Statistics, Data Analysis, and Machine Learning for Physicists

Timothy Brandt

Spring 2020

Lecture 1

This course will cover statistics, data analysis, and machine learning in the context of physics research. My own background is in astrophysics, so the examples will tend to come from that area, but these tools are general.

We'll start the course in earnest with an introduction to statistics and probability. As an example, let's take the case where there is a rare disease that affects a fraction 10^{-4} of the population, and there is a test for this disease that with a 100% true positive rate and a 1% false positive rate. In other words, if you have the disease, the test will always be positive. If you don't have the disease, the test will be positive 1% of the time. We'll look at Bayesian and frequentist interpretations of the following scenario: I feel fine, I take the test, and it comes back positive. What is the probability that I do, in fact, have the disease? It's not 99%. We'll look at both Bayesian and frequentist interpretations.

In the frequentist interpretation the question "What is the probability that I have the disease?" is nonsensical. Either I have the disease or I don't, and nothing about my beliefs (and no test) can change that. The probability of me getting this false positive is 10^{-2} if I do have it, and that is all I can say. The frequentist will only make statements about the average results of random, repeatable actions. To apply the frequentist model, I have to carefully rephrase the question:

- Given a person who is selected at random from the population and tested for this disease, what is the probability that I have selected someone who actually has the disease given the requirement that the test be positive?

In this phrasing, we only speak of random, repeatable events, so I can apply the frequentist approach. The probability of a random person being healthy and yet testing positive is the product

$$p(\mathrm{h},+) = (1-10^{-4})(10^{-2}) \approx 1\%. \tag{1}$$

The probability of a random person being sick and testing positive is

$$p(\mathrm{s},+) = (10^{-4})(1) = 0.01\%. \tag{2}$$

So, by the basic rules of probability, the conditional probability of a random person being sick conditioned on a positive test is

$$p(\mathrm{s}|+) = \frac{p(\mathrm{s},+)}{p(+)} = \frac{p(\mathrm{s},+)}{p(\mathrm{h},+)+p(\mathrm{s},+)} \approx \frac{0.01\%}{1\%+0.01\%} \approx 1\%. \tag{3}$$

So, there is actually only a $\sim$1% chance that a person selected at random from the population, who then happens to test positive, will actually represent a case of this disease. Notice how carefully I have to phrase that!

Now let's take the Bayesian approach. A Bayesian will indeed assign a probability to me having the disease, something the frequentist is not willing to do. My probability (belief) before the test was 10^{-4}. This is modified by the positive result of the test.

$$p(\mathrm{s}|+) = \frac{p(+|\mathrm{s})p(\mathrm{s})}{p(+)} \tag{4}$$

The denominator, $p(+)$, is the probability of my getting a positive test, the probability of my data. In Bayesian statistics it is generally treated as an (annoying) normalization constant, so that I have

$$p(\mathrm{s}|+) \propto p(+|\mathrm{s})p(\mathrm{s}). \tag{5}$$

Let's see how this works in practice. I can easily compute the right-hand side both given a model that I am sick s and healthy h:

$$p(\mathrm{s}|+) \propto (1)(10^{-4}) = 10^{-4} \tag{6}$$

and

$$p(\mathrm{h}|+) \propto (10^{-2})(1 - 10^{-4}) \approx 1\% \tag{7}$$

I can now compute the normalization constant as the sum of those two ($\approx 10^{-2}$) to turn the proportionalities into equalities and derive *posterior probabilities*. As before, I have

$$p(\mathrm{s}|+) = \frac{p(+|\mathrm{s})p(\mathrm{s})}{p(+)} \approx \frac{10^{-4}}{10^{-2}} = 1\%. \tag{8}$$

The result is the same, but the phrasing of the question was different. As a Bayesian, I made explicit use of *prior probabilities* to make a natural statement that a frequentist would say is nonsense.

Now we'll talk a little about random numbers and randomness. A random number is just what it sounds like: we cannot predict its value in advance. We still might know something about it, though. Random (or quasi-random) numbers show up all the time. How many photons from that star hit my detector over some time interval δt? How many muon-muon collisions took place in my reactor in that same time interval? etc. Even if we have no way of actually calculating this number, we can still estimate what it will be–just the average number of events we can observe over a long time interval. Often, a random number is something whose value we do not know, but whose parent distribution we do know. Back to photons, what is the *specific* probability of observing 0 photons, or one, etc.?

To get this and derive one of the most important distributions used in this class, we begin with the binomial distribution:

$$p(n) = p^n(1-p)^{N-n}\frac{N!}{(N-n)!n!} \tag{9}$$

$$= p^n(1-p)^{N-n}\frac{N(N-1)\times\cdots\times(N-n+1)}{n!} \tag{10}$$

Now we let $p \to 0$, $Np = \lambda =$ constant. Since p is very small for each individual event, my number of events n should be much, much less than my number of chances: $\frac{n}{N} \ll 1$. The limit $p \to 0$, $Np = \lambda =$ constant also implies that $\frac{n}{N} \to 0$, or that $N - n \approx N$. Taking this limit,

$$p(n) \approx p^n(1-p)^N\frac{N^n}{n!} \tag{11}$$

$$p(n) \approx (Np)^n(1-p)^N\frac{1}{n!} \tag{12}$$

$$\ln p(n) \approx n\ln Np + N\ln(1-p) - \ln n! \tag{13}$$

$$\ln p(n) \approx n\ln Np - Np - \ln n! \tag{14}$$

$$\ln p(n) \approx n\ln\lambda - \lambda - \ln n! \tag{15}$$

$$p(n) \approx \frac{\lambda^n e^{-\lambda}}{n!}. \tag{16}$$

This is the Poisson distribution, one of the two or three most important distributions you will encounter in physics. Any hands for the most important (Gaussian)? Any others that I might have in mind (χ^2)?

To refresh your memories about how probability distributions work, we can go ahead and derive the moments of the Poisson distribution. First, we'll verify that it is indeed normalized, so that

$$\sum_{n=0}^{\infty}\frac{\lambda^n e^{-\lambda}}{n!} = 1. \tag{17}$$

I changed the upper limit of summation from N to ∞, which is all right because the Poisson distribution is the limit in which $N \to \infty$ at fixed $\lambda = Np$. To prove that $p(n)$ is normalized, I first factor out the exponential:

$$\sum_{n=0}^{\infty} \frac{\lambda^n e^{-\lambda}}{n!} = e^{-\lambda} \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} = e^{-\lambda} e^{\lambda} = 1. \tag{18}$$

The sum is just the Taylor expansion for the exponential, so the two exponentials cancel.

Statistical distributions may be defined by their moments, and these are often quoted (see the main wikipedia page for any statistical distribution). For a discrete probability distribution like the Poisson, the first moment is defined by

$$\mu_1 = \sum_{n=0}^{\infty} n^1 \, p(n) \tag{19}$$

and the k-th moment, for $k > 1$, is defined by

$$\mu_k = \sum_{n=0}^{\infty} (n - \mu_1)^k \, p(n). \tag{20}$$

The zero-th moment is just the sum of the probability distribution over all possible values, which should be 1. The first moment is the mean, the average value of realizations of the distribution weighted by the probability of each realization. Higher moments measure the width and shape of the distribution about its mean. You should have the first and second moments memorized for the Poisson, Gaussian, and (eventually) the χ^2 distributions. For the Poisson distribution, the first moment, the mean, is:

$$\mu = \sum_{n=0}^{\infty} n \frac{\lambda^n e^{-\lambda}}{n!} = 0 \frac{\lambda^0 e^{-\lambda}}{0!} + \sum_{n=1}^{\infty} n \frac{\lambda^n e^{-\lambda}}{n!} \tag{21}$$

$$= \sum_{n=1}^{\infty} \frac{\lambda^n e^{-\lambda}}{(n-1)!} \tag{22}$$

$$= \lambda \sum_{n=1}^{\infty} \frac{\lambda^{n-1} e^{-\lambda}}{(n-1)!} \tag{23}$$

$$= \lambda \sum_{n=0}^{\infty} \frac{\lambda^n e^{-\lambda}}{n!} \tag{24}$$

$$= \lambda \tag{25}$$

In the last step, I used the fact that the distribution is already normalized. Finally, for the second moment,

usually called the variance:

$$\sigma^2 = \langle (n - \langle n \rangle)^2 \rangle = \langle n^2 - 2n\langle n \rangle + \langle n \rangle^2 \rangle = \langle n^2 \rangle - \langle n \rangle^2 \tag{26}$$

$$= \sum_{n=0}^{\infty} n^2 \frac{\lambda^n e^{-\lambda}}{n!} - \lambda^2 \tag{27}$$

$$= 0^2 \frac{\lambda^0 e^{-\lambda}}{0!} + \sum_{n=1}^{\infty} n^2 \frac{\lambda^n e^{-\lambda}}{n!} - \lambda^2 \tag{28}$$

$$= \sum_{n=1}^{\infty} n \frac{\lambda^n e^{-\lambda}}{(n-1)!} - \lambda^2 \tag{29}$$

$$= \lambda e^{-\lambda} \sum_{n=1}^{\infty} n\lambda^{n-1} \frac{1}{(n-1)!} - \lambda^2 \tag{30}$$

$$= \lambda e^{-\lambda} \sum_{n=1}^{\infty} \frac{d}{d\lambda}\left(\lambda^n\right) \frac{1}{(n-1)!} - \lambda^2 \tag{31}$$

$$= \lambda e^{-\lambda} \frac{d}{d\lambda} \left(\sum_{n=1}^{\infty} \frac{\lambda^n}{(n-1)!} \right) - \lambda^2 \tag{32}$$

$$= \lambda e^{-\lambda} \frac{d}{d\lambda} \left(\lambda \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} \right) - \lambda^2 \tag{33}$$

$$= \lambda e^{-\lambda} \frac{d}{d\lambda} \left(\lambda e^{\lambda} \right) - \lambda^2 \tag{34}$$

$$= \lambda^2 + \lambda - \lambda^2 \tag{35}$$

$$= \lambda. \tag{36}$$

There was a trick in there, writing $n\lambda^{n-1}$ as a derivative and then using the linearity of the derivative. So, the Poisson distribution has a mean number of λ events (as we would expect), and a variance of λ (standard deviation of $\sqrt{\lambda}$).

We'll also do a quick refresher on the Gaussian distribution, since it is probably the most important statistical distribution you encounter in physics. It shows up constantly. Published papers usually assume distributions to be Gaussian (whether or not their authors realize it). There is an important difference between the Poisson and the Gaussian distribution: the Poisson distribution is **discrete**, while the Gaussian distribution is **continuous**. The Poisson distribution could be described by the probability of getting exactly n events, with the normalization criterion being

$$\sum_{n=0}^{\infty} p(n) = 1. \tag{37}$$

The Gaussian distribution is defined by a *probability density* rather than simply a probability:

$$\frac{dp}{dx} = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2} \right] \tag{38}$$

The probability of getting *precisely* x is zero. The Gaussian distribution is normalized, with the normalization criterion being an integral rather than a sum:

$$\int_{-\infty}^{\infty} \frac{dx}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2} \right] = 1. \tag{39}$$

The standard trick to prove this is to use polar coordinates r, θ to show that

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} dx\, dy\, \exp\left[-\left(x^2 + y^2\right) \right] = \pi. \tag{40}$$

Using the volume element $rd\theta$ in polar coordinates, this goes like

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} dx\, dy\, \exp\left[-\left(x^2+y^2\right)\right] = \int_0^{2\pi} d\theta \int_0^{\infty} r\, dr\, \exp\left[-r^2\right] \tag{41}$$

$$= 2\pi \int_0^{\infty} r\, dr\, \exp\left[-r^2\right] \tag{42}$$

$$= 2\pi \left(-\frac{1}{2}\right) \exp\left[-r^2\right]\Big|_0^{\infty} \tag{43}$$

$$= 2\pi \left(0 - \frac{1}{2}\right) \tag{44}$$

$$= \pi. \tag{45}$$

And, since

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} dx\, dy\, \exp\left[-\left(x^2+y^2\right)\right] = \left(\int_{-\infty}^{\infty} dy\, \exp\left[-y^2\right]\right)\left(\int_{-\infty}^{\infty} dx\, \exp\left[-x^2\right]\right), \tag{46}$$

we have

$$\int_{-\infty}^{\infty} dy\, \exp\left[-y^2\right] = \sqrt{\pi} \tag{47}$$

or, equivalently,

$$\int_{-\infty}^{\infty} \frac{dy}{\sqrt{\pi}} \exp\left[-y^2\right] = 1. \tag{48}$$

The change of variables

$$u = \frac{y-\mu}{\sqrt{2\sigma^2}} \tag{49}$$

does the rest.

The first moment of the Gaussian distribution is given by

$$\mu = \int_{-\infty}^{\infty} \frac{x\, dx}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \tag{50}$$

$$= \int_{-\infty}^{\infty} \frac{(x-\mu+\mu)\, dx}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \tag{51}$$

$$= \int_{-\infty}^{\infty} \frac{(x-\mu)\, dx}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] + \mu \int_{-\infty}^{\infty} \frac{dx}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \tag{52}$$

$$= \int_{-\infty}^{\infty} \frac{s\, ds}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{s^2}{2\sigma^2}\right] + \mu \tag{53}$$

$$= \mu \tag{54}$$

where $s = x - \mu$ and the first integral vanishes by symmetry. The second moment of the distribution, with

the same definition of s and using integration by parts, is

$$\sigma^2 = \int_{-\infty}^{\infty} \frac{(x-\mu)^2\, dx}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \tag{55}$$

$$= \int_{-\infty}^{\infty} \frac{s^2\, ds}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{s^2}{2\sigma^2}\right] \tag{56}$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \left(\sigma^2 s\right) \left(\frac{s\, ds}{\sigma^2} \exp\left[-\frac{s^2}{2\sigma^2}\right]\right) \tag{57}$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \sigma^2 ds \, \exp\left[-\frac{s^2}{2\sigma^2}\right] \tag{58}$$

$$= \sigma^2 \int_{-\infty}^{\infty} \frac{ds}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{s^2}{2\sigma^2}\right] \tag{59}$$

$$= \sigma^2. \tag{60}$$

A few quick words about the Gaussian distribution. Due to the central limit theorem, real distributions are often nearly Gaussian. That said, the Gaussian distribution has extremely suppressed tails ($\exp\left[-x^2\right]$ goes to zero *very* fast at large $|x|$). Real distributions are almost never Gaussian out in their tails, say four or five sigma from their means. Sometimes this doesn't matter; sometimes it matters a great deal.

One of the core tasks for physicists, and one of the motivations for this course, is to compare some data with a model for that data. For this, we will introduce something called the likelihood. In Bayesian statistical inference, the *posterior probability* of my model is given by

$$p(\text{model}|\text{data}) \propto p(\text{data}|\text{model}) p(\text{model}). \tag{61}$$

This first term, the probability of my data given my model, is called the likelihood. In order to replace the proportionality with an equality, I would have to include a factor $p(\text{data})$—this is just a normalizing constant to a Bayesian. But there is nothing inherently Bayesian about the likelihood, and it is a very important function in frequentist statistics. A frequentist still can (and often does) choose the model that maximizes the likelihood; they just aren't willing to assign prior probabilities to models. The probability of getting a set of data given a model for generating that data remains a perfectly sensible quantity in a frequentist framework.

In the next lecture we will write down the likelihood function

$$\mathcal{L} = p(\text{data}|\text{model}), \tag{62}$$

for the specific case of Gaussian errors. We will then use it to derive a lot of very standard, and very useful, results.