There are other foundational theories of probability — one by the Italian mathematician and actuary Bruno de Finetti, and another by the American physicists Richard T. Cox (1898-1991) and Edwin T. Jaynes (1922-1998) — in which the probability function $P_{def}(A|B)$ or $P_{CJ}(A|B)$ has 2 inputs, not 1, so that conditional probability is the primitive concept, not unconditional 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. Example (Tay-Sachs) we actually computed not $P(\text{at least 1 t-s baby})$ but $P(\text{at least 1 t-s baby} \mid \text{family of 5, mother and father both carriers})$.

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 $P_k(B \mid A_0)$. 
Consequences of the conditional probability definition (theorems)

1. A, B events in C:
   \[ P(A \cap B) = P(B) \cdot P(A \mid B) \]
   \[
   \text{if } P(B) > 0 \text{ then }
   
   \text{and if } P(A) > 0
   
   \text{then } P(A \cap B) = P(A) \cdot P(B \mid A).
   
2. Direct generalization: if \( A_1, \ldots, A_n \) are events with \( P(A_1 \cap \ldots \cap A_{n-1}) > 0 \), then
   \[ P(A_1 \cap \ldots \cap A_n) = P(A_1) \cdot P(A_2 \mid A_1) \cdot P(A_3 \mid A_1, A_2) \]
   \[
   \ldots P(A_n \mid A_1, \ldots, A_{n-1})
   
   \text{chain rule for } n = \infty.
   
Recall previous definition: sample space; if you can find events \( B_1, \ldots, B_k \) in \( C \)
such that the $B_j$ are disjoint and exhaustive ($\bigcup_{j=1}^{n} B_i = S$), then you have found a partition $(B_1, ..., B_k)$ of $S$.

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

$$p(A) = \sum_{j=1}^{k} p(B_j) p(A | B_j)$$

This is the 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 \( P(B_j) \) of \( S \) and
2. \( A \) depends on \( B \) in such a way that the conditional probabilities \( P(A|B_j) \) are easier to compute than \( P(A) \) itself — then you can work out \( P(A) \) indirectly:

\[
P(A) = \sum_{j=1}^{k} P(B_j) P(A|B_j)
\]

(Bayesian mixture modeling)
Assuming all conditional probabilities are defined in what follows, if $C$ is in $\mathcal{C}$ then

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

**Definition.** Events $A$, $B$ are independent iff

\[ P(A \cap B) = P(A) \cdot P(B) \]

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

\[ P(A \mid B) = P(A) \quad \left\{ \begin{array}{l}
\text{Bayesian} \\
\text{and} \\
P(B \mid A) = P(B)
\end{array} \right. \]
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 $A_1, \ldots, A_k$, they are (mutually) independent if, for every subset $A_{i_1}, \ldots, A_{i_j}$ of $(A_1, \ldots, A_k)$ ($j = 2, \ldots, k$),

$$P(A_{i_1} \cap \ldots \cap A_{i_j}) = P(A_{i_1}) \cdots P(A_{i_j})$$
Interpretation of independence

$A$, $B$ independent if and only if

information about $A$ doesn't change the chances associated with $B$, and vice versa.

Definition

Another useful extension of independence

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

$$P(A_{i_1} \cap \cdots \cap A_{i_j} \mid B) = \prod_{k=1}^{j} P(A_k \mid B).$$
Example: There is a machine that can take an ordinary coin and produce IID tosses of the coin with \( P(H) = \theta \), and \( \theta \) can be set to any value in \([0, 1]\) with a dial on the machine's control panel. Someone sets the dial to a \( \theta \) value that's unknown to you and starts producing coin tosses \( I_1, I_2, \ldots \).

Suppose the first 10 tosses came out "bits" (binary digits) \( \overline{HHTHHHTHHH} \) (7 H's, 3 T's) (John Tukey).

Q: Is there information in these first 10 tosses that helps you to predict \( I_n \)?
Yes, definitely: it looks like $\Theta$ is around $\frac{7}{10}$, so you would predict $\mathbb{E}_n$ depends on $\Theta$, probabilistically. $\mathbb{E}_n = \frac{7}{10}$. Now, suppose instead that you watched the person with the machine set the dial to $\Theta = 0.81$, so that $\Theta$ is now known to you. The next 10 tosses came out $\text{HHHTHTHHHH}$ ($8\text{H}, 2\text{T}$). Is there information in these 10 tosses that helps you to predict the next toss?

No; you know that $\Theta = 0.81$, so there's no information in any of the $\mathbb{E}_n$. 
That helps you to predict any of the other $I_j$. Thus the $I_j$ are unconditionally dependent but conditionally independent given $\Theta$.

Bayes's Theorem

Suppose that the events $B_1, \ldots, B_k$ 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 $i = 1, \ldots, k$

$$P(B_i | A) = \frac{P(B_i \cap A)}{P(A)}$$
and, by the LTIP, this is

\[ p(B_i | A) = \frac{p(B_i) \cdot p(A | B_i)}{\sum_{j=1}^{k} p(B_j) \cdot p(A | B_j)} \]

How this theorem is used in Bayesian statistics: \( B_i \) 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 data: information that will help you identify the most probable \( B_i \).
Before the dataset A arrives, you have background information about the plausibility of the Bi that you can represent with prior probabilities \( P(B_i) \). After the dataset A arrives, you can use Bayes's Theorem to update your prior probabilities to posterior probabilities \( P(B_i | A) \).

The probabilities \( P(A | B_i) \) represent how likely the dataset A would be if \( B_i \) were the actual unknown state; this is often called likelihood information.
P(A) does not depend on the \( B_i \), and can therefore be regarded as a normalizing constant, put into Bayes's theorem to make all the \( P(B_i | A) \) add up to 1. Thus

\[
P(B_i | A) = \frac{P(B_i) \cdot P(A | B_i)}{P(A)}
\]

is interpreted as

\[
\text{(posterior)} = \text{(prior)} \cdot \text{(likelihood) \over \text{(normalizing constant)}}
\]
Random variables and their distributions

Example: Tay-Sachs Disease

<table>
<thead>
<tr>
<th>NNNNNN</th>
<th>0</th>
<th># of T-s alleles = 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>TNNNN</td>
<td>1</td>
<td>\text{Definition}</td>
</tr>
<tr>
<td>NTTNN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NNTNN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NNNTN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NNNNT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TNTNN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TNTTN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TNNTN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TNNNT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TNNNT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NTTTN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NTTTN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NTTNT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NNTTN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NNTTN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NNTTT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TTTTT</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In the T-s case study, the elements \( s \) of \( S' \) look like \( NNNTN \) and the RV \( I \) counts how many Ts they contain.
For instance, $\zeta(TNNTN) = 2$ and $\zeta(NNNTT) = (a+b)/2$ (i.e., $\zeta$ ignores the order of the children). We can use the following notation to simplify things.

**Notation**

$$P(\zeta = \gamma) = P(\{s : \zeta(s) = \gamma\})$$

For example, $P(\zeta = 1) = P(\{s : \zeta(s) = 1\}) = P(\{\text{TNNNN}, \text{NTNNN}, \text{NNTNN}, \text{NNNTN}, \text{NNNNNT}\})$. 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.

(7 Aug 17)
You can see that a rv $T$ is completely specified by two things: the values it can take on, and the probability for those values.

(see p. 23) Definition. The (probability) distribution of a random variable $T$ is the collection of all probabilities of the form $P(T \in A)$ for all sets $A$ of real numbers in the non-void collection $C_{\mathbb{R}}$ of subsets of the real number line $\mathbb{R}$. The rv $T$ in the $T$-s we study has a finite set of possible values —
A random variable has a **discrete distribution**, or equivalently it is a discrete RV, if the set of (distinct) possible values is finite or at most countably infinite; RVs for which the set of possible values is uncountable are called continuous random variables.

Example: 1. The RV $X = \begin{cases} 1 & \text{if } Y > 0 \\ 0 & \text{otherwise} \end{cases}$ (with $Y = \# T's$ below) is discrete, taking on only the values $\{0, 1\}$—such RVs are called dichotomous or binary.
Imagine a scale for weighing things that has a dial you can set to specify how many significant figures of precision you want. Buy a "1 pound" package of butter at your favorite market and weigh it.

If there's no conceptual limit to the number of significant figures you could get, a rv \( Y = (\text{the actual (true) weight of the package}) \) should be modeled as continuous, having values (eg.) on \((0, \infty)\), the positive part of \( \mathbb{R} \).

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 conceptually continuous (i.e., no limit in principle to the precision of measurement) as continuous.

Definition: Given a discrete rv $\mathbb{I}$, the probability function (pmf or pf) of $\mathbb{I}$ is the function of that keeps track of the probability associated with $\mathbb{I}: f_{\mathbb{I}}(y) = P(\mathbb{I} = y)$. The set $\{y : f_{\mathbb{I}}(y) > 0\}$ is called the support of (the distribution of) $\mathbb{I}$.

(As is almost unique in using "pf", nearly everybody talks about the pmf.)
A rv \( Y \) that only takes on the values \( \{0, 1\} \) — i.e., a binary rv — is said to have a Bernoulli distribution with parameter \( p \) — written Bernoulli(\( p \)) — if

\[
    f(y) = P(Y = y) = \begin{cases} 
        p & \text{for } y = 1 \\
        1 - p & \text{for } y = 0 \\
        0 & \text{else}
    \end{cases}
\]

Notation: \( Y \) follows a Bernoulli(\( p \)) distribution \( \iff Y \sim \text{Bernoulli}(p) \) or \( (Y \mid p) \).
Example: In the powerball lottery (see problem 2) 5 white balls are drawn at random without replacement from a bin with balls numbered \( \{1, 2, \ldots, 69\} \). Let \( W_i \) be the number of \( i \)th drawn ball (\( i = 1, 2, \ldots, 5 \)).

Let \( W_i \) be the number of \( i \)th drawn ball (\( i = 1, 2, \ldots, 5 \)).

Clearly \( p(W_i = w_i) = \begin{cases} \frac{1}{69} & \text{for } w_i = 1, 2, \ldots, 69 \\ 0 & \text{otherwise} \end{cases} \)

less clearly (but true) \( W_1, \ldots, W_5 \) follow the same distribution if nothing is known about the previous draws.

Definition: For any two integers \( a \leq b \), a rv \( X \) that's equally likely to be any of the values \( \{a, a+1, \ldots, b\} \) has the uniform distribution \( \text{Uniform}\{a, b\} \). Evidently...
If $f(y) = P(I = y) = \begin{cases} \frac{1}{b-a+1} & \text{for } y = a, \ldots, b \\ 0 & \text{else} \end{cases}$

is distributed as

$X \sim \text{Uniform } \{a, b\} \iff X \text{ chosen at random from } \{a, a+1, \ldots, b\}$

Definition: $\textbf{\text{Binomial distribution}}$: $n$ trials are performed, with each trial recorded as a success or failure $F$. If each trial is independent of all the others and the chance of success is constant across the trials, then $Y = \# \text{ of successes}$.
In shorthand \( Y \sim \text{Binomial} \left( n, p \right) \) or \( (Y \mid n, p) \).

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

for \( i = 1, \ldots, n \). Then under these assumptions, \( B_i \sim \text{Bernoulli} \left( p \right) \) and all the \( B_i \) are independent.

Notation: \( X_i \sim f(x_i) \)

means that all of the rvs \( X_1, X_2, \ldots \) are independent and identically distributed.

Thus with the success/failure trials, \( B_i \sim \text{Bernoulli} \left( p \right) \) and \( (i = 1, \ldots, n) \)

\( (Y = \sum_{i=1}^{n} B_i) \sim \text{Binomial} \left( n, p \right) \).
This is our first example of the distribution of the sum 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 decimal numbers carry about 7 significant figures of accuracy,

\[
\pi \approx 3.141592653589
\]

leading to roundoff error in the last digit;

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 uniform distribution is key to these calculations.

Consider first

Uniform \{0, 0.1, \ldots, 0.9\} and then \{0, 0.1, 0.2, \ldots, 0.99\}
In the limit with more & more right

This should go to $f(u) = \begin{cases} 1 & \text{if } 0 \leq u \leq 1 \\ 0 & \text{else} \end{cases}$, the continuous uniform distribution

Uniform $(0,1)$ on the unit interval.

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

$$f(u) = \begin{cases} 1 & \text{if } 0 \leq u \leq 1 \\ 0 & \text{else} \end{cases}.$$  

Definition

A random variable $Y$ has a continuous distribution if there exists a continuous non-negative function $f_Y$ defined on $\mathbb{R}$ such that for every interval $[a,b)$, $P(a \leq Y \leq b) = \int_a^b f_Y(y) \, dy$. 

An analogue of summation is integration.
In this definition, $a$ can be $-\infty$ and $b$ can be $\infty$.

**Definition** If $\mathcal{I}$ is a continuous rv, the function $f_{\mathcal{I}}$ in the previous definition is called the probability density function (pdf) $f_{\mathcal{I}}(y)$ of $\mathcal{I}$. The set $\{y : f_{\mathcal{I}}(y) > 0\}$ is called the support of (the distribution of) $\mathcal{I}$.

2. Clearly (9) $f_{\mathcal{I}}(y) \geq 0$ for all $y$

and (1) $\int_{-\infty}^{\infty} f_{\mathcal{I}}(y) \, dy = 1$.

What about individual points - singletons - $\{y\}$ on $\mathbb{R}$?

You'll recall from calculus that if $f_{\mathcal{I}}$ is continuous...
on its support, \( \int_{9}^{b} f_{\Xi}(\gamma) \, d\gamma \) can equally well stand for \( P(9 \leq \Xi \leq b) \) or \( P(9 < \Xi \leq b) \) or \( P(9 \leq \Xi < b) \) or \( P(9 < \Xi < b) \), because (e.g.) \( \int_{9}^{b} f_{\Xi}(\gamma) \, d\gamma = 0 \) if \( f_{\Xi} \) is continuous at \( \gamma = 9 \). Thus, importantly \( P(\Xi = \gamma) = 0 \) for all \( -\infty < \gamma < \infty \).

weirdly, this doesn’t mean that the value \( \gamma \) of \( \Xi \) is impossible, or it does with discrete rv; it just means that singletons have 0 probability (otherwise \( \int_{-\infty}^{\infty} f_{\Xi}(\gamma) \, d\gamma = +\infty \) not 1).
Definition: With $a$ and $b$ any two real numbers satisfying $a < b$, a random variable $Y$ is distributed as Uniform$(a, b)$ if

$$Y \sim \text{Uniform}(a, b) \iff f_Y(y) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq y \leq b \\ 0 & \text{else} \end{cases}$$

Definition: The indicator function (true/false) for any proposition $A$ is $I(A) = \begin{cases} 1 & \text{if } A \text{ true} \\ 0 & \text{if } A \text{ false} \end{cases}$.

People sometimes also write (with $A$ a set)

$$I_A(y) = \begin{cases} 1 & \text{if } y \in A \\ 0 & \text{else} \end{cases}.$$
with this definition, \( Y \sim \text{Uniform}(a, b) \)

\[ f_Y(y) = \frac{1}{b-a} \cdot I(a \leq y \leq b) = \frac{I_{[a, b]}(y)}{b-a}. \]

\begin{align*}
\text{Contrast} & \quad \text{\( Y \sim \text{Uniform}(a, b) \) continuous} \\
& \quad \text{versus} \\
& \quad \text{\( Y \sim \text{Uniform}\{a, b\} \) for \( a, b \) integers} \\
& \quad \text{with} \ a < b \\
\end{align*}

\( Y \) discrete and uniform on \( \{a, a+1, \ldots, b\} \).

Density and probability are not the same thing.

Density value \( f_Y(y) \) are themselves not probabilities; for example, they can easily be > 1 and even be too,
Density values define probability: \( P(a \leq Y \leq b) = \int_{a}^{b} f_{Y}(y) \, dy \).

For small \( \varepsilon > 0 \) you can see from this sketch that
\[
P\left(a - \frac{\varepsilon}{2} \leq Y \leq a + \frac{\varepsilon}{2}\right) = \int_{a-\frac{\varepsilon}{2}}^{a+\frac{\varepsilon}{2}} f_{Y}(y) \, dy
\]

Example
(triangular distribution)
\[
f_{Y}(y) \text{ line with slope } m = \frac{c}{b-a}
\]
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 \( \frac{c}{b-a} = m \) and passes through the point \((a, 0)\), so the equation of the line is

\[
y - 0 = m(x - a) \iff y = \frac{c}{b-a} (x-a)
\]

Density have to integrate to 1, so

\[
\int_a^b \frac{c}{b-a} (x-a) \, dx = 1 \iff c = \frac{2}{b-a}
\]

We integrate \( \frac{c}{b-a} (x-a) \) for \( x = a \) to \( b \).

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

\[
1 = \frac{1}{2} (b-a) c \quad \text{and} \quad c = \frac{2}{b-a}
\]

(25 Apr 19)
The triangular distribution that starts at \( y = a \) and rises linearly to \( y = b \) has density \( f(y) = \begin{cases} \frac{2(y-a)}{(b-a)^2} & \text{for } a \leq y \leq b \\ 0 & \text{else} \end{cases} \)

Integrate \( \frac{2(y-a)}{(b-a)^2} \) 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 \( (I) \) above, what's \( P(a \leq X \leq \frac{b-a}{2}) \)?

Hard(?): \[ \int_{a}^{b-\frac{b-a}{2}} \frac{2(x-a)}{(b-a)^2} \, dx = \frac{(3b-3a)^2}{4(b-a)^2} \]

Easy(?): \[ \text{area of triangle} = \ldots \]
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. 185)

\[ Y = \text{voltage in an electrical system} \]

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. Give us an example: the pdf

\[
f_Y(y) = \frac{1}{(1+y)} I(y > 0).
\]

Check:

\[
\int_0^{\infty} \frac{1}{(1+y)} \, dy = 1.
\]
You can check that \( \int_{-\infty}^{\infty} \frac{1}{1+x^2} \, dx = 1 \),

so the right tail beyond \( Z = 1000 \) 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. Example:

In medical clinical trials of people
with potentially fatal diseases, the
outcome variable \( Z \), for person \( j \) in
(say) the treatment group might be
$T$: survival time in days from beginning of the trial; however, one good thing is, some patients may still be alive at the time $T_{\text{end}}$, at which the trial finishes. Your model for $T_i$ would then have a continuous part for $0 \leq T_i \leq T_{\text{end}}$ and a discrete lump of probability $p$ at $T_i = T_{\text{end}}$,signifying $(T_i > T_{\text{end}})$ but we don't know what $T_i$ would have been if we could have observed it:

$$
\int_0^{T_{\text{end}}} f_{X_i}(\gamma) \, d\gamma = (1-p) \quad \text{and} \quad P(T_i > T_{\text{end}}) = p.
$$

(For example) (see p. 10 of doc. can notes)
Unifying idea connecting discrete & continuous rvs

Discrete $\iff$ p.f (pmf)
Continuous $\iff$ p.d.f
Mixed $\iff$ (p.f + p.d.f)

Q: Is there something that uniquely characterizes the distribution of $\Xi$, both when $\Xi$ is discrete & when it’s continuous & when it’s mixed?

A: Yes, the cumulative distribution function $F_\Xi(\gamma)$

Definition:

The cumulative distribution function (cdf) of a rv $\Xi$ is defined to be

$$F_\Xi(\gamma) = P(\Xi \leq \gamma) \text{ for all } -\infty < \gamma < \infty$$
Example: $X \sim \text{Bernoulli}(p)$

\[ P(X = y) = \begin{cases} p & \text{for } y = 1 \\ 1-p & \text{else} \\ 0 & \text{PMF} \end{cases} \]

Notice that there's a clever way to write this:

\[ P(X = y) = p^y (1-p)^{1-y} I_{\{0,1\}}(y) \]

The CDF of $X$ is 0 for $y < 0$; at $y = 0$ it jumps up to $(1-p)$ and stays there for $0 \leq y < 1$; and at $y = 1$ it jumps up to 1 and stays there for $y \geq 1$.

You can see that in general \[ 0 \leq F_X(y) \leq 1 \]

\[ F_X(y) = \begin{cases} 0 & y < 0 \\ 1-p & 0 \leq y < 1 \\ 1 & y \geq 1 \end{cases} \]
and it's also clear that a cdf \( F_\Xi(\gamma) \) has to be a non-decreasing function of \( \gamma \): if \( \gamma_1 < \gamma_2 \), then \( F_\Xi(\gamma_1) \leq F_\Xi(\gamma_2) \).

Furthermore, \( \lim_{\gamma \to -\infty} F_\Xi(\gamma) = 0 \) and \( \gamma \to +\infty \)

\[
\lim_{\gamma \to +\infty} F_\Xi(\gamma) = 1.
\]

CDFs can be continuous or continuous on all of \( \mathbb{R} \) but certainly don't have to be (see the cdf of the Bernoulli (\( p \)) distribution). Technical fact:

Def: \( F_\Xi(\gamma^-) = \lim_{\gamma^* \to \gamma^+} F_\Xi(\gamma^*) \)

limit from the left

\( \gamma^* \to \gamma \)

\( \gamma \) is the left endpoint of the interval. Technical fact:
Def. $F_Z(y^+) = \lim_{y^+ \to y} F_Z(y^*) = \lim_{y^- \to y} F_Z(y^-)$

**Technical fact:** $F_Z(y) = F_Z(y^+)$ for all $-\infty < y < \infty$.

People call this continuity from the right or continuity from above.

**Consequences of the cdf definition**

1. $P(Z > y) = 1 - F_Z(y)$
2. For all $y_1, y_2$ with $y_1 < y_2$
   
   $P(y_1 < Z \leq y_2) = F_Z(y_2) - F_Z(y_1)$.

If

$F_Z(y^-) = F_Z(y^+) = F_Z(y)$

then $F_Z$ is continuous.
Consequence 2 means that if \( Y \) is continuous, there's an intimate connection between \( F_X(y) \) and \( f_X(y) \): 

\[
F_X(y) = \int_{-\infty}^{y} f_X(t) \, dt
\]

If \( X \) is a continuous rv with pdf \( f_X(y) \) and cdf \( F_X(y) \) then

\[
F_X(y) = \int_{-\infty}^{y} f_X(t) \, dt
\]

and thus

\[
P(y_1 < X \leq y_2) = F_X(y_2) - F_X(y_1) = \int_{y_1}^{y_2} f_X(y) \, dy
\]
In other words, the derivative of $F_Z(y)$ is $f_Z(y)$ (and $F_Z(y)$ is an anti-derivative of $f_Z(y)$).

**Definition**

$Y$ follows an exponential distribution with parameter $\lambda > 0$ if

\[
f_Z(y) = \begin{cases} \lambda e^{-\lambda y} & y > 0 \\ 0 & y \leq 0 \end{cases}
\]

The exponential distribution 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: \( \gamma \)

\[ \gamma \text{ exponentially distributed (with parameter) } \quad \iff \quad \gamma \sim \text{Exponential}(\lambda) \]

\[ \lambda > 0 \quad \text{for } \gamma > 0 \]

\[ F_{\gamma}( \gamma ) = \int_{-\infty}^{\gamma} f_{\gamma}(t) \, dt = \int_{0}^{\gamma} 2e^{-\lambda t} \, dt \]

\[ = 2e^{-\lambda \gamma} \left[ \frac{-1}{\lambda} \right]_{0}^{\gamma} = 1 - e^{-\lambda \gamma} \]

\[ \Rightarrow F_{\gamma}( \gamma ) = \begin{cases} 0 & \text{for } \gamma < 0 \\ 1 - e^{-\lambda \gamma} & \text{for } \gamma \geq 0 \end{cases} \]

(30 Apr 19)
This inverse time: CDFs.
next joint time: distributions

\[ f_{\mathbf{X}}(x) \]

cont.

\[ P(\mathbf{X} = x_i) = 0 \]

1 rv

cont. \[ \frac{f_{\mathbf{X,Y}}(x,y)}{\text{cont.}} \]

2-dim.

\[ c e \]

\[ -(x^2 + y^2) \]

C E

2D contour plot of

\[ x \]

\[ y \]

perspective plot

symmetry = median

mean

point of mode

50% 50%
\[ P(Z \geq z) = \int_{z}^{\infty} \int_{\mathbb{R}} f_{X,Y}(x,y) \, dy \, dx \]
\[ = \int_{z}^{\infty} \int_{\mathbb{R}} \frac{1}{4} x^2 \, dy \, dx \]
Q: What's the place \( y_p \) on the positive part of \( \mathbb{R} \) where \( P(0 \leq \Xi \leq y_p) = p \)?

Well, \( P(0 \leq \Xi \leq y_p) = F(\gamma_p) = p \)

\[
\begin{align*}
\gamma_p &= 1 - e^{-2\gamma_p} \\
1 - p &= e^{-2\gamma_p} \\
\log(1 - p) &= -2\gamma_p \\
\gamma_p &= -\log(1 - p) / 2
\end{align*}
\]

So, \( \gamma_p = F^{-1}(p) \)

Def. \( \gamma_p \) is called the \( p \text{-th percentile} \) of the distribution of \( \Xi \).
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

$$ F_Y^{-1}(p) = \text{the smallest } y \text{ value such that } F_Y(y) \geq p $$

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

Measures of center for the distribution of a rv $Y$

One way to define the center of a distribution is to find the $50^{th}$ percentile.
Definition: The $\frac{1}{2}$ quantile is the 50th percentile of a distribution, is called the median of the dist.

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

Definition: The $\frac{1}{4}$ quantile = the 25th percentile, is the lower quantile; the $\frac{3}{4}$ quantile = the 75th percentile is the upper quantile; and $(F_X^{-1}(0.75) - F_X^{-1}(0.25)) = $ interquartile range (IQR)
Example: \( Y \sim \text{Uniform}(a, b); \) then

\[
F_Y(y) = \begin{cases} 
0 & \text{for } y \leq a \\
\frac{y - a}{b - a} & \text{for } a \leq y \leq b \\
1 & \text{for } y \geq b 
\end{cases}
\]

Easy to invert \( F_Y \):

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

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

Studying two random variables at a time:

**Def.** \( X, Y \) rvs: the joint (or bivariate) distribution of \((X, Y)\) is the collection \( P[(X, Y) \in C] \) of all probabilities for all sets \( C \subseteq \mathbb{R}^2 \) such that \( (X, Y) \in C \) isn't weird.
Case 1 \[ (X \text{ and } Y \text{ both discrete}) \]

Def. \( X, Y \text{ rv. } \Rightarrow \text{ If there are only finitely or countably infinitely many possible values } (x, y) \text{ for } (X, Y), \) \( X \) and \( Y \) have a discrete joint dist.

Def. The joint probability function \( f(x, y) \) of \( (X, Y) \) discrete is the function \( f(x, y) = P(X = x, Y = y) \).

The set \( \{(x, y) : f(x, y) > 0\} \) is the support of \( f(x, y) \).

Consequences

1. \[ \sum_{x, y} f(x, y) = 1 \]
2. For any set \( C \) of ordered pairs \( (x, y) \), \[ P((X, Y) \in C) = \sum_{(x, y) \in C} f_{X,Y}(x, y) \]
two rv $X$ and $Y$ have a continuous joint distribution

Case 2: if you can find a nonnegative function $f_{X,Y}(x,y)$ defined for all $(x,y) \in \mathbb{R}^2$ (the real plane) such that for every (non-weird) subset $C$ of the plane $P[\{X,Y\} \in C] = \iiint_C f_{X,Y}(x,y) \, dx \, dy$

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

Immediate Consequences:
1. For all $(x,y)$ in $\mathbb{R}^2$,

\[ f_{X,Y}(x,y) \geq 0 \quad \text{and} \quad \iint_{-\infty}^{\infty} \iint_{-\infty}^{\infty} f_{X,Y}(x,y) \, dx \, dy = 1 \]
2. If \((X,Y)\) have a continuous joint distribution, \(X\) and \(Y\) each have a continuous (marginal) univariate distribution when considered separately.

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

(a) Every individual point, and every countably infinite sequence of points, 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.
This means that the converse of (2) is (unfortunately) not true: if $X$ has a continuous distribution on $\mathbb{R} = \mathbb{R}^1$ and $Y \equiv X$, then both $X$ and $Y$ have continuous distributions but $P[(X, Y) \text{ lies on the line } y = x] = \frac{1}{2}$.

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

**Example**

Suppose that $(X, Y)$ have joint pdf

$$f_{X,Y}(x, y) = \begin{cases} \frac{2}{xy} & \text{for } 0 \leq x, y \leq 1 \\ 0 & \text{else} \end{cases}$$

Let's work out the normalizing constant.

The support of $f_{X,Y}$ is the shaded region in the diagram.
\[ 1 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{\mathcal{N}}(x, \gamma) \, dx \, d\gamma = \]

\[ = \iint \int_{\mathcal{N}} f_{\mathcal{N}}(x, \gamma) \, d\gamma \, dx \]

\[ = \int_{-1}^{1} \int_{-1}^{1} c x^2 \, d\gamma \, dx \]

\[ = \int_{-1}^{1} c x^2 \left( \int_{-1}^{1} \gamma \, d\gamma \right) \, dx \]

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

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

\[ = \frac{1}{2} c \left( \frac{x^3}{3} \bigg|_{-1}^{1} \right) - \frac{c}{2} \left( \frac{x^7}{7} \bigg|_{-1}^{1} \right) = \frac{4}{21} c = 1 \]
So \( c = \frac{21}{4} \)

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

\[
1 = \int_{S} f_{\mathbb{E}^2}(x, y) \, dx \, dy
\]

\[
= \int_{0}^{1} \int_{-\sqrt{y}}^{\sqrt{y}} c x^2 y \, dx \, dy
\]

\[
= \int_{0}^{1} \left[ \frac{1}{3} x^3 \right]_{-\sqrt{y}}^{\sqrt{y}} y \, dy
\]

\[
= \int_{0}^{1} \frac{1}{3} y \left( y^{3/2} - (-y^{3/2}) \right) \, dy
\]

\[
= c \int_{0}^{1} y \cdot \frac{1}{3} (y^{1/2} - y^{3/2}) \, dy
\]