INFN School of Statistics 2019
Paestum, Italy

Probability Theory - 2

Harrison B. Prosper

Department of Physics, Florida State University

June 3, 2019

Outline

Outline

In most realistic applications, it is convenient to think of probabilities as functions of the outcomes. To every possible outcome x, a probability is assigned via an unambiguous rule.

1. Probability Function

Any function $f(x)$ that assigns a probability $P(x)$ to outcome x is called a probability function. In the previous lecture, we considered outcomes that can be modeled as n-tuples, $(z_1, \cdots, z_n)$, with elements drawn from the set of natural numbers $\mathbb{N} = \{0, 1, \cdots, \aleph_0\}$. But, we can also model outcomes using n-tuples with elements drawn from the set of real numbers $\mathbb{R} = (-c, c)$[a].

[a] Georg Cantor (1845 - 1918), inventor of set theory, proved the astonishing theorem $c = 2^{\aleph_0}$, that the cardinalities c and $\aleph_0$ of sets $\mathbb{R}$ and $\mathbb{N}$, respectively, are related in this amazing way. We typically use the symbol ∞ instead of c.

2. Probability Mass Function

If x is from $\mathbb{N}$, then the probability function $f(x)$ is called a probability mass function (pmf).

3. Probability Density Function

If x is from $\mathbb{R}$, a continuous set, then the probability function $f(x)$ is called a probability density function (pdf) and is often written with a lower case letter, e.g., p. For continuous sets $f(x)$ does not *directly* assign a probability.

To do so, we integrate the pdf $f(x)$ over an interval that is at least as large as an infinitesimal dx[a]. In practice, we compute

$$P(x) = \int_{x_1}^{x_2} f(X)\, dX.$$

[a] An infinitely small non-zero number!

Random Variables

Formal books on statistics often distinguish between a random variable X, denoted with an upper case letter, and its outcomes x, denoted by lower case letters. However, I shall not make that distinction unless it makes things clearer.

Randomness

What do we mean when we say that an outcome is random?

Consider the time of decay of an excited atom. According to quantum mechanics, no rule exists that determines *when* such a thing happens. The time of decay is an example of a "causeless" effect!

On the other hand, we couldn't do our work without the use of strictly deterministic rules that simulate random outcomes! (Think `TRandom3`.)

Randomness...

Suppose there are 10^{100} particles in the known universe undergoing random changes of state on the average 10^{30} times per second. That amounts to 10^{130} random changes of microstates per second. Now, suppose that the universe, as we know it, will last 10^{20} seconds.

During this immense span of time, the universe will have undergone 10^{150} random changes of microstates.

Question: If the universe is a giant simulation governed by a universal random number generator with a Poincaré cycle[1] that is $> 10^{150}$, how could we distinguish that universe from one with causeless effects? Discuss, but, not now!

[1] A sequence of states that returns to the initial state after, typically, passing through an immense number of states.

More Definitions

There are several numbers that can be used to characterize a probability function. Here are a few.

4. Moments

The r^{th} moment $\mu_r(a)$ about a of $f(x)$ is defined by[a]

$$\mu_r(a) = \int_{S_x} (x-a)^r \, f(x) \, dx,$$

where S_x is the domain[b] of $f(x)$.

$\mu = \mu_1(0)$ is called the mean and is one measure of where the function lies; $\mathrm{Var}_x = \mu_2(\mu)$ is called the variance; $\sigma = \sqrt{\mathrm{Var}_x}$, the standard deviation, is one measure of the width of $f(x)$.

[a] For discrete distributions, we replace the integral by a sum.

[b] The domain of a function is the set of its "input" values. The range is the set of its "outputs".

Yet More Definitions

5. Quantile Function

The function

$$D(x) = \int_{X \leq x} f(X)\, dX$$

is called the cumulative distribution function (cdf) of $f(x)$, or sometimes simply the distribution function. (Here is an example where distinguishing between X and x is helpful.) The function $x = Q(P)$ which returns x, given $D(x) = P$, is called the quantile function and x is called the P-quantile of $f(x)$.

Sometimes it is convenient to distinguish between the left cdf $D_L(x) \equiv D(x)$ and the right cdf defined by

$$D_R(x) = \int_{X \geq x} f(X)\, dX.$$

And Yet More Definitions...!

6. Covariance, Correlation, Independence

The covariance of random variables x and y with probability function $f(x, y)$ is defined by

$$\mathrm{Cov}_{xy} = \int_{S_x} \int_{S_y} (x - \mu_x)(y - \mu_y) \, f(x, y) \, dx \, dy.$$

Cov_{xy} is a measure of the correlation between the variables x and y.

If the probability function $f(x, y)$ can be written as $f(x, y) = f(x) \, f(y)$ then the variables x and y are said to be independent in which case $\mathrm{Cov}_{xy} = 0$.

However, in general, $\mathrm{Cov}_{xy} = 0$ does not imply independence.

Outline

Example (2.1 The Binomial Distribution)

Consider m proton-proton collisions at the LHC of which r are successes. A success might be the creation of a Higgs boson that decays to four leptons. Suppose, we are able to record $n \leq m$ collision events of which $k \leq n$ are successes.

What is the probability $P(k, n|r, m)$ of recording exactly k successes and exactly $n - k$ failures in n trials, that is, collisions, given r successes and $m - r$ failures?

There is no solution unless we are prepared to make some assumptions, for example:

1. The problem is the same as drawing k blue balls and $n - k$ red balls from a box containing r blue balls and $m - r$ red balls.
2. Each draw of n objects of which k are successes is a single outcome.
3. Each outcome is equally probable.

Example (2.1 The Binomial Distribution)

Let's follow a counting strategy, like the one we used in the previous lecture.

1. Count the number of ways T to draw samples of size n from m collisions.
2. Count the number of ways S to draw exactly k successful collisions from r successes and exactly $n - k$ failed collisions from $m - r$ failures.
3. By assumption, outcomes are equally probable; therefore, the probability of the specific outcome (k, n) given (r, m) equals S/T.

Solution

1. How many samples of size n can be drawn from a sample of size m?

$$\binom{m}{n}$$

2. How many samples of size k can be drawn from r successes?

$$\binom{r}{k}$$

3. How many samples of size $n-k$ can be drawn from $m-r$ failures?

$$\binom{m-r}{n-k}$$

Solution contd.

$$
\begin{aligned}
P(k, n|r, m) &= \text{number of favorable outcomes} \\
&\quad / \text{ number of possible outcomes} \\
&= \binom{r}{k}\binom{m-r}{n-k} \Big/ \binom{m}{n} \\
&= \binom{n}{k} f(k, n, r, m), \\
\text{where } f(k, n, r, m) &= \frac{r!}{(r-k)!} \frac{(m-r)!}{(m-r-n+k)!} \Big/ \frac{m!}{(m-n)!}.
\end{aligned}
$$

But, what we really want to know is the probability $P(k, n)$ of k successes given n trials regardless of the values of r and m, which is just as well because we don't know r and we know m only approximately.

In order to compute $P(k, n)$, the probability rules require that we calculate the sum

$$P(k, n) = \sum_{r,m} P(k, n | r, m) \, P(r, m).$$

What is $P(r, m)$? It is the probability of the sequence of r successes in m collisions at the LHC.

Unfortunately, it is far from clear how to arrive at sensible values for $P(r, m)$! Since we don't know what to do with $P(r, m)$, we'll leave it unspecified.

But, note the consequence. Different choices for $P(r, m)$ will imply different values for the probability of k successes in n trials for the *same* data (k, n)! This seems unavoidable.

Let's press on regardless!

At the LHC, m is huge and generally $r << m$. For the Higgs boson at 13 TeV, theory predicts and experiment confirms that $z = r/m \approx 10^{-10}$. It therefore makes sense to consider the idealization $m \to \infty$ in the expression

$$P(k, n) = \sum_{r,m} P(k, n|r, m)\, P(r, m).$$

To that end, let's write the above in terms of the unknown relative frequency of success $z = r/m$:

$$P(k, n|z, m) = \binom{n}{k} f(k, n, z, m),$$

$$\text{where } f(k, n, z, m) = \frac{(zm)!}{(zm-k)!} \frac{(m-zm)!}{(m-zm-n+k)!} \Big/ \frac{m!}{(m-n)!},$$

and let $m \to \infty$ while keeping k and n fixed.

We can rewrite $f(k, n, z, m)$ as

$$f(k,n,z,m) = \frac{(zm)!}{(zm-k)!}\frac{(m-zm)!}{(m-zm-n+k)!}\Big/\frac{m!}{(m-n)!},$$
$$= z^k(1-z)^{n-k}$$
$$\times \frac{\left[\prod_{i=0}^{k-1}(1-i/(zm))\right]\left[\prod_{i=0}^{n-k-1}(1-i/(m(1-z)))\right]}{\prod_{i=0}^{n-1}(1-i/m)},$$
$$\to z^k(1-z)^{n-k} \quad \text{as } m \to \infty.$$

What happens to the probabilities $P(r, m)$? To see what happens, write $P(k, n)$ as a double sum over the rational numbers z and the integers m

$$P(k,n) = \sum_{r,m} P(k,n|r,m)\,P(r,m),$$
$$= \sum_z \sum_m P(k,n|zm,m)\,P(zm,m) \text{ with } z = r/m.$$

$$P(k,n) = \sum_z \sum_m P(k,n|zm,m)\, P(zm,m) \text{ with } z = r/m,$$
$$= \sum_z \binom{n}{k} z^k (1-z)^{n-k}\, \pi(z),$$

where $\pi(z) \equiv \sum_m P(zm,m)$ is called a prior density. As $m \to \infty$, the sum over the relative frequencies z converges to an integral and we obtain:

Bruno de Finetti's Representation Theorem

$$P(k,n) = \int_0^1 \text{binomial}(k,n,z)\, \pi(z)\, dz,$$

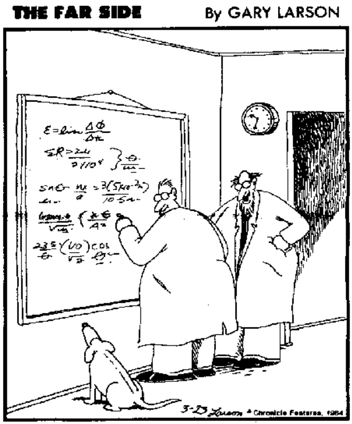
where $\text{binomial}(k,n,z) = \binom{n}{k} z^k (1-z)^{n-k}$.

The Binomial Distribution

What are we to make of the prior density $\pi(z)$? Well, we could ask our friendly theorist for a prediction of the relative frequency z of Higgs boson production at the LHC at 13 TeV. Suppose she accurately predicts that $z = p$.

We might consider modeling this information by setting $\pi(z) = \delta(z - p)$ in de Finetti's theorem. If we do so, we obtain the binomial distribution

$$\boxed{P(k, n) = \mathrm{binomial}(k, n, p)}$$



"Ohhhhhhh . . . Look at that, Schuster . . . Dogs are so cute when they try to comprehend quantum mechanics."

Example (2.2 Poisson Distribution)

The Poisson distribution is arguably the most important distribution in particle physics and astronomy. As you will see in subsequent lectures, the distribution features prominently in the counting of rare events.

The Poisson distribution can be derived from a stochastic model. It can also be derived from the binomial distribution.

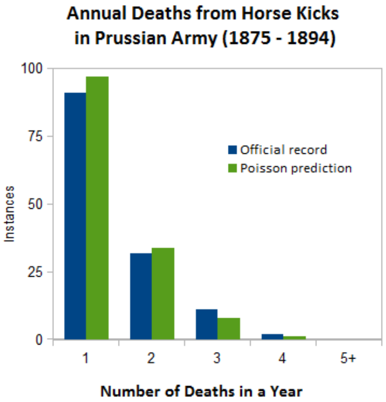
Exercise 3

Derive the Poisson distribution from the binomial distribution assuming that $p \to 0$, while $n \to \infty$ such that $a = np$ is constant.

Historical Aside

In 1898, the Russian economist von Bortkiewicz published a book in which he presented data on the number of deaths per annum in the Prussian Army from horse kicks.

von Bortkiewicz noted that the distribution of observed counts could be modeled by the distribution first described (in 1837) by Siméon Poisson (1781 - 1840).[2]



[2] https://mindyourdecisions.com/blog/2013/06/21/what-do-deaths-from-horse-kicks-have-to-do-with-statistics/

The Poisson Distribution as a Stochastic Process

Suppose that at time $t + dt$ we have recorded n counts and that in the time interval $(t, t + dt)$ only two things can happen:

1. we had n counts at time t and recorded none during $(t, t + dt)$, or
2. we had $n - 1$ counts at time t and recorded one count during $(t, t + dt)$.

Being kicked to death by a horse is rare as are Higgs bosons, so the chance of having more than one occur during a short time interval is assumed to be negligibly small.

Let's further suppose that the probability to get an event during the specified time interval of size dt is proportional to its size.

We can now assign probabilities.

The Poisson Distribution as a Stochastic Process

Here are the transition probabilities that define the Poisson process:

$$
\begin{aligned}
P_n(t+dt) &= \text{probability that the count is } n \text{ at time } t+dt \\
P_n(t) &= \text{probability that the count is } n \text{ at time } t \\
P_{n-1}(t) &= \text{probability that the count is } n-1 \text{ at time } t \\
q dt &= \text{probability to record 1 event during } (t, t+dt) \\
1 - q dt &= \text{probability to record 0 events during } (t, t+dt)
\end{aligned}
$$

In principle, q could depend on time, but we shall assume it does not. Using the probability rules, we can write

$$P_n(t+dt) = (1 - qdt)\, P_n(t) + qdt\, P_{n-1}(t),$$

or noting that $dP_n(t)/dt = [P_n(t+dt) - P_n(t)]/dt$ [3]

$$\frac{dP_n}{dt} = -q\, P_n + q\, P_{n-1}.$$

[3] Apparently, manipulating infinitesimals like this is legitimate mathematics!

The Poisson Distribution as a Stochastic Process

Equations such as

$$\frac{dP_n}{dt} = -q\,P_n + q\,P_{n-1},$$

can be solved recursively, noting that $P_{-1} = 0$.

Exercise 4

Show that

$$\boxed{P_n(t) = \mathrm{Poisson}(n, a) = \frac{e^{-a}a^n}{n!}},$$

where the mean count is $a = qt$. Also, show that $\mathrm{Var}_n = a$, an important fact about the Poisson distribution.

Outline

Gaussian Distribution

The probability density function of the Gaussian distribution, also known as the normal distribution, is

$$\boxed{\mathrm{N}(\mu,\sigma) \equiv \mathrm{Gaussian}(x,\mu,\sigma) = \frac{e^{-\frac{1}{2}(x-\mu)^2/\sigma^2}}{\sigma\sqrt{2\pi}}}$$

with mean μ and variance σ^2. Also note, where $z = (x-\mu)/\sigma$,

$$\begin{aligned} P(z \in [-1.00, 1.00]) &= 0.683 \\ P(z \in [-1.64, 1.64]) &= 0.900 \\ P(z \in [-1.96, 1.96]) &= 0.950 \\ P(z \in [-2.58, 2.58]) &= 0.990 \\ P(z \in [-3.29, 3.29]) &= 0.999 \\ P(z \in [5.00, \infty)) &= 2.7 \times 10^{-7} \end{aligned}$$

The Gaussian is the most important distribution in statistics...

Gaussian Distribution

...because all sensible probability distributions approach a Gaussian in some limit. The precise statement is the central limit theorem.

Example (2.3 The Central Limit Theorem)

Consider the average $t = \frac{1}{n}\sum_{i=1}^{n} x_i$, where $x_i \sim p(\mu, \sigma)$ and $p(\mu, \sigma)$ is any probability density with finite mean μ and standard deviation σ.

Define the variable $z = (t - \mu)/(\sigma/\sqrt{n})$. The mean of its probability density, $f(z)$, is 0 and its standard deviation is 1. The central limit theorem states that

$$\boxed{\lim_{n\to\infty} P(z < x) = \int_{-\infty}^{x} \mathrm{Gaussian}(X, 0, 1)\, dX}.$$

χ^2 **Distribution** Write $z = (x - \mu)/\sigma$, where $x \sim \mathrm{N}(\mu, \sigma)$ and consider the sum

$$t = \sum_{i=1}^{n} z_i^2.$$

What is the probability density function of t? For any well-behaved probability density function, $p(z_1, \cdots, z_n)$, the pdf of t, $p(t)$, is given by the random variable theorem[4]

$$\boxed{p(t) = \int dz_1 \cdots \int dz_n \, \delta\left(t - g(z_1, \cdots, z_n)\right) \, p(z_1, \cdots, z_n)},$$

where $g(z_1, \cdots, z_n)$ is the function, such as the sum above, that maps z_1 to z_n to t. Here's a brief proof of the theorem...

[4] *A theorem for physicists in the theory of random variables*, D. Gillespie, Am. J. of Phys. **51**, 520 (1983).

Theorem

$$p(t) = \int dz_1 \cdots \int dz_n \, \delta\left(t - g(z_1, \cdots, z_n)\right) \, p(z_1, \cdots, z_n).$$

Proof.

Step 1. Stare at the above for about a minute ...
Step 2. ... and conclude it is intuitively obvious and therefore true!
□

Having "proved" the theorem, let's apply it to our problem!

On the next slide, we shall use the amazingly useful formula

$$\boxed{\delta(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{i\omega x} \, d\omega},$$

which every one of you should memorize!

The z_j are independent random variables. Therefore, using the representation of the δ-function on the previous slide, we can write

$$
\begin{aligned}
p(t) &= \int dz_1 \cdots \int dz_n \, \delta\left(t - g(z_1, \cdots, z_n)\right) p(z_1, \cdots, z_n) \\
&= \frac{1}{2\pi} \int_{-\infty}^{\infty} d\omega \, e^{i\omega t} \prod_{j=1}^{n} \int_{-\infty}^{\infty} e^{-i\omega z_j^2} \, p(z_j) \, dz_j, \\
&= \frac{1}{2\pi} \int_{-\infty}^{\infty} d\omega \, e^{i\omega t} \left(\int_{-\infty}^{\infty} e^{-i\omega z^2} \frac{e^{-z^2/2}}{\sqrt{2\pi}} \, dz \right)^n, \\
&= \frac{1}{2\pi} \int_{-\infty}^{\infty} d\omega \, e^{i\omega t} \left(\frac{1}{\sqrt{2i\omega + 1}} \int_{-\infty}^{\infty} \frac{e^{-(2i\omega+1)z^2/2}}{\sqrt{2\pi}} \, d\sqrt{2i\omega + 1}z \right)^n, \\
&= \frac{1}{2\pi} \int_{-\infty}^{\infty} d\omega \, \frac{e^{i\omega t}}{(2i)^{n/2}} \frac{1}{(\omega - i/2)^{n/2}}.
\end{aligned}
$$

χ^2 Distribution

The integrand has a singularity in the complex plane at $\omega = i/2$

$$p(t) = \frac{1}{2\pi i} \int_{-\infty}^{\infty} d\omega \, \frac{i e^{i\omega t}}{(2i)^{n/2}} \, \frac{1}{(\omega - i/2)^{n/2}}.$$

Simple poles of order n that lie away from the real line can be handled using the following magical result,

$$\frac{1}{2\pi i} \int_{-\infty}^{\infty} d\omega \, F(\omega) \, \frac{1}{(\omega - \omega_0)^n} = \lim_{\omega \to \omega_0} \frac{1}{(n-1)!} \frac{d^{n-1}}{d\omega^{n-1}} F(\omega).$$

Alas, for odd n our singularity involves an annoying square-root.

No worries! We merely follow the time-honored strategy of physicists: solve a simpler problem then generalize its solution by inspection!

χ^2 Distribution Let's compute

$$p(t) = \frac{1}{2\pi i} \int_{-\infty}^{\infty} d\omega \, \frac{ie^{i\omega t}}{(2i)^m} \frac{1}{(\omega - i/2)^m},$$

for integer m. We have a pole singularity of order m and, therefore,

$$\frac{1}{2\pi i} \int_{-\infty}^{\infty} d\omega \, \frac{ie^{i\omega t}}{(2i)^m} \frac{1}{(\omega - i/2)^m} = \frac{1}{(m-1)!} \frac{i}{(2i)^m} (it)^{m-1} e^{-t/2},$$
$$= \frac{1}{\Gamma(m)} \frac{t^{m-1} e^{-t/2}}{2^m}.$$

The result is valid for non-integer values of m; it is therefore the solution for $m = n/2$. We conclude that the pdf of the sum of squares of n standardized Gaussian random variables is the χ^2 density ($t = \chi^2$) of n degrees of freedom,

$$\boxed{p(t) = \frac{1}{\Gamma(n/2)} \frac{t^{n/2-1} e^{-t/2}}{2^{n/2}}, \ \text{mean } n, \ \text{variance } 2n}$$

Cauchy Distribution

Let $x, y \sim \mathrm{N}(0,1) \equiv g(x) = \exp(-x^2/2)/\sqrt{2\pi}$. The pdf of $t = y/x$ is given by

$$\begin{aligned} p(t) &= \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy\, \delta(t - y/x)\, g(x)\, g(y), \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} dx\, e^{-\frac{1}{2}x^2} \left[\int_{-\infty}^{\infty} dy\, \delta(t - y/x)\, e^{-\frac{1}{2}y^2} \right]. \end{aligned}$$

The integral in the brackets is of the form

$$\int \delta(h)\, f(y)\, dy,$$

where the δ-function argument h is a function of y. A more tractable form of the δ-function can be obtained from the identity

$$\int \delta(h)\, dh = \int \delta(h) \left| \frac{dh}{dy} \right| dy.$$

$$\int \delta(h)\, dh = \int \delta(h) \left| \frac{dh}{dy} \right| dy.$$

The function $\delta(h)|dh/dy|$ is zero everywhere except at $y = y_0$, where y_0 is the solution of $h(y) = 0$. By definition, this function is $\delta(y - y_0)$.

Therefore, $\delta(h) = \delta(y - y_0)/|dh/dy|$ and the integral becomes

$$\int \delta(h)\, f(y)\, dy \rightarrow \int \delta(y - y_0)/|dh/dy|\, f(y)\, dy.$$

For our problem, $h(y) = t - y/x$, $|dh/dy|^{-1} = x$, $f(y) = \exp(-\frac{1}{2}y^2)$, and the solution of $h(y) = 0$ is $y_0 = xt$.

Therefore,

$$
\begin{aligned}
p(t) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} dx\, e^{-\frac{1}{2}x^2}\left[\int_{-\infty}^{\infty} dy\, \delta(t - y/x)\, e^{-\frac{1}{2}y^2}\right], \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty} dx\, e^{-\frac{1}{2}x^2}\left[\int_{-\infty}^{\infty} dy\, \delta(y - y_0)\, x\, e^{-\frac{1}{2}y^2}\right], \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty} dx\, e^{-\frac{1}{2}x^2}\, xe^{-\frac{1}{2}(xt)^2}, \\
&= \frac{1}{\pi(1+t^2)}\int_{0}^{\infty} e^{-x^2(1+t^2)/2}\, d(x^2(1+t^2)/2),
\end{aligned}
$$

from which we conclude that the pdf of the ratio of two standardized Gaussian random variables is the rather pathological Cauchy density,

$$
\boxed{p(t) = \frac{1}{\pi(1+t^2)}}
$$

Exercise 5 – Example of the Central Limit Theorem

Consider the sum $t = \sum_{j=1}^{n} x_j$, where the x_j are independent and identically distributed (iid) random variables from the uniform distribution with density $U(\mu, \sigma) = 1/2$ defined on the domain $x \in (-1, 1)$. We want the pdf of t, $p(t)$, to have unit variance.

1. What is σ^2 for U?
2. Show that $a = 3/n$ if the variance of $p(t)$ has unit variance.
3. Show that $p(t)$ can be written as
$$p(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{i\omega t} \mathrm{sinc}^n(a\omega)\, d\omega, \ \mathrm{sinc}(x) \equiv \sin x / x.$$
4. Then show that
$$\lim_{n \to \infty} p(t) = e^{-t^2/2} / \sqrt{2\pi}.$$

Outline

Summary

- According to Kolmogorov, probabilities are functions whose domain are suitable collections of sets and whose range is the unit interval.
- To be useful, however, probabilities must be interpreted. The most common interpretations are: relative frequency and degree of belief.
- If it is possible to decompose experimental outcomes (basically, a set of n-tuples) into outcomes that are considered equally probable, then the probability of an outcome may be taken to be the ratio of the number of outcomes with the desired characteristics to that of all possible outcomes.
- More generally, we use probability functions; probability mass functions for discrete distributions and probability density functions for continuous ones.