Continuity correction for normal approximation of the binomial

Kris Gunsalus

10/18/2021

We've talked about the idea that "continuity correction" needs to be performed when we approximate a (discrete) binomial distribution with a (continuous) normal distribution.

The correction mysteriously involves the addition or subtraction of 1/2 from the usual expression for a z-score. Let's see why this is empirically.

Binomial probabilities

Let's say we have a binomial distribution with the following properties:

```
n = 16
p = 0.5
```

The mean and SD for a binomial distribution are:

```
mean.binom = n*p
sd.binom = sqrt(n*p*(1-p))
mean.binom
```

```
## [1] 8
```

```
sd.binom
```

```
## [1] 2
```

The general rule of thumb is that we can use the normal approximation if $np \geq 5$ and $n(1-p) \geq 5$. Since $np > 5$, we can go ahead and use this approximation.

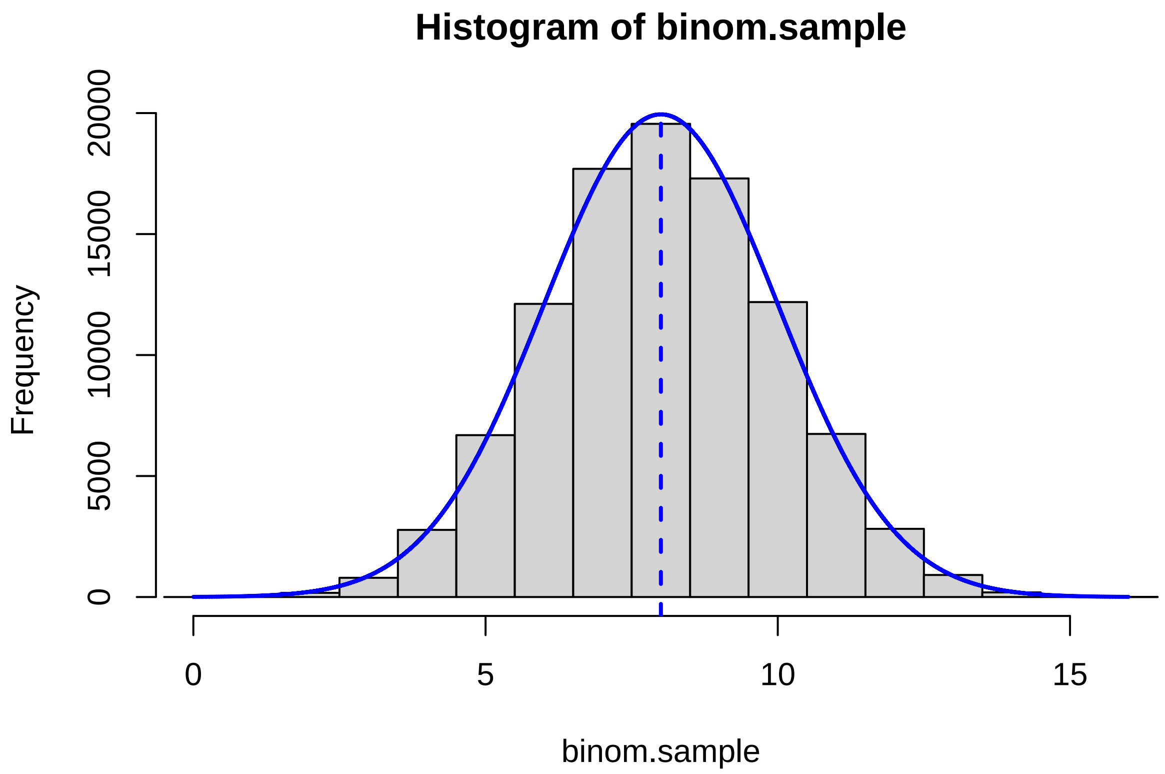
Let's take a large number of samples from this distribution draw a histogram of the results. Then we will add a line that shows a normal approximation with a normal approximation overlaid on top:

```
# sample from a binomial
samples=100000
binom.sample = rbinom(samples, n, p)
table(binom.sample)  # what happens if only sample 1000?
```

```
## binom.sample
##     1     2     3     4     5     6     7     8     9    10    11    12    13
##    27   173   795  2774  6690 12115 17695 19555 17299 12189  6740  2819   911
##    14    15    16
##   192    22     4
```

```
# sample from a normal with the same parameters
xfit = seq(0,n,length=samples)
norm.sample = dnorm(xfit,mean=mean.binom,sd=sd.binom)
yfit = norm.sample*samples

hist(binom.sample, xlim = range(0:n), breaks=seq(-0.5,n+0.5,1))
lines(xfit,yfit,col="blue",lwd = 2)
abline(v=mean.binom,col="blue", lwd=2, lty=2)
```



What is $Pr[X = 8]$? (This is the mean value of the distribution, so the most frequently occurring value). First, compute this using `dbinom`.

```
# dbinom gives exact probability for a given value
dbinom(mean.binom, n, p)
```

```
## [1] 0.1963806
```

Look at the height of the normal approximation for this distribution at X=8. Is this the same thing as $Pr[X = 8]$ for the normal distribution?

```
# this is the height of the curve (not area under the curve, so not a proper probability)
dnorm(mean.binom, mean=mean.binom, sd=sd.binom)
```

```
## [1] 0.1994711
```

Examine the total probability for this distribution $Pr[X \leq 8]$ and $Pr[X \leq 7]$ using `dbinom`. Then, take the difference between these.

The result should be exactly the same as using `dbinom` to get $Pr[X = 8]$. Is this what we see?

```
# Pr[X = mean.binom] using dbinom
dbinom(mean.binom, n, p)
## [1] 0.1963806

pbinom(mean.binom, n, p)      # Pr[X < mean.binom]
## [1] 0.5981903
pbinom(mean.binom - 1, n, p)  # Pr[X < mean.binom - 1]
## [1] 0.4018097

# Pr[X = mean.binom] using pbinom
pbinom(mean.binom, n, p) - pbinom(mean.binom - 1, n, p)
## [1] 0.1963806
```

Now repeat this exercise using `pnorm`.

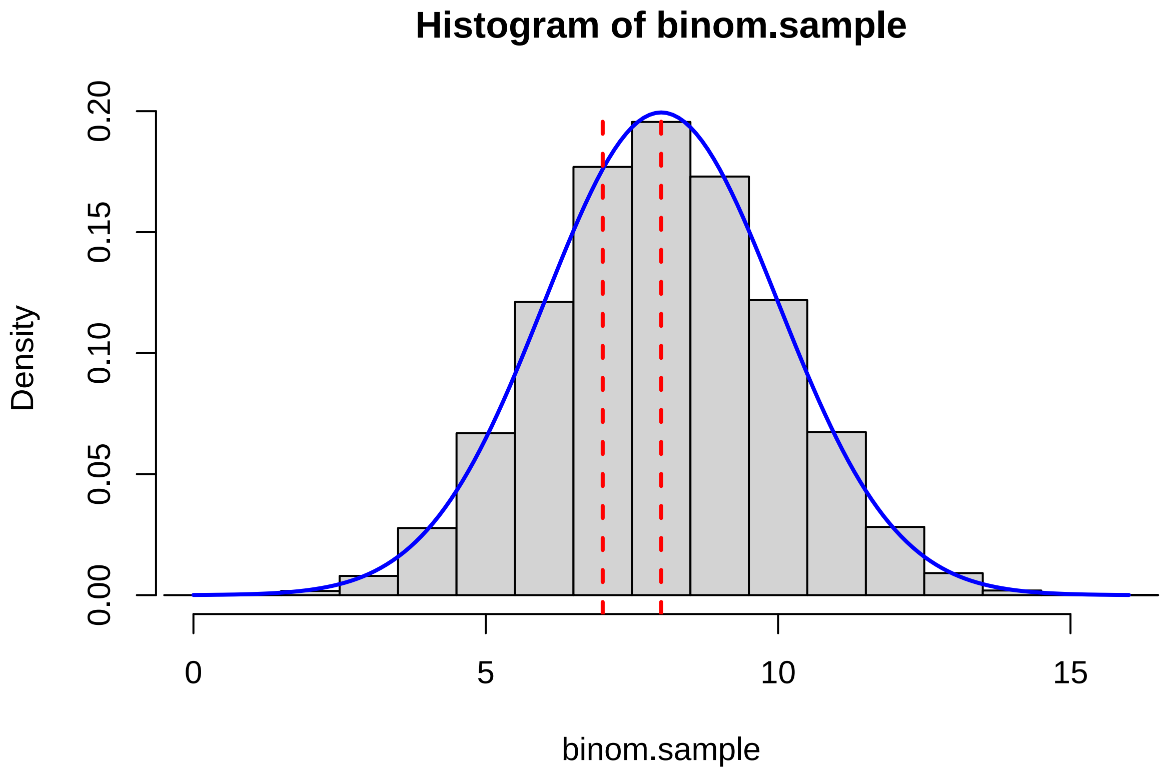
```
pnorm(mean.binom, mean=mean.binom, sd=sd.binom)      # Pr[X < mean.binom]
## [1] 0.5
pnorm(mean.binom - 1, mean=mean.binom, sd=sd.binom)  # Pr[X < mean.binom - 1]
## [1] 0.3085375

# Pr[X = mean.binom] using pnorm
pnorm(mean.binom, mean=mean.binom, sd=sd.binom) - pnorm(mean.binom - 1, mean=mean.binom, sd=sd.binom)
## [1] 0.1914625
```

What's going on here? We see that the total probability for the mean using the normal approximation is exactly 0.5 (as one would actually expect). But this is not quite right if we are interested in approximating the binomial distribution:

```
# normal density
norm.dens = dnorm(seq(0,n,by=0.1), mean=mean.binom, sd = sd.binom)

# overlay bins from normal approx onto density of binomial samples
hist(binom.sample, xlim = range(0:n), breaks=seq(-0.5,n+0.5,1), freq=FALSE)
lines(seq(0,n,by=0.1),norm.dens,col="blue",lwd = 2)
abline(v=mean.binom,col="red", lwd=2, lty=2)
abline(v=mean.binom-1,col="red", lwd=2, lty=2)
```



Is this close enough of an approximation? Let's do a continuity correction.

```
pnorm(mean.binom + 0.5, mean=mean.binom, sd=sd.binom)
```

```
## [1] 0.5987063
```

```
pnorm(mean.binom - 0.5, mean=mean.binom, sd=sd.binom)
```

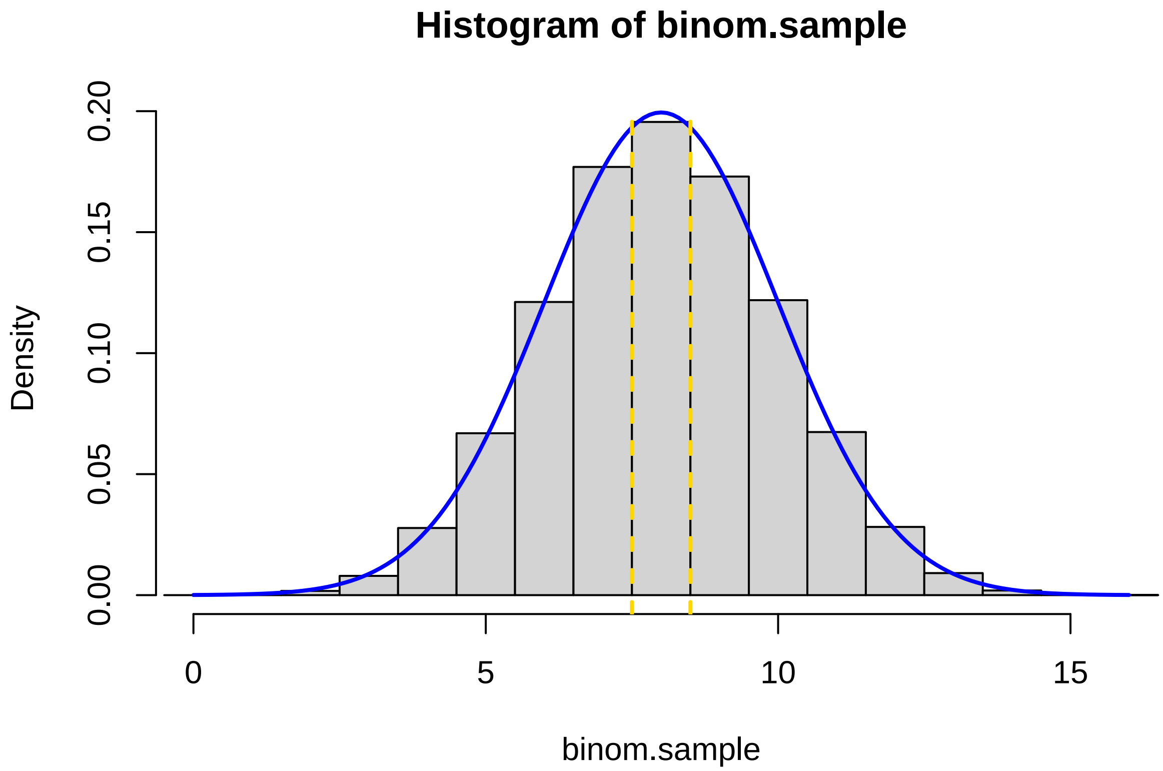
```
## [1] 0.4012937
```

```
pnorm(mean.binom + 0.5, mean=mean.binom, sd=sd.binom) - pnorm(mean.binom - 0.5, mean=mean.binom, sd=sd.b
```

```
## [1] 0.1974127
```

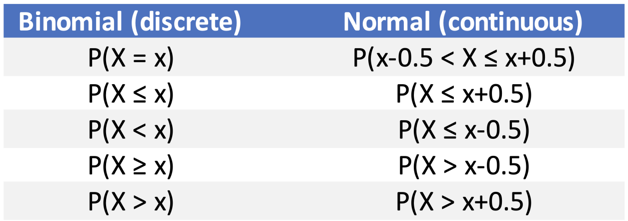
This is closer. We can see why this is more correct by looking at the boundaries of the bin that we just estimated the total probability for:

```
# overlay bins from normal approx onto density of binomial samples
hist(binom.sample, xlim = range(0:n), breaks=seq(-0.5,n+0.5,1), freq=FALSE)
lines(seq(0,n,by=0.1),norm.dens,col="blue",lwd = 2)
abline(v=mean.binom + 0.5,col="gold1", lwd=2, lty=2)
abline(v=mean.binom-0.5,col="gold1", lwd=2, lty=2)
```



Summary

To use a normal approximation of a binomial distribution, we need to use the following corrections for continuity:



Binomial (discrete)	Normal (continuous)
P(X = x)	P(x-0.5 < X ≤ x+0.5)
P(X ≤ x)	P(X ≤ x+0.5)
P(X < x)	P(X ≤ x-0.5)
P(X ≥ x)	P(X > x-0.5)
P(X > x)	P(X > x+0.5)

In this way, we can *include* or *exclude* the area under the curve that spans a discrete value of X, as needed for the particular question we are asking.