ME111 - Laboratório de Estatística

Aula 8 - Distribuição Binomial

Profa. Larissa Avila Matos

- n ensaios de Bernoulli (independência e mesma probabilidade de sucesso p);
- Resultado de cada ensaio: $X = 1$ se sucesso ou $X = 0$ se fracasso, com

$$\mathbb{P}(X_i = 1) = p, \quad \mathbb{P}(X_i = 0) = 1 - p;$$

- Seja $Y =$ número de sucessos nos n ensaios de Bernoulli, ou seja,

$$Y = X_1 + \cdots + X_n; \quad Y = 0, 1, 2, \cdots, n;$$

- Então, Y tem distribuição binomial com parâmetros n e p, $Y \sim Bin(n, p)$, com função de probabilidade dada por

$$\mathbb{P}(Y = y) = \left(\begin{array}{c} n \\ y \end{array} \right) p^y (1 - p)^{n-y}.$$

- Suponha o lançamento de uma moeda 10 vezes, onde a probabilidade de sair cara é 0.6. Seja Y = número de cara no 10 lançamentos, então $Y \sim Bin(10, 0.6)$.

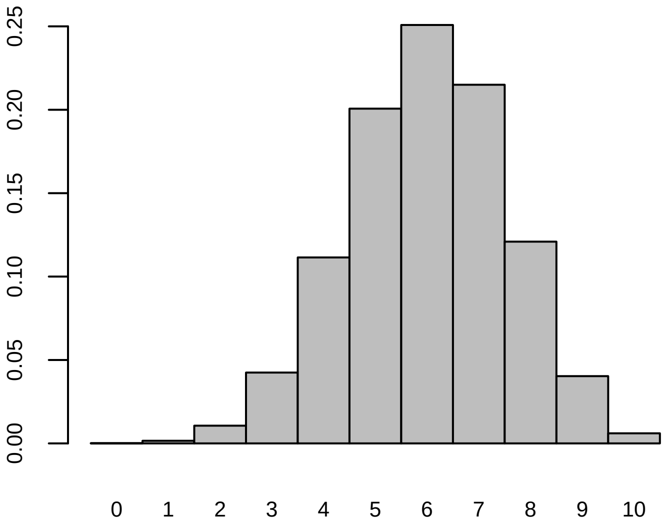
```
n <- 10
p <- 0.6
y <- 0:n
y
```

```
 [1]  0  1  2  3  4  5  6  7  8  9 10
```

```
prob <- dbinom(y,n,p)
prob
```

```
 [1] 0.0001048576 0.0015728640 0.0106168320 0.0424673280 0.1114767360
 [6] 0.2006581248 0.2508226560 0.2149908480 0.1209323520 0.0403107840
[11] 0.0060466176
```

```r
par(mfrow=c(1,1),mar=c(2, 2, 2, 2))
barplot(prob,cex.axis=0.7,names.arg=y,cex.names=0.7,space=0)
```



Valor esperado e variância

- Se $Y \sim Bin(n, p)$ (isto é, Y é uma variável aleatória com distribuição binomial), então:

  1. O valor esperado de Y é

  $$E(Y) = \sum_{y} y \, \mathbb{P}(Y = y) = np, \quad y = 0, 1, \cdots, n.$$

  2. A variância é

  $$\text{var}(Y) = \sum_{y} (y - E(Y))^2 \, \mathbb{P}(Y = y) = np(1-p), \quad y = 0, 1, \cdots, n,$$

  e o desvio padrão é

  $$\text{dp}(Y) = \sqrt{\text{var}(Y)} = \sqrt{np(1-p)}.$$

```r
media <- sum(y*prob)
media
```

```
[1] 6
```

```r
n*p
```

```
[1] 6
```

```r
variancia <- sum(((y-media)^2)*prob)
variancia
```

```
[1] 2.4
```

```r
n*p*(1-p)
```

```
[1] 2.4
```

Exercício

- Uma moeda honesta é lançada $n = 10$ vezes em idênticas condições.
- Determine a probabilidade de ocorrer cara entre 40% e 70% das vezes, inclusive.

- Seja X : número total de caras nos 10 lançamentos,

  $S =$ "Sucesso": ocorrência de cara,

  $p = P(S) = 0,5$ (moeda honesta),

  $X \sim Bin(10, 0.5)$

Exercício

- Uma moeda honesta é lançada $n = 10$ vezes em idênticas condições.
- Determine a probabilidade de ocorrer cara entre 40% e 70% das vezes, inclusive.

- Seja X : número total de caras nos 10 lançamentos,

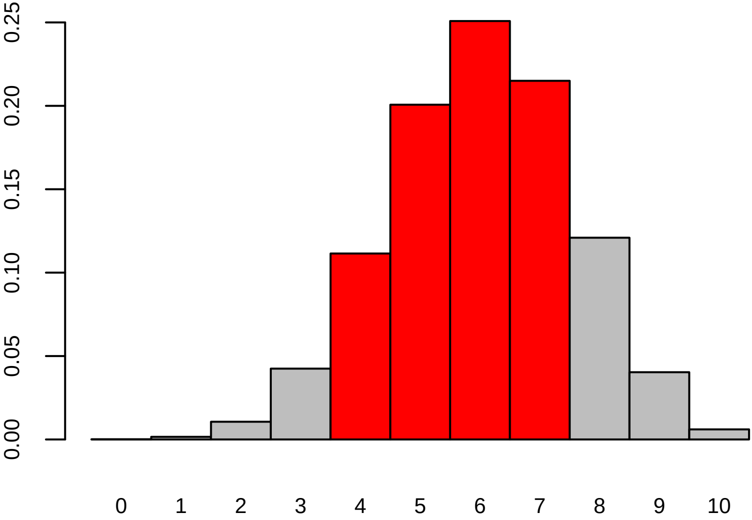
  $S =$ "Sucesso": ocorrência de cara,

  $p = P(S) = 0,5$ (moeda honesta),

  $X \sim Bin(10, 0.5)$

- Queremos calcular a probabilidade $\mathbb{P}(4 \leq X \leq 7)$.

- $\mathbb{P}(4 \leq X \leq 7)$



```r
n <- 10
p <- 0.5
prob <- dbinom(4,n,p)+dbinom(5,n,p)+dbinom(6,n,p)+dbinom(7,n,p)
prob
```

```
[1] 0.7734375
```

```r
prob <- pbinom(7,n,p)-pbinom(3,n,p)
prob
```

```
[1] 0.7734375
```

1. **`dbinom`**: calcula $\mathbb{P}(Y = y) = np^y(1-p)^{n-y}$;
2. **`rbinom`**: simula de uma distribuição binomial;
3. Obtenha uma amostra de tamanho B = 100 de uma binomial com $p = 0.3$ e $n = 10$ usando **`rbinom`**;
4. Construa um gráfico usando a amostra obtida.

- $Y \sim Bin(10, 0.3)$

```r
n <- 10
B <- 1000
p <- 0.3
amostra <- rbinom(B,size=n,prob=p)
table(amostra)/B # distribuição empirica
```

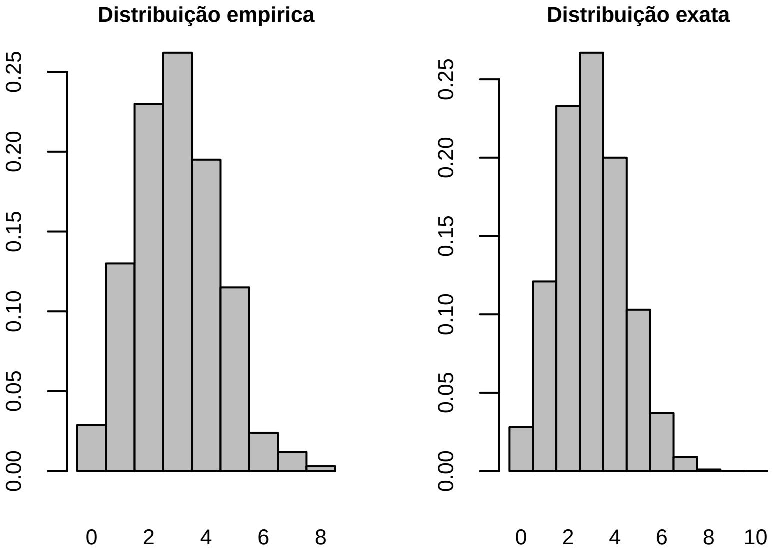
```
amostra
    0     1     2     3     4     5     6     7     8
0.029 0.130 0.230 0.262 0.195 0.115 0.024 0.012 0.003
```

```r
y<-0:10
dbinom(y,10,0.3) # distribuição exata
```

```
 [1] 0.0282475249 0.1210608210 0.2334744405 0.2668279320 0.2001209490
 [6] 0.1029193452 0.0367569090 0.0090016920 0.0014467005 0.0001377810
[11] 0.0000059049
```

```r
round(dbinom(y,10,0.3),3)
```

```
 [1] 0.028 0.121 0.233 0.267 0.200 0.103 0.037 0.009 0.001 0.000 0.000
```


Distribuição empirica
0.00
0.05
0.10
0.15
0.20
0.25
0
2
4
6
8
Distribuição exata
0.00
0.05
0.10
0.15
0.20
0.25
0
2
4
6
8
10

Simulando - Distribuição Binomial

- Repita 3-4 do slide anterior usando: $p = 0.1, 0.3, 0.5, 0.7, 0.9, 0.99$ e $n = 20$.
- Repita 3-4 do slide anterior usando: $p = 0.1$ e $n = 5, 10, 20, 50, 75, 100$.
- Repita 3-4 do slide anterior usando: $p = 0.5$ e $n = 5, 10, 20, 50, 75, 100$.
- Repita 3-4 do slide anterior usando: $p = 0.9$ e $n = 5, 10, 20, 50, 75, 100$.
- Acima use `mean`, `sd` para calcular a média e o desvio padrão amostral.
- O que você observa nestas simulações?

- Simetria e assimetria: $p = 0.1, 0.3, 0.5, 0.7, 0.9, 0.99$ e $n = 20$

```r
B <- 1000
n <- 20
p1 <- 0.1
amostra1 <- rbinom(B,size=n,prob=p1)
p2 <- 0.3
amostra3 <- rbinom(B,size=n,prob=p2)
p3 <- 0.5
amostra5 <- rbinom(B,size=n,prob=p3)
p4 <- 0.7
amostra7 <- rbinom(B,size=n,prob=p4)
p5 <- 0.9
amostra9 <- rbinom(B,size=n,prob=p5)
p6 <- 0.99
amostra99 <- rbinom(B,size=n,prob=p6)
```

```r
graph <- function(i){
  y <- table(eval(parse(text = paste('amostra',i,sep=''))))/B
  barplot(y,cex.main=0.7,cex.axis=0.7,cex.names=0.7,space=0,
          main=sprintf(paste('n=20, p=0.',i,sep='')))
}
par(mfcol=c(2,3),mar=c(2, 2, 2, 2))
x=mapply(graph,c(1,3,5,7,9,99))
```

Média

p	Empirica	Exata
0.1	2.068	2.0
0.3	5.984	6.0
0.5	9.822	10.0
0.7	13.953	14.0
0.9	17.996	18.0
0.99	19.803	19.8

Desvio

p	Empirica	Exata
0.1	1.3875523	1.3416408
0.3	2.0410561	2.0493902
0.5	2.2916299	2.2360680
0.7	1.9971930	2.0493902
0.9	1.3710813	1.3416408
0.99	0.4340269	0.4449719

- $p = 0.1$ e $n = 5, 10, 20, 50, 75, 100$

```
B <- 1000
p <- 0.1
n1 <- 5
amostra_n5 <- rbinom(B,size=n1,prob=p)
n2 <- 10
amostra_n10 <- rbinom(B,size=n2,prob=p)
n3 <- 20
amostra_n20 <- rbinom(B,size=n3,prob=p)
n4 <- 50
amostra_n50 <- rbinom(B,size=n4,prob=p)
n5 <- 75
amostra_n75 <- rbinom(B,size=n5,prob=p)
n6 <- 100
amostra_n100 <- rbinom(B,size=n6,prob=p)
```

```r
graph <- function(i){
  y <- table(eval(parse(text = paste('amostra_n',i,sep=''))))/B
  barplot(y,cex.main=0.7,cex.axis=0.7,cex.names=0.7,space=0,
          main=sprintf(paste('p=0.1, n=',i,sep='')))
}
par(mfcol=c(2,3),mar=c(2, 2, 2, 2))
x=mapply(graph,c(5,10,20,50,75,100))
```

Média

n	Empirica	Exata
5	0.493	0.5
10	1.046	1.0
20	1.999	2.0
50	5.018	5.0
75	7.485	7.5
100	9.878	10.0

Desvio

n	Empirica	Exata
5	0.6740960	0.6708204
10	0.9595857	0.9486833
20	1.4017717	1.3416408
50	2.1946509	2.1213203
75	2.6815977	2.5980762
100	3.0146649	3.0000000

- $p = 0.5$ e $n = 5, 10, 20, 50, 75, 100$

```
B <- 1000
p <- 0.5
n1 <- 5
amostra_n5 <- rbinom(B,size=n1,prob=p)
n2 <- 10
amostra_n10 <- rbinom(B,size=n2,prob=p)
n3 <- 20
amostra_n20 <- rbinom(B,size=n3,prob=p)
n4 <- 50
amostra_n50 <- rbinom(B,size=n4,prob=p)
n5 <- 75
amostra_n75 <- rbinom(B,size=n5,prob=p)
n6 <- 100
amostra_n100 <- rbinom(B,size=n6,prob=p)
```

```r
graph <- function(i){
  y <- table(eval(parse(text = paste('amostra_n',i,sep=''))))/B
  barplot(y,cex.main=0.7,cex.axis=0.7,cex.names=0.7,space=0,
          main=sprintf(paste('p=0.5, n=',i,sep='')))
}
par(mfcol=c(2,3),mar=c(2, 2, 2, 2))
x=mapply(graph,c(5,10,20,50,75,100))
```

Média

n	Empirica	Exata
5	2.479	2.5
10	4.970	5.0
20	10.142	10.0
50	25.051	25.0
75	37.479	37.5
100	50.109	50.0

Desvio

n	Empirica	Exata
5	1.115708	1.118034
10	1.563824	1.581139
20	2.345920	2.236068
50	3.575078	3.535534
75	4.306751	4.330127
100	4.939992	5.000000

- $p = 0.9$ e $n = 5, 10, 20, 50, 75, 100$

```
B <- 1000
p <- 0.9
n1 <- 5
amostra_n5 <- rbinom(B,size=n1,prob=p)
n2 <- 10
amostra_n10 <- rbinom(B,size=n2,prob=p)
n3 <- 20
amostra_n20 <- rbinom(B,size=n3,prob=p)
n4 <- 50
amostra_n50 <- rbinom(B,size=n4,prob=p)
n5 <- 75
amostra_n75 <- rbinom(B,size=n5,prob=p)
n6 <- 100
amostra_n100 <- rbinom(B,size=n6,prob=p)
```

```r
graph <- function(i){
  y <- table(eval(parse(text = paste('amostra_n',i,sep=''))))/B
  barplot(y,cex.main=0.7,cex.axis=0.7,cex.names=0.7,space=0,
          main=sprintf(paste('p=0.9, n=',i,sep='')))
}
par(mfcol=c(2,3),mar=c(2, 2, 2, 2))
x=mapply(graph,c(5,10,20,50,75,100))
```

	Média	
n	Empirica	Exata
5	4.503	4.5
10	8.979	9.0
20	17.977	18.0
50	45.137	45.0
75	67.451	67.5
100	90.087	90.0

	Desvio	
n	Empirica	Exata
5	0.6917149	0.6708204
10	0.9987775	0.9486833
20	1.3454669	1.3416408
50	2.0592405	2.1213203
75	2.6808181	2.5980762
100	2.9953970	3.0000000

- Seja $X \sim Bin(10, 0.8)$, queremos calcular a probabilidade aproximada de $Y = 7$ ($\mathbb{P}(Y = 7)$).
- Gere uma amostra de tamanho $B = 100$.

```
B=100
n <- 10
p <- 0.8
Y <- rbinom(B,n,p)
y <- 7
prob <- sum(ifelse(Y==y,1,0))/B
prob
```

```
[1] 0.24
```

```
dbinom(y,n,p) # comparando
```

```
[1] 0.2013266
```

- Repita os passos do slide anterior para calcular as probabilidades de $Y = 0$ a $Y = n$ e guarde os valores em um vetor.
- Repita o item anterior usando $B = 10000$.
- Resolva acima usando o comando table.
- Calcule as probabilidades exatas e compare com as probabilidades encontradas para $B = 100$ e $B = 10000$.
- O que você observa?

```r
B=100
n <- 10; p <- 0.8
Y <- rbinom(B,n,p)
prob1 <- NULL
for(i in 1:n){
  prob[i] <- sum(ifelse(Y==i,1,0))/B
}
prob
```

```
 [1] 0.00 0.00 0.00 0.00 0.03 0.09 0.23 0.29 0.19 0.17
```

```r
table(Y)/B
```

```
Y
   5    6    7    8    9   10 
0.03 0.09 0.23 0.29 0.19 0.17 
```

```r
comp <- round(dbinom(0:n,n,p),3) # comparando
names(comp) <- seq(0,10); comp
```

```
    0     1     2     3     4     5     6     7     8     9    10 
0.000 0.000 0.000 0.001 0.006 0.026 0.088 0.201 0.302 0.268 0.107 
```

```r
n <- 10; p <- 0.8
B1=100
Y1 <- rbinom(B1,n,p)
table(Y1)/B1
```

```
Y1
   4    5    6    7    8    9   10
0.01 0.04 0.07 0.21 0.29 0.28 0.10
```

```r
B2=10000
Y2 <- rbinom(B2,n,p)
table(Y2)/B2
```

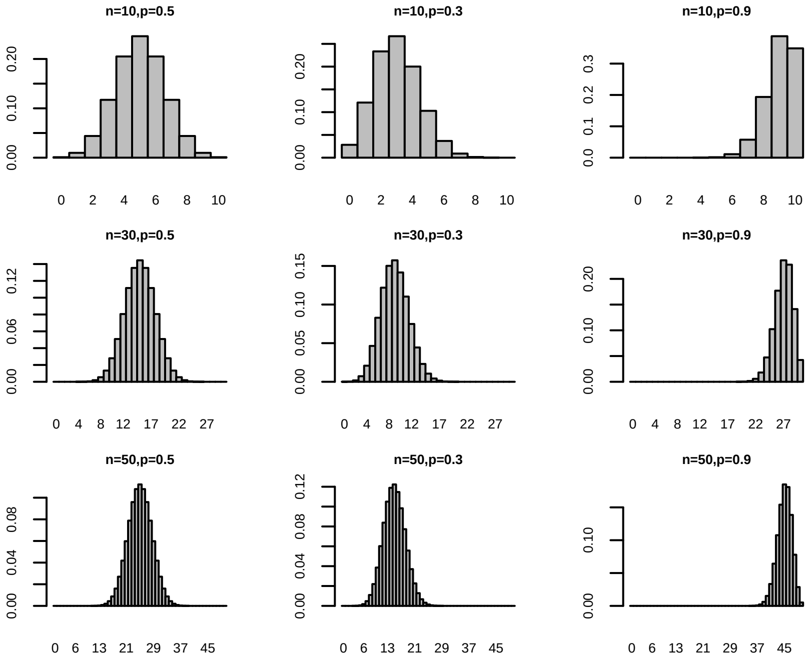
```
Y2
     2      3      4      5      6      7      8      9     10
0.0001 0.0012 0.0049 0.0253 0.0868 0.2079 0.2950 0.2690 0.1098
```

```r
comp <- round(dbinom(0:n,n,p),3) # comparando
names(comp) <- seq(0,10); comp
```

```
    0     1     2     3     4     5     6     7     8     9    10
0.000 0.000 0.000 0.001 0.006 0.026 0.088 0.201 0.302 0.268 0.107
```

Distribuições Binomiais (n, p)

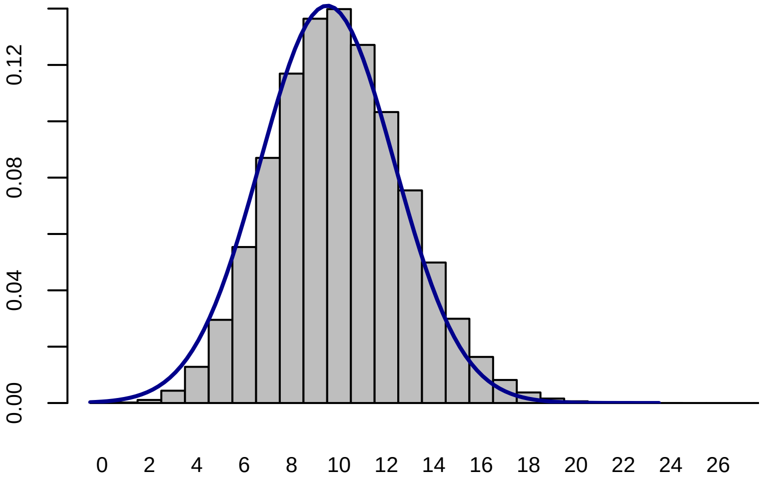
Distribuições exatas



n=10,p=0.5
n=10,p=0.3
n=10,p=0.9
n=30,p=0.5
n=30,p=0.3
n=30,p=0.9
n=50,p=0.5
n=50,p=0.3
n=50,p=0.9

- Para p fixado, à medida que n cresce, os histogramas vão se tornando mais simétricos e com a forma da curva $Normal$.

- Considere a binomial com $n = 50$ e $p = 0.2$, representada pelo histograma



- Sabemos que $\mathbb{P}(X = 13)$ é igual a área do retângulo de base unitária e altura igual a $\mathbb{P}(X = 13)$; similarmente, $\mathbb{P}(X = 14)$, etc ...

- Logo, $\mathbb{P}(X \geq 13)$ é igual à soma das áreas dos retângulos correspondentes.

- A idéia é aproximar tal área pela área sob uma curva normal, à direita de 13. (Qual curva normal?)

- Vimos que, se $X \sim Bin(n, p)$

$$\Rightarrow E(X) = np \quad \text{e} \quad var(X) = np(1-p).$$

- Parece razoável considerar a normal com média e variância iguais às da binomial, ou seja, aproximamos a distribuição de probabilidades de X pela distribuição de probabilidades de uma variável aleatória Y, sendo

$$Y \sim N(\mu_y, \sigma_y^2), \quad \text{com} \quad \mu_y = np \quad \text{e} \quad \sigma_y^2 = np(1-p)$$

.

- Portanto,

  1. $\mathbb{P}(a \leq X \leq b) \approx \mathbb{P}(a \leq Y \leq b)$,
  2. $\mathbb{P}(X \geq a) \approx \mathbb{P}(Y \geq a)$,
  3. $\mathbb{P}(X \leq b) \approx \mathbb{P}(Y \leq b)$,

com $Y \sim N(np, np(1-p))$.

- O cálculo da probabilidade aproximada é feito da forma usual para a distribuição normal:

$$\mathbb{P}(a \leq X \leq b) \approx \mathbb{P}(a \leq Y \leq b), \quad \text{com} \quad Y \sim N(np, np(1-p)).$$

- Temos que

$$Z = \frac{Y - np}{\sqrt{np(1-p)}} \sim N(0,1),$$

então

$$\begin{aligned}\mathbb{P}(a \leq Y \leq b) &= \mathbb{P}(\frac{a - np}{\sqrt{np(1-p)}} \leq \frac{Y - np}{\sqrt{np(1-p)}} \leq \frac{b - np}{\sqrt{np(1-p)}}) \\ &= \mathbb{P}(\frac{a - np}{\sqrt{np(1-p)}} \leq Z \leq \frac{b - np}{\sqrt{np(1-p)}}).\end{aligned}$$

Exercício

- Seja $X \sim Bin(225, 0.2)$, calcule:

1. $\mathbb{P}(39 \leq X \leq 48)$.

Exercício

- Seja $X \sim Bin(225, 0.2)$, calcule:

1. $\mathbb{P}(39 \leq X \leq 48)$.

Sabemos que $E(X) = 225 * 0.2 = 45$ e $\text{var}(X) = 225 * 0.2 * 0.8 = 36$. Então,

$$\mathbb{P}(39 \leq X \leq 48) \approx \mathbb{P}(39 \leq Y \leq 48), \quad \text{com} \quad Y \sim N(45, 36)$$

$$\begin{aligned} \Rightarrow \mathbb{P}(39 \leq Y \leq 48) &= \mathbb{P}\left(\frac{39-45}{\sqrt{36}} \leq \frac{Y-45}{\sqrt{36}} \leq \frac{48-45}{\sqrt{36}}\right) \\ &= \mathbb{P}(-1 \leq Z \leq 0.5) \\ &= \mathbb{P}(Z \leq 0.5) - \mathbb{P}(Z \leq -1) \end{aligned}$$

$$\Rightarrow \mathbb{P}(39 \leq X \leq 48) \approx \mathbb{P}(39 \leq Y \leq 48) = \mathbb{P}(Z \leq 0.5) - \mathbb{P}(Z \leq -1)$$

- $p = 0.2$, $np = 45$ e $n(1-p) = 180$

```
pbinom(48,225,0.2)-pbinom(38,225,0.2)
```

```
[1] 0.5852713
```

```
pnorm(48,mean=45,sd=6)-pnorm(39,mean=45,sd=6)
```

```
[1] 0.5328072
```

```
pnorm(0.5)-pnorm(-1)
```

```
[1] 0.5328072
```

2 $\mathbb{P}(X \geq 42)$.

2 $\mathbb{P}(X \geq 42)$.

$$\mathbb{P}(X \geq 42) \approx \mathbb{P}(Y \geq 42), \quad \text{com} \quad Y \sim N(45, 36)$$

$$\Rightarrow \mathbb{P}(Y \geq 42) = 1 - \mathbb{P}(Y \leq 42)$$

$$\Rightarrow \mathbb{P}(Y \geq 42) = \mathbb{P}\left(\frac{Y - 45}{\sqrt{36}} \geq \frac{42 - 45}{\sqrt{36}}\right) = \mathbb{P}(Z \geq -0.5),$$

onde

$$\mathbb{P}(Z \geq -0.5) = 1 - \mathbb{P}(Z \leq -0.5)$$

e

$$\mathbb{P}(Z \geq -0.5) = \mathbb{P}(Z \leq 0.5)$$

```r
1-pbinom(42,225,0.2)
```

```
[1] 0.6563933
```

```r
1-pnorm(42,mean=45,sd=6)
```

```
[1] 0.6914625
```

```r
1-pnorm(-0.5)
```

```
[1] 0.6914625
```

```r
pnorm(0.5)
```

```
[1] 0.6914625
```

Observações :

1. A aproximação da distribuição binomial pela normal é boa quando $np(1-p) \geq 3$.
2. A demonstração da validade desta aproximação é feita utilizando-se o Teorema Central do Limite (TCL).
3. A aproximação pode ser melhorada através do uso da "Correção de Continuidade".

Normalizando - Distribuição Binomial

1. Usando `rbinom` gere uma amostra de uma binomial com `n=10` e `p=0.3` e guarde em `Y`. Considere `B=1000` e `10000`.
2. Crie uma nova variável `Z=(Y-mean(Y))/sd(Y)`.
3. Calcule a média e desvio padrão de `Z`. Qual foi resultado?
4. Repita os itens 1-3 com $n = 15, 20, 50, 100$. O que você observa?