Introduction to probability theory

Discrete random variables

Gabriel T. Landi

Additional reading

- Ross, chapters 2 to 4
- Reif, chapter 1
- Tomé, de Oliveira, chapters 1 and 2.
- Salinas, chapter 1.

<u>Probability theory</u> <u>and</u> <u>statistical</u> <u>mechanics</u>

I find it quite amazing that the exact same theory we use to analyze data from the LHC experiment is also used in completely uncorrelated areas, such as medical trials or opinion polls. There is, of course, a reason for this: probability theory is what we could call a "stem theory", in analogy with stem cells, which adapt to the task at hand.

Let me explain what I mean. Take the silly example of rolling a dice. There are 6 possible outcomes

$$X = 1,\ 2,\ 3,\ 4,\ 5,\ 6. \tag{1}$$

Naively, we usually think that

"According to prob. theory, the prob. of obtaining each outcome is $1/6$".

But that is <u>not</u> true! "prob. theory" never said that. That is an assumption that <u>you</u> made: you assumed that the 6 outcomes were equally likely so, since the probabilities must add up to 1, each must be $1/6$.

The reason why prob. theory is a "stem theory" and the reason why it is so successful is because it <u>never</u> attributes probabilities to events. That part is up to you. What the theory does is, it tells you how to compute relevant quantities once the probabilities are given.

Figuring out what are the probabilities is usually a difficult task. The prime source of information is experiment. There are also some approximation techniques, like assuming the dice outcomes are equally likely or assuming that the uncertainties in an experiment follow a Gaussian distribution. But that is pretty much it.

The most remarkable feat of the discipline we call statistical mechanics is that, for certain classes of systems, we can infer these probabilities from first principles. That is, we can start with the fundamental laws of physics, such as Newton's law or Schrödinger's equations, and we may deduce what the probabilities will be, without having to perform the experiment.

This is something which is absolutely unique about physics. You will not find it anywhere else.

The most important result in statistical mechanics is known as the Gibbs formula or the canonical ensemble. It concerns systems that are in thermal equilibrium with a large heat bath kept at a temperature T. In this case the prob. of finding the system in any microscopic configuration is proportional only to the energy E of that state

$$\boxed{P = \frac{e^{-E/k_B T}}{Z}} \qquad (2)$$

where k_B is Boltzmann's constant and Z is a normalization constant called the partition function. This result uniquely determines the prob. of a state once you know its energy

Parts of this result were already contemplated by Maxwell and Boltzmann, but it was Josiah Willard Gibbs, a professor at Yale which, around 1902, realized its remarkable potential and scope. This is why he called Eq (2) canonical.

Eq (2) will be the basis for most of our course and we will spend quite a lot of time discussing it. But before we do so, I want to teach you about some important concepts in prob. theory.

Discrete random variables

Prob. theory is constructed in terms of events, which can be anything you want. For instance, when you say

" What is the prob. that winter is coming? "

The statement "winter is coming" is the event.

Here we will be interested in the situation where the event is a number. That is, we have some quantity x which may take on a series of values with certain probabilities. This quantity x is called a random variable (r.v.). For now we will assume that the values taken on by x are discrete. The continuous case will be treated in the next set of notes.

To each value k that X may take, we have an associated probability

$$\boxed{P_k = P(X = k) = \text{prob. that } X \text{ takes on the value } k} \tag{3}$$

These probabilities must always satisfy two fundamental properties:

$$\boxed{\begin{aligned} & P_k \in [0, 1] \\ & \sum_k P_k = 1 \end{aligned}} \tag{4}$$

The last condition means that the probabilities are normalized

The Bernoulli distribution

The simplest example of a discrete r.v. is one which may take on the values $X = 0$ or $X = 1$. The probabilities may then be parametrized as

$$\boxed{\begin{aligned} P_1 &= P(X=1) = p \\ P_0 &= P(X=0) = q = 1-p \end{aligned} \qquad p \in [0,1]} \tag{5}$$

The distribution is completely characterized by the parameter p. In the statistics literature it is common to write

$$\boxed{X \sim \mathrm{Bern}(p)} \tag{6}$$

the symbol "$\sim$" means "is distributed according to". X is a r.v. It does not have a specific value. So all we can do is specify <u>how</u> it is distributed.

The Bernoulli distribution appears often in physics as representing the occupation of some site. When $X = 0$ we say the site is empty and when $X = 1$ we say it is occupied.

Note also how we may use $\mathrm{Bern}(p)$ to represent yes/no events. For instance, we may associate

$$\begin{aligned} &\text{Winter is coming} : X = 1 \\ &\text{Winter is not coming} : X = 0 . \end{aligned} \tag{7}$$

Usually, $X = 1$ is called a "<u>success</u>" and $X = 0$ is called a failure

Another r. v. that we encounter frequently in physics is one which takes on the values +1 and −1. this usually represents a spin system, where

$$+1 = \uparrow = \text{spin is pointing up}$$
$$-1 = \downarrow = \text{spin is pointing down.} \tag{8}$$

We can define this r. v. starting from Bern(p) as

$$Y = 2X - 1, \qquad X \sim \text{Bern}(p) \tag{9}$$

then

$$X = 0 \rightsquigarrow Y = -1$$
$$X = 1 \rightsquigarrow Y = +1 \tag{10}$$

As for the probabilities, we have

$$\begin{aligned} P(Y=1) &= P(2X-1=1) = P(2X=2) \\ &= P(X=1) = p \\ P(Y=-1) &= P(2X-1=-1) = P(2X=0) \\ &= P(X=0) = q = 1-p. \end{aligned} \tag{11}$$

The Binomial distribution

The Binomial distribution describes the number of successes in n Bern(p) trials. That is, we roll the Bern(p) n times and see how many "1"s we get. The distribution is

$$\boxed{P_k = P(X=k) = \binom{n}{k} q^{n-k} p^k} \qquad k = 0, 1, \dots, n \qquad (12)$$

Again, $q = 1-p$ and

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!} \qquad (13)$$

is the Binomial coefficient. You may understand (12) intuitively

$$p^k = \text{prob. of } k \text{ successes}$$

$$(1-p)^{n-k} = \text{prob. of } n-k \text{ failures}$$

$$\binom{n}{k} = \text{does not matter the exact order in which the successes occurred.}$$

Usually, we read $\binom{n}{k}$ as "n choose k".

We say that

$$\boxed{X \sim \text{Bin}(n, p)} \qquad (14)$$

The normalization of Eq (12) may be checked using the Binomial theorem

$$\sum_{k=0}^{M} P_k = \sum_{k=0}^{M} \binom{M}{k} q^{M-k} p^k = (q+p)^M = (1)^M = 1 \qquad (15)$$

Ex: Ehrenfest model

Suppose you have a gas with M particles. What is the prob. of finding k particles in the left side of the container?

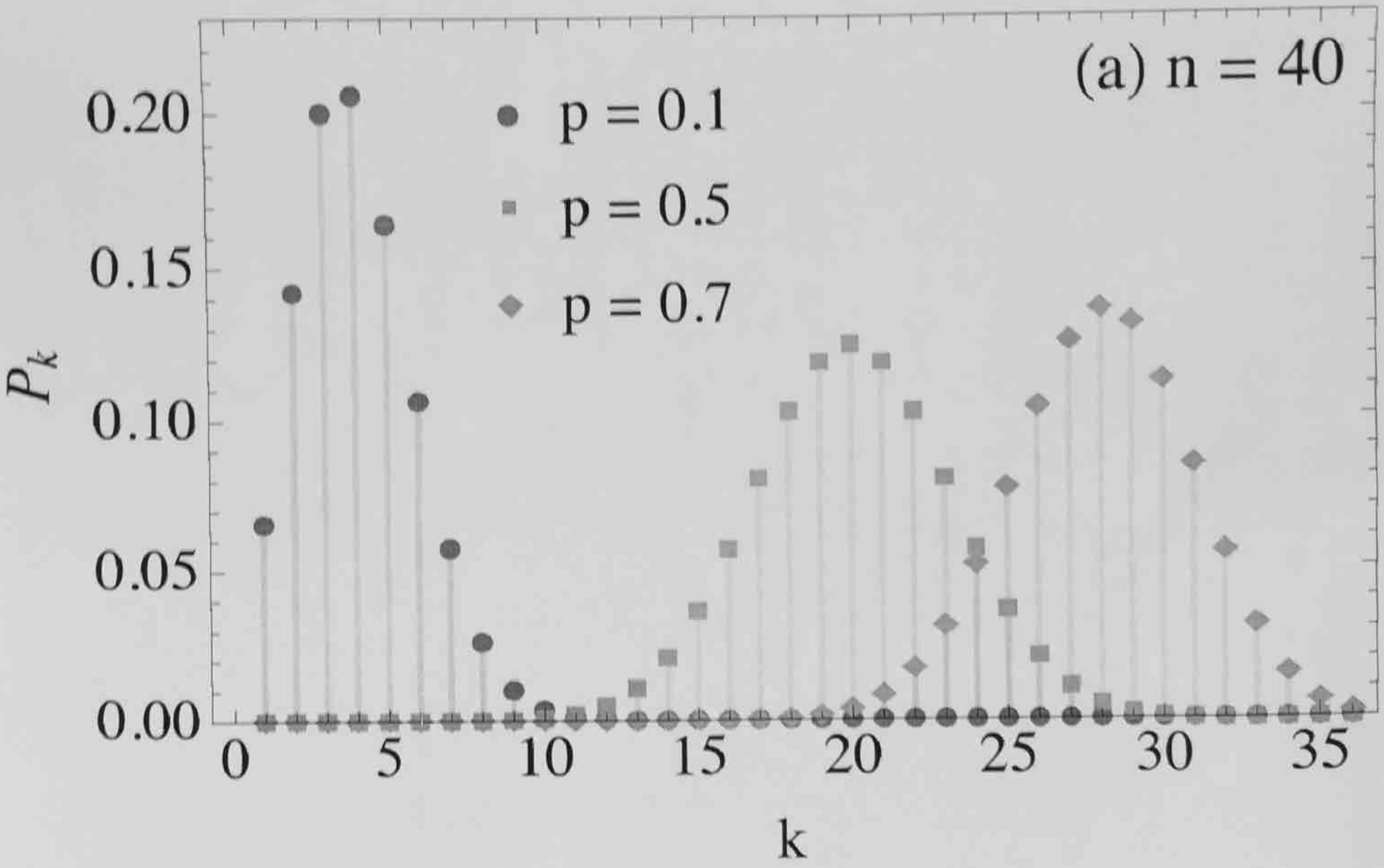
Assuming that the left and right parts have equal volumes, we will have that

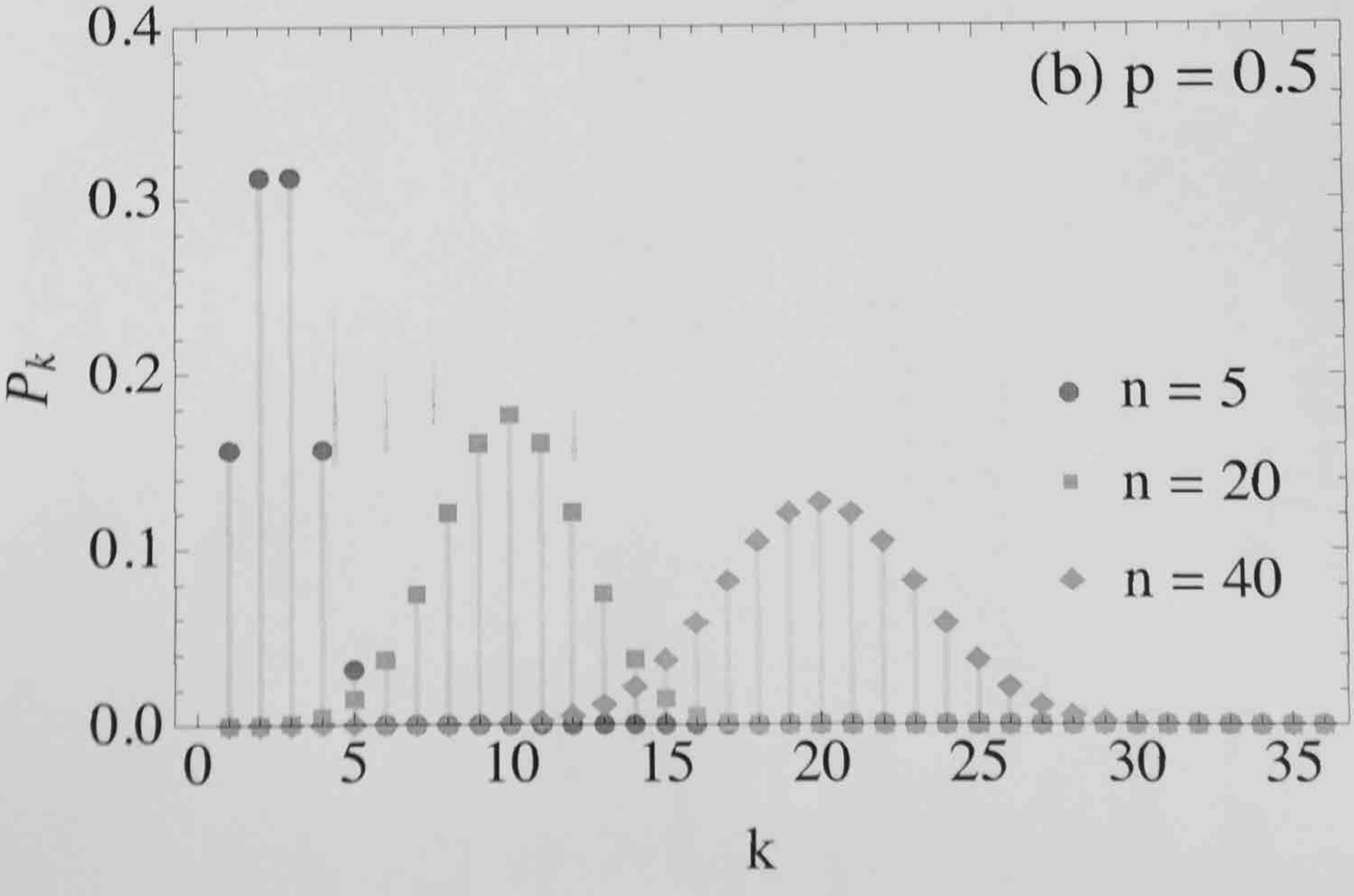
x = number of particles on the left side $\sim Bin(M, 1/2)$

since $p = 1/2$. We then have, from (12), with $p = q = 1/2$

$$P(X=k) = \binom{M}{k} \frac{1}{2^M} \qquad (16)$$

Plots of the Binomial distribution.





The mean/average

The prob. distribution contains all the information about a r.v. But sometimes we want just a few numbers to get a rough description. The most important such number is the mean:

$$\boxed{\langle X \rangle = \sum_k k P_k = \sum_k k P(X=k)} \tag{17}$$

Please note that different books use different notations

$$\langle X \rangle = E(X) = \overline{X} \tag{18}$$

Example: $X \sim \text{Bern}(p)$

$$\langle X \rangle = \sum_k k P_k = (0) P_0 + (1) P_1 = (0)(1-p) + (1)p$$

$$\therefore \quad \boxed{X \sim \text{Bern}(p): \quad \langle X \rangle = p} \tag{19}$$

This is the coolest feature about Bern(p): the mean is the actual probability.

The mean satisfies the following properties

$$\langle cX \rangle = c\langle X \rangle \tag{20}$$
$$\langle X + c \rangle = \langle X \rangle + c$$

where c is a constant. Please take a second to make sure you understand them

Example: mean of Bin(n,p)

Intuition: recall that $X \sim \text{Bin}(n,p)$ is the joint outcome of n Bern(p) trials. In each trial the average success is p, Eq (19). Thus we expect that

$$\boxed{X \sim \text{Bin}(n,p) : \qquad \langle X \rangle = np} \tag{21}$$

Now we do the calculation using Eq (17):

$$\langle X \rangle = \sum_{k=0}^{n} k P_k = \sum_{k=1}^{n} k \binom{n}{k} q^{n-k} p^k$$

↑ Sum may start at 1 since $k=0$ is zero anyway.

Now we write

$$k \binom{n}{k} = k \frac{n!}{k!(n-k)!} = k \frac{n(n-1)!}{k(k-1)!((n-1)-(k-1))!}$$

$$= n \frac{(n-1)!}{(k-1)!((n-1)-(k-1))!}$$

$$= n \binom{n-1}{k-1}$$

We then get

$$\langle x \rangle = \sum_{k=1}^{m} m \binom{m-1}{k-1} q^{(m-1)-(k-1)} p^{(k-1)} p$$

$k' = k-1$

$$= mp \sum_{k'=0}^{m-1} \binom{m-1}{k'} q^{(m-1)-k'} p^{k'}$$

Binomial theorem

$$= mp \underbrace{(q+p)^{m-1}}_{1}$$

$$= mp$$

which is (21).

The Poisson distribution

The Poisson distribution is characterized by a single parameter λ and has the distribution

$$\boxed{X \sim \text{Pois}(\lambda): \quad P_k = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k = 0, 1, 2, \ldots} \tag{24}$$

this is the most important discrete distribution. It appears in the strangest situations.

The normalization of Pois(λ) follows from the Taylor series expansion of e^{λ}:

$$\sum_{k=0}^{\infty} P_k = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda} e^{\lambda} = 1 \tag{25}$$

When people think about the Poisson distribution, they usually think about $e^{-\lambda}$. But we see that $e^{-\lambda}$ is just a normalization constant. The actual distribution is

$\lambda^k / k!$

The mean of Pois(λ) is

$$\langle x \rangle = \sum_{k=0}^{\infty} k\, e^{-\lambda} \frac{\lambda^k}{k!} = e^{-\lambda} \lambda \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} \qquad k' = k-1$$

$$= e^{-\lambda} \lambda \sum_{k'=0}^{\infty} \frac{\lambda^{k'}}{k'!}$$

$$= e^{-\lambda} \lambda\, e^{\lambda}$$

$$\therefore \quad \boxed{X \sim \text{Pois}(\lambda): \quad \langle x \rangle = \lambda} \tag{26}$$

This is a very special result: Poi (λ) is fully characterized by just a single parameter λ, which also happens to be the mean. Thus, if $x \sim$ Poi(λ) and you want to find out what is λ, all you need to know is $\langle x \rangle$.

Let me also show you another cool way of finding $\langle x \rangle$. We have

$$\langle x \rangle = e^{-\lambda} \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} \tag{27}$$

Now we start with the Taylor series of e^λ

$$e^\lambda = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}$$

Differentiate both sides with respect to λ

$$e^\lambda = \sum_{k=0}^{\infty} k \frac{\lambda^{k-1}}{k!}$$

This is almost (27). we are just missing a λ. So multiply by λ on both sides:

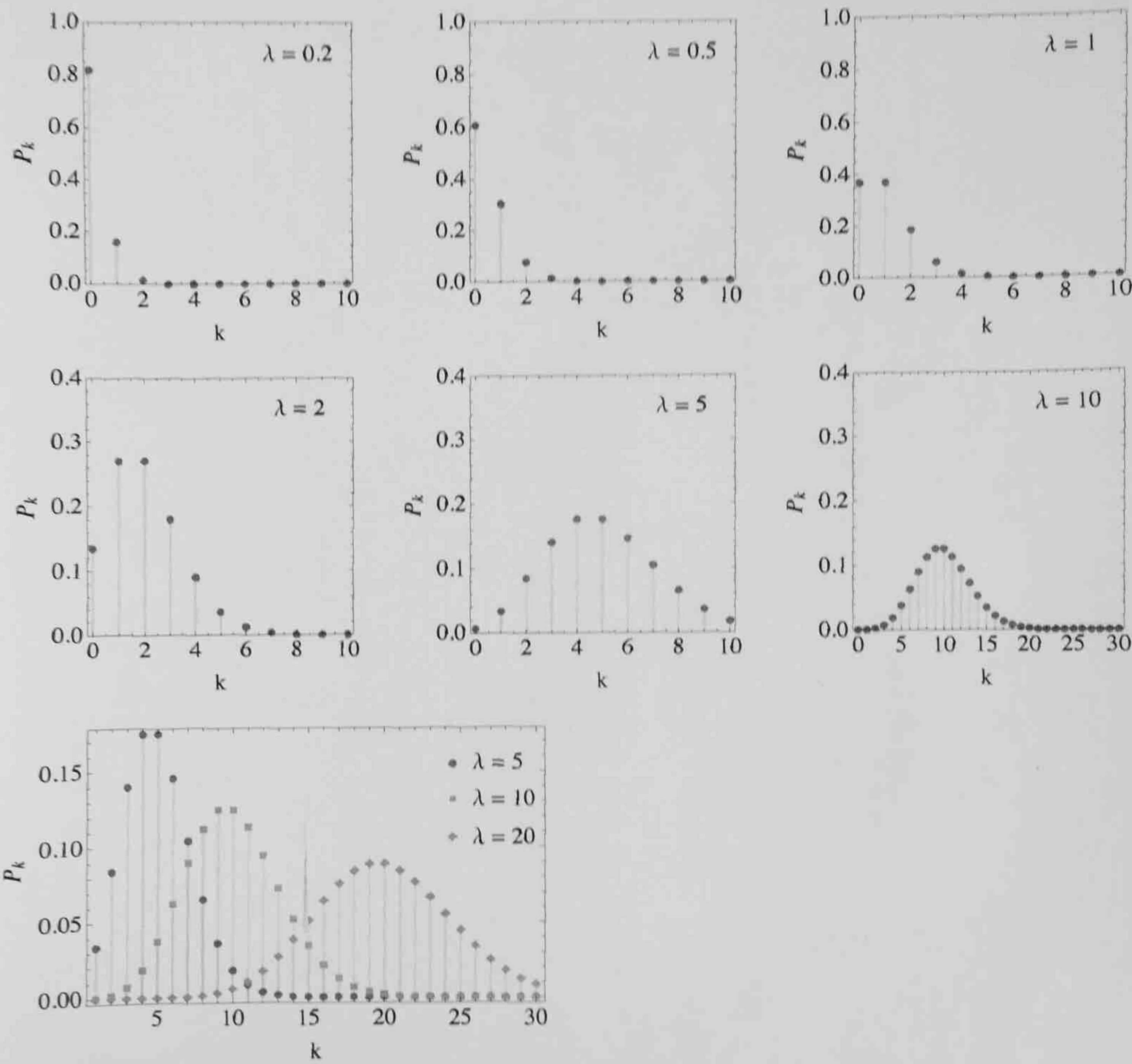
$$\lambda e^\lambda = \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} \tag{28}$$

We have replenished the missing λ. Now we have exactly (27). Thus

$$\langle x \rangle = e^{-\lambda} \lambda e^\lambda = \lambda$$

which is (26).

Plots of Pois λ



Poisson paradigm

The Poisson distribution appears approximately whenever you are analyzing the number of "successes" of a very large number of trials, where each trial has a very small chance of success.

An important example is radioactive decay. Consider a sample full of radioactive atoms and suppose we are measuring the number of decays per minute. The prob. that a given atom decays during that minute is very small. But there are many many atoms. So you get some finite result.

If you know that, on average $\lambda = \langle X \rangle$ atoms decay per minute, then X will be approximately $\mathrm{Pois}(\lambda)$.

As another example, let X denote the number of goals in a world cup match. The prob. that one player scores is very small, but there are many players, so $X \sim \mathrm{Pois}(\lambda)$ (approximately, of course). We know that the average number of goals is $\langle X \rangle = 2.5$. If we assume that $X \sim \mathrm{Pois}(\lambda)$, then we may estimate the full distribution

$$P(k \text{ goals in a match}) = e^{-2.5} \frac{(2.5)^k}{k!} \qquad (29)$$

We may even make a table

# of goals	0	1	2	3	4	5	6	7	8
prob.	0.082	0.205	0.257	0.213	0.133	0.067	0.028	0.01	0.0031

We were here in 2014

Other examples of weird situations that are approximately Poisson include

- the number of e-mails you receive per hour
- the number of chocolate chips in a cookie
- the number of earthquakes in a certain region
- the number of rain drops falling in a $1 cm^2$ area
- the number of patients in the ER during 1 h.

Binomial converges to the Poisson

Let $X \sim Bin(n,p)$ and assume that $n \to \infty$ but $p \to 0$, in such a way that $\lambda = np$ remains fixed. That is, there are many trials, but the chance of success in each is small.

Substituting $p = \lambda / n$ in the distribution $P(X=k)$ of the Binomial, Eq (12), we get

$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} \frac{\lambda^k}{n^k} \left(1 - \frac{\lambda}{n}\right)^{n-k}$$

We now write

$$\frac{n!}{(n-k)!\, n^k} = \frac{n(n-1)\dots(n-k+1)}{n^k} \approx 1 \text{ when } n \to \infty$$

Consequently, we get

$$P(X=k) \approx \frac{1}{k!} \lambda^k \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k}$$

When $n \to \infty$ we then get

$$\left(1 - \frac{\lambda}{n}\right)^{-k} \to 0$$

$$\left(1 - \frac{\lambda}{n}\right)^{n} \to e^{-\lambda}$$

the last limit is the definition of the exponential (if you leave your money in the bank, it grows exponentially). We therefore get

$$P(X=k) \approx \frac{e^{-\lambda} \lambda^k}{k!}$$

Thus, when $n \to \infty$, $p \to 0$ but $\lambda = np$ remains fixed, the Binomial converges to the Poisson. This is the origin of Poisson's paradigm

The Variance

The mean is known as the first moment of the distribution. Similarly, we define the second moment as

$$\langle x^2 \rangle = \sum_k k^2 P_k \tag{30}$$

More generally, if $f(x)$ is an arbitrary function, then

$$\boxed{\langle f(x) \rangle = \sum_k f(k) P_k} \tag{31}$$

The mean provides a measure of the position of the distribution. It is also interesting to have a measure of the width. A good measure is the variance

$$\boxed{\text{Var}(x) = \langle (x - \langle x \rangle)^2 \rangle = \langle x^2 \rangle - \langle x \rangle^2} \tag{32}$$

I wrote two definitions, but they are actually equal:

$$\begin{aligned} \langle (x - \langle x \rangle)^2 \rangle &= \langle x^2 - 2\langle x \rangle x + \langle x \rangle^2 \rangle \\ &= \langle x^2 \rangle - 2\langle x \rangle \langle x \rangle + \langle x \rangle^2 \\ &= \langle x^2 \rangle - \langle x \rangle^2 \end{aligned}$$

The variance is always positive, which is evident from the first definition. It is zero only when $x = \langle x \rangle$, that is, when x is a deterministic variable.

Example: $X \sim \text{Bern}(p)$

First we compute

$$\langle X^2 \rangle = (0)^2 P_0 + (1)^2 P_1 = p \tag{33}$$

In fact, for Bern (p)

$$\langle X^n \rangle = p \, , \quad n = 1, 2, 3, \dots \tag{34}$$

Now we compute the variance as $\langle X^2 \rangle - \langle X \rangle^2$:

$$\boxed{X \sim \text{Bern}(p): \quad \text{Var}(X) = p - p^2 = p(1-p)} \tag{35}$$

—//—

The variance satisfies

$$\boxed{\begin{aligned} \text{Var}(cX) &= c^2 \, \text{Var}(X) \\ \text{Var}(X+c) &= \text{Var}(X) \end{aligned}} \tag{36}$$

The first is reassuring: even if $c<0$, you always get a non-negative variance. The second shows that $\text{Var}(X)$ is indeed a good measure of the spread of the distribution: If you shift it by c, the variance remains unaltered. I leave it to you to verify (36).

If X has units of meter, then $\text{Var}(X)$ will have units of m^2. It is therefore useful to have a definition of width in meters. For this reason we define the <u>standard deviation</u>

$$\boxed{SD(X) = \sqrt{\text{Var}(X)}} \tag{37}$$

Example: variance of the Poisson

We first compute $\langle x^2 \rangle$ using the differentiation trick:

$$\langle x^2 \rangle = e^{-\lambda} \sum_{k=0}^{\infty} k^2 \frac{\lambda^k}{k!} \qquad (38)$$

We have:

$$e^{\lambda} = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}$$

$\frac{d}{d\lambda}$:

$$e^{\lambda} = \sum_{k=0}^{\infty} \frac{k\,\lambda^{k-1}}{k!}$$

Replenish λ:

$$\lambda\, e^{\lambda} = \sum_{k=0}^{\infty} \frac{k\,\lambda^{k}}{k!}$$

$\frac{d}{d\lambda}$:

$$(1+\lambda)\, e^{\lambda} = \sum_{k=0}^{\infty} \frac{k^2\,\lambda^{k-1}}{k!}$$

Replenish again:

$$\lambda(1+\lambda)\, e^{\lambda} = \sum_{k=0}^{\infty} \frac{k^2\,\lambda^{k}}{k!}$$

This is (38) so

$$\langle x^2 \rangle = e^{-\lambda}\, \lambda(1+\lambda)\, e^{\lambda} = \lambda + \lambda^2 \qquad (39)$$

The variance is then

$$\begin{aligned} \mathrm{var}(x) &= \langle x^2 \rangle - \langle x \rangle^2 \\ &= \lambda + \lambda^2 - \lambda^2 \\ &= \lambda \end{aligned}$$

$$\therefore \quad \boxed{x \sim \mathrm{Pois}(\lambda): \quad \mathrm{var}(x) = \lambda} \qquad (40)$$

Statistical independence ~~and conditional prob.~~

Let X and Y be two random variables. We denote

$$P(X=k, Y=l) = \text{Prob. that } \underline{\text{both}}\ X=k \text{ and } Y=l \qquad (41)$$
$$= \text{"Joint distribution"}$$

We say that X and Y are <u>statistically</u> <u>independent</u> when

$$\boxed{P(X=k, Y=l) = P(X=k)\,P(Y=l)} \qquad (42)$$

Remember this:

> " If two quantities are independent, the prob. is a <u>product</u> " (43)

Statistical independence implies the following:

$$\langle XY \rangle = \sum_{k,l} k\,l\,P(X=k, Y=l)$$
$$= \sum_{k,l} k\,l\,P(X=k)\,P(Y=l)$$
$$= \left[\sum_{k} k\,P(X=k)\right]\left[\sum_{l} l\,P(Y=l)\right]$$
$$= \langle X \rangle \langle Y \rangle$$

$$\therefore \quad \boxed{\langle XY \rangle = \langle X \rangle \langle Y \rangle} \quad \text{when } X, Y \text{ are stat. indep.} \qquad (44)$$

Now let $Z = X + Y$. By linearity

$$\langle Z \rangle = \langle X \rangle + \langle Y \rangle \tag{45}$$

As for the second moment, we have

$$\begin{aligned}
\langle Z^2 \rangle = \langle (X+Y)^2 \rangle &= \langle X^2 + 2XY + Y^2 \rangle \\
&= \langle X^2 \rangle + 2\langle XY \rangle + \langle Y^2 \rangle \\
&\overset{\text{stat indep.}}{=} \langle X^2 \rangle + 2\langle X \rangle\langle Y \rangle + \langle Y^2 \rangle
\end{aligned}$$

The variance is then

$$\begin{aligned}
\mathrm{Var}(Z) &= \langle Z^2 \rangle - \langle Z \rangle^2 \\
&= \left[\langle X^2 \rangle + 2\langle X \rangle\langle Y \rangle + \langle Y^2 \rangle \right] - \left[\langle X \rangle^2 + 2\langle X \rangle\langle Y \rangle + \langle Y \rangle^2 \right] \\
&= \langle X^2 \rangle - \langle X \rangle^2 + \langle Y^2 \rangle - \langle Y \rangle^2
\end{aligned}$$

Thus

$$\boxed{\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)} \quad \text{when } X, Y \text{ are stat indep} \tag{46}$$

This is a quite remarkable result. No such structure holds for $\langle Z^2 \rangle$. But for the variance we find a perfect cancelation. It is possible to find combinations of higher powers of Z that also satisfy this property of additivity. These quantities are called <u>cumulants</u> and they play a very important role in statistical mechanics and quantum field theory.

The interesting thing about cumulants is that they are extensive: if you have two independent parts, the cumulant is simply the sum of the individual cumulants. For example, the heat capacity is a cumulant.

Warning: Eq (45) holds even if X and Y are dependent. But Eq (46) is true only if they are independent.

To make sense out of this, suppose $Y = X$ (this is an example of extreme dependence). Then by (36)

$$\mathrm{Var}(X+Y) = \mathrm{Var}(2X) = 4\,\mathrm{Var}(X)$$

which disagrees with (46)

Example: variance of $\mathrm{Bin}(n,p)$

Recall that $\mathrm{Bin}(n,p)$ is composed out of n $\mathrm{Bern}(p)$ trials. Let

$$X_i \sim \mathrm{Bern}(p) \quad , \quad i = 1, \dots, n \tag{47}$$

We say the X_i are independent, identically distributed (iid) r.v.s. Then

$$Y = X_1 + \dots + X_n \sim \mathrm{Bin}(n,p) \tag{48}$$

We found in (35) that $Var(X_i) = p(1-p)$. Thus, by (46) we will have

$$\boxed{Y \sim Bin(n,p): \quad Var(Y) = np(1-p)} \tag{49}$$

You may also find this directly from the definition, but it is much harder.

Random walk

Suppose you are drunk and you are trying to get home from a party. At each time you either take a step to the left or to the right. Assume also that the steps always have the same length, which we define as "1" in some weird (drunk) system of units. Then each step may be described by a random variable $X_i = \pm 1$.

After t steps your position will be

$$Y_t = X_1 + X_2 + \dots + X_t. \quad (50)$$

Let $P(X_i = +1) = p$ and $P(X_i = -1) = 1 - p$. Then

$$\langle X_i \rangle = (+1)\,p + (-1)(1-p) = 2p - 1 \quad (51)$$

Consequently, after t steps your average position will be

$$\langle Y_t \rangle = t\,(2p - 1) \quad (52)$$

If you are really really really drunk then $p = 1/2$ (you have no idea where you are going) and consequently

$$\langle Y_t \rangle = 0$$

In this case you don't really go anywhere.

Next let us compute the variance. We have

$$\langle x_i^2 \rangle = (+1)^2 p + (-1)^2 (1-p) = 1$$

Thus

$$\begin{aligned} \mathrm{Var}(x_i) &= \langle x_i^2 \rangle - \langle x_i \rangle^2 \\ &= 1 - (2p-1)^2 \\ &= 1 - 4p^2 + 4p - 1 \\ &= 4p(1-p) . \end{aligned}$$

Since the x_i are independent, by Eq (46) we will have

$$\mathrm{Var}(Y_t) = \sum_{i=1}^{t} \mathrm{Var}(x_i) = 4tp(1-p) \tag{53}$$

If $p = 1/2$ (super drunk) we get

$$\mathrm{Var}(Y_t) = t \tag{54}$$

So even though on average you are not going anywhere, the width of your distribution increases with the number of steps.

Conditional probability

Let X and Y be two (not necessarily independent) r.v.s. We define the conditional probability as

$$\boxed{P(X=k \mid Y=\ell) = \text{Prob. of } X=k \text{ \underline{given that} } Y=\ell} \qquad (55)$$

Mathematically, we define it from

$$\boxed{\begin{aligned} P(X=k, Y=\ell) &= P(X=k \mid Y=\ell)\, P(Y=\ell) \\ &= P(Y=\ell \mid X=k)\, P(X=k) \end{aligned}} \qquad (56)$$

Please take a second to think about the intuition behind this formula.

In terms of conditional probabilities, we may also formulate the definition of statistical independence in an intuitive way:

$$\boxed{P(X=k \mid Y=\ell) = P(X=k)} \quad \text{If } X, Y \text{ are stat. indep} \qquad (57)$$

In words: whatever outcome you got for Y has no influence on X.

Conditional probabilities may also be used to simplify certain problems. For instance, consider this: let

$$X_i \sim \mathrm{Pois}(\lambda)$$

Show that $Y = X_1 + X_2 \sim \mathrm{Pois}(2\lambda)$. We may do this using something called the law of total probability

$$\boxed{P(Y=k) = \sum_{l} P(Y=k \mid X_1 = l)\, P(X_1 = l)} \tag{58}$$

we condition $Y=k$ on all possible outcomes of X_1 and sum over the result. But since X_1 and X_2 are independent

$$P(Y=k \mid X_1 = l) = P(X_1 + X_2 = k \mid X_1 = l)$$
$$= P(X_2 = k-l) \quad \leftarrow \text{I used independence.}$$

this is true if $l \leq k$. Otherwise we get zero.

$$P(Y=k) = \sum_{l=0}^{k} P(X_2 = k-l)\, P(X_1 = l)$$
$$= e^{-\lambda} e^{-\lambda} \sum_{l=0}^{k} \frac{\lambda^{k-l}}{(k-l)!} \frac{\lambda^{l}}{l!}$$
$$= e^{-2\lambda} \lambda^{k} \sum_{l=0}^{k} \frac{1}{l!\,(k-l)!}$$
$$= e^{-2\lambda} \frac{\lambda^{k}}{k!} \underbrace{\sum_{l=0}^{k} \binom{k}{l}}_{= 2^k \text{ by the Binomial theorem}}$$
$$= e^{-2\lambda} \frac{(2\lambda)^{k}}{k!}$$

thus we see that, indeed, $Y \sim \mathrm{Pois}(2\lambda)$.