Cross Validated

What is the intuition behind beta distribution?

Asked 6 years, 8 months ago Active 6 months ago Viewed 167k times

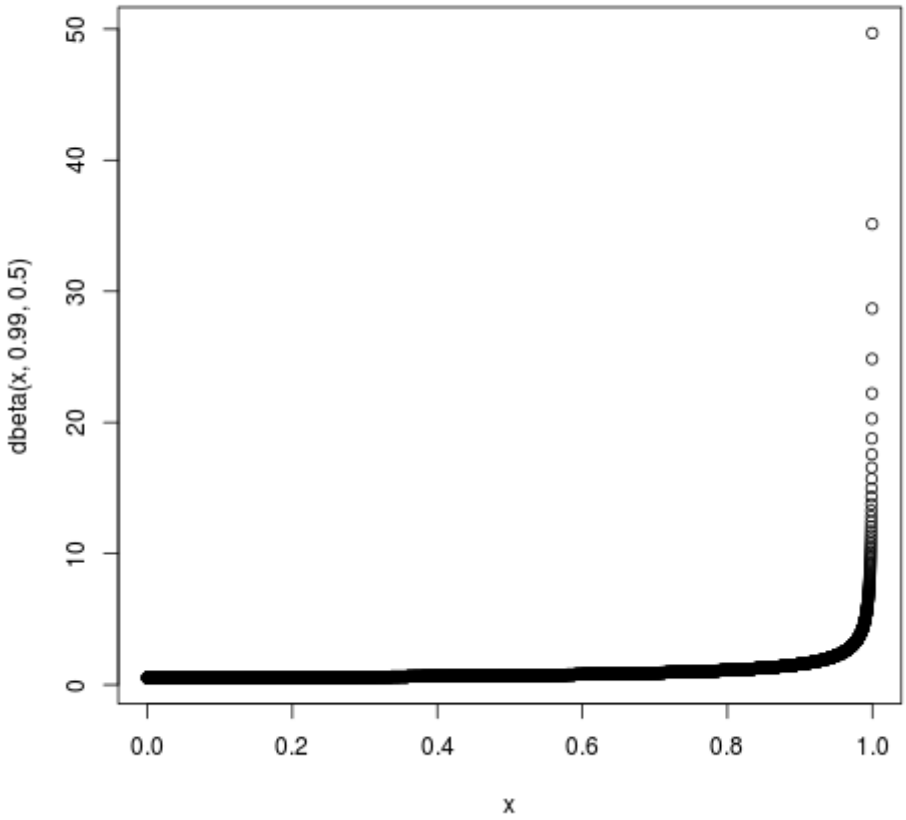
434

457

Disclaimer: I'm not a statistician but a software engineer. Most of my knowledge in statistics comes from self-education, thus I still have many gaps in understanding concepts that may seem trivial for other people here. So I would be very thankful if answers included less specific terms and more explanation. Imagine that you are talking to your grandma :)

I'm trying to grasp the **nature** of **beta distribution** - what it should be used for and how to interpret it in each case. If we were talking about, say, normal distribution, one could describe it as arrival time of a train: most frequently it arrives just in time, a bit less frequently it is 1 minute earlier or 1 minute late and very rarely it arrives with difference of 20 minutes from the mean. Uniform distribution describes, in particular, chance of each ticket in lottery. Binomial distribution may be described with coin flips and so on. But is there such **intuitive explanation** of **beta distribution**?

Let's say, $\alpha = .99$ and $\beta = .5$. Beta distribution $B(\alpha, \beta)$ in this case looks like this (generated in R):



But what does it actually mean? Y-axis is obviously a probability density, but what is on the X-axis?

I would highly appreciate any explanation, either with this example or any other.

distributions beta-distribution intuition beta-binomial

edited Feb 15 '17 at 9:14 — Tim ♦ — 65.4k 11 148 246

asked Jan 15 '13 at 15:31 — ffriend — 5,460 4 18 27

11 The y-axis is not a probability (which is obvious, because by definition a probability cannot lie outside the interval $[0, 1]$, but this plot extends up to 50 and--in principle--to ∞). It is a probability density: a probability per unit of x (and you have described x as a rate). – whuber ♦ Jan 15 '13 at 16:40

4 @whuber: yeah, I understand what PDF is - that was just mistake in my description. Thanks for a valid note! – ffriend Jan 15 '13 at 19:32

1 I'll try and find the reference but I know some of the more bizarre shapes for the generalized Beta distribution with form $a + (b - a)Beta(\alpha_1, \alpha_2)$ have applications such as physics. Also, you can fit it to expert data (min, mode, max) in data-poor environments and it is often better than using a Triangular distribution (unfortunately often used by IEs). – SecretAgentMan Sep 20 '18 at 1:40

You've obviously never traveled with the railway company Deutsche Bahn. You'd be less optimistic. – henning

617

+100

The short version is that the Beta distribution can be understood as representing a distribution *of probabilities*- that is, it represents all the possible values of a probability when we don't know what that probability is. Here is my favorite intuitive explanation of this:

Anyone who follows baseball is familiar with batting averages- simply the number of times a player gets a base hit divided by the number of times he goes up at bat (so it's just a percentage between `0` and `1`). `.266` is in general considered an average batting average, while `.300` is considered an excellent one.

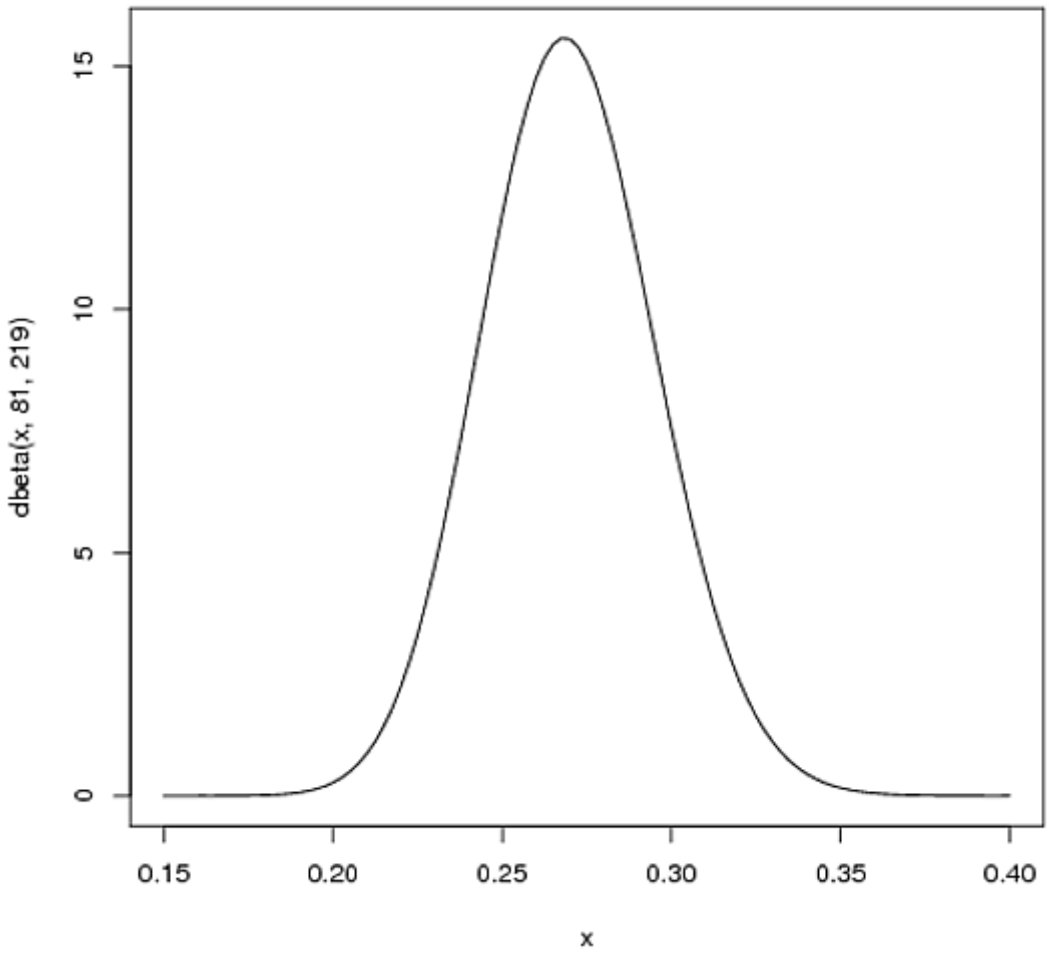
Imagine we have a baseball player, and we want to predict what his season-long batting average will be. You might say we can just use his batting average so far- but this will be a very poor measure at the start of a season! If a player goes up to bat once and gets a single, his batting average is briefly `1.000`, while if he strikes out, his batting average is `0.000`. It doesn't get much better if you go up to bat five or six times- you could get a lucky streak and get an average of `1.000`, or an unlucky streak and get an average of `0`, neither of which are a remotely good predictor of how you will bat that season.

Why is your batting average in the first few hits not a good predictor of your eventual batting average? When a player's first at-bat is a strikeout, why does no one predict that he'll never get a hit all season? Because we're going in with *prior expectations.* We know that in history, most batting averages over a season have hovered between something like `.215` and `.360`, with some extremely rare exceptions on either side. We know that if a player gets a few strikeouts in a row at the start, that might indicate he'll end up a bit worse than average, but we know he probably won't deviate from that range.

Given our batting average problem, which can be represented with a binomial distribution (a series of successes and failures), the best way to represent these prior expectations (what we in statistics just call a prior) is with the Beta distribution- it's saying, before we've seen the player take his first swing, what we roughly expect his batting average to be. The domain of the Beta distribution is `(0, 1)`, just like a probability, so we already know we're on the right track- but the appropriateness of the Beta for this task goes far beyond that.

We expect that the player's season-long batting average will be most likely around `.27`, but that it could reasonably range from `.21` to `.35`. This can be represented with a Beta distribution with parameters $\alpha = 81$ and $\beta = 219$:

```
curve(dbeta(x, 81, 219))
```



I came up with these parameters for two reasons:

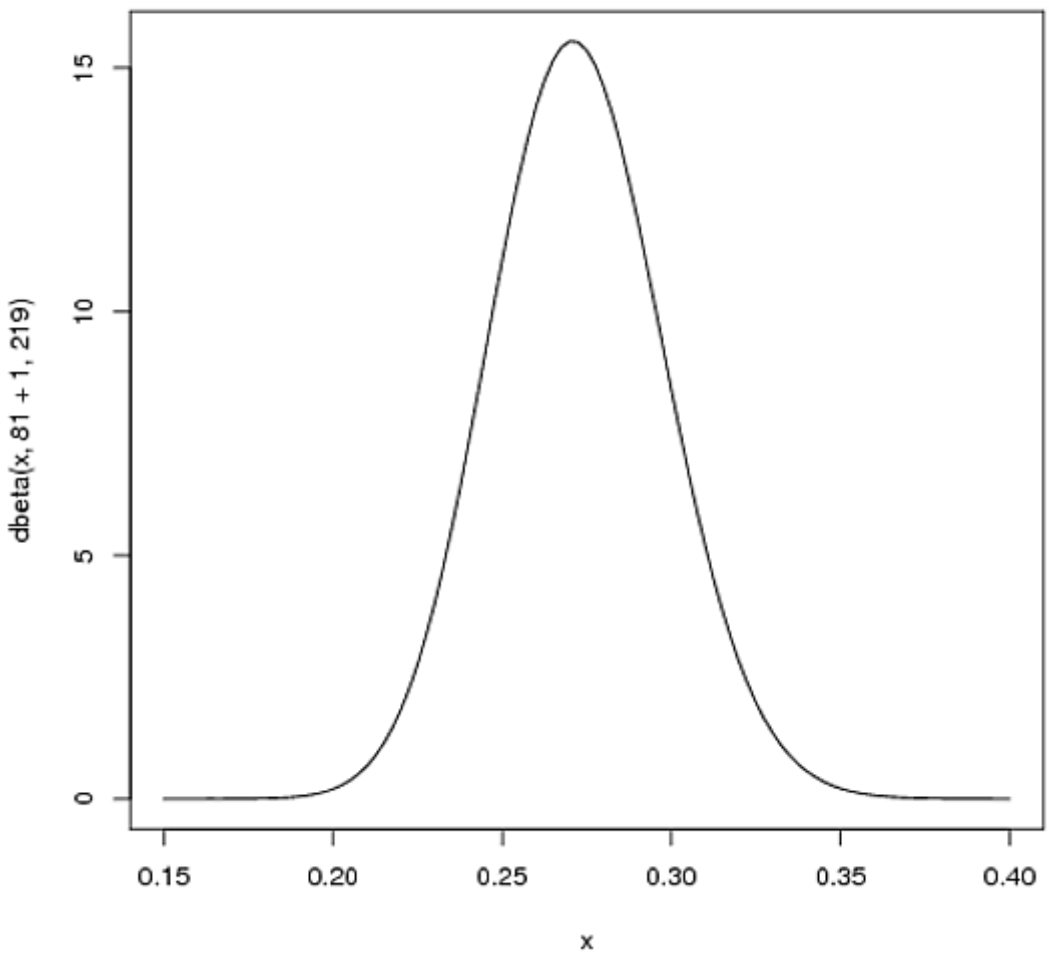
- The mean is $\frac{\alpha}{\alpha+\beta} = \frac{81}{81+219} = .270$
- As you can see in the plot, this distribution lies almost entirely within `(.2, .35)` - the reasonable range for a batting average.

You asked what the x axis represents in a beta distribution density plot- here it represents his batting average. Thus notice that in this case, not only is the y-axis a probability (or more precisely a probability density), but the x-axis is as well (batting average is just a probability of a hit, after all)! The Beta distribution is representing a probability distribution *of probabilities*.

But here's why the Beta distribution is so appropriate. Imagine the player gets a single hit. His record for the season is now `1 hit; 1 at bat`. We have to then *update* our probabilities- we want to shift this entire curve over just a bit to reflect our new information. While the math for proving this is a bit involved (it's shown here), the result is *very simple*. The new Beta distribution will be:

$$\text{Beta}(\alpha_0 + \text{hits}, \beta_0 + \text{misses})$$

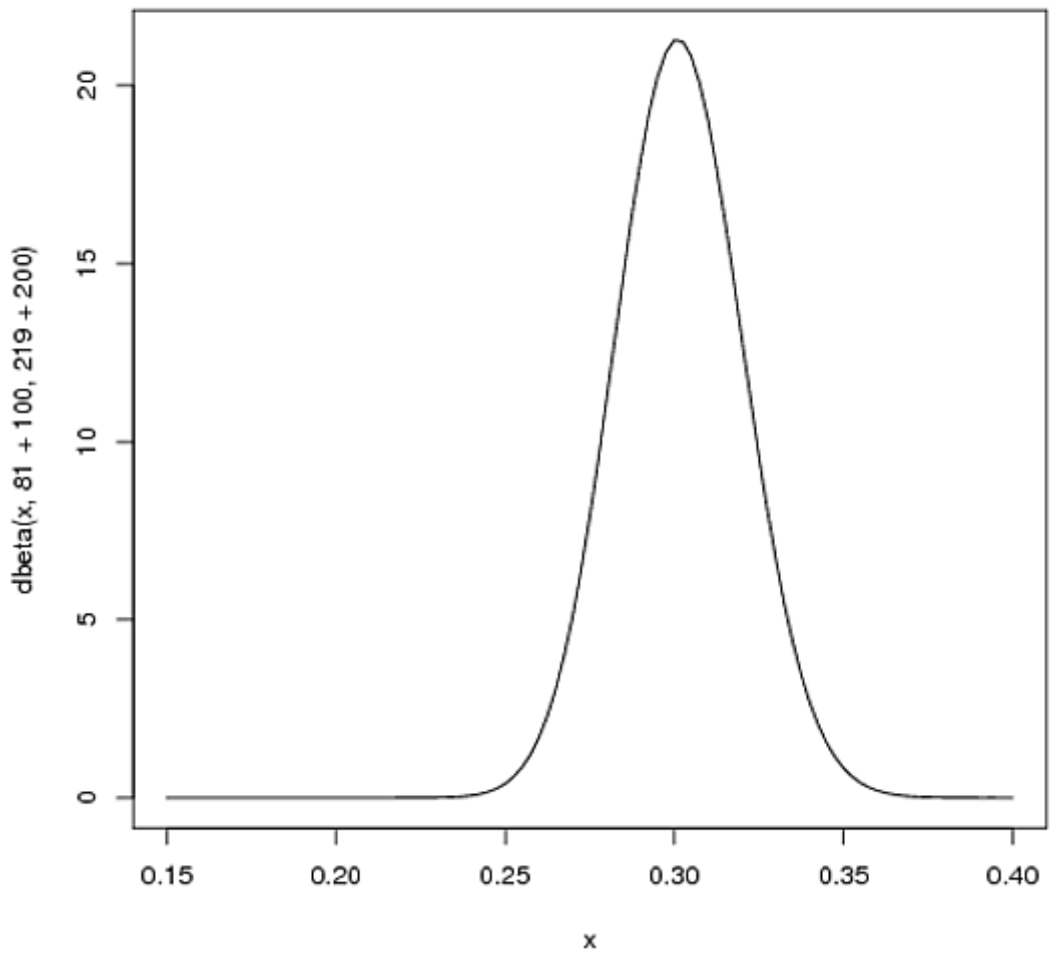
```
curve(dbeta(x, 82, 219))
```



Notice that it has barely changed at all- the change is indeed invisible to the naked eye! (That's because one hit doesn't really mean anything).

However, the more the player hits over the course of the season, the more the curve will shift to accommodate the new evidence, and furthermore the more it will narrow based on the fact that we have more proof. Let's say halfway through the season he has been up to bat 300 times, hitting 100 out of those times. The new distribution would be $\text{Beta}(81 + 100, 219 + 200)$, or:

```
curve(dbeta(x, 81+100, 219+200))
```



Notice the curve is now both thinner and shifted to the right (higher batting average) than it used to be- we have a better sense of what the player's batting average is.

One of the most interesting outputs of this formula is the expected value of the resulting Beta distribution, which is basically your new estimate. Recall that the expected value of the Beta distribution is $\frac{\alpha}{\alpha+\beta}$. Thus, after 100 hits of 300 *real* at-bats, the expected value of the new Beta distribution is $\frac{81+100}{81+100+219+200} = .303$- notice that it is lower than the naive estimate of $\frac{100}{100+200} = .333$, but higher than the estimate you started the season with ($\frac{81}{81+219} = .270$). You might notice that this formula is equivalent to adding a "head start" to the number of hits and non-hits of a player- you're saying "start him off in the season with 81 hits and 219 non hits on his record").

Thus, the Beta distribution is best for representing a probabilistic distribution *of probabilities*- the case where we don't know what a probability is in advance, but we have some reasonable guesses.

edited Apr 8 '17 at 0:11

rolando2

answered Jan 15 '13 at 16:41

David Robinson