**Real Statistics Using Excel**
*Everything you need to do real*
*statistical analysis using Excel*

## Beta Distribution

**Definition 1**: For the binomial distribution the number of successes $x$ is the random variable and the number of trials $n$ and the probability of success $p$ on any single trial are parameters (i.e. constants). Instead we would like to view the probability of success on any single trial as the random variable, and the number of trials $n$ and the total number of successes in $n$ trials as constants.

Let $\alpha$ = # of successes in $n$ trials and $\beta$ = # of failures in $n$ trials (and so $\alpha + \beta = n$). The probability density function (pdf) for $x$ = the probability of success on any single trial is given by

$$f(x) = \frac{(n-1)!}{(\alpha-1)!\,(\beta-1)!} x^{\alpha-1}(1-x)^{\beta-1}$$

This is a special case of the beta function

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$
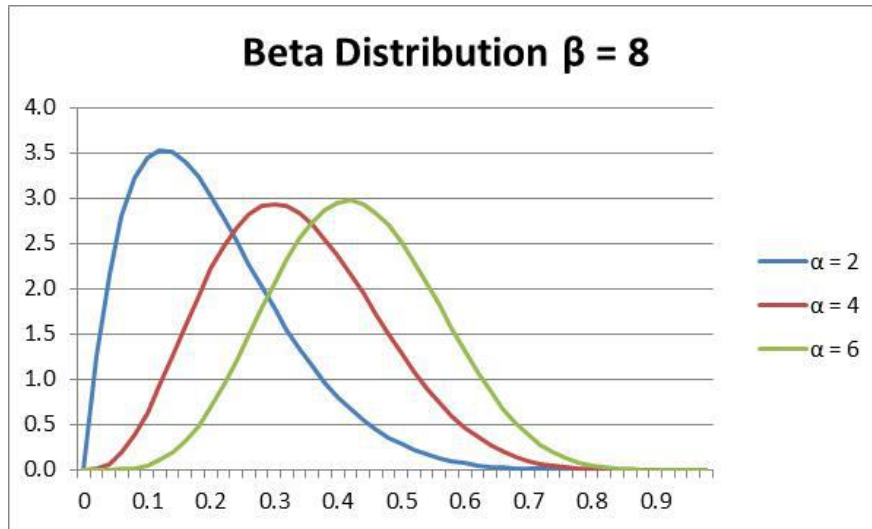
where $\Gamma$ is the [gamma function](#).



**Figure 1 – Beta Distribution**

**Excel Functions**: Excel provides the following functions:

**BETADIST**($x$, $\alpha$, $\beta$) = the cumulative distribution function $F(x)$ at $x$ for the pdf given above.

**BETAINV**($p$, $\alpha$, $\beta$) = $x$ such that BETADIST($x$, $\alpha$, $\beta$) = $p$. Thus BETAINV is the inverse of BETADIST.

Excel 2010/2013/2016 provide the following two additional functions: **BETA.INV** which is equivalent to BETAINV and **BETA.DIST**(*x, α, β, cum*) where *cum* takes the value TRUE or FALSE and BETA.DIST(*x, α, β*, TRUE) = BETADIST(*x, α, β*) while BETA.DIST(*x, α, β*, FALSE) is the pdf of the beta distribution at *x* (as described above).

**Example 1**: A lottery organization claims that at least one out of every ten people wins. Of the last 500 lottery tickets sold 37 were winners. Based on this sample, what is the probability that the lottery organization's claim is true: namely players have at least a 10% probability of buying a winning ticket? What is the 95% confidence interval?

To answer the first question we use the cumulative beta distribution function as follows:

BETADIST(.1, 37, 463) = 98.1%

This represents that organization's claim is false (i.e. less than 10% probability of success). Thus the probability that the organization's claim is true is 100% − 98.1% = 1.9%

The lower bound of the 95% confidence interval is

BETAINV(.025, 37, 463) = 5.3%

The upper bound of the 95% confidence interval is

BETAINV(.975, 37, 463) = 9.8%

Since 10% is not in the 95% confidence (5.3%, 9.8%), we conclude (with 95% confidence) that the lottery's claim is not accurate.

## 36 Responses to *Beta Distribution*

**Raj** *says:*
January 3, 2018 at 10:30 pm

Thank you Charles ..
However May i know what is the application of this charts

regards
Raj

Reply

**Charles** *says:*
January 3, 2018 at 10:39 pm

Raj,
To provide a sense of what the distribution looks like and what is the effect of changing a parameter value.
Charles

---

**Raj** *says:*

How did we get the graphs in the Beta distribution. can you please explain

**Charles** *says:*

Raj,

Enter 2 in cell B1, 4 in cell C1 and 6 in cell D1.

Next enter 0 in cell A2 and the formula =A2+.02 in cell A3. Highlight the range A3:A51 and press Ctrl-D.

Now insert the formula =BETA.DIST(B2,B$1,8,FALSE) in cell B2. Highlight range B2:D51 and press Crtl-R and then Ctrl-D.

This completes all the data. You now create the chart by highlighting range and selecting Insert > Charts > Line Chart.

You can then add titles and make other modifications as described on the webpage

Excel Charting

Charles

---

**Lam** *says:*

I used Exact Excel method to calculate the two sided CI (95% Confidence Level) as below. It is equivalent to "Clopper Pearson" method.

The upper bound is BETAINV(0.975, 38, 463) =10.1%. The lower bound is BETAINV(.025, 37, 464) = 5.3%.

In this case, 10% is in the CI (5.3%, 10.1%), we can conclude that the lottery's claim is accurate.

In sum, the exact method uses $\alpha$=38 instead of 37 used in the upper bound. In contrast, using the exact method draws a different conclusion although the two set of confidence limits are very close. There is statistical significance but no practical significance.

---

**Fred** *says:*

Charles,

First, thanks for devoting time to set up that website, it is a goldmine!

In your lottery example (nber 1) above, could we also use the Binomial Distribution to compute the probability to observe 37 or less winning tickets if the underlying probability is 10%?

So, =BINOM.DIST(37,500,0.1,1)?

Many thanks in advance Charles,

Fred

**Charles** *says:*
October 14, 2017 at 5:56 pm

Fred, BINOM.DIST(37,500,.1,1) = .02743, which means that the probability of getting 37 or fewer winning tickets out of 500 is 2.743% when the probability of getting a winning ticket (on any one ticket) is 10%. This doesn't answer question 1.
Charles

**Fred** *says:*
October 15, 2017 at 2:26 pm

Thank you! I was looking at it from the point of view of hypothesis testing, We would reject H0: p=0.1 because p-value=0.02743 <0.05?
You should really self-publish an e-book version of the website on Amazon.
Thanks again!
Fred

**R Singh** *says:*
February 28, 2017 at 7:04 pm

Hello there,
For some reason, the equations on all webpages are not displayed. May be some background maintenance is underway; just wanted to bring to your attention.
Thanks!

**Charles** *says:*
March 1, 2017 at 7:55 am

Thanks for bringing this to my attention.
I don't know why the equations were not displayed. I see that they are displayed now (at least on my computer).
Charles

**Karl Buennagel** *says:*
March 21, 2016 at 6:19 pm

Nice job Charles. That is a sweet explanation. I saved the webpage. Thanks! If this is a book, I need a copy!

**Charles** *says:*
March 21, 2016 at 8:54 pm

Karl,
There will be a series of books shortly.
Charles

**Adam** *says:*

I am a little confused how Bayesian updating is carried out for the following set-up:

I have developed a logistic regression model which provides % probability of a 1 or 0. The 1 or 0 is then compared to what actually happened and a Beta Distribution is updated for each new success or failure which provides % probability of successes (1s).

I now which to use my Logistic Regression model which new data to provide a new % estimate of a 1. This in turn now needs to be updated with the previously calculated probability from the Beta distribution calculated on the previous time step.

How should I use Bayes to update the current Logistic regression probity with the actual % wins/losses? I am confused as which one is the prior and psoterior?

Thanks for the help.

Adam

> **Charles** *says:*
>
> Sorry Adam, but I really don't understand the scenario that you are describing.
> Charles
>

> > **Anthony** *says:*
> >
> >
> > I believe Adam was referring to how the Beta distribution is often used iteratively in Bayesian stats. It can be explained as the probability of probabilities. E.g. what are the chances of having a final sample mean of 0.3 if the current sample mean is 0.27 with 2 out of 9 iterations remaining.
> >
> > The usual example is baseball, see http://varianceexplained.org/statistics/beta_distribution_and_baseball/, but your lottery example works.
> >
> > In summary, if we approach the goal at a faster rate than expected then we are more likely to meet or exceed that goal. And if we fall behind the less likely we are to are to meet it.
> >
> >

> > > **Charles** *says:*
> > >
> > >
> > > Anthony,
> > > Thanks for your explanation.
> > > Charles
> > >
> > >

**mitrananda** *says:*

how the alpha and beta values are to be taken for a problem.
please explain me.
thank you

Reply

> **Charles** *says:*
>
>
> For the problem described on the referenced webpage, α = # of successes in n trials and β = # of failures in n trials (and so α + β = n).
> Charles
>
> Reply

---

**John Vanneck** *says:*

Hi,

When there are 0 successes or n successes (out of n trials), the formula returns #NUM!

Do you have a work around for this? As in the situation where we had 100 trials and 0 successes we could safely assume that the probability of success was close to 0 – but this formula would not give an interval.

Reply

> **John Vanneck** *says:*
>
>
> oops – I forgot to mention that I am talking about estimating confidence intervals using the beta.inv(p,alpha,beta) formula. Thank you
>
> Reply

> **Charles** *says:*
>
>
> John,
>
> Note that BETA.DIST(p, s, n-s, True) = 1 – BINOM.DIST(s-1, n-1, p, True).
>
> Thus, BETA.DIST(p, 0, 100, True) = 1 – BINOM.DIST(-1, 99, p, True).
>
> This means that the problem you are addressing is equivalent to asking (the binomial distribution question) what is the probability of getting -1 successes in 99 trials (for any value of p = probability of success on any single trial)? This is clearly 0 with a confidence interval consisting only of the point zero. Thus the confidence interval you are looking for is [1,1], i.e. the point 1.
>
> Charles

**Rick** *says:*

"Thus the probability that the organization's claim is true is 100% − 98.1% = 1.9%"

This sure sounds like the inverse probability error.

$Pr(.074 \mid \pi = .1) = .019$
does not mean
$Pr(\pi = .1 \mid .074) = .019$

Would it not be better to say that obtaining p < 37/500 = .074 would occur only 1.9% of the time if the organization's claim (H0: $\pi = .1$) were true?

Rick

**Charles** *says:*

Rick,

While you are right to be cautious, I believe the logic used is correct.

$Pr(\pi < .1 \mid$ # of successes in 500 trials = 37 and # of failures in 500 trials) = BETADIST(.1, 37, 463) = 98.1% Thus $Pr(\pi >= .1 \mid$ # of successes in 500 trials = 37 and # of failures in 500 trials) = 1 − 98.1% = 1.9%

Charles

**Charles** *says:*

Rick,

While you are right to be cautious, I believe the logic used is correct.

$Pr(\pi < .1 \mid$ # of successes in 500 trials = 37 and # of failures in 500 trials) = BETADIST(.1, 37, 463) = 98.1% Thus $Pr(\pi >= .1 \mid$ # of successes in 500 trials = 37 and # of failures in 500 trials) = 1 − 98.1% = 1.9%

Charles

**Matthias** *says:*

I noticed that in this example, assuming 50 wins and 450 losses, so exactly 10%, yields:
BETADIST(.1, 50, 450) = 51.6%
and increasing the sample size to e.g. 5000 out of 45000 or higher, only leads to values closer and closer to

50%.
With the stated logic there would always be at least a 50% chance the claimed chance of 10% wins is a fraud.

Could it be that the correct calculation for 50 out 500 is the following:
2*(51.6% − 50%) = 3.2% chance of a lie
And with your example 37 out of 463:
2*(98.1% − 50%) = 96.2% chance of a lie

This way the whole range from 0% to 100% truth or lie is possible and entering the promised 10% wins approaches 0% chance of a lie with increasing sample size.

Not sure about this, though.
Thank you for the helpful article!

**Charles** *says:*

Matthias,

I don't know what you mean by "exactly 10%" in your first sentence. I also don't know what you mean by "a lie". Sorry, but I also don't understand the calculations. Let me review a few things.

BETADIST(.1, 50, 450) is the cumulative probability distribution function and is equal to 51.59%. This means that given that we have observed 50 winners in 500 lottery tickets, the probability that any single ticket has up to a 10% chance of winning is 51.59%. Note that I have said "up to a 10% chance of winning" and not "exactly a 10% chance of winning". All this is difficult to interpret since we have two levels of probability.

The chance that any single ticket has more than a 10% of winning is therefore 48.41%. Thus, it is more likely that any single ticket is a winner less than 10% of the time than more than 10% of the time, but the difference is close (and so my be due to chance).

Note that if you increase the sample size to 5000, but keep the 50 fixed, then BETADIST(.1,50,4950) = 1, which is what you would expect since it seems pretty unlikely that 10% of the tickets are winners given that only 50 out of 5,000 were winners. In fact, note that BETADIST(.01,50,4950) = 51.86%.

Charles

**Justin** *says:*

Good and simple explanation of beta distribution compared to 99% sites I looked.Could you tell why we place (n-1)! in the numerator when the number of trials will always be an whole number.It might be certain to have (\alpha-1)! and (\beta-1)! in the denominator since and might be real numbers.Also why isn't the power of x and (x-1) in the numerator \alpha and \beta instead of (\alpha-1) and (\beta-1) given in the expression?

**Charles** *says:*

Justin,
The only time I need to use the beta distribution on the website is when the alpha and beta values are integers, although the beta distribution is used for many other purposes, including cases where the alpha and beta parameters are not integers. The distribution uses the gamma function. You see a number of instances of some integer minus 1. The reason for this is that Gamma(n) = (n-1)! when n is a positive integer.
Charles

Reply

**Richard** *says:*
October 27, 2015 at 4:05 pm

I like your derivation (whenever I have seen this done previously, it has been through repeated application of integration by parts on the cdf, and is much more complicated), but I have a similar problem to Justin. To go from the pdf
f(x) = n!/k!/(n-k)!*x^k*(1-x)^(n-k)
I would have to set a-1=k and b-1=n-k, which gives
f(x) = n!/(a-1)!/(b-1)!*x^(a-1)*(1-x)^(b-1)
and is therefore not the same as the first of your equations above (by a factor of n). Do you see where I'm going wrong?

Reply

**Charles** *says:*
November 3, 2015 at 4:29 pm

Richard,

First note that Gamma(m) = (m-1)! for any positive integer m. Thus for any positive integers alpha and beta, (alpha-1)! = Gamma(alpha) and (beta-1)! = Gamma(beta). Now let n = alpha + beta. Then (n-1)! = Gamma(n) = Gamma(alpha+beta).

Putting the pieces together yields (n-1)!/((alpha-1)!*(beta-1)!) = Gamma(alpha+beta)/(Gamma(alpha)*Gamma(beta)).

This is not the same thing as deriving the cdf from the pdf.

Charles

Reply

**Sadam Hussain Channa** *says:*
December 26, 2017 at 6:34 pm

I used Beta Distribution Equation in Excel but result not same of Beta.Dist...Why

**Charles** *says:*
December 27, 2017 at 9:49 am

Sadam,
BETADIST calculates the cumulative distribution function (cdf) F(x), while the formula is for the probability density function (pdf) f(x).
BETA.DIST(x,alpha,beta,FALSE) will calculate the pdf.

**Rob** *says:*

September 4, 2013 at 8:50 am

It appears a typo propagated in the examples, surely you did not mean 163 for β – that gives a P of ~.9999 of tickets being at least 10% winners.

Neat site, BTW.

Rob

Reply

> **Charles** *says:*
>
> September 4, 2013 at 10:20 pm
>
> Hi Rob,
> There is a typo. The value for beta should actually be 463 Thanks for catching the error. I have just made the corrections to the website.
> Charles
>
> Reply

**m.r.sh** *says:*

September 4, 2013 at 8:10 am

Hi PRO!
If α + β = n then 37+163=200 but in your example 1 "…Of the last 500 lottery tickets 37 were winners…"! are you sure this is true?!

Reply

> **Charles** *says:*
>
> September 4, 2013 at 10:18 pm
>
> There is a typo. The value for beta should be 463. Thanks for catching the error. I have just made the corrections to the website.
> Charles
>
> Reply