Statistics for Particle Physicists
LPC, Fermilab

Probability - Part 2

Harrison B. Prosper

Department of Physics, Florida State University

September 20, 2017

Outline

Outline

In the previous lecture, we talked about probabilities as individual numbers. But in most applications, it is convenient to think of them as functions of the outcomes: to every possible outcome x, we specify the probability via an unambiguous rule.

1. Probability Function

A function $f(x)$ that gives a rule for assigning a probability $P(x)$ to outcome (event) x is called a probability function. So far, we have talked only about outcomes that can be modeled with n-tuples, $(z_1, \cdots, z_n)$, whose elements are drawn from the set of natural numbers $\mathbb{N} = \{0, 1, \cdots, \aleph_0\}$. But, outcomes can also be modeled 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). But, we physicists hardly ever use this jargon. Notice that $P(x) = f(x)$, i.e., $f(x)$ is a probability.

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. Notice that $f(x)$ is not a probability.

To get a probability, we must integrate the pdf over an interval whose size is at least as large as an infinitesimal dx[a]. More usefully, we must compute

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

[a] An infinitely small non-zero number!

Random Variables

Formal books on statistics distinguish between what they call a random variable X, denoted with an upper case letter, and its outcomes x, denoted by lower case letters. However, we shall typically not make that distinction unless it is useful. (See previous slide.)

Randomness

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

Consider the time of decay of an atom or a B meson. According to quantum mechanics, no rule exists that determines *when* such a thing happens; its 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 mimic random outcomes! (Think `TRandom3`)

Something to Ponder...

Suppose that the 10^{100} particles in the known universe do random things on the average 10^{30} times per second. That is 10^{130} random things per second. Now, suppose that the universe, as we know it, will last 10^{20} seconds.

The universe will therefore undergo 10^{150} random things during its lifetime. What if the universe is a giant simulation governed by a universal random number generator with a Poincaré cycle[a] that is $> 10^{150}$?

How could we distinguish that universe from one with causeless effects?

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

Discuss! But, not now!

More Definitions

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

4. Moments

The r^{th} moment $\mu_r(a)$ about a of a probability distribution with probability function $f(x)$ 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; $\text{Var}_x = \mu_2(\mu)$ is called the variance; $\sigma = \sqrt{\text{Var}_x}$, 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 $f(x)$. (Here is another example where distinguishing between X and x is helpful.) The function $x = Q(P)$ that returns x, where $D(x) = P$, 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 More Definitions ... Enough Already!

6. Covariance, Correlation, Independence

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

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

It 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 variables x and y are said to be independent in which case $\mathrm{Cov}_{xy} = 0$.

However, in general, (essentially, for all non-Gaussian distributions!) $\mathrm{Cov}_{xy} = 0$ does not imply independence.

Consider yourself warned!

Outline

Example (2.1 The Binomial and the LHC)

In m proton-proton collisions we have r successes, say the creation of a Higgs boson. However, we can only record $n \leq m$ collision events of which $k \leq n$ are successes.

What is the probability $P(k, n|r, m)$ of getting k successes and $n - k$ failures in n trials given that we draw at random from a "box" (called the LHC) containing r unknown successes and $m - r$ unknown failures?

Here's a possible solution strategy (assuming that the order of collisions is irrelevant):

1. Count the number of ways to draw samples of size n from m collisions regardless of whether a collision is a success or a failure.
2. Count the number of ways to draw exactly k successful collisions from r successes and count the number of ways to draw exactly $n - k$ failed collisions from $m - r$ failures.
3. Then do something with those counts!

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?

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

4. How are these counts related to $P(k, n|r, m)$?

Answer:

$$P(k,n|r,m) = \binom{r}{k}\binom{m-r}{n-k}\Big/\binom{m}{n}$$
$$= \binom{n}{k} f(k,n,r,m),$$
$$\text{where } f(k,n,r,m) = \frac{r!}{(r-k)!}\frac{(m-r)!}{(m-r-n+k)!}\Big/\frac{m!}{(m-n)!}.$$

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 them.

Unfortunately, the probability rules require that we calculate the sum

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

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

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

But, since we do not know these counts, we do not know $P(r,m)$! The best we can hope to do is either to assign those probabilities by consulting our inner oracle or perhaps by asking our friendly neighborhood theorist for a prediction.

This means that for the same data (k,n), which we do know, you may assign different probabilities than I do and may therefore get different answers for the probability of k successes in n trials.
Agh...the horror, the horror!

No matter, let's press on regardless. After all, we are physicists; jacks of all trades and masters of none!

At the LHC, m is huge and generally $r << m$. For the Higgs boson at 13 TeV, $r \approx 10^{-10}\, m$. 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 is $z = r/m$:

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

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

and let $m \to \infty$ while keeping k and n fixed. It's as if we've used up our disk space quota even as the LHC continues to run!

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

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

What happens to the probabilities $P(r, m)$ that we don't know? To see what happens, write $P(k, n)$ as a double sum over the relative frequencies z and the integers m

$$
\begin{aligned}
P(k, n) &= \sum_{r,m} P(k, n|r, m)\, P(r, m), \\
&= \sum_{z}\sum_{m} P(k, n|zm, m)\, P(zm, m) \text{ with } z = r/m.
\end{aligned}
$$

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

where $\pi(z) \equiv \sum_m P(zm,m)$ is called a prior density. As $m \to \infty$, the number of relative frequencies (which are rational numbers) grows to $\aleph_0$, the sum converges to an integral, and we obtain:

Bruno de Finetti's Representation Theorem

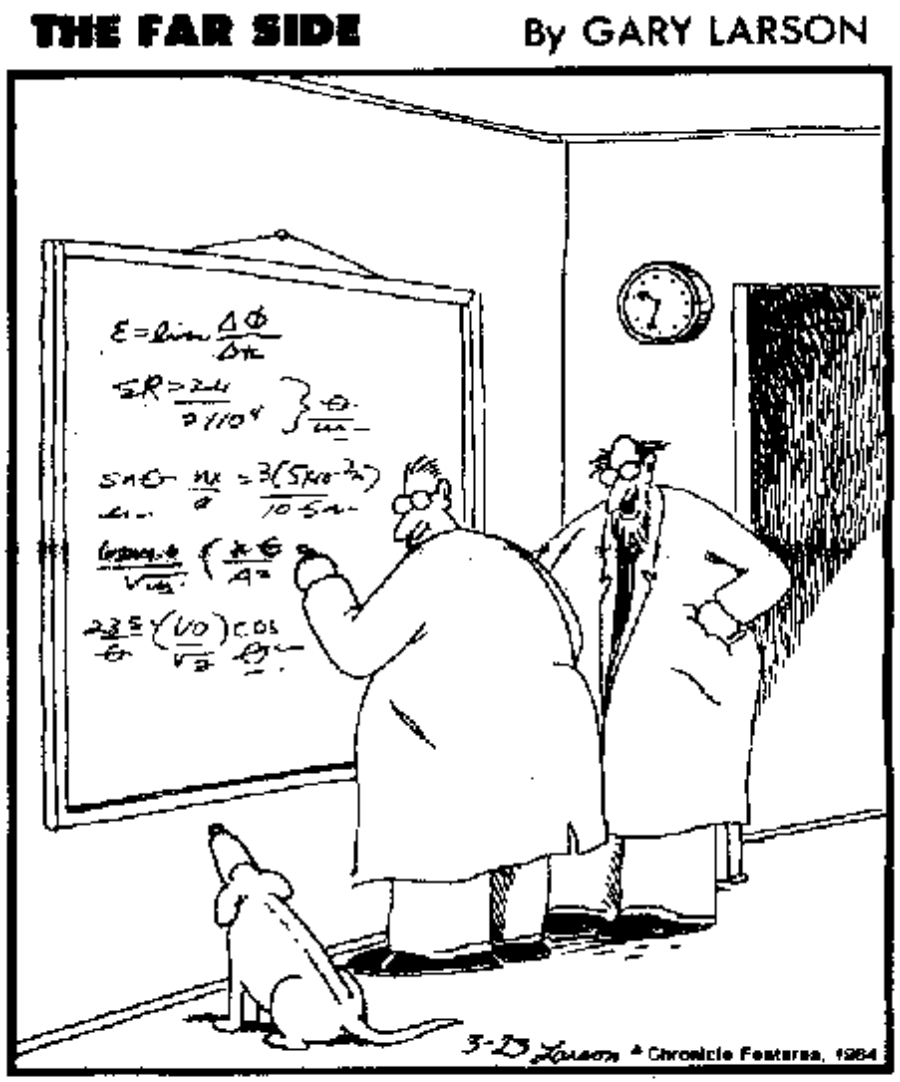
$$P(k,n) = \int_0^1 \text{binomial}(k,n,z)\, \pi(z)\, dz,$$
$$\text{where binomial}(k,n,z) = \binom{n}{k} z^k (1-z)^{n-k}.$$

The Binomial Distribution

What are we to make of the prior density $\pi(z)$? Let's try this: ask our friendly theorist for a prediction of the relative frequency of Higgs boson production at the LHC. Suppose she makes an accurate prediction that it is p.

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

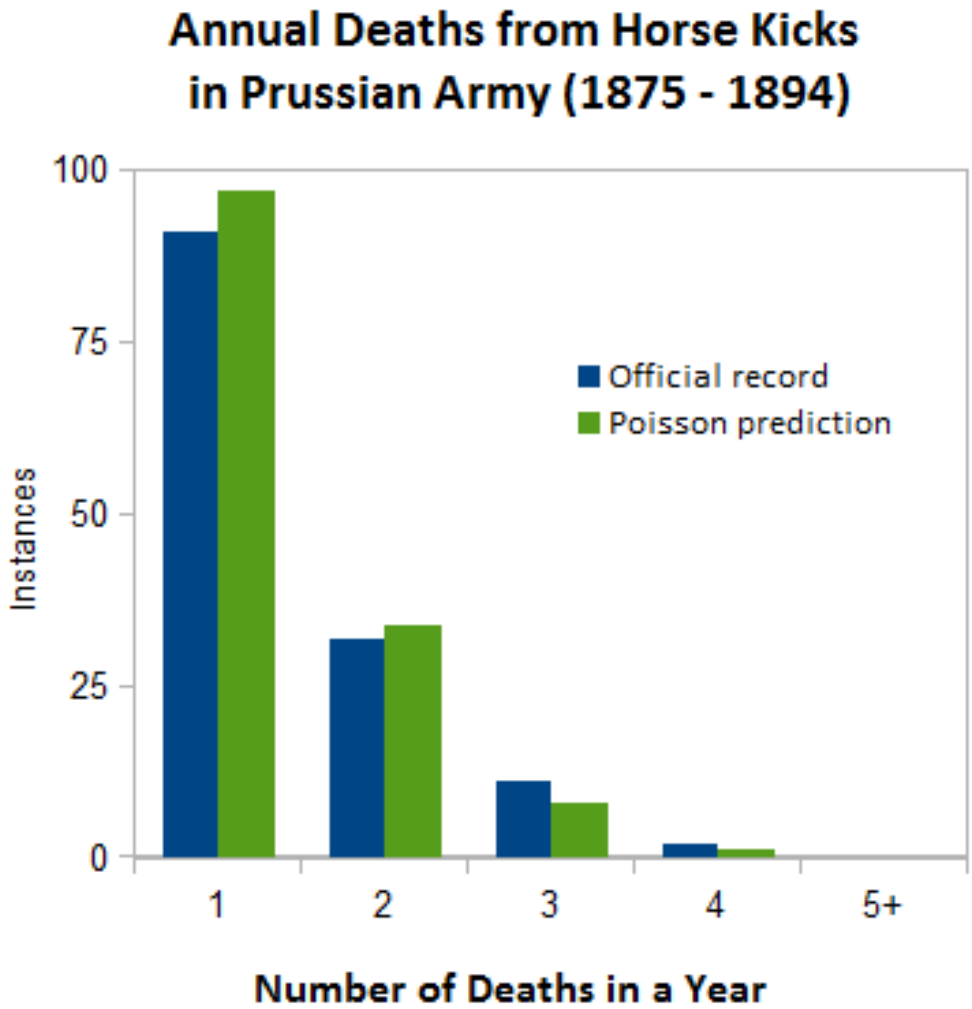
$$\boxed{P(k, n) = \text{binomial}(k, n, p)}$$



The Poisson Distribution

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).



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

The Poisson Distribution

The Poisson distribution can be derived from a stochastic model.

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 $(t, t + dt)$.

After all, horse kicks and Higgs bosons are rare so the chance of having more than one occur during the time interval is, to a very good approximation, zilch![2]

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

We can now assign probabilities.

[2] A technical term.

The Poisson Distribution

Here are the transition probabilities that define the Process model:

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

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

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

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 legit mathematics!

The Poisson Distribution

Equations such as

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

can be solved recursively. Try it and show that

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

where the mean count is $a = qt$. Also, show that $\text{Var}_n = a$, an important fact about the Poisson distribution that allows us to make the statement that for a mean count a we would expect counts n to fluctuate by roughly $\pm\sqrt{a}$.

Outline

Gaussian Distribution

The probability density function of the Gaussian distribution is

$$\boxed{\text{Gauss}(x, \mu, \sigma) = \frac{e^{-\frac{1}{2}(x-\mu)^2/\sigma^2}}{\sigma\sqrt{2\pi}}}.$$

It has mean μ and variance σ^2. The other oft used properties are the probability contents of various intervals. Let $z = (x - \mu)/\sigma$. Then

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

The Gaussian is the most important distribution in statistics. You will explore an example of why this is so in this week's homework assignment.

Gaussian Distribution

A bumper sticker: All sensible probability distributions approach a Gaussian in some limit. The precise statement is the central limit theorem.

Example (2.2 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 standardized variable $z = \sqrt{n}(t - \mu)/\sigma$. The mean of the probability density of z, $p(z)$, is 0 and its standard deviation is 1. The central limit theorem states

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

When measurement errors can be modeled as the sum of a large number of random contributions, we expect, and this is borne out in practice, the probability density of these errors to be roughly Gaussian.

χ^2 Distribution

Write $z = (x - \mu)/\sigma$, where $x \sim \text{Gaussian}(\mu, \sigma)$ 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. The δ-function imposes the constraint $t = g(z_1, \cdots, z_n)$. Let's apply this to our problem. But, first, a very 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 it for a while ...

Step 2. ... and realize that it is obviously true! QED.

□

Having "proved" it, let's proceed!

First note that $p(z_1, \cdots, z_n) = p(z_1)p(z_2)\cdots p(z_n)$ and

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

Putting together the pieces and (being physicists) brazenly shuffling the order of integration, we get

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

χ^2 Distribution

... almost there!

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

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

We can handle it using the following absolutely magical result for pole singularities that do not lurk on the real line,

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

Well, not quite! Our singularity involves an annoying square-root.

No matter, we avail ourselves of a time-honored strategy of physicists: solve a simpler problem then generalize its solution by inspection!

χ^2 Distribution

... there is light at the end of the tunnel!

Writing $m = n/2$, we have

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

Now let's solve this for integer m. In that case, we have a nice pole singularity of order m and life is good; witness

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

$$= \frac{1}{\Gamma(m)} \frac{t^{m-1} e^{-t/2}}{2^m}.$$

This result remains valid for non-integral values of m. Therefore, the pdf of the sum of n standardized Gaussian variates is $(t = \chi^2)$

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

Cauchy Distribution

Let $x, y \sim \text{Gaussian}(0,1) \equiv g(x)$. What is the pdf of $t = y/x$?
It is given by

$$p(t) = \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy\, \delta(t - y/x)\, g(x)\, g(y),$$
$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy\, \delta(t - y/x)\, e^{-\frac{1}{2}(x^2+y^2)}.$$

This integral is positively begging us to use polar coordinates, $y = r \sin\theta$, $x = r\cos\theta$ and $dx\, dy \to r\, dr\, d\theta$, so that we can write

$$p(t) = \frac{1}{\pi} \left(\int_0^{\infty} e^{-\frac{1}{2}r^2}\, r\, dr/2 \right) \int_0^{2\pi} \delta(t - \tan\theta)\, d\theta,$$
$$= \frac{1}{\pi} \int_0^{2\pi} \delta(t - \tan\theta)\, d\theta.$$

At first glance, the odd looking beast

$$p(t) = \frac{1}{\pi} \int_0^{2\pi} \delta(t - \tan\theta)\, d\theta,$$

looks tricky! But, recall that $\delta(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{i\omega x}\, d\omega$. You can use it to confirm the formula $\delta(h(\theta)) = \delta(\theta - \theta_0)/|dh/d\theta|$, where θ_0 is the root of $h(\theta)$.

For this problem, $h(\theta) = t - \tan\theta = 0$ and $1/|dh/d\theta| = \cos^2\theta$. Therefore,

$$\begin{aligned} p(t) &= \frac{1}{\pi} \int_0^{2\pi} \delta(\theta - \theta_0) \cos^2\theta\, d\theta, \\ &= \frac{1}{\pi} \cos^2\theta_0. \end{aligned}$$

But, $\tan\theta = t \implies \cos\theta = 1/\sqrt{1 + t^2}$. Therefore,

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

Outline

Summary

- According to Kolmogorov, probabilities are functions defined on suitable sets, have range [0, 1], and follow simple rules.
- The two 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 considered equally likely, then the probability of an outcome may be taken to be the ratio of the number of favorable outcomes to that of all possible outcomes.
- More generally, we use probability functions; probability mass functions for discrete distributions and probability densities for continuous ones.
- And **no**, the probability distributions we use do not come from Nature! We create them through mathematical reasoning. Our hope, however, is that the distributions we create are adequate models of the data generation mechanisms.