MAD-CB

UNIFESP
Universidade Federal de São Paulo
1933

Inferência - 2

Motivação – Sondagens Políticas

- Quando IBOPE diz que um candidato está em frente do outro por 52% a 48% com uma *margem de erro* de 4%, o que quer dizer essa margem de erro?
- Conceito de margem de erro implica que as variáveis são aleatórias
- Daí pode tratar dos assuntos de:
  - Intervalos de confiança
  - Valor p

Tirando Conclusões das Proporções – Exemplo

- Uma cidade tem exatamente 1.000.000 eleitores
  - 504.000 Republicans
  - 496.000 Democrats
- Pesquisador chega para fazer uma sondagem
  - Questão – Quantas Democrats tem a cidade?
- Não sabe o valor da população (49,6%)
- Quer estimar este valor através amostras

A Sondagem

- Sonda afiliação partidária de uma amostra de 1000 eleitores aleatórios

```
poll <- sample(cidade, npoll, replace = TRUE)
table(poll)
```

```
## poll
##   D   R
## 498 502
```

- Previsão da sondagem é vitoria para os Republicans
- Mas, esta amostra representa a população?
- A *estimativa* do resultado reflete a realidade?

Variáveis Aleatórias

- Os resultados dos processos aleatórios
- A sondagem selecionou 1% dos eleitores aleatoriamente
- O que acontece se fazemos isso várias vezes (5)
- Vamos contar os Democrats em 5 sondagens
- Resultados de 5 sondagens: 479, 493, 509, 492, 513
  - Em algumas, os Democrats ganham
- Pode ver que resultados variam bastante
  - Variância *aleatória*
- Para entender os resultados, precisa entender **modelos de amostragem**

Modelos de Amostragem

- Qual valor podemos esperar de nossa sondagem original?
  - Probabilidade de ser Democrat (p = 0.496) x tamanho de amostra (npoll = 1000)

$$E(Dem) = 1000p$$

```
evDems <- p * npoll
```

- Valor Esperado ($E(Dem)$) = 496

Erro Padrão

- Mostra tamanho do erro aleatório
- Erro Padrão dos valores

$$\mathrm{SE}(Dem) = \sqrt{1000p(1-p)}$$

```
seDem <- sqrt(npoll * p * (1 - p))
```

- Erro padrão dos valores = 15.811
  - Erro fica mais ou menos 496 $\pm$ 15.811

Versão Normalizada

- Pode normalizar esses valores controlando para tamanho de amostra
- Valor esperado da proporção na amostra

$$E(Dem/1000) = p$$

- Este implica que $Dem/1000$ mais um erro aleatório igualará à p

Erro Padrão da Proporção

- Dá um tamanho mais exato a correção necessário na amostra

$$\text{SE}(Dem/1000) = \frac{\sqrt{p(1-p)}}{\sqrt{N}}$$

```
sePad <- sqrt(p * (1 - p)/sqrt(npoll))
```

- $\text{SE}(Dem/1000) = 0.089$

Erros e Tamanho de Amostra

$$\text{SE}(Dem/1000) = \frac{\sqrt{p(1-p)}}{\sqrt{N}}$$

- O que acontece se aumentamos o tamanho de amostra (N)?

Estimativas

- $Dem/1000$ é nossa estimativa de p
- Notação $\hat{p} \approx p$
- O valor esperado exato depende do valor de p que não sabemos
- Melhor aproximação para p é $\hat{p}$
- Assim, podemos dizer que

```
p_hat <- mean(poll == "D")
se <- sqrt(p_hat * (1 - p_hat)/1000)
cat("Nossa estimativa da proporção dos Democrats\né", p_hat,
    "mais ou menos", round(se, 5))
```

```
## Nossa estimativa da proporção dos Democrats
## é 0.498 mais ou menos 0.01581
```

Distribuição de Probabilidade para as Variáveis Aleatórias

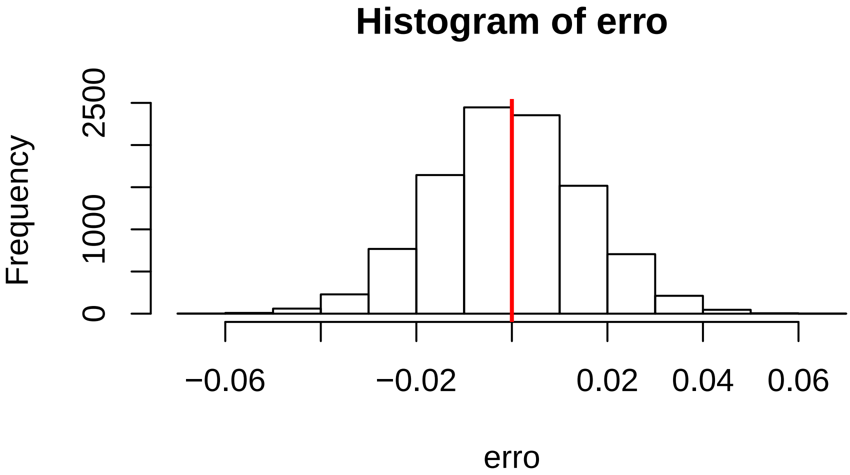
- O “mais ou menos” não é muito útil
- Podemos calcular a probabilidade que $\hat{p}$ fica dentro de 1% do verdadeiro p?
- Vamos começar com uma simulação de nossas eleitores
- Medir a distribuição dos erros $\hat{p} - p$

```
trials = 10^4
erro <- replicate(trials, {
  X <- sample(cidade, npoll, replace = TRUE)
  mean(X == "D") - p
})
```

```r
mean(abs(erro) > 0.01581) ## erros maiores que o SE
```

```
## [1] 0.3246
```

```r
hist(erro)
abline(v = 0.0, col = "red", lwd = 2)
```

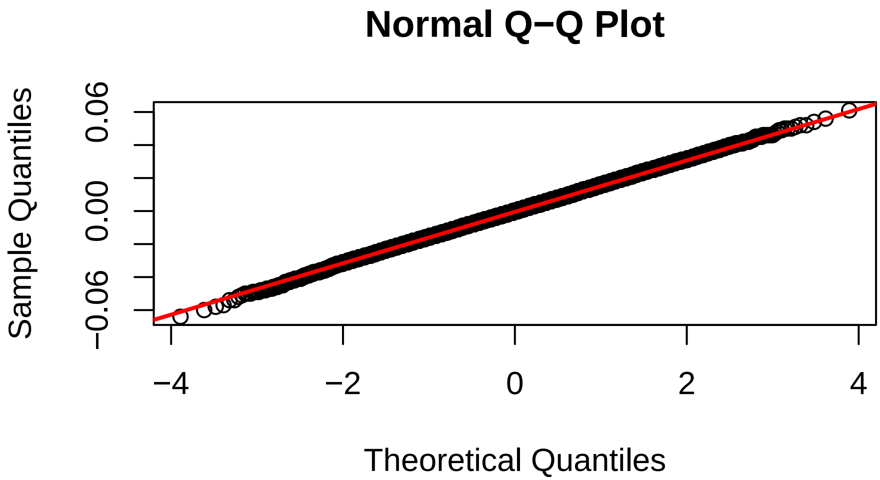


Implicações da Histograma

- Esta é a distribuição de probabilidade de nossa sondagem
- Distribuição da $\hat{p}$ parece perto a normal
- Centro da distribuição em 0
  - Confirma que valor esperado de $\hat{p}$ é p

Confirmação de Aproximação à Normal

```
qqnorm(erro)
qqline(erro, col = "red", lwd = 2)
```



Comparação dos Dados com a Distribuição Normal

- Comparar % dos erros maior que o SE

```
cat("Proporção verdadeira: ", mean(abs(erro) > 0.01581))
```

```
## Proporção verdadeira:  0.3246
```

- à proporção prevista pela distribuição normal

```
cat("Proporção teorica: ", pnorm(-1) + (1 - pnorm(1)))
```

```
## Proporção teorica:  0.3173105
```

Conclusão

- Podemos dizer em conclusão:

  Com só uma sondagem, podemos dizer que nossa estimativa da proporção de Democrats é $\hat{p}$ *e há uma chance de 32% que nosso erro fica maior de que 1.581%*

Intervalos de Confiança

- Varição aleatória faz a sondagem não acerta o valor correto 32% das vezes
  - Para uma empresa de sondagens, não muito bom
- Se falamos de um intervalo que acerta 95% das vezes, estamos bem pensados no mercado
- Podemos construir um intervalo $[A, B]$ em que:

$$\Pr(A \leq p \text{ and } B \geq p) \geq 0.95$$

Nota de Rodapé – Porque 0.95

- **Costume Tradição**
- Não tem mágica teórica

Variável Aleatória Z

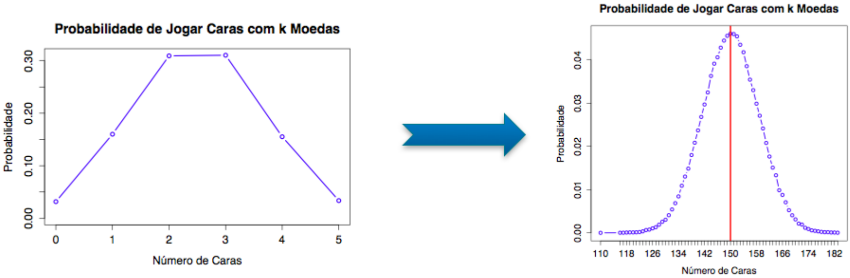
- Como escolhemos A e B para fazer este intervalo tão pequeno quanto possível?
- Sabemos que $\hat{p}$ segue uma distribuição normal (por causa da CLT) com valor esperado de p e erro padrão de $\sqrt{\hat{p}(1-\hat{p})}/\sqrt{N}$
- Esse implica a variável aleatória seguinte (Z):

$$Z = \sqrt{N}\frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})}}$$

- Z é aproximadamente normal com
  - Valor esperado de 0
  - Desvio padrão de 1

Teorema de Limite Central (CLT) - Repeteco

- 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)



Para CLT Funcionar – Premissas Requisitadas

- Amostras são aleatórias
- Observações são independentes
  - Nenhum tem relação com nenhum outra
- Dados são corretos
  - Neste caso, as pessoas falam a verdade; não mentem
  - Grande problema com sondagens politicas
  - Também auto-descrições das sintomas por pacientes

O Sábio Dr. House

- Todos Mentem

IC – Exemplo

- Mais uma jogada com moedas
  - Jogar 1 moeda 1.000 vezes
  - Usando uma simulação Monte Carlo
- Queremos descobrir a verdadeira, mas desconhecida probabilidade (p) de jogar CARA
  - Fazer com números: CARA $= 1$; COROA $= 0$

```
set.seed(1); n <- 1000; k <-  1; prob <- 0.5
tiras <- rbinom(n, k, prob)
(caras <- sum(tiras)) ## número de CARAS
```

```
## [1] 480
```

- Fazemos estimativa de p com a amostra de 1.000 jogadas $\hat{p} = 480/1000 = 48\%$
- Com qual grau de confiança podemos dizer que o valor da população p é realmente perto a nosso estimativa da proporção das CARAS?

Equações para Intervalo de Confiança das Proporções

- Sabemos a média (valor esperado) e variância da proporção estimada - $\hat{p}$

$$E(\hat{p}) = \frac{480}{1000} = 0.516$$

$$Var(\hat{p}) = \frac{p(1-p)}{n}$$

- E, por causa da CLT, sabemos que

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

Equações para Intervalo de Confiança das Proporções – 2

- Podemos converter os valores em uma contagem Z
  - Normalizar os valores em termos da média e desvio padrão

$$z_i = \frac{x_i - \bar{x}}{s}$$

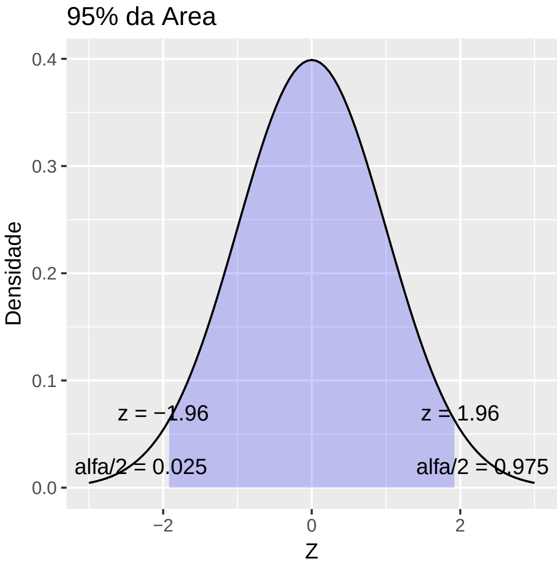
- Contagem Z vem da distribuição normal padronizada, que tem $\mu = 0$ e $\sigma = 1$

$$z = N(0, 1)$$

- Podemos substituir nossos valores nessas equações

$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \approx N(0, 1)$$

Distribuição Normal Padronizada



O Que Significa Isso?

- Para 95% das amostras, z vai ficar entre -1.96 e 1.96
- Valores mais extremos que esses vão ocorrer só 5% das vezes
- Região em que estamos confiantes que nosso valor $\hat{p}$ representa o valor da população verdadeira
  - -1.96 é o limite inferior
  - 1.96 é o limite superior -Temos 95% confiança que o valor verdadeiro desconhecido de p fica dentro deste intervalo
- $\therefore$ "Intervalo de Confiança"
- Probabilidade que nosso $\hat{p}$ cai fora deste intervalo é só 5% ou menos
- 19 de 20 amostras vai ter um p que cairia dentro do intervalo e só 1 vai ter um valor fora

Formula para Intervalo de Confiança

95% Intervalo de confiança de p:

$$\left[\hat{p} - 1,96\sqrt{\frac{p(1-p)}{n}}, \quad \hat{p} + 1,96\sqrt{\frac{p(1-p)}{n}}\right]$$

3 Elementos para Calcular Intervalo

95% Intervalo de confiança de p:

$$\left[\hat{p} - 1,96\sqrt{\frac{p(1-p)}{n}},\ \hat{p} + 1,96\sqrt{\frac{p(1-p)}{n}}\right]$$

Proporção Estimada

```
R> phat <- sum(flips)/1000
```

Valor Crítico $Z_{\alpha/2}$

```
R> z = qnorm(nivel/2, mean=0, sd=1, lower.tail=FALSE)
```

Margem de Erro

```
R> marg.erro = z * sqrt(phat*(1-phat)/1000)
```

Usamos esses 3 elementos no cálculo do intervalo

Calcular Um IC para Proporção

```
phat <- sum(tiras)/1000
nivel <- 0.05
z <-  qnorm(nivel/2, mean = 0, sd = 1, lower.tail = FALSE)
marg.erro  <-  z * sqrt(phat*(1 - phat)/1000)
(ci <- phat + c(-marg.erro, +marg.erro))
```

```
## [1] 0.4490351 0.5109649
```

- Nosso estimativa de $\hat{p}$ (0.48) cai dentro do intervalo. Serve como boa estimativa

Calcular um IC Usando Pacote binom

- Facilita cálculos com a distribuição binomial

```
##        method   x    n mean     lower     upper
## 1 asymptotic 480 1000 0.48 0.4490351 0.5109649
```