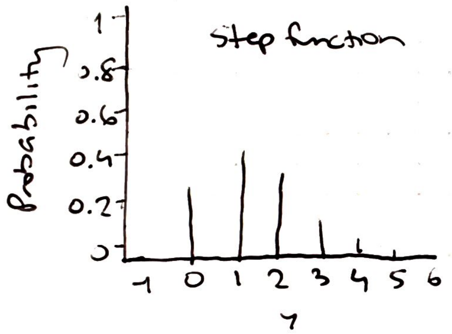
5/2/19

Lecture 10

Email will be sent out later today about office hours for Friday, Saturday, and Sunday

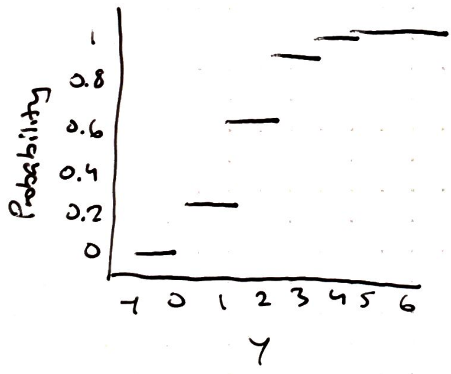
You must know how to do pmf, pdf, cdf, and inverse cdf

Binomial PMF w/ n=5 and p=0.25

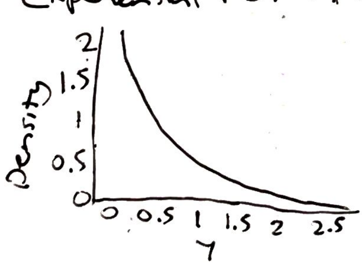


Important distinction between pmfs + pdfs: PMFs for discrete rvs have probability on the vertical scale but with PDFs for continuous rvs it's looking at density values which are essentially probability per tiny interval on the horizontal axis.

Binomial CDF w/ n=5 and p=0.25

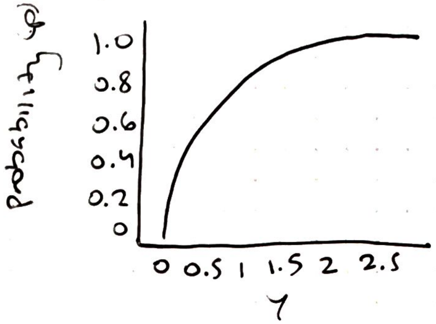


Exponential PDF w/ lambda=2



Notice that the vertical scale is density, which represents the amount of concentration of probability at or near a particular point of interest to you.

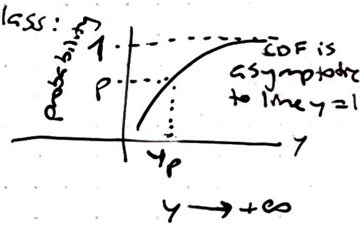
Exponential CDF with lambda = 2

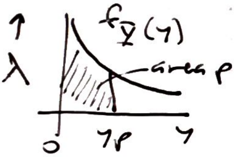


Because the pdf is continuous, the CDF will be continuous also.

Looking back at CDF from last class:

$$F_{\mathbb{Y}}(y) = \begin{cases} 0 & \text{for } y \le 0 \\ 1 - e^{-\lambda y} & y > 0 \end{cases}$$





Q: What's the place y_p on the positive part of $\mathbb{R}$ where $P(0 \le \mathbb{Y} \le y_p) = p$?

For $y_p > 0$, $P(0 \le \mathbb{Y} \le y_p) = \boxed{F_{\mathbb{Y}}(y_p) = p}$

$= 1 - e^{-\lambda y_p} = p$ so $\boxed{y_p = F^{-1}_{\mathbb{Y}}(p)}$

$1 - p = e^{-\lambda y_p}$

$\log(1-p) = -\lambda y_p$

$$y_p = -\frac{\log(1-p)}{\lambda}$$

Def: y_p is called the pth quantile or the 100th pth percentile of (the distribution of) $\mathbb{Y}$

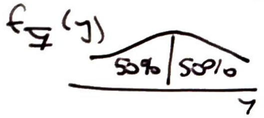
Some care is required when Y is discrete or mixed

<u>Def:</u> Y rv with CDF $F_Y(y)$

For all $0<p<1$ define $F_Y^{-1}(p)$ as the smallest y value such that $F_Y(y) \geq p$

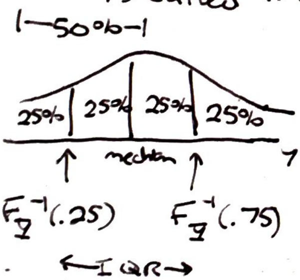
Then $F_Y^{-1}(p)$ is the pth quantile of Y and F_Y^{-1} is the quantile function

Measures of <u>center</u> for the distribution of a rv Y



One way to define the center of a distribution is to find the 50th percentile.

<u>Def:</u> The $\frac{1}{2}$ quantile AKA the 50th percentile of a distribution is called the median of the distribution



Measures of spread for the distribution of a rv Y

One way to define the spread of a dist. is to see how far apart its 75th and 25th percentiles are

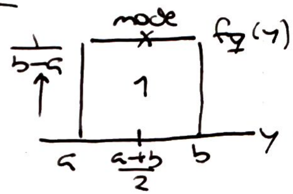
Def: the $\frac{1}{4}$ quantile $= F_Y^{-1}(0.25)$

the 25th percentile is the lower quartile

the $\frac{3}{4}$ quantile $F_Y^{-1}(.75)$ = the 75th percentile is the upper quartile

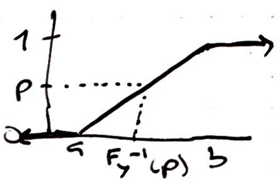
$F_Y^{-1}(.75) - F_Y^{-1}(0.25)$ = interquartile range (IQR)

Ex: $Y \sim \text{Uniform}(a,b)$



$$F_Y(y) = \begin{cases} 0 & \text{for } y \le a \\ \frac{y-a}{b-a} & a \le y \le b \\ 1 & y \ge b \end{cases}$$

CDF of uniform distribution



Easy to invert F_Y: $F_Y^{-1}(p) = (1-p)a + pb$ for $0 \le p < 1$

and the median is $(a+b)/2$

Studying two rvs at a time

Def: X Y rvs: the joint (or bivariate) distribution of (X, Y) is the collection $P[(X,Y) \in C]$ of all probabilities for all sets $C \in \mathbb{R}^2$ such that $(X,Y) \in C$ isn't weird

(Probability that the point (X, Y) is in the set on the two dimensional plane)

Case 1: (X and Y both discrete)

Def: X, Y rv If there are only finitely or countably infinitely many possible values (x, y) for (X, Y), X and Y have a discrete joint distribution.

Def: The joint probability mass function (joint pmf) of (X, Y) discrete is the function

$$f_{X,Y}(x, y) = P(X=x, Y=y)$$

↑and

The set $\{(x,y): f_{XY}(x,y) > 0\}$ is the support of f_{XY}

Consequences

1) $\sum_{\text{all}(x,y)} f_{X,Y}(x,y) = 1$

2) For any (non-weird) set C of ordered pairs (x, y),

$$P((X,Y)\in C) = \sum_{(x,y)\in C} f_{X,Y}(x,y)$$

Case 2: X, Y both continuous

Def: Two rv X and Y have a continuous joint distribution if you can find a nonnegative function $f_{XY}(x,y)$ defined for all $(x,y) \in \mathbb{R}^2$ (the real plane) such that for every (non-weird) subset C of the plane $P[(X,Y)\in C] = \iint_C f_{XY}(x,y)\,dx\,dy$

$f_{XY}(x,y)$ is the joint pdf of (X, Y)

The set $\{(x,y) : f_{XY}(x,y) > 0\}$ is the support of (the dist.) of (X, Y)

Immediate Consequences

① For all (x,y) in $\mathbb{R}^2$

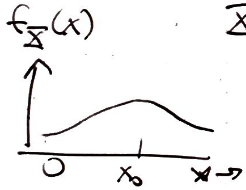
The bivariate pdf $f_{XY}(x,y) \geq 0$

and if you integrate the bivariate pdf over the entire real plane that has to integrate to 1

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx\,dy = 1$$

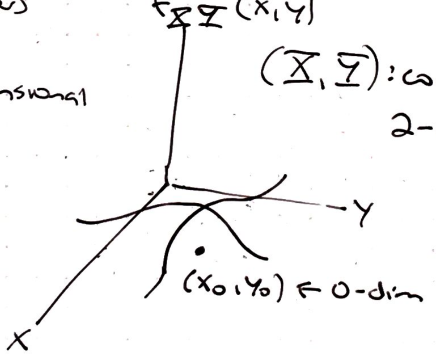
The link between probabilities and densities is that if you want to know how much probability there is in a region you integrate the joint density function over that region.

PDF of X



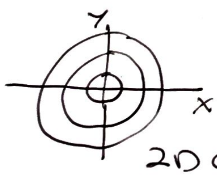
X: continuous
1 rv
1-dimensional

$P(X = x_0) = 0$
↑
0-dimensional



(X, Y): continuous
2-dim

3D perspective plot



2D contour plot of ↑

2) If $(\underline{X}, \underline{Y})$ have a continuous joint distribution, then $\underline{X}$ and $\underline{Y}$ each have a continuous univariate (one variable) marginal distribution when considered separately

3) For all continuous pdfs $f_{\underline{X}\underline{Y}}(x,y)$

(a) Every individual point, and every countably infinite sequence or set of points in $\mathbb{R}^2$ (0-dimensional) has probability 0 under $f_{\underline{X}\underline{Y}}$

(b) If g is a continuous function of one real variable defined on (a,b), then the sets

$\{(x,y): y = g(x), a < x < b\}$ and

$\{(x,y): x = g(y), a < x < b\}$ also have probability 0 (1-dimensional)

4) So (2) is not true: If $\underline{X}$ has a continuous distribution on $\mathbb{R} = \mathbb{R}^1$ and $\underline{Y} = \underline{X}$ then both $\underline{X}$ and $\underline{Y}$ have continuous distributions but $P\left[\begin{array}{l}(x,y) \text{ lies} \\ \text{on the line } y=x\end{array}\right] = \begin{array}{c}1? \\ 0?\end{array}$

So $(\underline{X}, \underline{Y})$ can't have a continuous joint distribution on $\mathbb{R}^2$

Joint distributions can lead to tricky integrals.

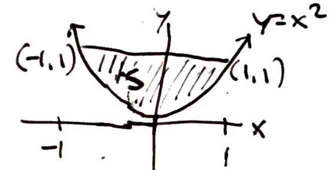
Bivariate densities are harder to manipulate than univariate densities

Ex: Suppose that $(\mathbb{X}, \mathbb{Y})$ have joint pdf

$$f_{\mathbb{X}\mathbb{Y}} = \begin{cases} cx^2y & \text{for } 0 \le x^2 \le y \le 1 \\ \infty & \text{else} \end{cases}$$

work out the normalizing constant

The support of $f_{\mathbb{X}\mathbb{Y}}$ is the shaded region (S)



$$\iint_S cx^2y \, dy \, dx$$

$$\int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} cx^2y \, dy \right) dx$$

$$\int_{-1}^{1} \left(\int_{x^2}^{1} cx^2y \, dy \right) dx = \frac{4}{21}c = 1$$

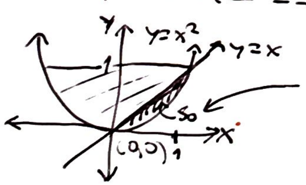
$$c = \frac{21}{4}$$

search the double integral calculator from wolfram alpha to solve

$$\int_0^1 \left[\int_{-\sqrt{y}}^{\sqrt{y}} x^2y \, dx \right] dy \qquad x^2 \le y \qquad -\sqrt{y} \le x \le \sqrt{y}$$

Comes out as $c = \frac{21}{4}$ too

Ex: Compute $P(\mathbb{X} \ge \mathbb{Y})$



The relevant part S_0 of S where $x \ge y$ is sketched here, so

$$P(\mathbb{X} \ge \mathbb{Y}) = \iint_{S_0} f_{\mathbb{X}\mathbb{Y}}(x, y) \, dy \, dx$$

$$P(X \geq Y) = \iint_{S_0} f_{XY}(x,y)\,dy\,dx$$

$$= \int_0^1 \left[\int_{x^2}^{x} \frac{21}{4} x^2 y \, dy \right] dx = \frac{3}{20}$$

You can have bivariate distributions in which one of (X, Y) is discrete and the other is continuous

<u>Case 3:</u> X: discrete Y: continuous

Suppose you can find a function $f_{XY}(x,y)$ defined on $\mathbb{R}^2$ such that for every pair of (non-weird) subsets A and B of $\mathbb{R}$ (assume integral exists). $P(X \in A \text{ and } Y \in B) = \int_B \sum_{x \in A} f_{XY}(x,y)\,dy$
Then f_{XY} is the joint pmf/pdf of (X, Y)

<u>Immediate Consequence:</u>
If X take on values $x_1, x_2, \ldots$ then

$$\int_{-\infty}^{\infty} \sum_{i=1}^{\infty} f_{XY}(x_i, y)\,dy = 1$$

<u>Ex:</u> Randomized controlled (clinical) trial
Patients in (T) get a treatment, patients in (C) get a placebo. Outcome is success (e.g. cancer goes into remission) or failure

$$\text{let } X_i = \begin{cases} 1 & \text{if patient } i \text{ in (T) is a success} \\ 0 & \text{else} \end{cases}$$

and let θ (unknown) be the proportion of patients in the population of all patients who might get the treatment who would have no relapse if they had been in the study. Then our uncertainty about Θ is continuous on interval $(0,1)$ and (X_i, Θ) has a mixed bivariate distribution

If you model $(X \mid \Theta)$ as Bernoulli(θ) and $\Theta \sim$ uniform$(0,1)$ then the joint pmf/pdf of (X, Θ) would be

$$f_{X,\Theta}(x,\theta) = \begin{cases} \theta^x (1-\theta)^{1-x} & \left(\begin{array}{l} \text{for } x = 0,1 \\ 0 < \theta < 1 \end{array} \right) \\ 0 & \text{else} \end{cases}$$

Then (e.g.) $P(X = 1) = P(X = 1$ and Θ is anything between 0 and 1$)$

$$= \int_0^1 \theta^1 (1-\theta)^{1-1} \, d\theta = \int_0^1 \theta \, d\theta = \frac{1}{2}$$

Bivariate cdfs

Def: The joint cdf of two rvs X and Y is the function $F_{XY}(x,y)$ satisfying $F_{XY} = P(X \le x \text{ and } Y \le y)$ for all $-\infty < x < \infty$ and $-\infty < y < \infty$