An intuitive interpretation of the beta distribution

November 15, 2013

Share

First of all this text is not just about an intuitive perspective on the beta distribution but at least as much about the idea of looking behind a measured empirical probability and thinking of it as a product of chance itself. Credits go to David Robinson for approaching the subject from a baseball angle and to John D. Cook for establishing the connection to Bayesian statistics. In this article I want to add a simulation which stochastically prooves the interpretation.

Ratings of sellers on an e-commerce platform

Let's assume you would like to buy a product and you can choose from two sellers A and B. The only relevant difference is the ratings – which a customer can give to a seller after a purchase choosing between 'good' and 'bad'. Seller A received 10 ratings of which 9 are good and seller B received so far 500 ratings of which 400 (80%) happen to be 'good'.

The naive approach would be to just assume that the ratio of good ratings is by itself the best indicator of a seller's quality. Though I guess most people would automatically take the number of invoveld ratings into account – but lets ignore the psychological aspect and focus on a statistical approach.

The "true" but hidden rating level

When we make a choice for a seller in above situation based on a average rating we implicitely assume that the rating represents an inherent feature of the seller. But when we think about it a bit, it becomes clear that the observed average rating is based on chance and that even a seller with a true but invisible rating level of 50% will yield thanks to luck 80% for a while. In the long run the "law of large numbers" will bring it back to where it belongs but even with 50/50 chance of receiving a good rating (due to the true but hidden quality level) it is possible to end up with 9 of 10 positive ratings.

But of course we don't know the "true" rating of a seller but when we think of the rating process as a Bernoulli experiment (1 = good, 0 = bad), we can simulate a large number of independent Bernoulli experiments for a finite sequence of probabilities and see how likely it is to end up with the observed number of positive ratings for a given total number of ratings.

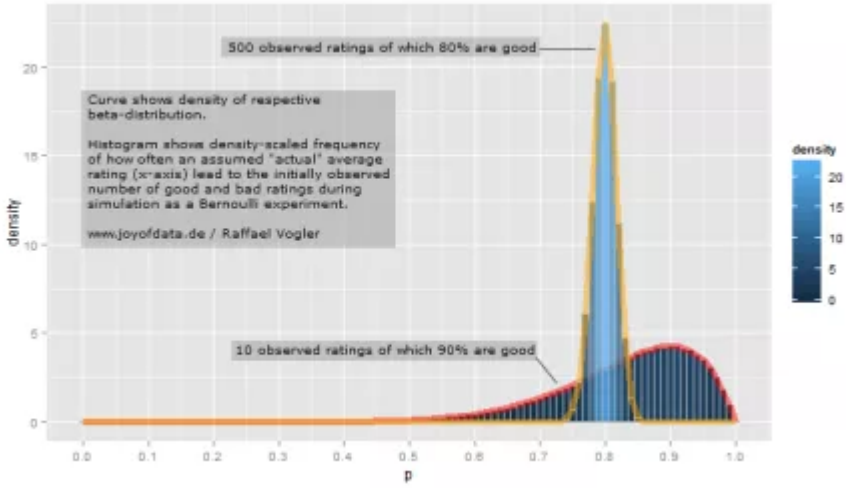
```
# simulates 10^5 Bernoulli experiments of length 500 for success probabilities
# from 0 to 1 in steps of size 0.01
# m will have 101 columns for every probability and 10^5 rows keeping
# the resulting sucess totals

m <- sapply(0:100/100, function(prob) rbinom(10^5, 400+100, prob))
```

```
# turns m into boolean matrix with true for element is 400 and else false.
# then we form sum all booleans (true as 1 and false as 0) and we know how
# often for a given probability (assumed rating) we observed 400 good ones.

v <- colSums(m == 400)

# v stores the number of succesful replications for every probability. But
# to feed it to hist() or geom_histogram() we need a vector keeping a probability
# as often as it leaded to a success.

df_sim <- data.frame(p = rep(0:100/100, v))
```

Let's make a complete R program of this idea and with ggplot2 we can see the resulting histograms for seller A (the flat shaped histogram) and seller B (the spiked shaped histogram). And now we finally get to the beta distribution whose density curve is plotted on top in red and orange colors.

Here comes the beta distribution into play

So the central observation is that the beta distribution f.x. with parameters α=400+1 and β=100+1 simply describes the probability that a certain true rating of seller B led to 400 positive ratings and 100 negative ratings.



A look at the confidence intervals

Now that we know that we can calculate a 95% confidence intervals for seller A and B for their true unknown rating level.

```
# 95%-CI for seller A (9 of 10 are good)

> qbeta(c(0.025,0.975),9+1,1+1)
[1] 0.5872201 0.9771688
```

```
# 95%-CI for seller B (400 of 500 are good)

> qbeta(c(0.025,0.975),400+1,100+1)
[1] 0.7626715 0.8326835
```

This tells us that the interval [58%, 98%] captures the true quality of seller A in terms of ratings with a chance of 95% and the interval [76%, 84%] captures the true quality of seller B (in terms of ratings) with a chance of 95%.

Possible applications

So one reasoning you could justify with those confidence intervals is that if a bad experience like a broken product due to bad packaging or late shipping is especially non-desirable you would choose seller B because his lower bound is about 20 p.p. higher than that of seller A. In another case – different scenario you might be more interested in tapping full potential, then you might go for seller A because with a 95% chance you can get up to 98% of true quality level.

For example the helpfulness ratings of product ratings (“34 people of 53 found this product rating to be helpful”) might be weighted using the lower bound of the 95% confidence interval or its span.

The full script

```
library(ggplot2)

# 90% positive of 10 ratings
o1 <- 9
o0 <- 1
M <- 100
N <- 100000

m <- sapply(0:M/M,function(prob)rbinom(N,o1+o0,prob))
v <- colSums(m==o1)
df_sim1 <- data.frame(p=rep(0:M/M,v))
df_beta1 <- data.frame(p=0:M/M, y=dbeta(0:M/M,o1+1,o0+1))

# 80% positive of 500 ratings
o1 <- 400
o0 <- 100
M <- 100
N <- 100000

m <- sapply(0:M/M,function(prob)rbinom(N,o1+o0,prob))
v <- colSums(m==o1)
```

```
df_sim2 <- data.frame(p=rep(0:M/M,v))
df_beta2 <- data.frame(p=0:M/M, y=dbeta(0:M/M,o1+1,o0+1))

ggplot(data=df_sim1,aes(p)) +
    scale_x_continuous(breaks=0:10/10) +

    geom_histogram(aes(y=..density..,fill=..density..),
        binwidth=0.01, origin=-.005, colour=I("gray")) +
    geom_line(data=df_beta1 ,aes(p,y),colour=I("red"),size=2,alpha=.5) +

    geom_histogram(data=df_sim2, aes(y=..density..,fill=..density..),
        binwidth=0.01, origin=-.005, colour=I("gray")) +
    geom_line(data=df_beta2,aes(p,y),colour=I("orange"),size=2,alpha=.5)
```