

MATH 203: PRINCIPLES OF STATISTICS 1
FALL 2024

USING THE CENTRAL LIMIT THEOREM

The Central Limit Theorem can be used to approximate the distributions of sums and means of independent and identically distributed (i.i.d.) random variables. If $X_1, \dots, X_n$ are i.i.d. (that is, they are a random sample) from a distribution with expectation μ and variance σ^2 , then

$$S_n = \sum_{i=1}^n X_i \quad \text{and} \quad \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

are random variables with that are approximately Normally distributed. That is,

$$S_n \rightsquigarrow \text{Normal}(n\mu, n\sigma^2) \quad \text{and} \quad \bar{X} \rightsquigarrow \text{Normal}(\mu, \sigma^2/n)$$

Therefore,

$$P(S_n \leq s) \approx \Phi\left(\frac{s - n\mu}{\sqrt{n\sigma^2}}\right) \quad P(\bar{X} \leq x) \approx \Phi\left(\frac{x - \mu}{\sigma/\sqrt{n}}\right)$$

where $\Phi(\cdot)$ is the standard Normal cdf.

Example: Suppose X records the score when a die is rolled once. Then X has a discrete Uniform distribution on the values $\{1, 2, 3, 4, 5, 6\}$ and

$$P(X = x) = \frac{1}{6} \quad x = 1, 2, 3, 4, 5, 6$$

with zero probability for any other value of x . We have that

$$E(X) = \sum_{x=1}^6 xp(x) = \sum_{x=1}^6 x \times \frac{1}{6} = \frac{21}{6} = 3.5$$

and

$$E(X^2) = \sum_{x=1}^6 x^2 p(x) = \sum_{x=1}^6 x^2 \times \frac{1}{6} = \frac{91}{6} = 15.1667$$

so that

$$\text{Var}(X) = E(X^2) - \{E(X)\}^2 = \frac{91}{6} - \left(\frac{21}{6}\right)^2 = 2.9167.$$

Suppose now we roll the die $n = 50$ times independently, with $X_1, \dots, X_n$ recording the individual scores. Then the total score across all rolls is S_n , and the Central Limit Theorem says that

$$S_n \rightsquigarrow \text{Normal}(n\mu, n\sigma^2) = \text{Normal}(50 \times 3.5, 50 \times 2.9167) = \text{Normal}(175.0, 145.8333)$$

and for the sample mean score $\bar{X}$

$$\bar{X} \rightsquigarrow \text{Normal}(\mu, \sigma^2/n) = \text{Normal}(3.5, 2.9167/50) = \text{Normal}(3.5, 0.0583)$$

Then for example, using the Z-transform approach, we have

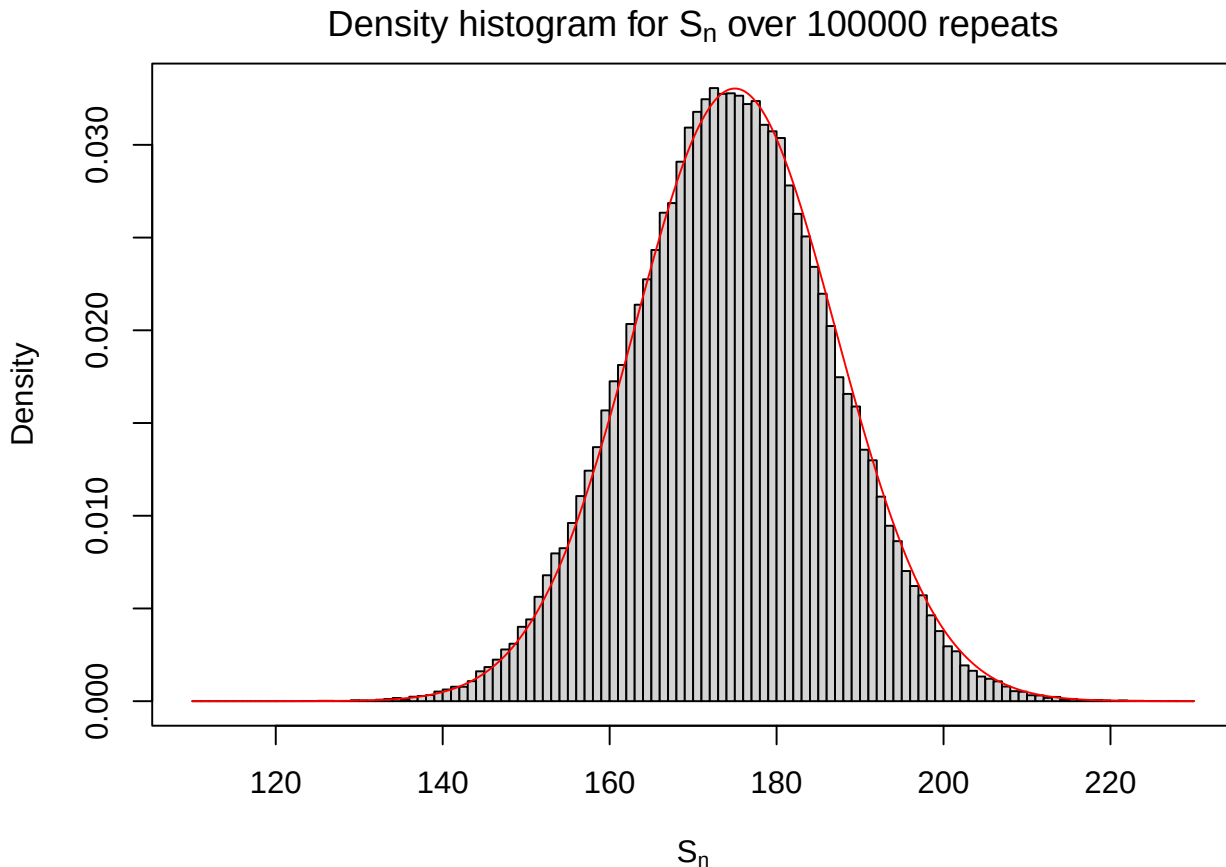
$$P(S_n > 190) \approx P\left(\frac{S_n - 175}{\sqrt{145.8333}} > \frac{190 - 175}{\sqrt{145.8333}}\right) = P(Z_n > 1.2421)$$

and so

$$P(S_n > 190) = 1 - P(Z_n \leq 1.2421) = 1 - 0.8929 = 0.1071$$

We can test this in simulation, using $N = 100000$ repeat runs of the experiment. The following histogram summarizes the observed values of S_n over the repeat runs, with the computed Normal approximation.

```
set.seed(2)
N<-100000;n<-50
par(mar=c(4,4,2,2))
Sn<-replicate(N,sum(sample(1:6,size=n,rep=T)))
tty<-expression(paste('Density histogram for ', S[n], ' over 100000 repeats'))
hist(Sn,br=seq(110,230,by=1),main=tty,
      freq=FALSE,xlab=expression(S[n]));box()
xv<-seq(110,230,by=0.01);yv<-dnorm(xv,n*mu,sqrt(n*sigsq));lines(xv,yv,col='red')
legend(110,0.30,c('Normal approximation'),col='red')
```



In the repeat runs, the number of times S_n exceeds 190 is 9926, so the estimated probability of exceeding 190 is 0.09926, which is fairly close to the claimed value 0.1071.

In this problem, the sum random variable S_n is **discrete**; there are 6^{50} possible sequences of rolls, and the resulting sum can take any whole number value between 60 and 300. If we consider the required probability, then it is evident that in the exact version

$$P(S_n > 190) = P(S_n \geq 191) = 1 - P(S_n \leq 190)$$

If we use the Central Limit Theorem to compute the two equivalent versions, we get two slightly different answers:

```
n<-50
mu<-sum(x)/6
sigsq<-sum(x^2/6)-(sum(x)/6)^2
1-pnorm(190,n*mu,sqrt(n*sigsq))

## [1] 0.1070965

1-pnorm(191,n*mu,sqrt(n*sigsq))

## [1] 0.09259787
```

In such cases, a *continuity correction* is sometimes applied, and we compute instead

$$P(S_n > 190.5)$$

adding on 0.5 to the y value

```
1-pnorm(190.5,n*mu,sqrt(n*sigsq))

## [1] 0.09965464
```

Example: Suppose that $Y \sim \text{Binomial}(n, p)$; recall that Y can be viewed as the total number of successes on n independent binary trials with success probability p . We may therefore write

$$Y = \sum_{i=1}^n X_i$$

where X_i is the result of the i th trial. We may then apply the Central Limit Theorem to Y (as it is the sum of independent and identically distributed random variables) and say that

$$Y \dot{\sim} \text{Normal}(np, npq)$$

as $E(X_i) = p$ and $\text{Var}(X_i) = p(1 - p) = pq$. For a numerical illustration, suppose $n = 500$ and $p = 0.4$. We can compute exactly that $P(Y > 210) = 0.1688395$.

```
n<-500
p<-0.4
q<-1-p
y<-210
1-pbinom(y,n,p)

## [1] 0.1688395
```

For a Central Limit Theorem approximation

$$P(Y > 210) = P\left(\frac{Y - np}{\sqrt{npq}} > \frac{210 - np}{\sqrt{npq}}\right) = P\left(\frac{Y - 500 \times 0.4}{\sqrt{500 \times 0.4 \times 0.6}} > \frac{210 - 500 \times 0.4}{\sqrt{500 \times 0.4 \times 0.6}}\right)$$

or

$$P(Y > 210) = P(Z > 0.9128709) = 1 - P(Z \leq 0.9128709) \approx 1 - 0.8193 = 0.1807.$$

This value is not a very close approximation. Using the continuity correction

$$P(Y > 210) \approx P\left(\frac{Y - np}{\sqrt{npq}} > \frac{210 + 0.5 - np}{\sqrt{npq}}\right) = P\left(\frac{Y - 500 \times 0.4}{\sqrt{500 \times 0.4 \times 0.6}} > \frac{210.5 - 500 \times 0.4}{\sqrt{500 \times 0.4 \times 0.6}}\right)$$

or

$$P(Y > 210) \approx P(Z > 0.9585145) = 1 - P(Z \leq 0.9585145) \approx 1 - 0.8311 = 0.1689.$$

This value is much closer to the exact value. Similarly, suppose we wish to compute

$$P(Y \leq 195)$$

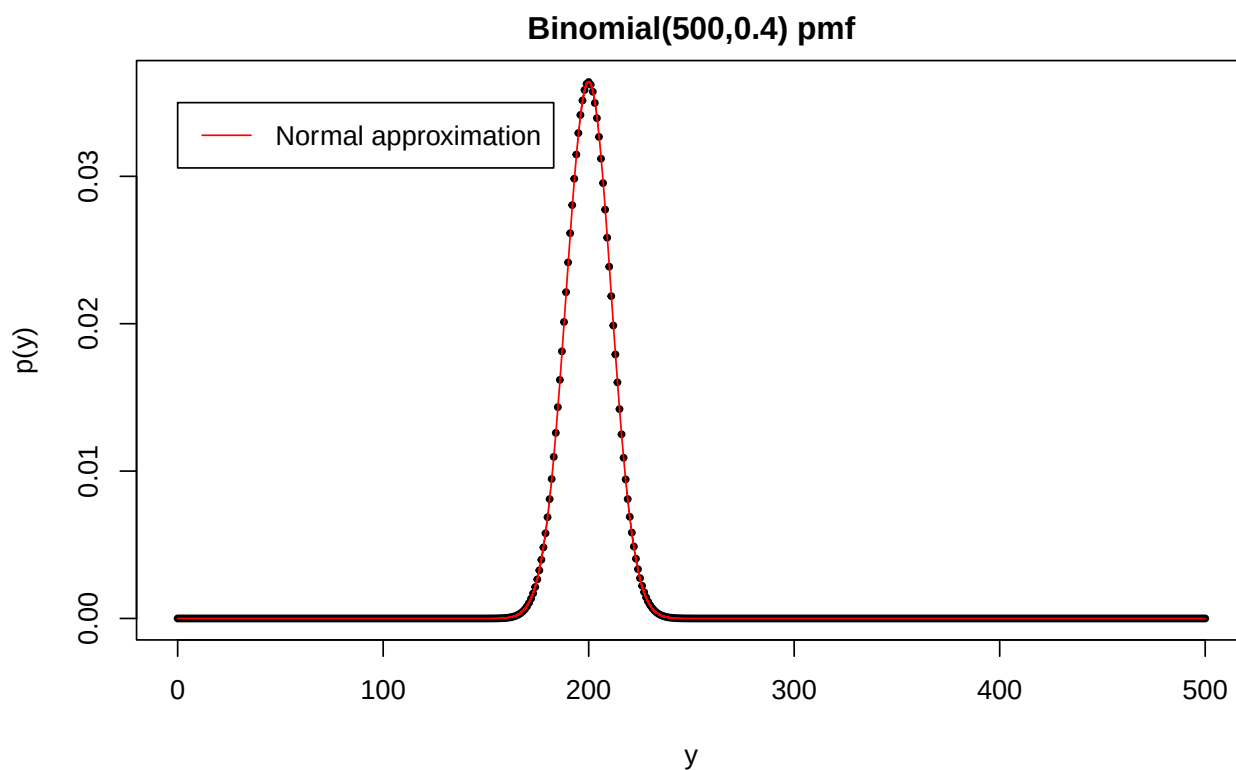
which in the exact version equals 0.3415634.

```
pnorm(195,n*p,sqrt(n*p*q))    #Without continuity correction
## [1] 0.3240384

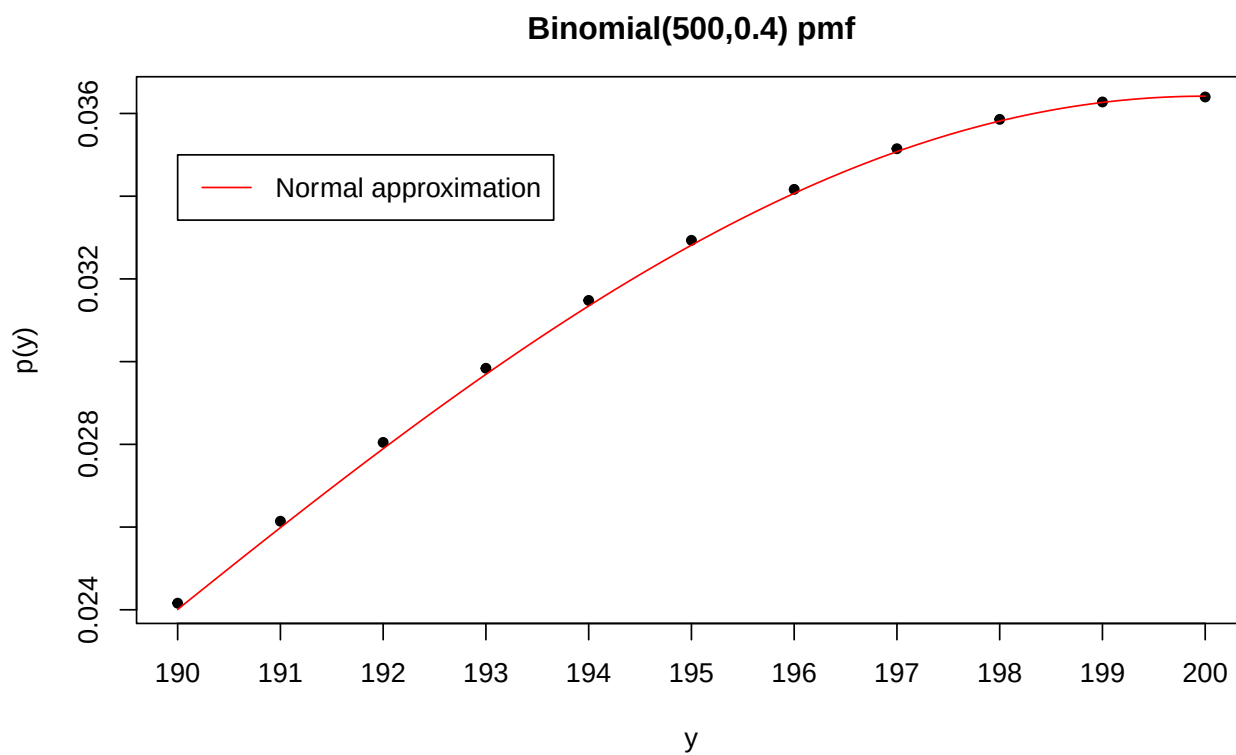
pnorm(195.5,n*p,sqrt(n*p*q))  #With continuity correction
## [1] 0.3406126
```

The logic behind the continuity correction stems from the following diagram, which plots the binomial probabilities and the Normal approximation:

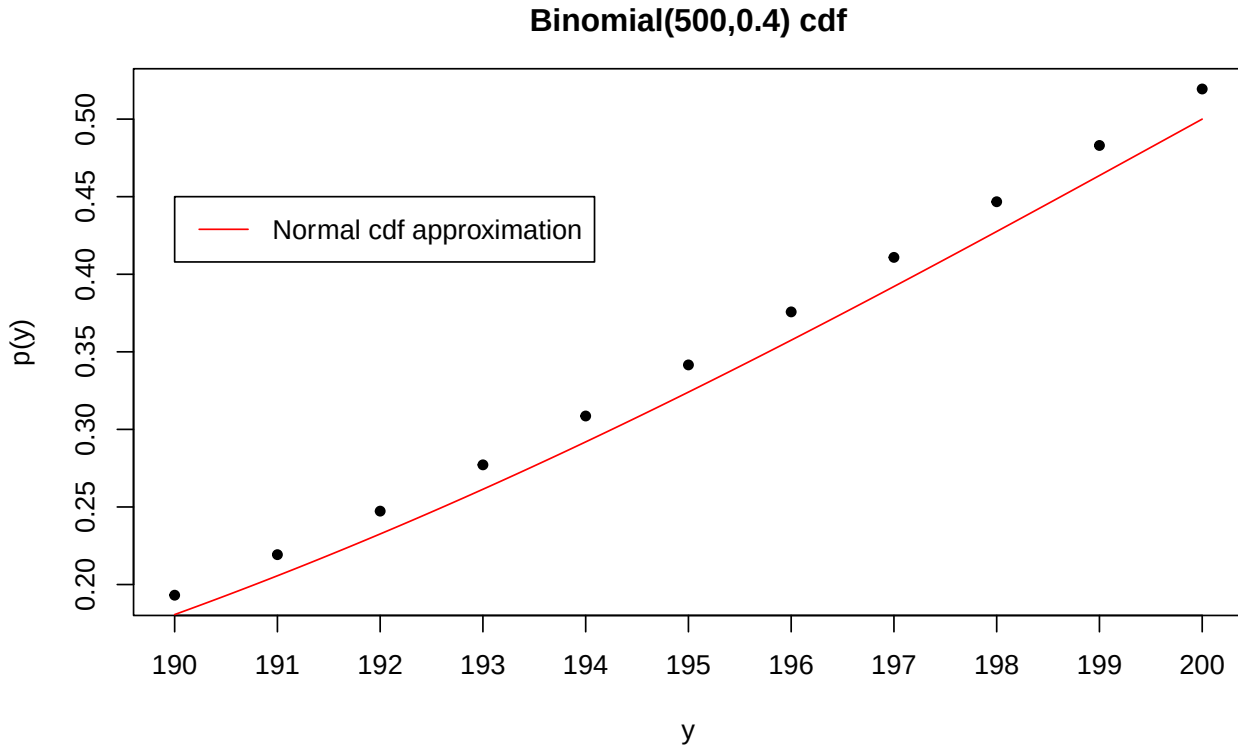
```
yv<-seq(0,n,by=1)
pyv<-dbinom(yv,n,p)
par(mar=c(4,4,2,2))
plot(yv,pyv,pch=19,cex=0.5, xlab='y', ylab='p(y)')
title('Binomial(500,0.4) pmf')
xv<-seq(0,n,by=0.001)
fxv<-dnorm(xv,n*p,sqrt(n*p*q))
lines(xv,fxv,col='red')
legend(0,0.035,c('Normal approximation'),col='red',lty=1)
```



Overall, the Normal approximation to the Binomial does a good job. However, if we zoom into the region near the target value $y = 190$ we see that the Normal approximation slightly underestimates the discrete probabilities



If we look at the cumulative probabilities, the picture is clearer:



Therefore, to approximate the required cumulative probability, we need to increase the y value by a small amount when using the Normal approximation.

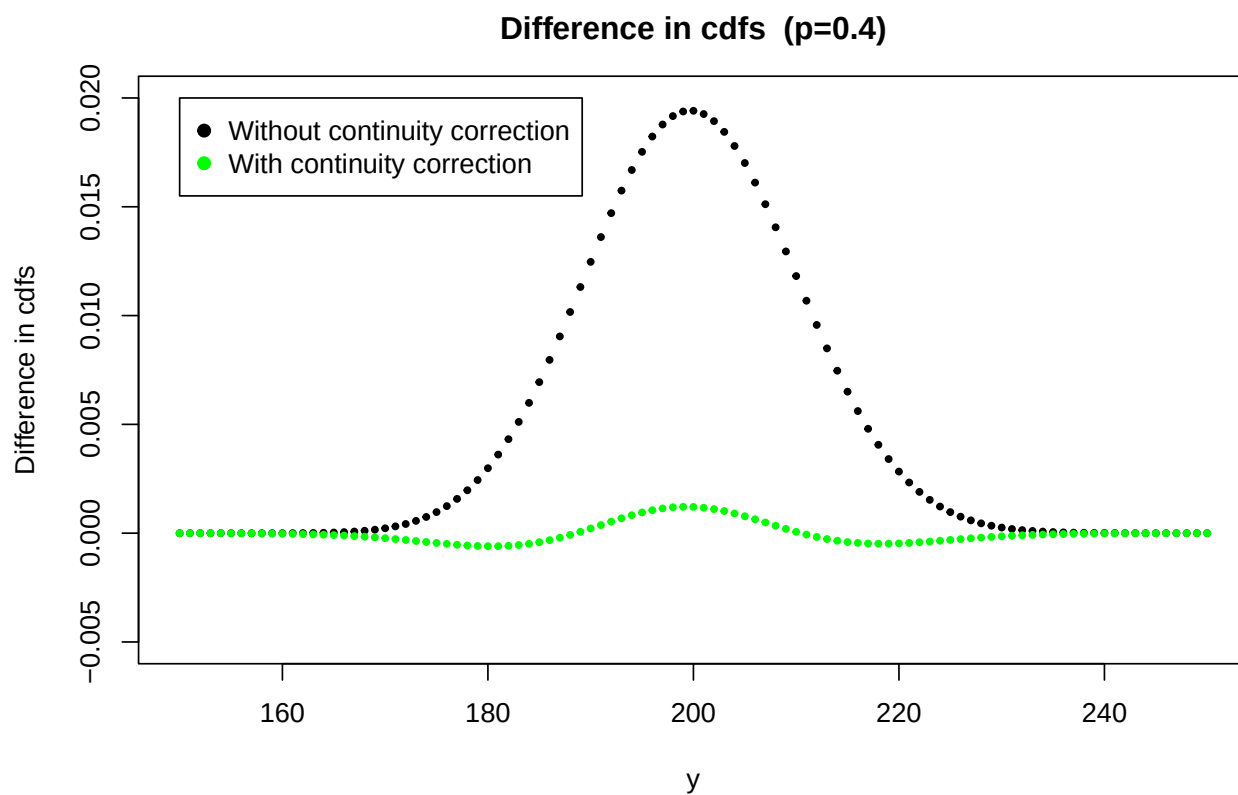
As a final comparison, if we compute the difference

$$F(y) - \Phi\left(\frac{y - np}{\sqrt{npq}}\right)$$

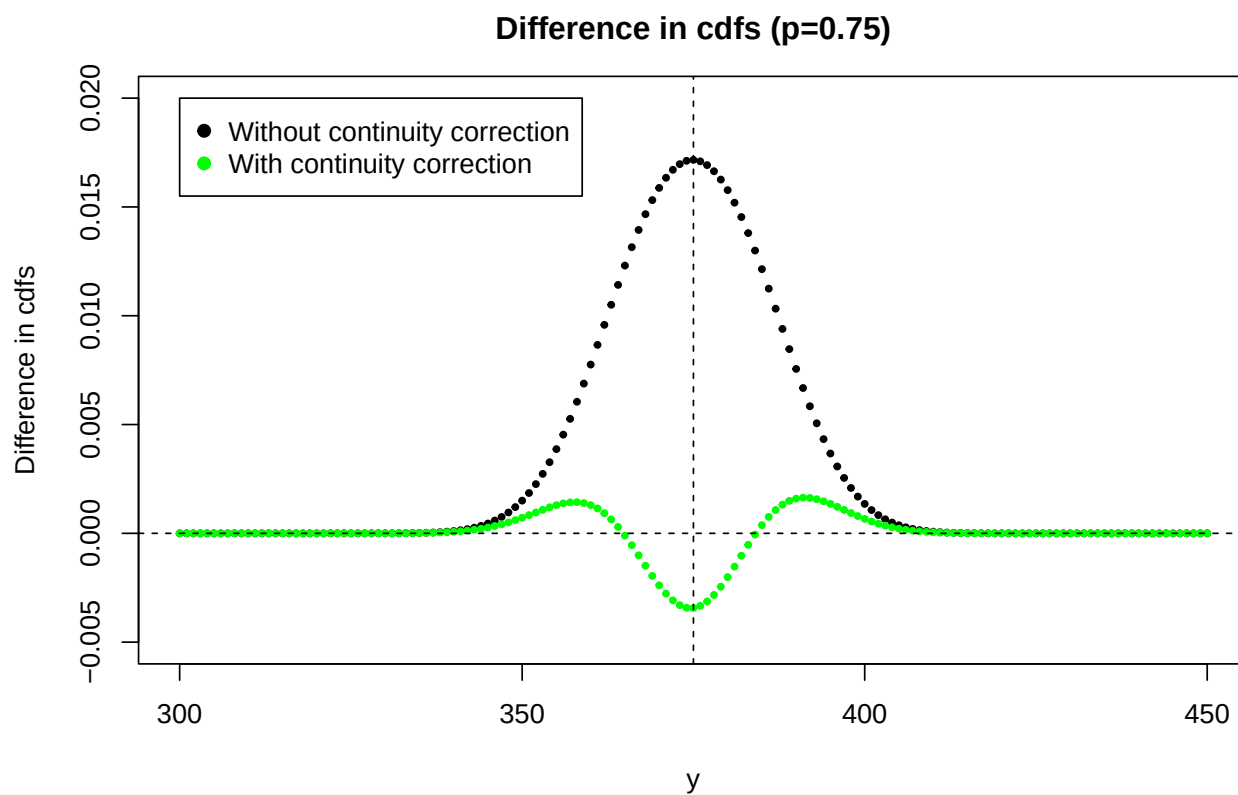
between the $Binomial(n, p)$ cdf and the $Normal(np, npq)$ cdf, we see that the difference is always positive, and peaks near $\mu = np$, whereas if we use the continuity correction

$$F(y) - \Phi\left(\frac{y + 0.5 - np}{\sqrt{npq}}\right)$$

then the difference is smaller in magnitude, and hence the approximation is better. We plot this below for y in the range 150 to 250.

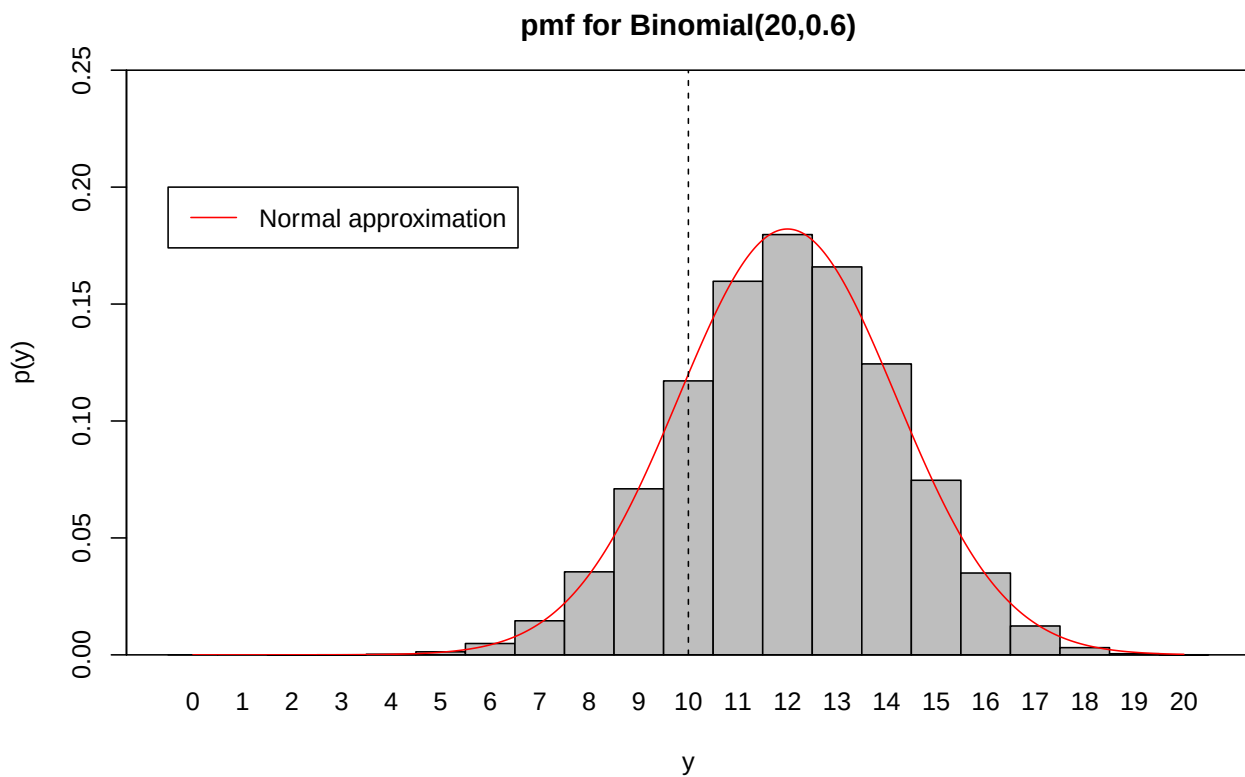


The same correction works well for other choices of p , provided the value is not very close to zero or one.



Example: Suppose $Y \sim \text{Binomial}(n, p)$ with $n = 20$ and $p = 0.6$. Here n is much smaller, and we can investigate the continuity correction in more detail. If we draw the pmf of Y as a bar plot, we obtain the following:

```
n<-20
p<-0.6; q<-1-p
y<-0:n; py<-dbinom(y,n,p)
mu<-n*p; sigsq<-n*p*q
par(mar=c(4,4,3,2))
barplot(py ~ y,col='gray', xlab="y", ylab="p(y)", space=0,
        main="pmf for Binomial(20,0.6)",ylim=range(0,0.25)); box()
xv<-seq(0,20,by=0.01); yv<-dnorm(xv,mu,sqrt(sigsq))
lines(xv+0.5,yv, col='red')
abline(v=10.5,lty=2); legend(0,0.20,c('Normal approximation'),col='red',lty=1)
```



The bar is plotted with bars centered at the labelled y values. The Normal approximation here would be $Normal(np, npq)$, or $Normal(12, 4.8)$, which is overlaid on the bar plot in red. We can see from the plot that there is always part of the each bar that is above the normal approximation curve; to compute $P(Y \leq 10)$, we use the Normal approximation up to y value 10.5.

```
pbinom(10,n,p)           #Exact binomial probability
## [1] 0.2446628
pnorm(10,n*p,sqrt(n*p*q)) #Normal approx. without continuity correction
## [1] 0.1806552
pnorm(10.5,n*p,sqrt(n*p*q)) #Normal approx. with continuity correction
## [1] 0.2467814
```