Probability distributions

Part 2

Beta distribution

We model a single event **showing an impression and getting or not getting a click** with a *Bernoulli(p)* distribution, where p is the unknown CTR.

Then, we model a sequence of such events with

Binomial(n, p) where p is the same parameter.

We observe k clicks.

What can we say about p?

E.g., if we show 100 impressions and get 6 clicks, what can we say about the CTR?

We are speaking of the *true CTR*, not of the empirical frequency, which is 6%.

The CTR is itself a random variable, which a peculiar distribution which is called *Beta*.

In our example, its distribution is

Beta(7, 95)

The first parameter is the number of hits, the second the number of failures, both increased by 1 (it is an useful mathematical convention).

The probability that the CTR is in the range [0.01, 0.05) is

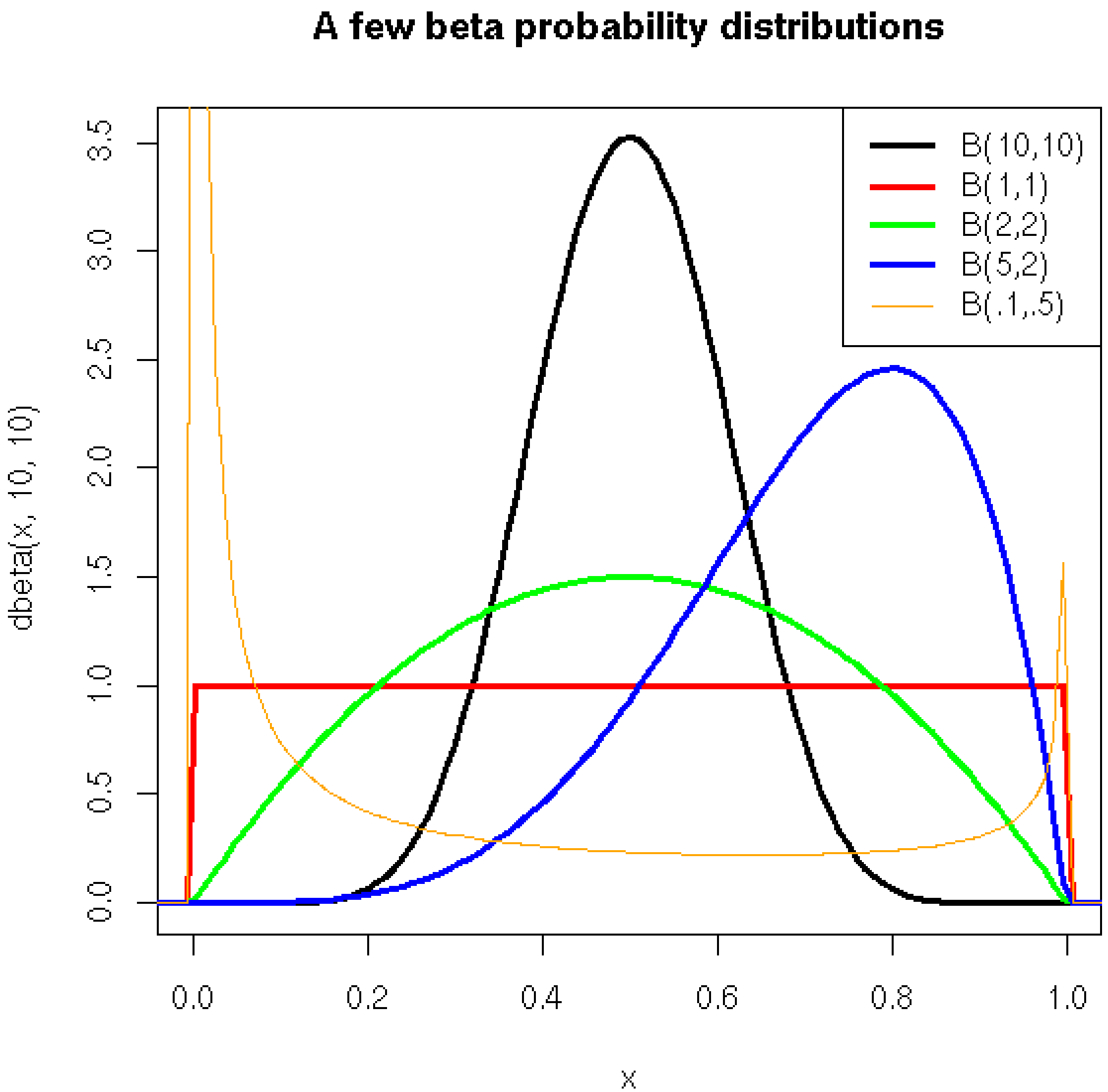
Prob(0.01 <= x < 0.05) with X ~ Beta(7, 95)

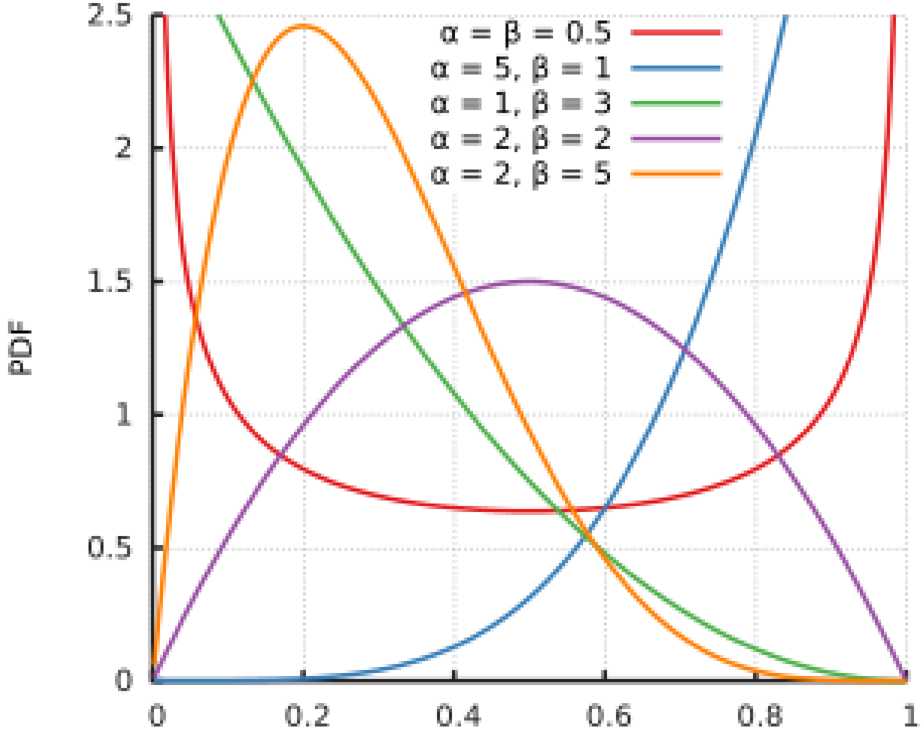
We know this probability is computable.

cdf(0.05) = 88.9% and cdf(0.01) = 24.1%

Prob(0.01 <= x < 0.05) = 88.9% - 24.1% = 64.8%

[cdf is the cumulative distribution function, cdf(z) ≡ Prob(x <= z)]


A few beta probability distributions
B(10,10)
B(1,1)
B(2,2)
B(5,2)
B(.1,.5)
dbeta(x, 10, 10)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0.0
0.2
0.4
0.6
0.8
1.0
x



PDF
2.5
2
1.5
1
0.5
0
0
0.2
0.4
0.6
0.8
1
α = β = 0.5
α = 5, β = 1
α = 1, β = 3
α = 2, β = 2
α = 2, β = 5

Gist

We model impressions delivering with a Binomial distribution.

We can use observed number of impressions n and number of clicks k to build a Beta distribution with parameters $k + 1$ (hits + 1) and $n - k + 1$ (failures + 1).

The Beta distribution describes the probability that the true CTR falls inside a certain region, e.g. between 1.0% and 1.1%.