INFN School of Statistics 2024
Paestum, Italy

Probability Theory
Part 2

Harrison B. Prosper

Kirby W. Kemper Endowed Professor of Physics
Department of Physics, Florida State University

13 May, 2024

Outline

Outline

Outline

Probability Function A function $f(x)$ that gives a rule for assigning a probability $P(x)$ to outcome x is called a probability function.

In the previous lecture we talked about outcomes that can be modeled with n-tuples, $(z_1, \cdots, z_n)$ with elements 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 $\mathbb{R} = (-c, c)$[1].

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

Notice that $f(x)$ is a probability.

[1] We typically use the symbol ∞ instead of c. Georg Cantor (1845 - 1918), inventor of set theory, proved the astonishing theorem $c = 2^{\aleph_0}$.

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

Probability Density Function If x is from $\mathbb{R}$, then the probability function $f(x)$ is called a probability density function (pdf) and is often written with a lower case letter.

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

To get a probability, the pdf must be integrated over an interval whose size is at least as large as an infinitesimal dx[2]. More usefully we compute

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

The functions are often referred to as probability distributions.

[2]An infinitely small non-zero number!

Random Variables

Formal books on statistics distinguish between a random variable X denoted with an upper case letter from its outcomes x denoted by lower case letters.

However, most physicists typically do not make this distinction.

Probability Distribution Function: $F(x) = \mathbb{P}(X \leq x)$.

Note that if $x \in \mathbb{R}$ then

$$f(x) = \frac{\partial F}{\partial x}$$

$F(x)$ is also known as the cumulative distribution function.

Several quantities are used to characterize probability distributions. Here are a few.

Moments

The r^{th} moment $\mu_r(a)$ about a of a probability distribution with probability function $f(x)$ is defined by[3]

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

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

$\mu = \mu_1(0)$ is called the mean and is a measure of the location of the function $f(x)$; $\mathrm{V}(x) = \mu_2(\mu)$ is called the variance and $\sigma = \sqrt{\mathrm{V}}$ is the standard deviation, which is a measure of the width of $f(x)$.

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

Quantile Function As noted earlier, the function

$$F(x) = \mathbb{P}(X \leq x) = \int_{X \leq x} f(X)\, dX$$

is called the cumulative distribution function (cdf) of $f(x)$. (Here distinguishing between X and x turns out to be helpful!) The function $x = Q(P)$ that returns x given $P = F(x)$ 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 $F_L(x) \equiv F(x)$ and the right cdf defined by

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

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

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

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 $\text{Cov}(x, y) = 0$.

Note however that, in general, $\text{Cov}(x, y) = 0$ does *not* imply independence.

Outline

Example (2.1 The Binomial Distribution)

Consider m proton-proton collisions at the LHC and suppose we have r successes, say the creation of a Higgs boson. However, we are able to record only $n < m$ collision events of which $k \leq n$ are successes.

Problem What is the probability $P(k, n|r, m)$ to get k successes and $n - k$ failures in n trials given that they are drawn at random from a "box" called the LHC containing r unknown successful collisions and $m - r$ unknown failures?

Assumptions

1. The order of proton-proton collisions is irrelevant.
2. Every sample of collisions of size n is equally probable.

This problem is exactly the same as drawing k red balls and $n - k$ blue balls at random from a box with r red balls and $m - r$ blue balls.

And, like the LHC, the drawing of red and blue balls, that is, Higgs boson and non-Higgs boson events, is done *without* replacement.

Example (2.1 The Binomial Distribution)

Solution Plan:

1. Determine the number of ways T to get n collisions from m collisions regardless of whether a collision is a success or a failure.
2. Determine the number of ways S to get exactly k successful collisions from r successful collisions.
3. Determine the number of ways F to get exactly $n - k$ failed collisions from $m - r$ failed collisions.
4. Since successes and failures are assumed to be independent, the number of samples of size n with k successes and $n - k$ failures is $N = S \times F$.
5. Therefore, $P(k, n|r, m) = S \times F/T$.

1. How many samples T of n collisions can be drawn from m collisions?
$$\binom{m}{n}$$
2. How many samples S of k successes can be drawn from r successes?
$$\binom{r}{k}$$
3. How many samples F of $n-k$ failures can be drawn from $m-r$ failures?
$$\binom{m-r}{n-k}$$

$$\begin{aligned} P(k, n | r, m) &= \frac{S\, F}{T}, \\ &= \binom{r}{k} \binom{m-r}{n-k} \Big/ \binom{m}{n} \end{aligned}$$

This probability can be rewritten as

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

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

We now ask what is the probability of k, n irrespective of r, m?

This requires that we consider all values of r and m that are possible *a priori* and sum the probability $P(k, n|r, m)$ weighted by the probability $P(r, m)$ of r and m.

That is, we need to compute the sum

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

The elimination of quantities like r and m that are not of current interest is an example of a common procedure in probability theory called marginalization.

We can already see a potential problem. It is far from clear what we should put for $P(r, m)$. But let's nevertheless continue and see where this leads.

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

Let's rewrite the expression above in terms of the *unknown* relative frequency of success, $z = r/m$:

$$\begin{aligned} P(k, n) &= \sum_{z,m} P(k, n|zm, m)\, P(zm, m), \\ &= \binom{n}{k} \sum_{z,m} f(k, n, z, m)\, P(zm, m) \text{ from slide 15.} \end{aligned}$$

At the LHC, m, the number of proton-proton collisions, is huge.

Therefore, let's consider the idealization $m \to \infty$ while keeping k and n fixed.

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

Exercise 2.1

show that $f(k,n,z,m) \to z^k(1-z)^{n-k}$ as $m \to \infty$.

What about the probabilities $P(zm, m)$?

To see what happens, write $P(k, n)$ as

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

$$P(k, n) = \sum_z \binom{n}{k} z^k (1-z)^{n-k} \sum_m P(zm, m),$$

As $m \to \infty$, the sum converges to an integral and we obtain:

Bruno de Finetti's Representation Theorem

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

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

$$\pi(z) = \lim_{m\to\infty} \sum_m P(zm, m).$$

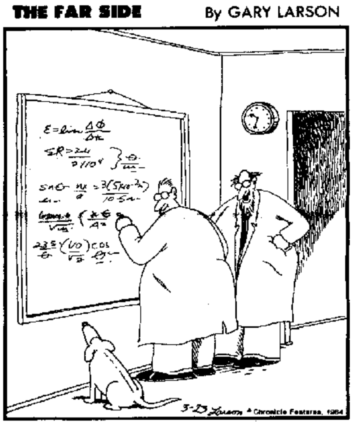
$\pi(z)$ is an example of a prior density.

The Binomial Distribution What are we to make of the prior density $\pi(z)$?

We ask our friendly theorist for a prediction of the relative frequency of Higgs boson production at the LHC. She predicts that it is p.

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

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



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

The Poisson Distribution

There are many situations in which the count n in the binomial distribution

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

is large and the probability p is small. For example, at 13 TeV only one in 10^{10} proton-proton collisions leads to a Higgs boson event. Consider, therefore, the limit $n \to \infty$ and $p \to 0$ while $a = pn$ and k remain fixed.

Exercise 2.2

Show that in this limit the binomial distribution becomes the Poisson distribution, $\text{Poisson}(k, a) = a^k \exp(-a)/k!$

The Poisson Distribution

The more fundamental definition of a Poisson distribution is via a stochastic model.

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

- no event occurred during $(t, t + dt)$ or
- one event occurred during $(t, t + dt)$.

We further suppose that the probability to get an event during the time interval $(t, t + dt)$ is proportional to its size dt.

We can now assign probabilities.

The Poisson Distribution

Here are the transition probabilities that define the Poisson model:

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

In principle, q could depend on time.
Using the probability rules, we can write

$$P_k(t+dt) = (1 - q dt)\, P_k(t) + q dt\, P_{k-1}(t),$$

or noting that $dP_k(t)/dt = [P_k(t+dt) - P_k(t)]/dt$,

$$\frac{dP_k}{dt} = -q\, P_k + q\, P_{k-1}.$$

The Poisson Distribution

Equations such as

$$\frac{dP_k}{dt} = -q\,P_k + q\,P_{k-1},$$

can be solved recursively.

Exercise 2.3

Show that

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

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

A widely used model in particle physics, astronomy, and cosmology is the multi-Poisson model defined by

$$P(\mathbf{k}|\mathbf{a}) = \prod_{m=1}^{M} \frac{a_m^{k_m} e^{-a_m}}{k_m!}.$$

This is the standard statistical model for binned data when the counts are conditionally independent.

In particle physics, analyses that use this model are sometimes referred to as shape analyses.

Exercise 2.4

Show that

$$P(\mathbf{k}|\mathbf{a}) = \text{Poisson}(k, a)\,\text{multinomial}(k_1, \cdots, k_m, p_1, \cdots, p_m) \text{ where}$$

$$k = \sum_{m=1}^{M} k_m, \quad a = \sum_{m=1}^{M} a_m, \quad p_m = \frac{a_m}{a}, \quad \sum_{m=1}^{M} p_m = 1,$$

and the multinomial distribution is given by

$$\text{multinomial}(k_1, \cdots, k_m, p_1, \cdots, p_m) \equiv \binom{k}{k_1, \cdots, k_m} \prod_{m=1}^{M} p_m^{k_m}$$

When bin counts are large, or when they have a large dynamic range typical of steeply falling spectra, it is sometimes convenient to drop the Poisson term and rely solely on the multinomial.

Outline

Gaussian Distribution The Gaussian, or normal, distribution is the most important distribution in statistics. Its probability density function is

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

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

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

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)

Let $t = \frac{1}{n}\sum_{i=1}^{n} x_i$, where x_i is sampled from $p(x, \mu, \sigma)$ and p is *any* probability density with mean μ and finite standard deviation σ.

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

$$\boxed{\lim_{n\to\infty} \int_{X\leq z} f(X)\, dX = \int_{-\infty}^{z} \text{Gauss}(X, 0, 1)\, dX}.$$

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 $x \sim \text{Gaussian}(\mu, \sigma)$ ($\sim$ means "is sampled from" and consider the sum

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

What is the pdf of t? Given the probability density function, $p(z_1, \cdots, z_n)$, the pdf of 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 $t = g(z_1, \cdots, z_n)$.

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

χ^2 Distribution First note that $p(z_1, \cdots, z_n) = p(z_1)p(z_2)\cdots p(z_n)$ and

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

(*Burn this formula into your brain...it's one of the most useful in physics and statistics*!) Putting together the pieces and shuffling the order of integration, we get

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

χ^2 Distribution Writing $m = n/2$, and performing the integral for integer m, we find

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

This result remains valid for non-integral values of m. Therefore, the pdf of the sum of the square of n standardized Gaussian random variables 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 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

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

At first glance, the odd looking beast

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

looks tricky! But, recall that $\delta(h(\theta)) = \delta(\theta - \theta_0)/|dh/d\theta|_{\theta_0}$, where θ_0 is the root of $h(\theta) = t - \tan\theta = 0$ and $1/|dh/d\theta| = \cos^2\theta$. In fact, on the domain $[0, 2\pi]$, there are two roots separated by π. Therefore,

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

Since $\tan\theta_0 = t$, $\cos\theta_0 = 1/\sqrt{1 + t^2}$, which yields the Cauchy pdf,

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

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.

Examples/exercises at:
`https://github.com/hbprosper/INFN-SOS/`