There are other foundational theories of probability – one by the Italian mathematician and actuary (deF) Bruno de Finetti (1906–1985), and another by the American physicists Richard T. Cox (1898–1991) and Edwin T. Jaynes (1922–1998) (CJ) – in which the probability function $P_{deF}(A \mid B)$ or $P_{CJ}(A \mid B)$ has 2 inputs, not 1, so that <u>conditional</u> probability is the primitive concept, not <u>unconditional</u> probability as with Kolmogorov's $P_K(A)$. deF and CJ

were responding to the reality that

in practice, all probabilities are conditional on background Assumptions, Information and Judgments (AIJ)

Example (Tay-Sachs)

we actually computed not

P(at least 1 T-S baby) but

P(at least 1 T-S baby | family of 5, and mother and father both carriers)

This impulse, to be explicit about your AIJ, is Bayesian; Kolmogorov worked in the frequentist paradigm; in this course, focusing on $P_K(B)$, we need to remember that it should really be $P_K(B|AIJ)$.

Consequences of the conditional probability definition (theorems)

① A, B events in $\mathbb{C}$:

if $P(B) > 0$ then

$$P(A \cap B) = P(B)\,P(A|B)$$

and if $P(A) > 0$

then $P(A \cap B) = P(A)\,P(B|A)$.

---

② Direct generalization: if $A_1, \dots, A_n$ are events with $P(A_1 \cap \dots \cap A_{n-1}) > 0$;

then

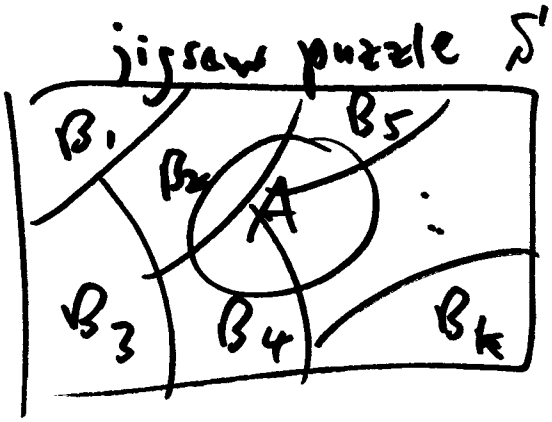
$\boxed{\text{chain rule for } \cap = \text{and}}$

$$P(A_1 \cap \dots \cap A_n) = P(A_1)\,P(A_2|A_1)\,P(A_3|A_1 \cap A_2) \cdots P(A_n|A_1 \cap \dots \cap A_{n-1})$$

Recall previous Definition

S sample space; if you can find events $B_1, \dots, B_k$ in $\mathbb{C}$

such that the B_j are disjoint and <u>exhaustive</u> $\left(\bigcup_{i=1}^{n} B_i = S\right)$, then you have found a <u>partition</u> $(B_1, \ldots, B_k)$ of S.



③ If $(B_1, \ldots, B_k)$ is a partition of S with $P(B_j) > 0$ for all $j = 1, \ldots, k$, then for any event A in C

$$P(A) = \sum_{j=1}^{k} P(B_j)\, P(A \mid B_j) \text{ —}$$

this is the $\boxed{\text{Law of Total Probability}}$

LTP

When is the LTP useful?

You're trying to compute $P(A)$ and you find it hard to compute directly. If you can find some aspect B of the world satisfying 2 properties —

① B defines a partition $\{B_1, \ldots, B_k\}$ with known $P(B_j) > 0$ of S and ② A <u>depends</u> on B in such a way that the conditional probabilities $P(A \mid B_j)$ are easier to compute than $P(A)$ itself — then you can work out $P(A)$ indirectly:

$$P(A) = \sum_{j=1}^{k} \overbrace{P(B_j)P(A \mid B_j)}^{P(A \cap B_j)}.$$

(Bayesian mixture modeling)

(18 Apr 19)

Extension of the LTP (4)

Assuming all conditional probabilities are defined in what follows, if C is in $\mathbb{C}$ then

$$P(A|C) = \sum_{j=1}^{k} P(B_j|C)\, P(A|B_j \cap C).$$

Definition

Events A, B are <u>independent</u> iff

(freq)

$$P(A \cap B) = P(A) \cdot P(B),$$

which (as long as $P(A) > 0$, $P(B) > 0$) is equivalent to

$$\left.\begin{aligned} P(A|B) &= P(A) \\ \text{and} \quad P(B|A) &= P(B) \end{aligned}\right\} \leftarrow \text{(Bayesian)}.$$

Consequences of the definition of independence

① If A and B are independent, then so are A and B^c, A^c and B, and A^c and B^c.

② Extension of the definition to more than 2 events:

Definition:

Given events $A_1, \dots, A_k$, they are <u>(mutually) independent</u> if, for every subset $A_{i_1}, \dots A_{i_j}$ of $(A_1, \dots, A_k)$ $(j = 2, \dots, k)$,

$$P(A_{i_1} \cap \dots \cap A_{i_j}) = P(A_{i_1}) \cdot \dots \cdot P(A_{i_j})$$

(Bayesian) Interpretation of independence

A, B independent $\overset{\text{iff}}{\longleftrightarrow}$ information about A doesn't change the chances associated with B, and vice versa.

Definition

Another useful extension of independence

Events $\{A_1, \dots, A_k\}$ are <u>conditionally</u> independent given event B if for every subset $\{A_{i_1}, \dots, A_{i_j}\}$ of $\{A_1, \dots, A_k\}$ $(j = 2, \dots, k)$

product

$$P(A_{i_1} \cap \dots \cap A_{i_j} \mid B) = \prod_{\ell=1}^{j} P(A_{i_\ell} \mid B).$$

Statistical Example

Suppose that there is a machine that can take an ordinary coin and produce IID tosses of the coin with $P(H) = \theta$ (any one toss), and θ can be set to any value in $[0,1]$ with a dial on the machine's control panel.

Someone sets the dial to a θ value that's <u>unknown</u> to you and starts producing coin tosses $I_1, I_2, \ldots$

Suppose the first 10 tosses come out

1 0 1 1 1 0 0 1 1 1 ← "bits" (binary digits)

H T H H H T T H H H (7 H, 3 T) (John Tukey)

$\boxed{Q:}$ Is there information in these first 10 tosses that helps you to predict I_{11}?

$\boxed{A:}$ Yes, definitely: it looks like θ is around $\frac{7}{10}$, so you would predict $I_{11} = H$. Thus I_{11} depends on $I_1, \ldots, I_{10}$ probabilistically.

Now, suppose instead that you <u>watched</u> the person with the machine set the dial to $\theta = 0.81$, so that θ is now <u>known</u> to you. The next 10 tosses come out $\boxed{HHHTHTHHHH}$ (8H, 2T). $\boxed{Q:}$ Is there information in these 10 tosses that helps you to predict the next toss?

$\boxed{A:}$ <u>No</u>; you know that $\theta = 0.81$, so there's no information in any of the I_v

that helps you to predict any of the other I_j.

given θ, the I_i are $\boxed{\text{indep.}}$

Thus the I_i are unconditionally dependent but conditionally independent given θ.

~~(14 April~~ ~~(4 Aug 17)~~

Bayes's Theorem ~~for events~~

Suppose that the events $B_1, \ldots, B_k$ (a finite partition) partition the sample space in such a way that $P(B_j) > 0$ for all $j = 1, \ldots, k$. If A is an event with $P(A) > 0$, then for each $i = 1, \ldots, k$

$$P(B_i \mid A) = \frac{P(B_i)\, P(A \mid B_i)}{P(A)}$$

and, by the LTP, this is

$$P(B_i \mid A) = \frac{P(B_i) \cdot P(A \mid B_i)}{\sum_{j=1}^{k} P(B_j) \cdot P(A \mid B_j)}.$$

How this theorem is used in Bayesian statistics: The B_i represent unknown states of the world: they're all possible — $P(B_i) > 0$ — and only one of them is true, but you don't know which one. A represents <u>data</u>: information that will help you identify the most probable B_i.

Before (a priori) the dataset A arrives, you have background information about the plausibility of the B_i that you can represent with prior probabilities $P(B_i)$. After (a posteriori) the dataset A arrives, you can use Bayes's Theorem to update your prior probabilities to posterior probabilities $P(B_i \mid A)$.

The probabilities $P(A \mid B_i)$ represent how likely the dataset A would be if B_i were the actual unknown state; this is often called likelihood information.

(the denominator)
$P(A)$ does not depend on the B_i, and can therefore be regarded as a <u>normalizing constant</u>, put into Bayes's Theorem to make all the $P(B_i|A)$ add up to 1. Thus

$$P(B_i|A) = \frac{P(B_i)\,P(A|B_i)}{P(A)}$$

is interpreted as

$$\begin{pmatrix}\text{posterior}\\ \text{information}\end{pmatrix} = \frac{\begin{pmatrix}\text{prior}\\ \text{information}\end{pmatrix}\cdot\overset{\text{(data)}}{\begin{pmatrix}\text{likelihood}\\ \text{information}\end{pmatrix}}}{\begin{pmatrix}\text{normalizing}\\ \text{constant}\end{pmatrix}}.$$

Random variables and their distributions

Example: Tay-Sachs Disease | T = T-S baby, N = not

$\mathcal{S}$ 2

NNNNN	0
TNNNN	
NTNNN	
NNTNN	1
NNNTN	
NNNNT	
TTNNN	
TNTNN	
TNNTN	
TNNNT	
NTTNN	2
NTNTN	
NTNNT	
NNTTN	
NNTNT	
NNNTT	
⋮	⋮
TTTTT	5

← # of T-S babies = Y

Definition

Given a sample space $\mathcal{S}$ for an experiment $\mathcal{E}$, a (real-valued) random variable (rv) is a function from the non-weird collection $\mathcal{C}$ of subsets of $\mathcal{S}$ to the real number line $\mathbb{R}$.

In the T-S case study, the elements s of $\mathcal{S}$ look like NNNTN and the rv Y counts how many Ts they contain

For instance, $I(TNNTN) = 2$ and $I(NNNTT) = 0 \text{ or } 2$ (i.e., I ignores the order of the children).

We can use the following notation to simplify things.

Notation $\quad \underbrace{P(I = y)}_{\text{"P of proposition"}} \triangleq P(\underbrace{\{s : I(s) = y\}}_{\text{set}})$

For example, $P(I = 1) = P(\{s \in S : I(s) = 1\})$
$= P(\{TNNNN, NTNNN, NNTNN, NNNTN, NNNNT\})$.

In general the values a random variable takes on could be just about anything, but in this course all of our rvs will be real-valued

In the T-S case study the rv I can only take on the values $0, 1, \ldots, 5$.

(7 Aug 17)

y	$P(Y=y)$
0	0.237
1	0.396
2	0.264
3	0.088
4	0.015
5	0.001

You can see that a <u>rv Y</u> is completely specified by two things: the values it can take on, and the probabilities for those values.

(see p. 23)

Definition The (<u>probability</u>) <u>distribution</u> of a random variable Y is the collection of all probabilities of the form $P(Y \in A)$ for all sets A of real numbers in the non-weird collection $\mathcal{C}_{\mathbb{R}}$ of subsets of the real number line $\mathbb{R}$.

The rv Y in the T-S case study has a finite set of possible values —

this is true of some, but not all, rvs.

Definition A random variable X has a <u>discrete distribution</u>, or equivalently X is a <u>discrete</u> rv, if the set of possible (distinct) values of X is finite or at most countably infinite; rvs for which the set of possible values is uncountable are called <u>continuous random variables</u>.

Examples ① The rv $X = \begin{cases} 1 & \text{if } X > 0 \\ 0 & \text{otherwise} \end{cases}$ (with $X = \#$ T-s tossing) is discrete, taking on only the values $\{0, 1\}$ {yes, no} {1, 0} – such rvs are called <u>dichotomous</u> or <u>binary</u>.

② Imagine a scale for weighing things that has a dial you can set to specify how many significant figures (sigfigs) of precision you want. Buy a "1 pound" package of butter at your favorite market and weigh it.

possible weights (ounces)
16
16.0
15.99
15.9930
15.9928
⋮

If there's no <u>conceptual</u> limit to the number of sigfigs you could get, a rv $X = \left(\begin{array}{c}\text{the actual (true)} \\ \text{weight of the package}\end{array}\right)$ should be modeled as continuous, having values (e.g.) on $(0, \infty)$, the positive part of $\mathbb{R}$.

$\boxed{\text{Reality check:}}$ Infinite precision is impossible in practice;

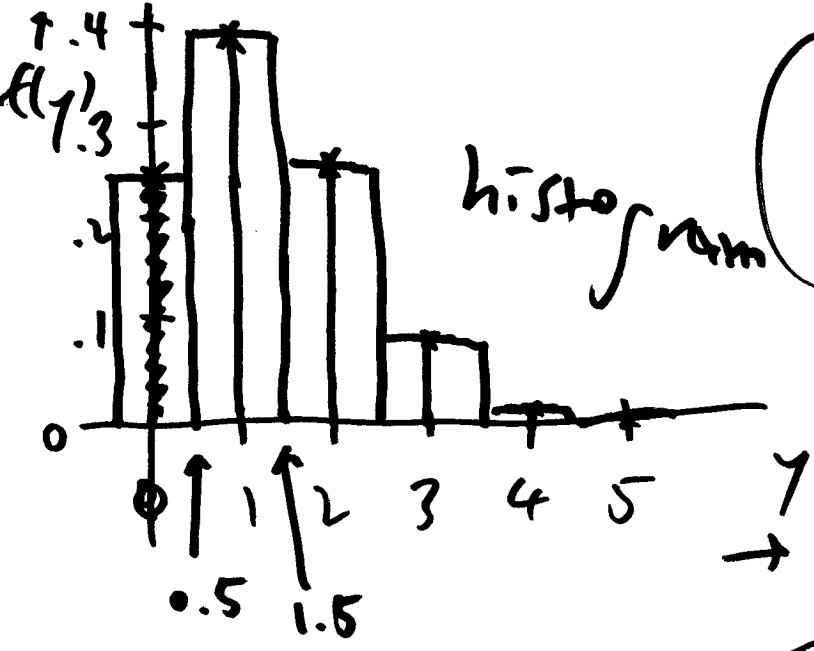
every measurement you ever make is in actuality discrete, but it's useful to regard rvs that are <u>conceptually</u> continuous (i.e., no limit in principle to the precision of measurement) as continuous.

(23 Apr 19)

Definition Given a discrete rv Y, the <u>probability</u> (mass) <u>function</u> (pmf or pf) of Y is the function f_Y that keeps track of the probabilities associated with Y: $f_Y(y) = P(Y = y)$.

The set $\{y : f_Y(y) > 0\}$ is called the <u>support</u> of (the distribution of) Y.

(DS is almost unique in using "pf", nearly everybody talks about the pmf.)



histogram

pf of I in the T-S case study

Definition

A rv I that only takes on the values {0, 1} — i.e., a <u>binary</u> rv — is said to have a <u>Bernoulli</u> distribution with <u>parameter</u> p — written Bernoulli(p) —

James Bernoulli Swiss (1655–1705)

if

$$f_I(y) = p(I = y) = \begin{cases} p & \text{for } y = 1 \\ 1-p & y = 0 \\ 0 & \text{else} \end{cases}$$

$$= p^y (1-p)^{1-y}$$

Notation (I follows a Bernoulli(p) distribution) $\leftrightarrow$ I is distributed as $\sim$ Bernoulli(p) or (I | p)

Example In the powerball lottery (see take home test problem 2) 5 white numbered balls are drawn at random <u>without</u> replacement from a bin with balls numbered $\{1, 2, \ldots, 69\}$.

Let W_i = # on i^{th} drawn white ball. $(i = 1, \ldots, 5)$

Clearly $p(W_1 = w_1) = \begin{cases} \frac{1}{69} & \text{for } w_1 = 1, 2, \ldots, 69 \text{ (or)} \\ 0 & \text{otherwise} \end{cases}$.

Less clearly (but true) $W_2, \ldots, W_5$ follow the same distribution if nothing is known about the previous draws.

Definition For any two integers $a \le b$ $(<)$, a rv Y that's equally likely to be any of the values $\{a, a+1, \ldots, b\}$ has the <u>uniform distribution</u> Uniform$\{a, b\}$. Evidently

its pf is $f_Y(y) = P(Y=y) = \begin{cases} \frac{1}{b-a+1} & \text{for } y = a, \dots, b \\ 0 & \text{else} \end{cases}$

~~"is distributed as"~~ $a \quad a+1 \quad \cdots \quad b$

$Y \sim \text{Uniform}\{a, b\} \leftrightarrow Y$ chosen at random from $\{a, a+1, \dots, b\}$.

Definition (h) random trials are performed, with each trial recorded as a success S or failure F. If each trial is independent of all the others and the chance (p) of success is constant across the trials, then Y = # of successes has the Binomial distribution

not sample space

pf

$$f_Y(y) = P(Y=y) = \begin{cases} \binom{n}{y} p^y (1-p)^{n-y} & \text{for } y = 0, 1, \dots, n \\ 0 & \text{else} \end{cases}$$

(with parameters n and p)

In shorthand $Y \sim \text{Binomial}(n, p)$.
or $(Y \mid n, p)$

Let $B_i = \begin{cases} 1 & \text{if trial } i \text{ is a success} \\ 0 & \quad \ldots \quad \text{failure} \end{cases}$

for $i = 1, \ldots, n$; then under these assumptions

$B_i \overset{IID}{\sim} \text{Bernoulli}(p)$ and all the B_i are

independent. $\boxed{\text{Notation}}$ $X_i \overset{IID}{\sim} f_{X_i}(x_i)$

means that all of the rvs $X_1, X_2, \ldots$

are independent and identically distributed

draws from the distribution with pf

$f_{X_i}(x_i)$. Thus with the success/failure

trials, $B_i \overset{IID}{\sim} \text{Bernoulli}(p)$ and
$(i = 1, \ldots, n)$

$\left(Y = \sum_{i=1}^{n} B_i\right) \sim \text{Binomial}(n, p)$.

This is our first example of the distribution of the <u>sum</u> of a bunch of IID rvs, a topic we'll examine in detail later.

Continuous random variables

Example

(Round-off error in computer science)

Single-precision floating point numbers carry about 7 sigfigs of accuracy, leading to <u>roundoff error</u> in the last digit;

3.1415922 | 653589

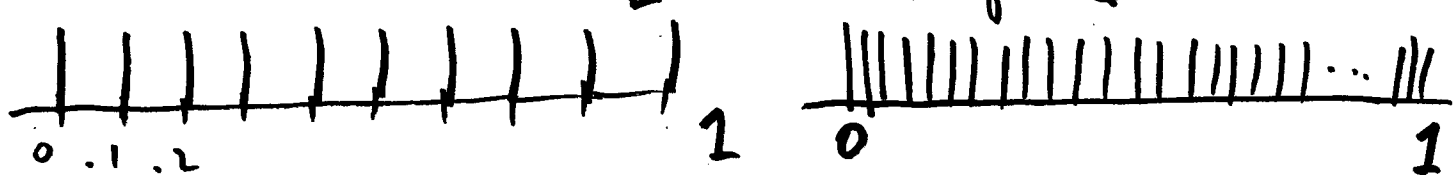
3 ~ .04 error

it's important to study how these errors accumulate as the number of steps in a calculation increases.

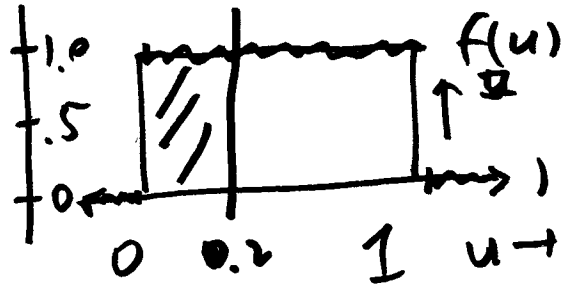
Since there's no reason one decimal digit would be favored over another in rounding, the <u>uniform</u> distribution is key to these calculations.

Consider first Uniform$\{0, .1, \ldots, .9\}$ and then $\{0, .01, .02, \ldots, .99\}$

← discrete pf →



In the limit with more & more sig figs 93

this should go to 

the

<u>continuous</u> <u>uniform</u> distribution

Uniform (0,1) on the unit interval.

The analogue of the discrete pf in this case is the smooth continuous function

$$f_Y(u) = \begin{cases} 1 & \text{for } 0 \le u \le 1 \\ 0 & \text{else} \end{cases}.$$

analogue of summation is integration

Definition

A random variable

(Y has a continuous distribution)

$\leftrightarrow$ (Y is a continuous rv) if there exists a continuous nonnegative function f_Y defined on $\mathbb{R}$ such that for every interval $[a,b]$, $P(a \le Y \le b) = \int_a^b f_Y(y)\,dy$.

In this definition, a can be $-\infty$ and b can be $+\infty$.

Definition If Y is a continuous rv, the function f_Y in the previous definition is called the <u>probability density function</u> (PDF) (pdf) of Y. The set $\{y : f_Y(y) > 0\}$ is called the <u>support</u> of (the distribution of) Y.

Clearly (a) $f_Y(y) \geq 0$ for all y

and (b) $\int_{-\infty}^{\infty} f_Y(y)\,dy = 1$.

What about individual points — <u>singletons</u> — $\{y\}$ on $\mathbb{R}$?

You'll recall from calculus that if f_Y is continuous

on its support, $\int_a^b f_Y(y)\,dy$ can equally well stand for $P(a \le Y \le b)$ or $P(a < Y \le b)$ or $P(a \le Y < b)$ or $P(a < Y < b)$, because (e.g.) $\int_a^a f_Y(y)\,dy = 0$ if f_Y is continuous at $y = a$. Thus, importantly, $\boxed{P(Y = y) = 0 \text{ for all } -\infty < y < \infty}$

weirdly, this doesn't mean that the value y of Y is impossible, as it does with discrete rv; it just means that singletons have to have 0 probability (otherwise $\int_{-\infty}^{\infty} f_Y(y)\,dy = +\infty$ not 1).

Definition With a and b any two real numbers satisfying $a < b$,

$Y \overset{\text{is distributed as}}{\sim} \text{Uniform}(a, b) \overset{\text{iff}}{\longleftrightarrow} P\left(\begin{array}{c} Y \text{ is in} \\ \text{any subinterval} \\ \text{of } (a,b) \end{array}\right)$

$= \text{the length of the subinterval} \overset{\text{iff}}{\longleftrightarrow}$

$$f_Y(y) = \left\{ \begin{array}{ll} \frac{1}{b-a} & \text{for } a \le y \le b \\ 0 & \text{else} \end{array} \right\}.$$

Definition The <u>indicator function</u>

for any (true/false) proposition A is $I(A) = \left\{ \begin{array}{ll} 1 & \text{if } A \text{ true} \\ 0 & \text{if false} \end{array} \right\}$

People sometimes also write (with A a set) $\subset \mathbb{R}$

$$I_A(y) = \left\{ \begin{array}{ll} 1 & \text{if } y \in A \\ 0 & \text{else} \end{array} \right\}.$$

with this definition, $Y \sim \text{Uniform}(a,b)$ (77)

$$\leftrightarrow f_Y(y) = \frac{1}{b-a} I(a \le y \le b) = \frac{I_{[a,b]}(y)}{b-a}.$$

Contrast

versus

$Y \sim \text{Uniform}(a,b)$ continuous and uniform on (a,b) or $(a,b]$ or $[a,b)$ or $[a,b]$

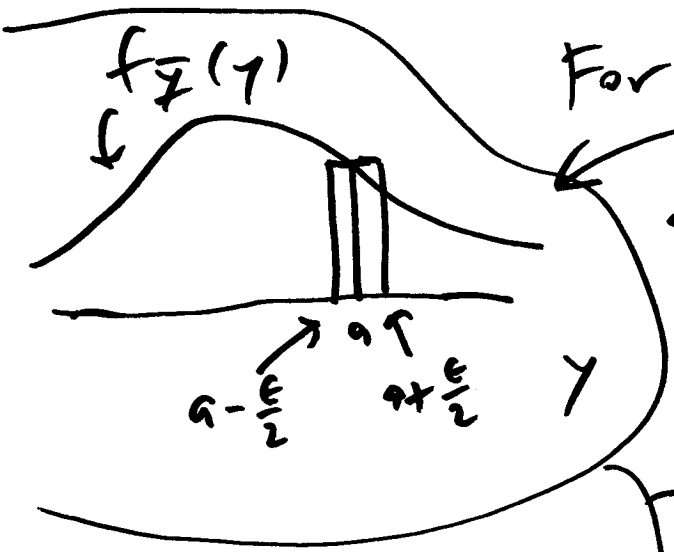
$Y \sim \text{Uniform}\{a,b\}$ for a, b integers with $a < b \leftrightarrow Y$ discrete and uniform on $\{a, a+1, \ldots, b\}$.

Density and probability are not the same thing

Density values $f_Y(y)$ are themselves not probabilities; for example, they can easily be > 1 & can even be $+\infty$,

as we'll see later. | Density values

define probabilities: $P(a \le Y \le b) = \int_a^b f_Y(y)\,dy$.



For small $\epsilon > 0$ you can see from this sketch that

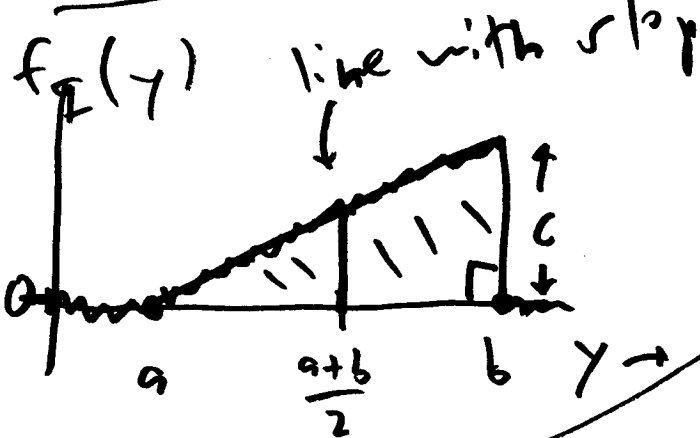
$$P\left(a-\frac{\epsilon}{2} \le Y \le a+\frac{\epsilon}{2}\right) = \int_{a-\frac{\epsilon}{2}}^{a+\frac{\epsilon}{2}} f_Y(y)\,dy$$

$\doteq$ area of rectangle

$= \epsilon \cdot f_Y(a)$.

Example

(triangular distribution)



Can a continuous rv Y have a pdf that looks like a triangle? Let's see what, if any, restrictions would be needed.

The line in the sketch has slope $\frac{c}{b-a} = m$ (79)

and passes through the point $(x_1, y_1) = (a, 0)$, so

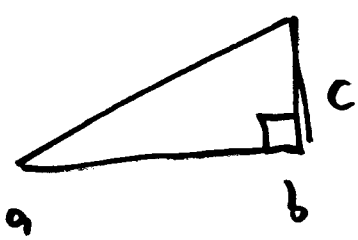
the equation of the line is

$$y - y_1 = m(x - x_1) \leftrightarrow y = \frac{c}{b-a}(x - a) = \cancel{f_X(x)}\ f_Y(y)$$

Densities have to integrate to 1,

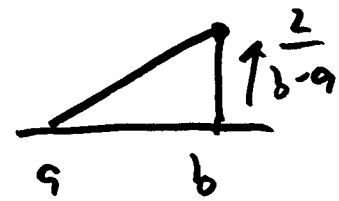
so $\displaystyle\int_a^b \frac{c}{b-a}(x-a)\,dx = 1 \leftrightarrow c = \frac{2}{b-a}$

Wd integrate $c(x-a)/(b-a)$ for $x = a$ to b



Easier way: area of a triangle is $\frac{1}{2}$(base)(height), so

$1 = \frac{1}{2}(b-a)c$ and $c = \frac{2}{b-a}$



(25 Apr 18)

Thus the triangular distribution that (8)

starts at $y = a$ (0 for) and rises linearly to

$\frac{2}{b-a}$ at $y = b$ has density (pdf) $f_X(y) = \begin{cases} \frac{2(y-a)}{(b-a)^2} & a \le y \le b \\ 0 & \text{else} \end{cases}$ (*)

(wd) integrate $2(y-a)/(b-a)^2$ for y from a to b

You can see that calculating probabilities with continuous rvs requires you to dust off your integral calculus.

Example | with the triangular distribution (X) (*) above, what's $P(a \le X \le \frac{b-a}{2})$?

Hard(?) way: $\int_a^{\frac{b-a}{2}} \frac{2(x-a)}{(b-a)^2} dx = \frac{(3a-b)^2}{4(b-a)^2}$

Easy(?) way:

area of triangle = ...

Sometimes it's mathematically convenient to work with unbounded continuous rvs, just as was true in the Poisson case study for discrete rvs.

Example (DS p. 105)

Y = voltage in an electrical system: in practice Y cannot be infinite, but you may not know ahead of time what its maximum practical value is, so model it as unbounded but without much probability for extremely large values. DS give as an example the pdf

$$f_Y(y) = \frac{1}{(1+y)^2} I(y > 0).$$

check:

$$\int_0^\infty \frac{1}{(1+y)^2}\,dy = 1$$

$$= \frac{(1+y)^{-1}}{-1}\Bigg|_0^\infty = -1(0-1) = 1 \checkmark$$

You can check that $\int_{1000}^{\infty} \frac{1}{(1+x)^2}\,dx = \frac{1}{1001} \doteq .001$,

so the right tail beyond $Y = 1000$ has almost no probability, matching the correct qualitative behavior.

Sometimes a rv will be neither discrete nor continuous; people then say that it has a mixed (discrete/continuous) distribution.

Definition.

Example:

In medical clinical trials of people with potentially fatal diseases, the outcome variable Y_i for person i in (say) the treatment group might be

Y = survival time in days from the beginning of the trial; however, and a good thing too, some patients may still be alive at the time T_{end} at which the trial finishes. Your probability model for Y_i would then have a continuous part for $0 \le Y_i \le T_{end}$ and a discrete lump of probability p at $Y_i = T_{end}$ signifying $(Y_i > T_{end})$ but we don't know what Y_i would have been if we could have observed it: (right-censoring)

$$\int_0^{T_{end}} f_{Y_i}(y)\,dy = (1-p) \quad \text{and} \quad P(Y_i > T_{end}) = p.$$

(Boy example) (see p. 10 of doc. com notes)

Unifying idea connecting discrete & continuous rvs

Discrete $\leftrightarrow$ pf (pmf)

Continuous $\leftrightarrow$ pdf

~~Mixed~~ $\leftrightarrow$ (pf + pdf)

$\boxed{Q:}$ Is there something that uniquely characterizes the distribution of X, both when X is discrete & when it's continuous & when it's mixed?

$\boxed{A:}$ Yes, the cumulative distribution (cdf) function $F_X(y)$

Definition:

The cumulative distribution function (cdf) of a rv X is defined to be

$$F_X(y) = P(X \le y) \text{ for all } -\infty < y < \infty$$

Example: $Y \sim \text{Bernoulli}(p)$

$$p(Y=y) = \begin{cases} p & \text{for } y=1 \\ 1-p & \quad\;\; 0 \\ 0 & \text{else} \end{cases}.$$

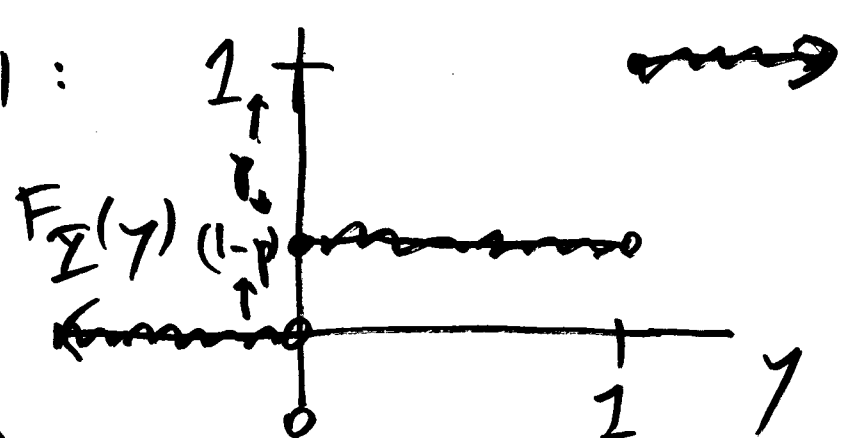
(PMF)

Notice that there's a clever way to write this

pf:

$$p(Y=y) = p^{y}(1-p)^{1-y} I_{\{0,1\}}(y)$$

The cdf of Y is 0 for $y<0$; at $Y=0$ it jumps up to $(1-p)$ and stays there for $0 \le y < 1$; and at $Y=1$ it jumps up to 1 & stays there for $y \ge 1$:



You can see that in general

$$0 \le F_Y(y) \le 1$$

and it's also clear that a cdf $F_Y(y)$ (86)

has to be a <u>non-decreasing</u> function

of y: if $y_1 < y_2$ then $F_Y(y_1) \le F_Y(y_2)$.

Furthermore, $\lim_{y \to -\infty} F_Y(y) = 0$ and

$\lim_{y \to +\infty} F_Y(y) = 1$. (~~when Y is continuous~~)

CDFs F_Y can be continuous on all of $\mathbb{R}$ but certainly don't have to be (see the CDF of the Bernoulli(p) distribution).

Technical fact:

Def: $F_Y(y^-) \triangleq \lim_{y^* \to y} F_Y(y^*) \triangleq \lim_{y^* \uparrow y} F_Y(y^*)$

limit from the left

$y^* < y$

(y^* goes to y from below)

Def. $F_Y(y^+) \triangleq \lim_{y^* \to y} F_Y(y^*) \triangleq \lim_{y^* \downarrow y} F_Y(y^*)$

limit from right

$y^* > y$ (y^* jogs to y from above)

technical fact: $F_Y(y) = F_Y(y^+)$ for all $-\infty < y < \infty$

people call this <u>continuity from the right</u> or <u>continuity from above</u>

If $F_Y(y^-) = F_Y(y^+) = F_Y(y)$ then F_Y is continuous at y

Consequences of the CDF definition

① $P(Y > y) = 1 - F_Y(y)$

② For all y_1, y_2 with $y_1 < y_2$

$P(y_1 < Y \le y_2) = F_Y(y_2) - F_Y(y_1)$.

Consequence ② means that if Y is continuous, there's an intimate connection between $F_Y(y)$ (cdf) and $f_Y(y)$ (pdf):

Y continuous: $y_1 < y_2$

$$P(y_1 < Y \le y_2) = \boxed{F_Y(y_2) - F_Y(y_1) = \int_{y_1}^{y_2} f_Y(y)\,dy}$$

and thus

Theorem If Y is a continuous rv, with pdf $f_Y(y)$ and CDF $F_Y(y)$ then

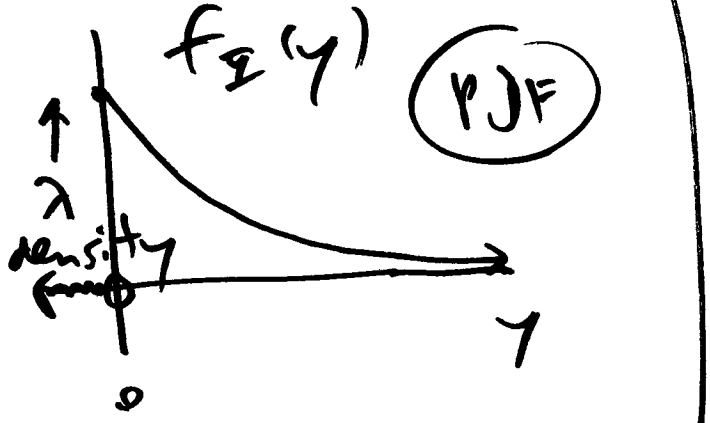
$$F_Y(y) = \int_{-\infty}^{y} f_Y(t)\,dt \quad \text{and} \quad \frac{d}{dy}F_Y(y) = f_Y(y).$$

at all continuity points of f

In other words Y continuous $\leftrightarrow$ the derivative of $\underset{\text{CDF}}{F_Y(y)}$ is $\overset{\text{pdf}}{f_Y(y)}$ (and $F_Y(y)$ is an anti-derivative (integral) of $f_Y(y)$.

Definition Y follows an <u>exponential distribution</u> $\mathcal{E}(\lambda)$ (rate) with parameter $\lambda > 0$

$$\leftrightarrow \underset{\text{pdf}}{f_Y(y)} = \begin{cases} \lambda e^{-\lambda y} & y > 0 \\ 0 & y \le 0 \end{cases}$$



The exponential dist. has a fundamental connection to the Poisson distribution in Poisson processes that we'll explore later.

It's easy to calculate the CDF of an exponential distribution:

Notation

Y exponentially distributed with parameter $\lambda > 0$ $\leftrightarrow$ $(Y \mid \lambda)$ $\downarrow$ $Y \sim \text{Exponential}(\lambda)$

$\Updownarrow$

$Y \sim \mathcal{E}(\lambda)$

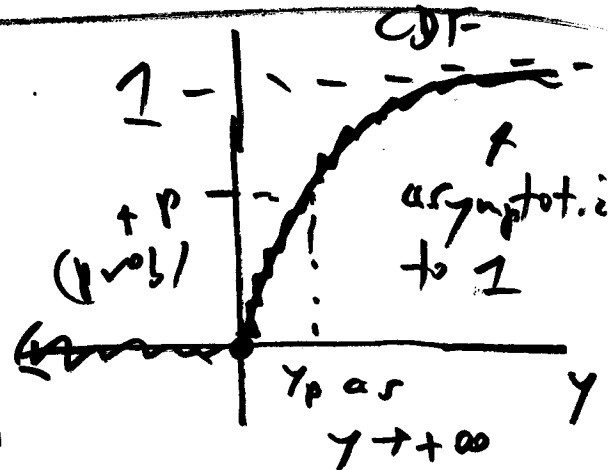
for $y > 0$

$$F_Y(y) = \int_{-\infty}^{y} f_Y(t)\,dt = \int_0^y \lambda e^{-\lambda t}\,dt$$

$$= \lambda \left. \frac{e^{-\lambda t}}{-\lambda} \right|_0^y = -1\left(e^{-\lambda y} - 1\right) = 1 - e^{-\lambda y}$$

So

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



(30 Apr 19)

AMS131
2 May 19

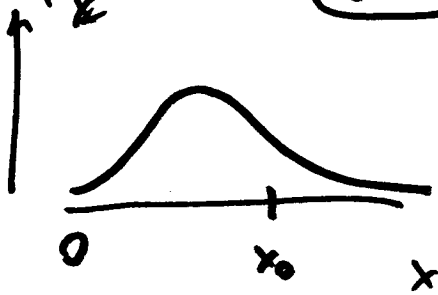
①

This time: inverse CDFs;
next time: joint distributions

(syllabus: readings)

mode
= point of symmetry
= mean
= median

$f_X(x)$ X cont. PDF of X



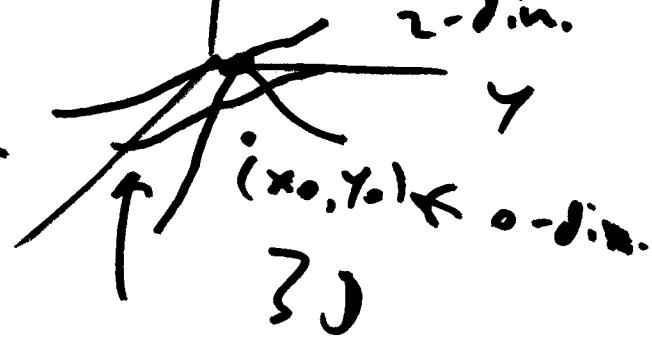
$P(X = x_0) = 0$

0-dimensional

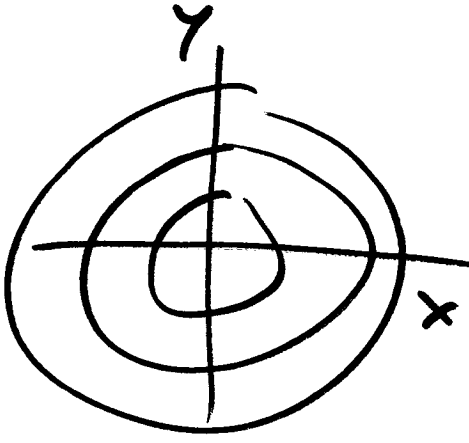
1-dimensional rv

$f_{XY}(x, y)$ cont. (X, Y) 2-dim.

$c\, e^{-(x^2+y^2)}$



3D perspective plot



2D contour plot of

$$\int_{-\infty}^{\infty} \boxed{\int_{-\infty}^{\infty} c\,x^2 y \, dy} \, dx = \iint_S c x^2 y \, dy \, dx$$

$$= \int_{-1}^{+1} \boxed{\int_{x^2}^{1} c\,x^2 y \, dy} \, dx = \cancel{\frac{4}{21}}$$

$$\int_{-1}^{+1} \int_{x^2}^{1} x^2 y \, dy \, dx = \frac{4}{21}$$

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

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

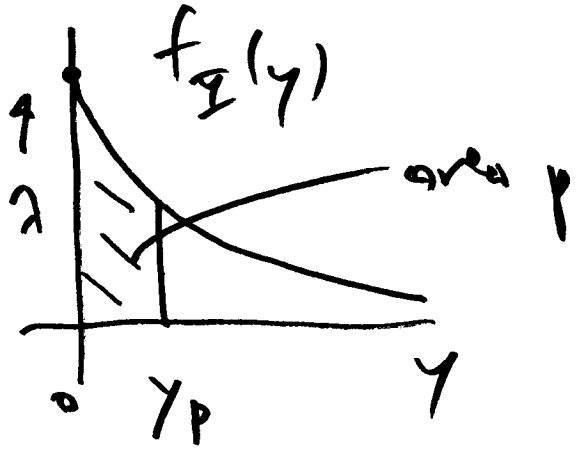
$$x^2 \le y$$

$$-\sqrt{y} \le x \le +\sqrt{y}$$

③

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

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



for $(y_p > 0)$

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

well, $P(0 \le Y \le y_p) = \boxed{F_Y(y_p) = p}$

$= \boxed{1 - e^{-\lambda y_p} = p}$ so $y_p = F_Y^{-1}(p)$

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

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

$y_p = -\dfrac{\log(1-p)}{\lambda} = F_Y^{-1}(p)$

Def. y_p is called the pth <u>quantile</u> or the $100p^{th}$ <u>percentile</u> of (the distribution of) Y.

Some care is required when Y is discrete or mixed.

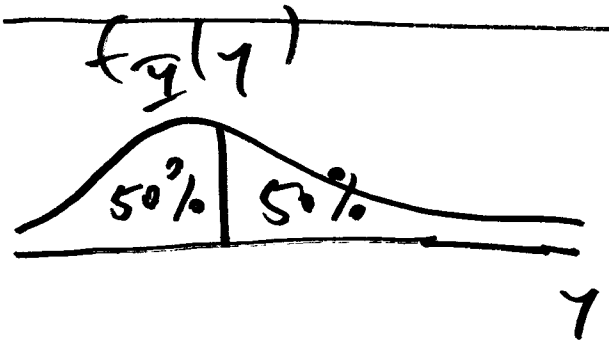
General definition

Y rv with CDF $F_Y(y)$.

For all $0 < p < 1$ define

$F_Y^{-1}(p)$ = the <u>smallest</u> y value such that $F_Y(y) \geq p$

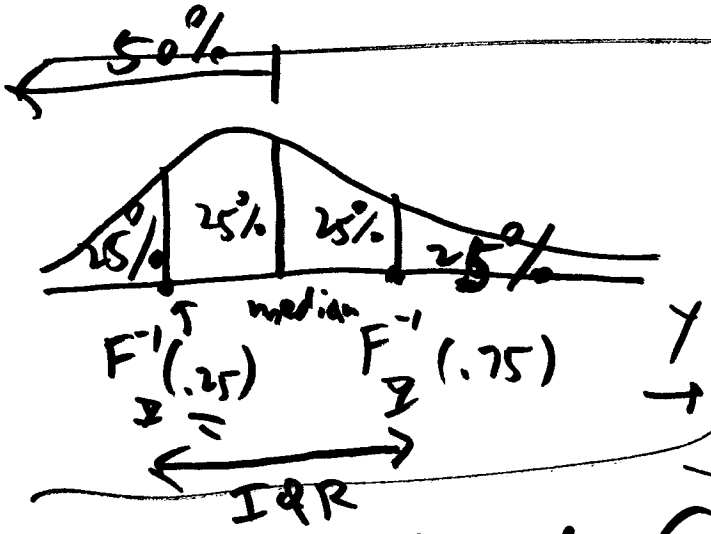
Then $F_Y^{-1}(p)$ is the pth quantile (a $100p^{th}$ percentile) of Y and F_Y^{-1} is the <u>quantile function</u>.



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

One way to define the <u>center</u> of a distribution is to find the 50th percentile.

Definition The $\frac{1}{2}$ quantile (0.5) $\triangleq$ the 50th percentile of a distribution is called the <u>median</u> of the dist.

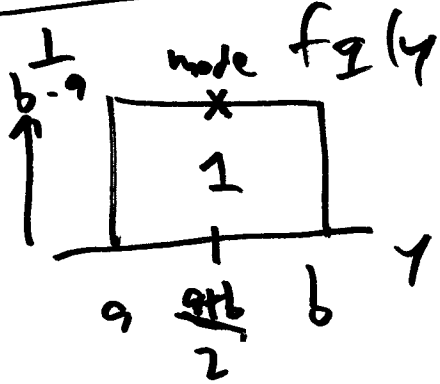


measures of <u>spread</u> for the distribution of a rv Y

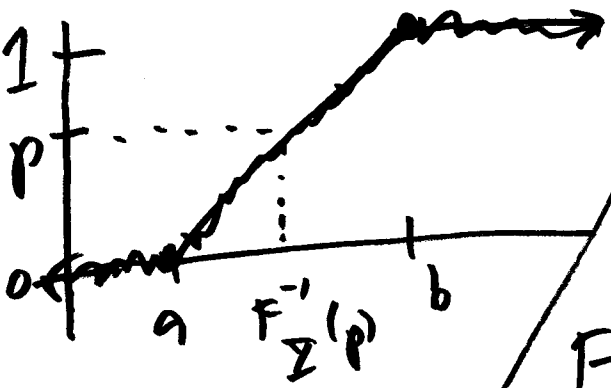
One way to define the <u>spread</u> of a dist. is to see how far apart its 75th and 25th percentiles are.

Definition The $\frac{1}{4}$ quantile (0.25) = the 25th percentile $F_Y^{-1}(.25)$ is the <u>lower quartile</u>; the $\frac{3}{4}$ quantile (0.75) = the 75th percentile $F_Y^{-1}(.75)$ is the <u>upper quartile</u>, and $(F_Y^{-1}(.75) - F_Y^{-1}(.25))$ = <u>interquartile range</u> (IQR)

Example $Y \sim \text{Uniform}(a, b)$; then



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



Easy to invert F_Y:

$$F_Y^{-1}(p) = (1-p)a + pb \quad \text{for } 0 < p < 1$$

And (no surprise) the median is $\frac{a+b}{2}$.

Studying two random variables at a time

Def. X, Y rvs: the <u>joint</u> (or <u>bivariate</u>) 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.

Case 1 (X and Y both discrete)

Def. X, Y r.v. → If there are only finitely or countably infinitely many possible values (x, y) for (X, Y), X and Y have a <u>discrete joint dist.</u>

Def The <u>joint probability (mass) function</u> (joint pmf) of (X, Y) discrete is the function $f_{X,Y}(x, y) = P(X = x, Y = y)$.

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

Consequences

① $\sum_{\text{all } (x,y)} f_{XY}(x, y) = 1$. (omit comma)

② For any (non-weird) set C of ordered pairs (x, y), $P[(X, Y) \in C] = \sum_{(x,y) \in C} f_{XY}(x, y)$

Def Two rv X and Y have a <u>continuous joint distribution</u> if you can find a nonnegative function $f_{X,Y}(x,y)$ defined for all $(x,y) \in \mathbb{R}^2$ (the real plane)

Case 2: X, Y both continuous

such that for every (non-weird) subset C of the plane $P[(X,Y) \in C] = \iint_C f_{XY}(x,y)\,dx\,dy$.

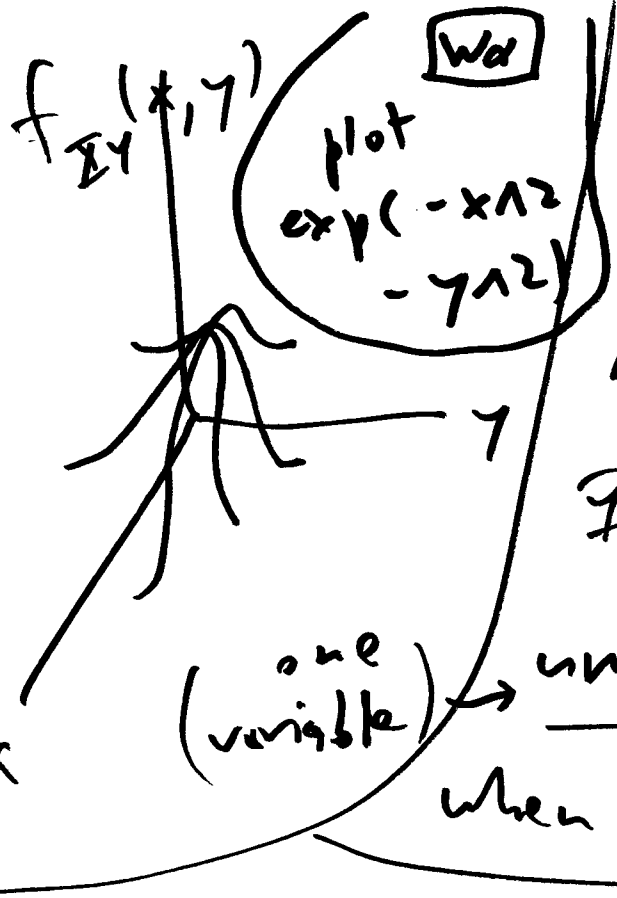
~~$f_{XY}(x,y)$ is the joint pdf of (X,Y).~~

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

Immediate Consequences

① For all (x,y) in $\mathbb{R}^2$,

$f_{XY}(x,y) \geq 0$, and $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx\,dy = 1$



② If (X, Y) have a continuous joint distribution, then X and Y each have a continuous (marginal) univariate distribution when considered separately.

(one variable)

---

③ For all continuous pdfs $f_{XY}(x,y)$,

(a) Every individual point, and every countably infinite sequence (or set) of points (0-dimensional) in $\mathbb{R}^2$ has probability 0 under f_{XY}. (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 $\{(x,y) : x = g(y),\ a < y < b\}$ also have probability 0.

(1-dimensional)

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

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

Joint distributions can lead to tricky integrals

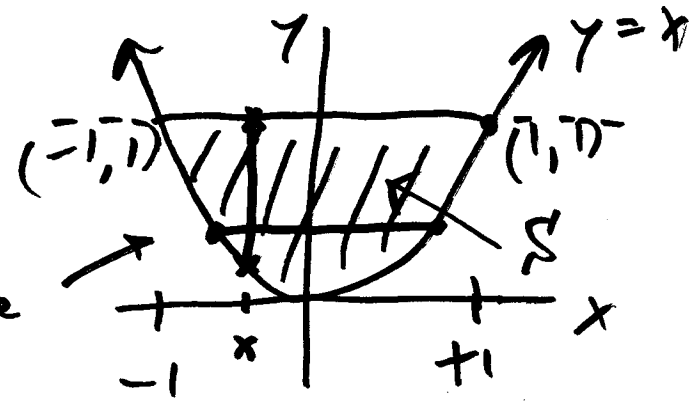
Example

Suppose that (X,Y) have joint pdf

$$f_{XY}(x,y) = \begin{cases} cx^2y & \text{for } 0 \le x^2 \le y \le 1 \\ 0 & \text{else} \end{cases}$$

Let's work out the normalizing constant.

The support of f_{XY} is the shaded region here



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

$$= \iint_{S} f_{XY}(x,y)\,dy\,dx$$

$$= \int_{-1}^{+1}\int_{x^2}^{1} c\,x^2 y\,dy\,dx$$

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

$$= \int_{-1}^{1} c\,x^2\left(\frac{y^2}{2}\Big|_{x^2}^{1}\right)dx$$

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

$$= \frac{1}{2}c\int_{-1}^{1} x^2\,dx - \frac{1}{2}c\int_{-1}^{1}\frac{x^6}{2}\,dx$$

$$= \frac{1}{2}c\left(\frac{x^3}{3}\Big|_{-1}^{1}\right) - \frac{c}{2}\left(\frac{x^7}{7}\Big|_{-1}^{1}\right) = \frac{4}{21}c = 1$$

Search on "iterated integrals in wolfram alpha"

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

The other way to parameterize the support S is to let y go from 0 to 1 while x goes from $-\sqrt{y}$ to $\sqrt{y}$:

$$1 = \iint_S f_{XZ}(x,y)\,dx\,dy$$

$$= \int_0^1 \int_{-\sqrt{y}}^{\sqrt{y}} c x^2 y \,dx\,dy$$

$$= \int_0^1 cy \left(\int_{-\sqrt{y}}^{\sqrt{y}} x^2\,dx \right) dy$$

$$= \int_0^1 cy \left(\frac{x^3}{3} \Big|_{-\sqrt{y}}^{\sqrt{y}} \right) dy$$

$$= c \int_0^1 y \cdot \frac{1}{3} \left(y^{3/2} - -y^{3/2} \right) dy$$