Probability distributions

Part 1

Random variables

A random variable can take on a set of possible different values (similarly to other mathematical variables), each with an associated probability.

Its possible values are possible outcomes of a random event, i.e. an "experiment" whose value is uncertain.

If you toss a coin, the outcome can assume two possible values, Head and Tail. We associate these values with two numerical values, say 0 and 1 (the association is arbitrary).

Intuitively, we define a random variable defined in the set {0, 1}with this *probability distribution:*

$$P(0) = 0.5,\ P(1) = 0.5$$

[Note: we are assuming the coin is "fair".]

If we roll a fair dice, the probability distribution is

P(1) = 1/6, P(2) = 1/6, …, P(6) = 1/6

The choice of values 1, 2, …, 6 is arbitrary, chosen only because natural. Every choice of 6 distinct numerical values would work.

If we have an urn with 6 white balls, 3 black balls and 1 grey ball, the probability distribution is

P(white) = 0.9, P(black) = 0.3, P(grey) = 0.1

We have written names instead of numbers for clarity.

All these examples are about *finite discrete probability distributions*: we have a finite collection of possible values which are natural numbers, or representable as natural numbers.

A distribution can be infinite and can be *continuous*, i.e. can assume infinite real values.

We will see both discrete and continuous probability distributions.

Bernoulli distribution

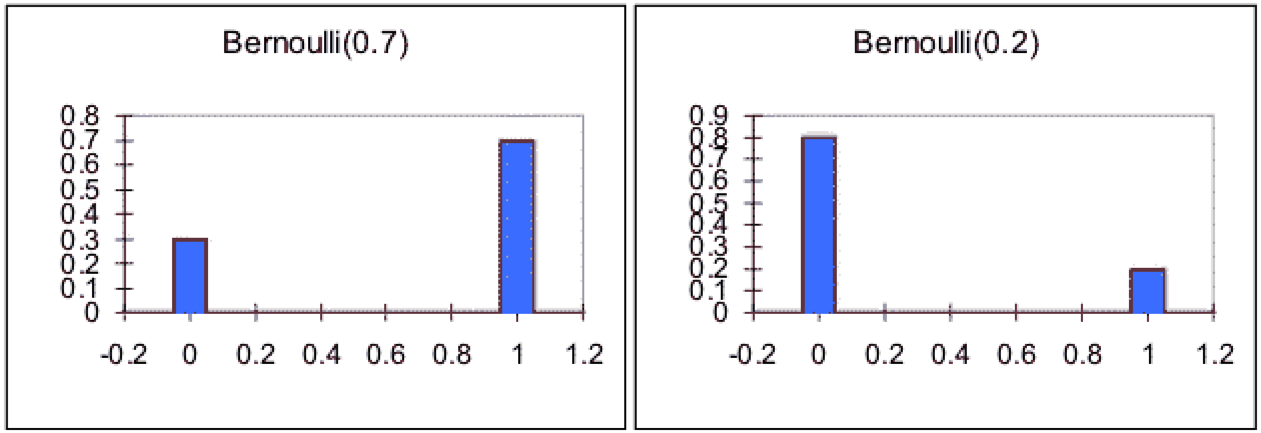
The Bernoulli distribution is a model for a random event with two possible outcomes.

Conventionally, its values are in {0, 1} and the distribution is

$$\text{Bernoulli}(p): P(1) = p,\ P(0) = 1 - p$$

If $p = 0.5$, this is a model for a fair coin toss. If $p \neq 0.5$, it is a model for an unfair coin toss.

Note: the sum of single probabilities must be 1. This is true for every discrete distribution.



Two graphical representations of Bernoulli distribuitons.

X-axis: possible values (here only 0 and 1).

Y-axis: probability of the x value.

For a discrete distribution, the probability of a certain value is named *probability mass function* and written as *pmf*:

pmf(1) ≡ P(outcome = 1) = 0.5

where the symbol ≡ means "is defined as".

If we want to model the event "an announce is clicked or not", we can use a Bernoulli distribution whose parameter *p* is the CTR (click through rate), i.e. the ratio

clicks / impressions

Pay attention. The interesting point here is that we do not know what is *p*.

We are building a model: we think of the process of an user clicking or not clicking an ad in terms of an abstract concept, a random variable. We assume it is a Bernoulli variable.

The CTR is a virtual quantity, something we imagine as a click generator.

The *observed CTR* is a different thing: it is the ratio of clicks versus impressions which happened.

The observed CTR is modeled as a realization of a random variable.

Another view is that the observed CTR is a "sample" extracted from a population of possible realizations of the random variable.

If we know the parameter *p* of the *Bernoulli(p)*, of course we know that

$$P(1) = p \text{ and } P(0) = 1 – p$$

This is trivial because implicit in the definition. Yet, the inverse problem is far from trivial.

We want to estimate the unknown parameter of a Bernoulli random variable observing the outcome of an experiment.

Before delivering an impression, what can we say about *p*, the CTR? The intuitive answer is: nothing. It can be every number in the real segment [0, 1] *with the same probability*.

Note: there are infinite possible real values for *p*.

The CTR is a random variable with a continuous probability distribution! The CTR distribution is much more complex than a Bernoulli.

An impression was delivered, the user did not click, i.e. the outcome was 0. The observed CTR is 0/1 = 0.

What can we say **now** about *p*, the CTR?

First: the "true" CTR cannot be 1. If it was 1, necessarily it would have given outcome 1 (1 is click, 0 is not click).

The observed CTR is not compatible with the hypothesis that the true CTR is 1.

Excluding 1 is not very useful: infinite values in the interval [0, 1) are still possible.

Yet, we really know something more than "it is not 1".

Are you more confident in the assertion

The "true" CTR is in the interval [0.85, 0.95], or with

The "true" CTR is in the interval [0.05, 0.15] ?

Of course, a CTR value around 10% is now more credible than a value around 90%.

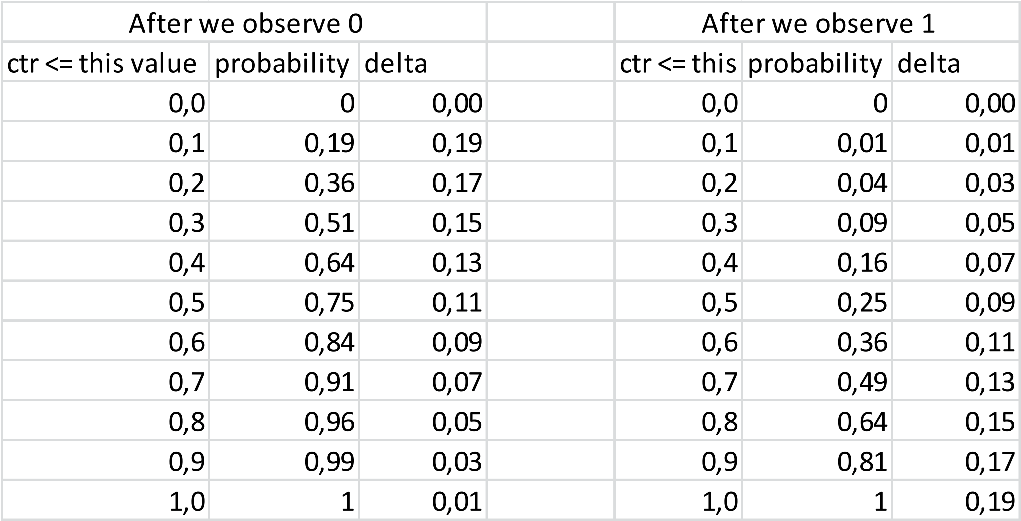
In Bayesian terms: we started with no preference, but after observing events we are a certain opinion about the unknown CTR. *The values around 10% are more likely than values around 90%, once we have observed the outcome non-click.*

Is it possible to estimate probabilities like

$P(a < x < b)$ where $x \sim \text{Bernoulli}(p)$?

(the symbol $\sim$ means x is an observed outcome of a Bernoulli distribution of parameter p).

The answer is yes.



After we observe 0				After we observe 1		
ctr <= this value	probability	delta		ctr <= this	probability	delta
0,0	0	0,00		0,0	0	0,00
0,1	0,19	0,19		0,1	0,01	0,01
0,2	0,36	0,17		0,2	0,04	0,03
0,3	0,51	0,15		0,3	0,09	0,05
0,4	0,64	0,13		0,4	0,16	0,07
0,5	0,75	0,11		0,5	0,25	0,09
0,6	0,84	0,09		0,6	0,36	0,11
0,7	0,91	0,07		0,7	0,49	0,13
0,8	0,96	0,05		0,8	0,64	0,15
0,9	0,99	0,03		0,9	0,81	0,17
1,0	1	0,01		1,0	1	0,19

After having observed 0, we can say that the unknown ctr is less or equal to 0.4 with probability 0.64. Prob(0.4 < ctr <= 0.5) = 0.11.

Note how the distribution of possible ctr values are "anti-mirrored".

After observing 1, Prob(ctr < 0.6) = 0.36, which is 1 – 0.36.

Why? Because 0.6 = 1 – 0.4.

[Computation with MS Excel]

Algorithms exist able to compute the *cumulative distribution function*, which is

$$\text{cdf}(z) \equiv \text{Prob}(x <= z)$$

the probability that a random value is less than or equal to z.

If you want to compute the probability that the unknown ctr is between 0.2 and 0.5, you can compute

$$\text{cdf}(0.5) – \text{cdf}(0.2)$$

The cdf is available in Excel and many other software tools.

Binomial distribution

The Bernoulli distribution models a single event 0/1, for us an impression which can give us a click or not.

Drawing inference from a single impression is not interesting in itself. We work with thousand or million impressions.

We need to understand not a single event but a long list of events, each one modeled as a Bernoullian event. Say, a repeated Bernoulli.

The Binomial distribution is what we need.

Binomial(n, p) is the distribution of probability of outcomes when we have n outcomes each one Bernoulli(p).

We toss a fair coin twice. The possible outcomes are

00 01 10 11

(choose 1 for Head and 0 for Tail or vice versa).

What is the probability of getting 0 successes? 0.25.

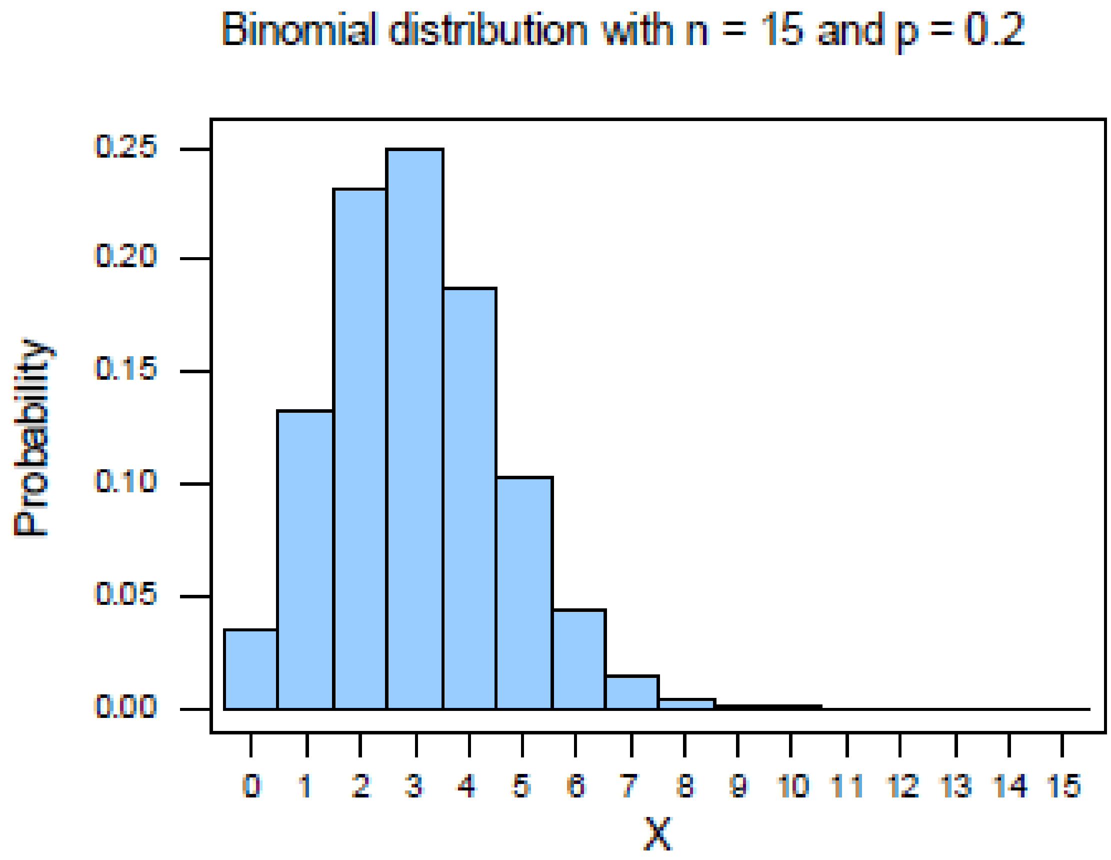
P(0) = 0.25

P(1) = 0.50

P(2) = 0.25

We are not interested in the sequence, only in the number of successes.

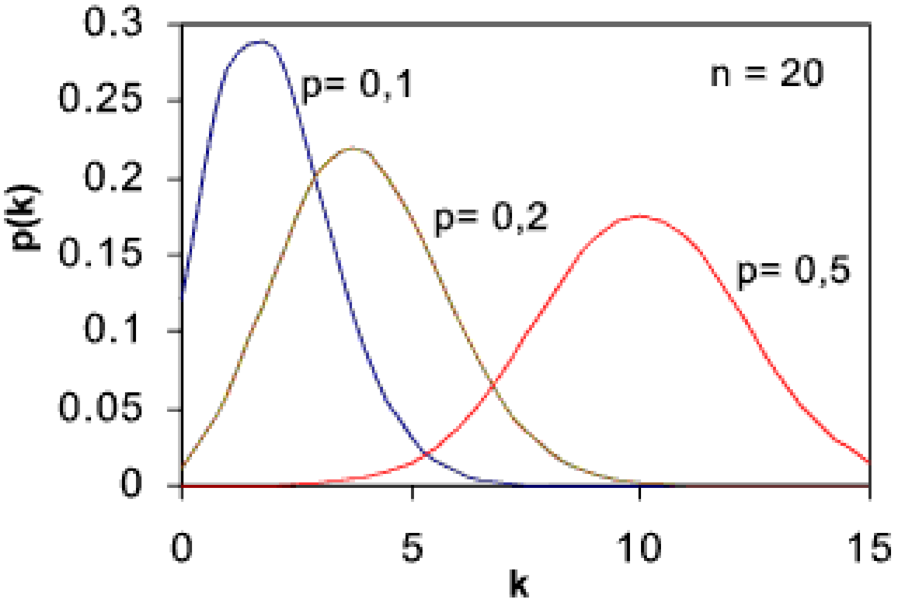
This is a Binomial(2, 0.5), two trials with probability of success 0.5.



We do 15 trial with success probability 20%.

Prob(3 successes) = 0.25, Prob(6 successes) = 0.04.

Prob(more than 10 successes) close to zero.



Binomial Probability

When the ctr increases, the probability distribution moves rightward.

High numbers of hits become more likely.

The curve becomes flatter moving to 0.5. After 0.5 it mirrors low values.

The curve for p = 0.8 mirrors that for 0.2, on the right side.

The flattest curve is for the "equilibrium" value.

Intuitively: p = 0.5 gives the max uncertainty.

Again, the problem we are mainly interested in is not predicting what can happen in 1,000,000 impressions if the CTR is 1%.

We do not know the CTR.

We have the inverse problem: observed an outcome, make inference about the unknown parameter.

We are able to compute Prob(parameter <= threshold).
It is the cdf, cumulative density function.

After we observe 5 clicks on 100 imps		
ctr <= this value	probability	delta
0,01	0%	0,00
0,02	2%	0,02
0,03	8%	0,07
0,04	22%	0,13
0,05	39%	0,17
0,06	57%	0,18
0,07	72%	0,15
0,08	83%	0,11
0,09	90%	0,07
0,10	95%	0,05
0,11	97%	0,03

Prob(ctr <= 0.06) = 57%

Prob(0.03 < ctr <= 0.07) = 72% - 8% = 64%

Indeed, having observed 5 / 100 we feel that the ctr must be around 5%.

Prob(ctr > 0.11) = 3% very unlikely.

If so, we were very unlucky, indeed.

Gist

To model the single event

an impression is delivered which may give a click

we use a distribution *Bernoulli(p)*, where *p* is the probability of getting a click, i.e. the CTR.

Observing the outcome of a single event (0 or 1 click) we can estimate

Prob(the unknown CTR is in the range a..b)

for each *a* and *b* between 0 and 1.

To model the event **many impressions are delivered** we use a distribution *Binomial(n, p)*, where *n* is the number of impressions and *p* is the CTR, again.

If we can estimate *p*, we can forecast the number of clicks we will get, i.e.

Prob(we get x clicks out of n impressions)

Vice versa, if we observe *k* clicks out of *n* impressions, we can estimate the unknown CTR, i.e.

Prob(the unknown CTR is in the range a..b)