array}\right\} \text{image}$$

$$\frac{y}{\nu} = e^{xt} \to \log\left(\frac{y}{\nu}\right) = xt \to x = h^{-1}(y) = \frac{1}{t}\log\left(\frac{y}{\nu}\right)$$

$$\frac{d}{dy}\,\frac{1}{t}\log\left(\frac{y}{\nu}\right) = \frac{1}{t}\left(\frac{y}{\nu}\right)^{-1}\cdot\frac{1}{\nu} = \frac{1}{ty}$$

$$f_{\mathbb{Y}}(y) = \begin{cases} \dfrac{3\left[1-\frac{1}{t}\log\left(\frac{y}{\nu}\right)\right]^2}{ty} & \nu<y<\nu e^{t} \\ 0 & \text{else} \end{cases}$$

Functions on 2 or more rvs

Case 1: discrete : n rvs $\mathbb{X}_1, \dots, \mathbb{X}_n$ discrete joint dist with joint pmf $f_{\underset{\sim}{\mathbb{X}}}(\underset{\sim}{x})$;

$$\text{define}\left\{\begin{array}{l} \mathbb{Y}_1 = h_1(\mathbb{X}_1, \dots, \mathbb{X}_n) \\ \vdots \\ \mathbb{Y}_m = h_m(\mathbb{X}_1, \dots, \mathbb{X}_n) \end{array}\right\} (m \geq 1)$$

real-valued

$(h_j : \mathbb{R}^n \to \mathbb{R})$