MAD-CB

UNIFESP

UNIVERSIDADE FEDERAL DE SÃO PAULO

1933

Projeto Final – Mudanças

Projeto – Gapminder

- Para melhor comportar com nosso progresso em estatística e R
- Usaremos "Gapminder" como fonte de dados para os projetos
- Gapminder
  - Fundação sueca que trata de questões sociais e econômicos importantes
  - É um site de dados e ferramentas para a apresentação deles
  - Baseado na premissa que pessoas podem entender dados importantes melhor se são acompanhados com um pouco de drama

Projeto – Passo 1

- Formar grupos (max 3 pessoas)
- Escolher uma base de dados de Gapminder que o grupo acha interessante
  - www.gapminder.org/data/
- Entregue email para mim antes no 10 de março (sexta que vem)
  - Nomes dos membros do grupo
  - Base de dados que o grupo vai analisar

Projeto – Passo 2

- Converter a base de Excel para R
- Preparar os dados para a analise
  - Limpeza; tidy data
- Fazer uma análise exploratória da base
  - Estatísticas
  - Visualizações
- Data limite: 31 de março

Projeto – Passo 3

- Fazer apresentação na aula do projeto
  - Quais questões tratadas nos dados vocês podem responder
  - Quais técnicas de análise usaram para tirar essas conclusões
  - Quais duvidas restam a ser resolvidos
- Escrever um relatório conciso resumindo o projeto e as conclusões
  - Data limite: 12 de maio

Inferência – Usando Probabilidade para Entender Dados

Teorema de Limite Centrale

Jogar Moeda

- Jogar uma moeda de um real 10.000 vezes e ver quantos CARAS resultam
- Assumimos que temos uma moeda justa
  - p(CARA) = p(COROA)
- Para simular as jogadas, usamos a distribuição binomial
- Distribuição Binomial
  - 2 resultados possíveis (Sim/Não, V/F, Cara/Coroa)
  - Pode repetir o experimento n vezes
  - Experimentos "Bernoulli"
- Estimador da verdadeira probabilidade de jogar CARA

$$p(Cara) \approx \hat{p}(Cara) = \frac{\#\ de\ CARAS\ observadas}{\#\ de\ jogadas}$$

Jogada com 2 Moedas

- Vamos começar com a jogada de 2 moedas
- Cada experimento vai ser a soma do número das Caras (H)
- Possibilidades
  - $\{T, T\} = 0$
  - $\{H,T\}, \{T,H\} = 1$
  - $\{H,H\} = 2$
- Código para produzir 10.000 resultados:

```
pr <- .5 # moeda justa: 50/50 chance de jogar um CARA
n <- 10000 # experimentos
k <- 2 # 2 moedas
moedaProb <- rbinom(n, k, pr)

## r<dist>() - função para criar números aleatórios
## dado o número de experimentos (n) e das moedas(k)
```

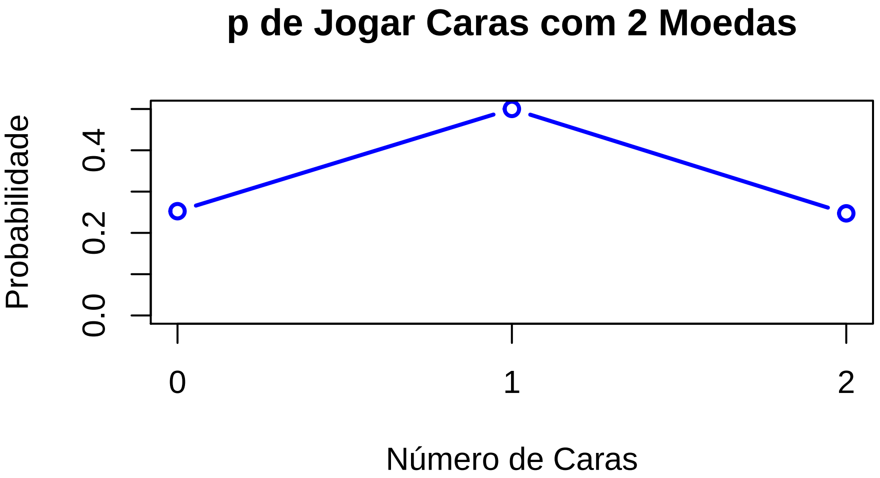
Início de moedaProb – Número de Caras por Experimento

```
moedaProb[1:100]
```

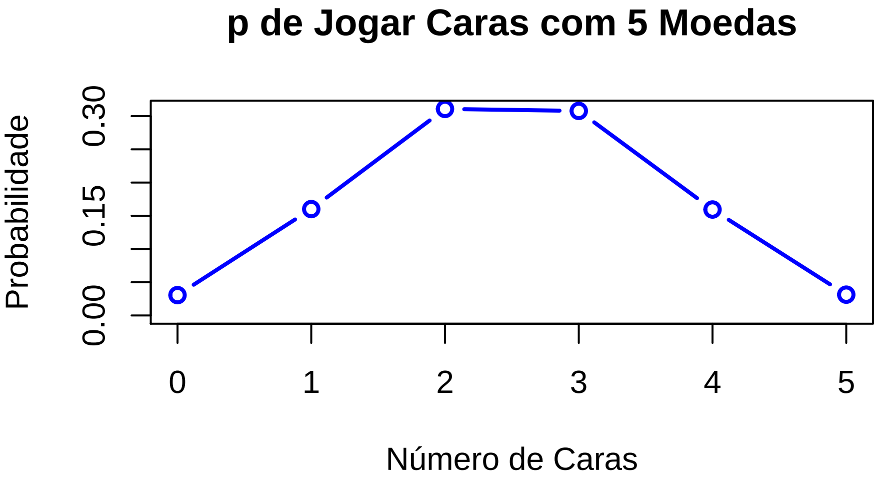
```
##    [1] 2 0 0 2 0 1 1 2 2 2 2 0 0 2 0 1 0 1 2 0 1 0 2 1 1 1 0 0 2 0 2 1 1
##   [36] 0 1 0 1 2 0 1 1 1 1 1 1 1 1 1 2 0 1 0 1 2 2 1 1 1 2 1 2 1 1 2 0 1
##   [71] 1 1 2 1 1 2 0 0 2 1 2 1 0 1 2 2 0 0 2 0 2 1 1 1 1 0 2 1 0 1
```

Probabilidade com 2 Moedas

```
titleline <- paste("p de Jogar Caras com", k, "Moedas")
plot(table(moedaProb)/n, type = "b", col = "blue",
     xlab = "Número de Caras", ylab = "Probabilidade",
     main = titleline)
```

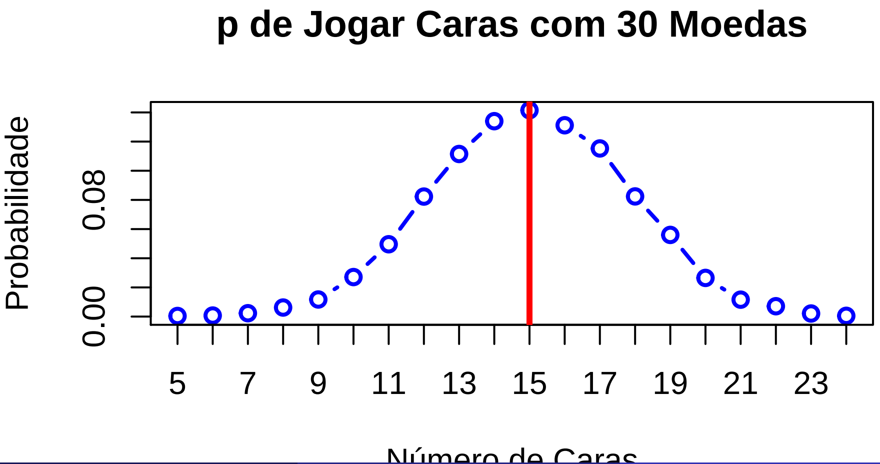


```r
titleline <- paste("p de Jogar Caras com", k, "Moedas")
plot(table(moedaProb)/n, type = "b", col = "blue",
     xlab = "Número de Caras", ylab = "Probabilidade",
     main = titleline)
```



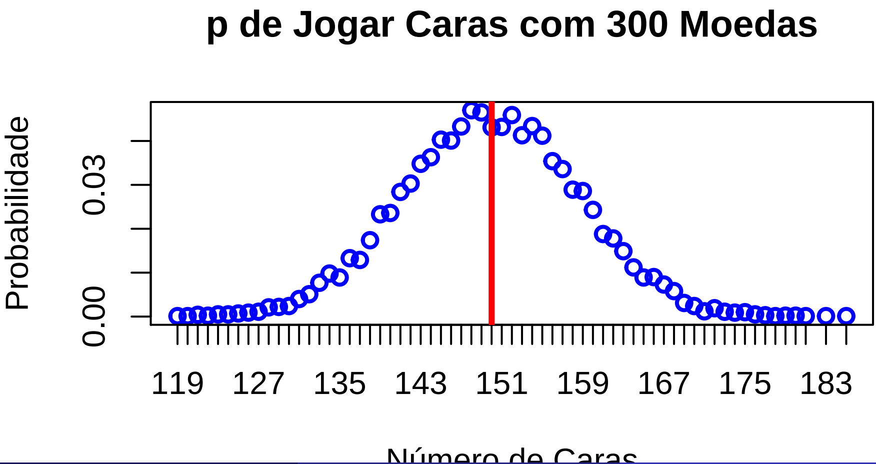
Probabilidade com 30 Moedas

```
titleline <- paste("p de Jogar Caras com", k, "Moedas")
plot(table(moedaProb)/n, type = "b", col = "blue",
     xlab = "Número de Caras", ylab = "Probabilidade",
     main = titleline)
abline(v = 15, col = "red", lwd = 3)
```



Probabilidade com 300 Moedas

```
titleline <- paste("p de Jogar Caras com", k, "Moedas")
plot(table(moedaProb)/n, type = "b", col = "blue",
     xlab = "Número de Caras", ylab = "Probabilidade",
     main = titleline)
abline(v = 150, col = "red", lwd = 3)
```



Probabilidade de Jogar Cara (geral)

```
sum(rbinom(100000, 300, pr))/100000/300
```

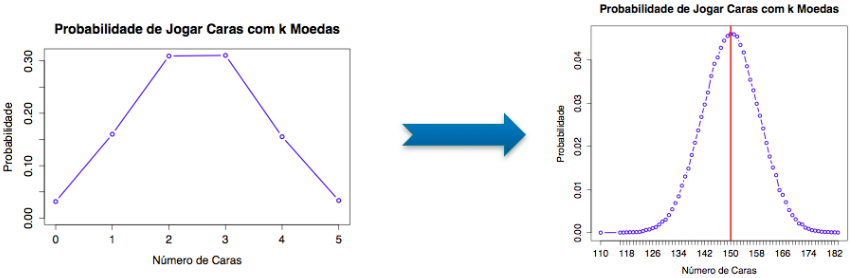
```
## [1] 0.5000429
```

$$p(Cara) \approx \hat{p}(Cara) = \frac{\# \; de \; CARAS \; observadas}{\# \; de \; jogadas}$$

- Lei dos Grandes Números – Lembrete
- Maior o número dos experimentos ("*Bernoulli trials*"), a média dos resultados irá convergir no valor esperado (a) do experimento
- Valor esperado (probabilidade teórico) das CARAS – 1/2

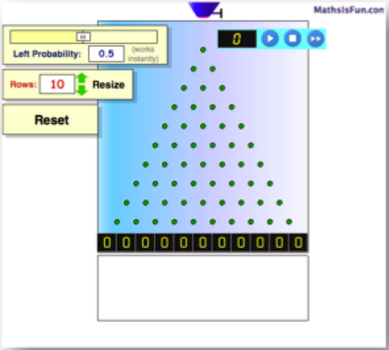
Teorema de Limite Central (CLT)

- Se repetimos um experimento muitas vezes, a probabilidade do resultado médio irá convergir a uma distribuição normal (curva de sino)
- Permite que usamos a distribuição normal como base da maioria de nossos testes estatísticas (paramétricas)

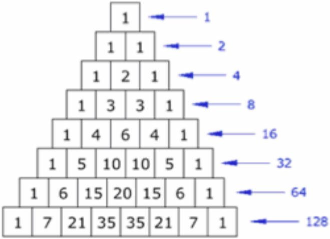


Quincunx (Quincunce)

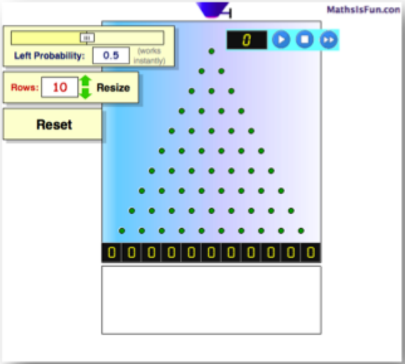
- Jogo/brinquedo em que bolas caem, batem contra pinos, pulam para direto ou esquerdo e continua caindo.
- Parecido com Pachinko mas todos os pinos estão em forma regular para que um pino forma um triangulo equilateral com os dois para baixo.
- Um paralelo com bolas do Triangulo de Pascal (com números)
- Quando cada bola bate contra um pino, tem só dois resultados possíveis
- Pode ir para esquerda ou a direta
- Segue as mesmas regras de uma variável binomial que as moedas.
- Com 1.000 bolas que pulam 10 vezes (10 fileiras), tem um total de 10.000 experimentos
- Vários sites têm exemplos do jogo.
- `http://www.mathsisfun.com/data/quincunx.html`

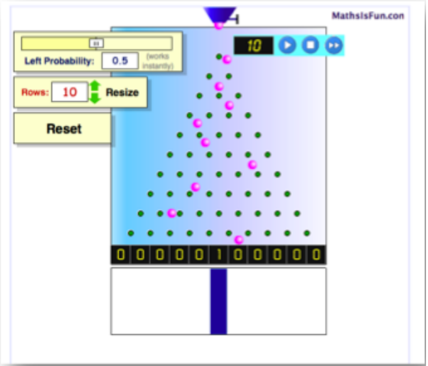


Triangulo de Pascal

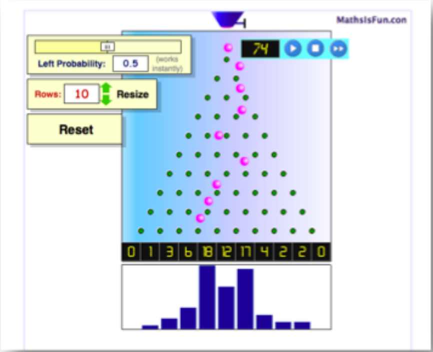


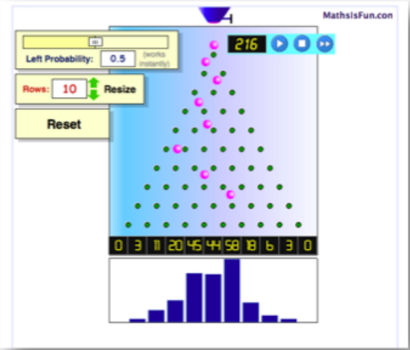
Quincunx – Início



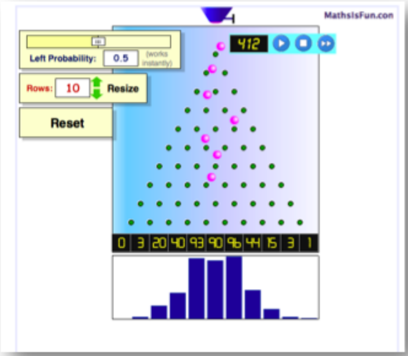


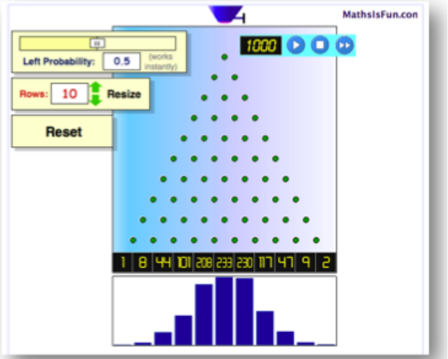
Quincunx – 2


MathsIsFun.com
Left Probability: 0.5
(works instantly)
74
Rows: 10
Resize
Reset
0 1 3 6 18 12 17 4 2 2 0


MathsIsFun.com
Left Probability: 0.5
(works instantly)
216
Rows: 10
Resize
Reset
0 3 11 20 45 44 58 18 6 3 0

Quincunx – 3


MathsIsFun.com
Left Probability:
0.5
(works instantly)
Rows:
10
Resize
Reset
412
0 3 20 40 93 90 96 44 15 3 1


MathsIsFun.com
Left Probability:
0.5
(works instantly)
Rows:
10
Resize
Reset
1000
1 8 44 101 208 233 230 117 47 9 2

Quincunx – Como Funciona

- Definir n para ser o número de fileiras e k para ser o número de vezes a bola pula para esquerda.
- Assim, *(n – k)* deve ser a probabilidade que a bola pula para direta.
- Probabilidade da bola pula para esquerda em cada fileira: p
- Probabilidade que a bola pula a esquerda k vezes é
  - p^k : $(p_1 * p_2 \ldots * p_k = p^k)$
  - Probabilidade que a bola pula a direta (n – k) vezes é $(1 - p)^{(n - k)}$
- Assim, qualquer uma das 11 lugares finais tem a probabilidade de $p^k(1 - p)^{(n - k)}$
- Por causa da lei de multiplicação da interseção para eventos independentes

Probabilidade de Cair em Qualquer Posição Final

- Existem muitos caminhos para chegar na última fileira – a posição final
  - *Quantos*?
- Cada caminho representa uma das possíveis **combinações**
- Com uma probabilidade de cair igualmente para direta ou para esquerda ($p = 0{,}5$) e 10 fileiras ($n = 10$), quantas maneiras a bola tem para cair na *quarta* posição da esquerda?
- Existem 210 caminhos possíveis de cair na quarta posição

```r
n <- 10; k <- 4; p <- 0.5
choose(n, k) # função para combinações
```

```
## [1] 210
```

Probabilidade de Cair em Qualquer Posição Final – 2

- Probabilidade de cair neste posição é
- Número de maneiras pode cair * a probabilidade :

$$p(k; n; p) = \binom{n}{k} p^k (1 - p)^{n - k}$$

- Esta formula descreve a função binomial – a probabilidade que um evento vai ocorrer dado 2 resultados possíveis

Nosso Exemplo

```r
n <- 10; k <- 4; p <- 0.5
choose(n, k) # função para combinações
```

```
## [1] 210
```

```r
dbinom(k, n, p)
```

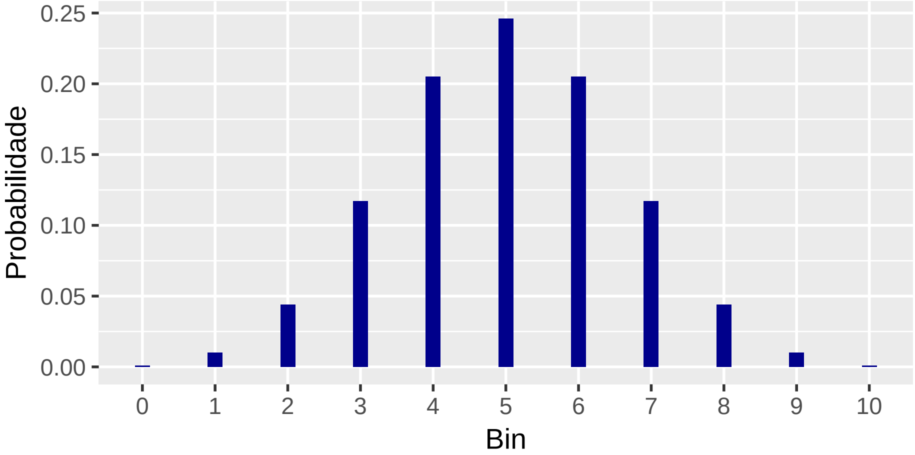
```
## [1] 0.2050781
```

```r
(quincunx <- round(dbinom(0:10, n, p), 3))
```

```
##  [1] 0.001 0.010 0.044 0.117 0.205 0.246 0.205 0.117 0.044 0.010 0.001
```

Probabilidade dos Bins de um Quinqunce

p = 0.5



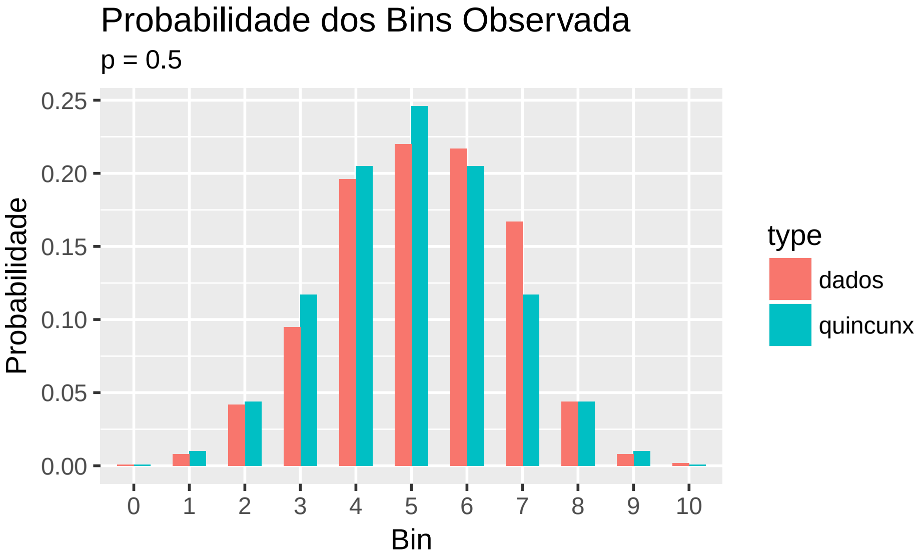
Comparar Resultados com a Teoria

```r
# Teoria
(quincunx <- round(dbinom(0:10, n, p), 3))
```

```
##  [1] 0.001 0.010 0.044 0.117 0.205 0.246 0.205 0.117 0.044 0.010 0.001
```

```r
## Dados do jogo quincunx
data.qq <- c(1, 8, 44, 101, 208, 233, 230, 177, 47, 9, 2)
(prob.qq <- round(data.qq / sum(data.qq), 3))
```

```
##  [1] 0.001 0.008 0.042 0.095 0.196 0.220 0.217 0.167 0.044 0.008 0.002
```



Probabilidade dos Bins Observada
p = 0.5
Probabilidade
0.25
0.20
0.15
0.10
0.05
0.00
0
1
2
3
4
5
6
7
8
9
10
Bin
type
dados
quincunx

Distribuição de Amostra

- O que observamos é uma distribuição de amostra
- Nosso trabalho é avaliar a congruência dela com uma distribuição teórica
- Valores observados variam de amostra em amostra
- Esta variabilidade se chama: variância amostral
- Podemos fazer várias amostras e criar uma distribuição das médias ($\bar{x}$)
- Distribuição das amostras terá uma média e variância também

Valor Esperado e Variância da Distribuição de Amostras

- Esses existem por causa da Teorema de Limite Central

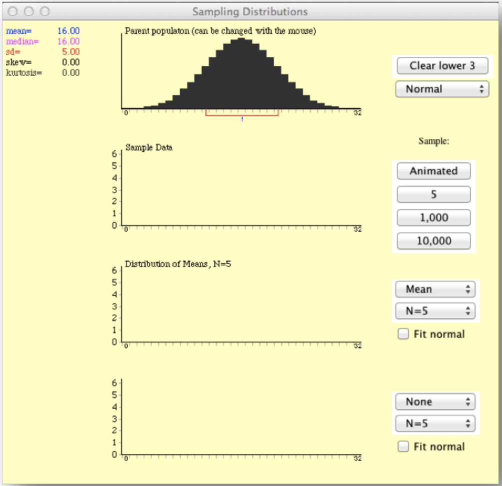
$$E(\bar{X}) = \mu$$

$$Var(\bar{X}) = \frac{\sigma^2}{n}$$

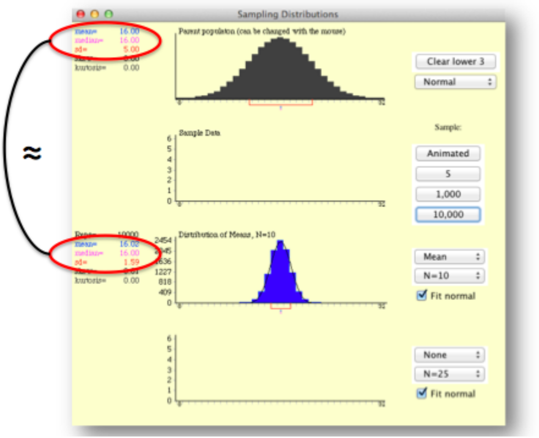
$$DP(\bar{X}) = \sqrt{Var(\bar{X})} = \frac{\sigma}{\sqrt{n}}$$

Comparar Estatísticas das Amostras a População

- Rice University – Applet das Distribuições Amostrais -Site: `http://onlinestatbook.com/stat_sim/sampling_dist/index.html`

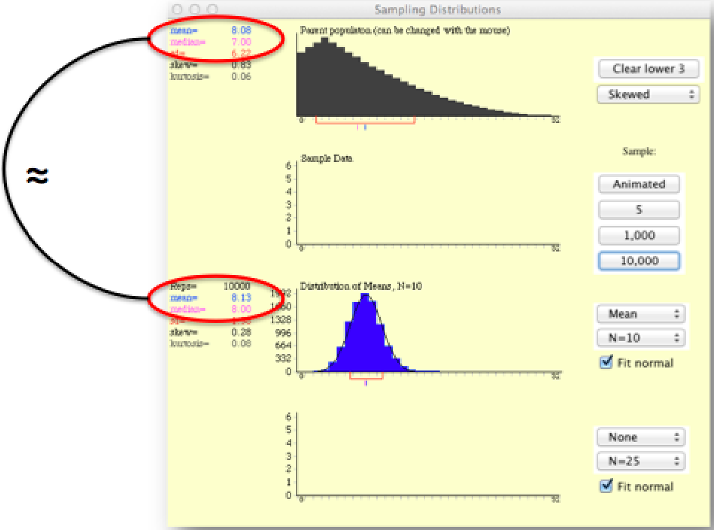


Distribuição Normal

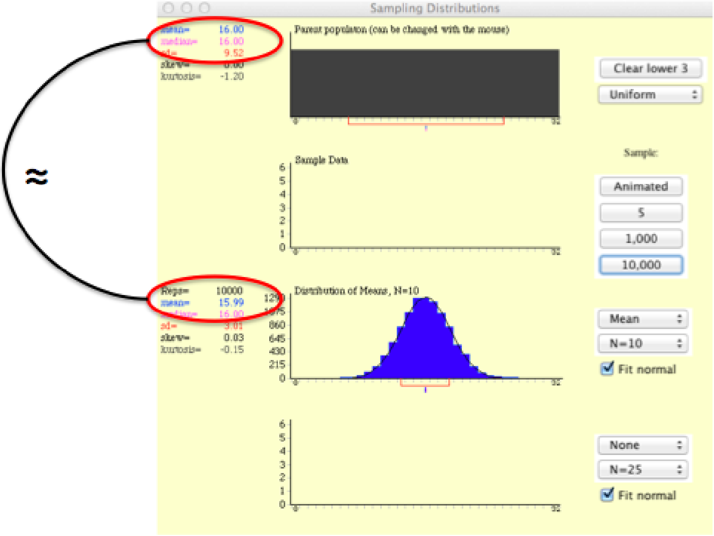


$$E(\bar{X}) = \mu;\ 16,02 \approx 16,00$$
$$DP(\bar{X}) = \frac{\sigma}{\sqrt{n}};\ \frac{5.00}{\sqrt{10}} = 1.58 \approx 1.59$$

Distribuição Assimétrica



Distribuição Uniforme



Resumo - Distribuição Amostral – Proporções

- Teorema de Limite Central (CLT)
- Estudamos amostras e comparar nossa amostra a todas as amostras possíveis
- Distribuição Amostral de proporção binomial

$$\hat{p} \approx N(p, \frac{p(1-p)}{n})$$

```
- Distribuição Amostral da Média
```

$$\bar{X} \approx N(\mu, \frac{\sigma^2}{n})$$

- N.B. $N(\mu, \sigma^2)$ quer dizer distribuição normal com média de μ e variância de σ^2

Vamos Imaginar que Temos uma Garrafa Cheio de Contas

- 2 Cores – Vermelho e Azul
- Não sabemos a proporção de cada cor
- Podemos fazer um experimento
  - Tirar 25 contas da garrafa e contar as cores para estimar a proporção verdadeira
  - Pode repetir isso múltiplas vezes (**muitas**!!) para estimar a proporção na garrafa
  - Usar a função `sample` em R
- *Simulação Monte Carlo*
  - Simular com o computador um evento e repetir muitas vezes
  - Estimação do valor de população
  - Aproveita da Lei de Grandes Números

- Vamos criar as contas com `rep`
    - Vai criar um vetor com todos as contas na garrafa
    - Não vou mostrar aqui
- Vamos selecionar 1 conta da garrafa

```
## [1] "azul"
```

- De novo

```
## [1] "vermelho"
```

Repetir Multiplas Vezes – com `replicate`

- Muitas vezes – 10.000

```
trials <- 10000
set.seed(1)
eventos <- replicate(trials, sample(conta, 1))
head(eventos)
```

```
## [1] "vermelho" "vermelho" "azul"     "azul"     "vermelho"
```

Determinar o Resultado da Simulação

- Usar funções `table` e `prop.table`
  - `table` – tabula os resultados
  - `prop.table` – calcula as proporções dos resultados

```
(tab <- table(eventos))
```

```
## eventos
##     azul vermelho
##     4704     5296
```

```
prop.table(tab)
```

```
## eventos
##     azul vermelho
##   0.4704   0.5296
```

Proporções Verdadeiras

- Divulgação das proporções verdadeiras
  - **azul** – 0.474
  - **vermelho** – 0.526

Com e Sem Substituição

- `replicate` funciona *com substituição*
  - Tirar a conta da garrafa e repor depois
- *Sem substituição* quer dizer que não repormos a conta
  - Fica permanentemente perdido para as tabulações futuras

Distribuições de Probabilidade

- Distribuições dos números e das probabilidades são vinculados
  - Ex: Quincunce
- *Função densidade de probabilidade* $f(x) = c$
  - probabilidade que a distribuição assume um valor específico
- *Função de probabilidade cumulativa* $F(x) \leq c$
  - proporção dos valores na distribuição que ficam abaixo ou igual a um valor específico

Aplicar para Proporção das Contas Azuis

- Converter as cores em números ("azul" $= 1$)

```
contnum <- as.numeric(conta == "azul")
```

- Podemos acertar que o função cumulativa para "azul" (1)
  - $F(1) = \frac{474}{1000} = 0.474$
- Para "vermelho" (0)
  - $F(0) = \frac{526}{1000} = 0.526$

Para Variáveis Categoricas – Distribuição Cumulativa Não Intuitiva

- Melhor fazer o que fizemos com os números
- Define probabilidade de todos os estados possíveis da variável $\Pr(\text{vermelho}) = 0.526$ e $\Pr(\text{azul}) = 0.474$.