ME111 - Laboratório de Estatística

Aula 9 - Distribuição Binomial - IC e TH

Profa. Larissa Avila Matos

- Suponha que queremos determinar a proporção de crianças com idade inferior aos 4 anos na cidade de Campinas que sofrem de bronquite. Podemos definir uma v.a. X como sendo

$$X = \begin{cases} 1, & \text{se a criança tem bronquite;} \\ 0, & \text{se a criança não tem bronquite.} \end{cases}$$

- Então, temos que X é uma v.a. discreta, com distribuição de Bernoulli tal que

$$\mu = \mathbb{E}(X) = p,$$

e

$$\sigma^2 = \text{Var}(X) = p(1-p).$$

- Suponha que retiramos uma amostra aleatória $X_1, X_2, \ldots, X_n$ sem reposição de tamanho n dessa população, onde definimos Y como sendo o total de criança com bronquite nessa amostra, vimos que

$$Y \sim \text{ Bin}(n, p),$$

ou seja,

$$\mathbb{P}(Y = y) = \left(\begin{array}{c} n \\ y \end{array} \right) p^y (1-p)^{n-y}, \quad y = 0, 1, \ldots, n.$$

- Agora, considere $\hat{p}$ como a proporção de criança com bronquite, ou seja,

$$\hat{p} = \frac{Y}{n}.$$

- Então, temos que

$$\mathbb{P}(Y = y) = \mathbb{P}(Y/n = y/n) = \mathbb{P}(\hat{p} = y/n),$$

ou seja, a distribuição amostral de $\hat{p}$ é obtida da distribuição de Y.

- Temos que

$$Y = X_1 + X_2 + \ldots + X_n,$$

onde cada X_i tem distribuição de Bernoulli com média $\mu = p$ e variância $\sigma^2 = p(1-p)$ com p sendo um parâmetro desconhecido e os X_i's sendo independentes.

- Assim,

$$Y = \sum_{i=1}^{n} X_i = n \sum_{i=1}^{n} \frac{X_i}{n} = n\overline{X}.$$

- Mas, como visto na aula passada, $Y = n\overline{X}$ tem distribuição aproximadamente normal, com média np e variância $np(1-p)$, ou seja

$$Y = n\overline{X} \sim N\left(np, np(1-p)\right).$$

- Então,

$$\overline{X} \sim N\left(p, \frac{p(1-p)}{n}\right).$$

- Podemos observar que $\overline{X}$, na expressão acima, é a própria variável $\hat{p}$ e, desse modo, para n grande podemos considerar a distribuição amostral de p como aproximadamente normal

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right).$$

Exemplo

- Suponha que queremos saber a porcentagem de crianças com menos de 4 anos que tem bronquite na cidadde de Campinas. Como não temos recursos suficientes para checar todas as crianças com menos de 4 anos, vamos estimar esta porcentagem baseados em alguns dados disponíveis.

- Suponha que temos dados sobre 20 crianças de Campinas:

$$
\begin{array}{cccccc}
X_1 = 1, & X_2 = 1, & X_3 = 0, & X_4 = 0, & X_5 = 0, & X_6 = 0, \\
X_7 = 1, & X_8 = 1, & X_9 = 0, & X_{10} = 0, & X_{11} = 0, & X_{12} = 0, \\
X_{13} = 0, & X_{14} = 0, & X_{15} = 0, & X_{16} = 0, & X_{17} = 0, & X_{18} = 0, \\
X_{19} = 1, & X_{20} = 0. & & & &
\end{array}
$$

- Isto é, por exemplo, a primeira criança tem bronquite.

- Então, o número de crianças com bronquite entre estas crianças é

$$\sum_{i=1}^{20} X_i = 5.$$

- E, a proporção estimada (probabilidade) de crianças com bronquite é

$$\hat{p} = \frac{5}{20} = 0,25.$$

- No R:

```
amostra<-c(1,1,0,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,1,0)
amostra
```

```
 [1] 1 1 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 1 0
```

```
length(amostra)
```

```
[1] 20
```

```
sum(amostra)
```

```
[1] 5
```

```
p.hat=mean(amostra)
p.hat
```

```
[1] 0.25
```

- Note que para a distribuição binomial, se sabemos a real probabilidade de crianças com bronquite, p, poderíamos calcular a probabilidade de termos $\hat{p} = 0,25$ baseados em uma amostra de tamanho 20.

- Quando $n = 20$, esta é justamente a probabilidade de observamos 4 crianças com bronquite, ou seja,

$$\mathbb{P}(Y = 5) = \left(\begin{array}{c} 20 \\ 5 \end{array} \right) p^5 (1 - p)^{15}.$$

- Se, por exemplo, $p = 0,4$, então

```
dbinom(5,20,0.4)
```

```
[1] 0.07464702
```

$\Rightarrow \mathbb{P}(Y = 5) = 0.075.$

- Ou seja, a probabilidade de tomarmos $\hat{p} = 0,25$ é 0.075.

Estudo de simulação

- Suponha que a proporção de crianças com bronquite da cidadde de Campinas (população) é $p = 0,35$.

- Imagine 1000 equipes de pesquisadores e suponha que cada equipe estima a proporção de crianças com bronquite baseada em dados de 20 crianças. Neste caso, diferentes equipes de pesquisadores conseguirão resultados diferentes. Por exemplo, a primeira equipe consegue $\hat{p} = 0,5$, a segunda equipe consegue $\hat{p} = 0,1$, e assim por diante.

- A distribuição amostral de $\hat{p}$ se refere a distribuição dos valores de $\hat{p}$ que as equipes de pesquisadores conseguiriam ao conduzir o mesmo estudo.

Estudo de simulação

- Suponha que a proporção de crianças com bronquite da cidadde de Campinas (população) é $p = 0,35$.

- Imagine 1000 equipes de pesquisadores e suponha que cada equipe estima a proporção de crianças com bronquite baseada em dados de 20 crianças. Neste caso, diferentes equipes de pesquisadores conseguirão resultados diferentes. Por exemplo, a primeira equipe consegue $\hat{p} = 0,5$, a segunda equipe consegue $\hat{p} = 0,1$, e assim por diante.

- A distribuição amostral de $\hat{p}$ se refere a distribuição dos valores de $\hat{p}$ que as equipes de pesquisadores conseguiriam ao conduzir o mesmo estudo.

```
B<-1000
n<-20
p<-0.35
sim<-rbinom(B,n,p)/n
```

- Na tabela a seguir, temos a distribuição de $\hat{p}$.

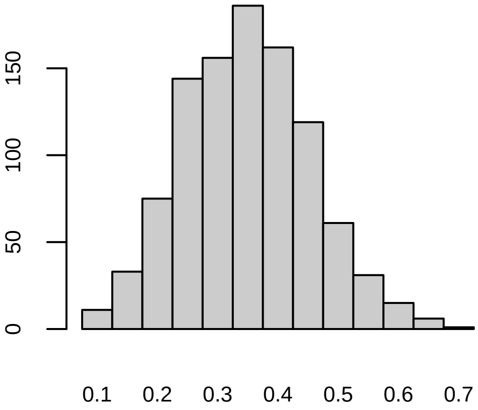
```r
table(sim)
```

```
sim
0.05  0.1 0.15  0.2 0.25  0.3 0.35  0.4 0.45  0.5 0.55  0.6 0.65  0.7 0.75
   5   16   33   81  126  172  163  170  109   71   37   12    2    2    1
```

- Na tabela a seguir, temos a distribuição de $\hat{p}$.

```r
table(sim)
```

```
sim
0.05  0.1 0.15  0.2 0.25  0.3 0.35  0.4 0.45  0.5 0.55  0.6 0.65  0.7 0.75
   5   16   33   81  126  172  163  170  109   71   37   12    2    2    1
```



- Com esses dados, podemos concluir que $\mathbb{E}(\hat{p})$ e $\text{Var}(\hat{p})$ são dados por

```r
mean(sim)
```

```
[1] 0.34845
```

```r
sd(sim)^2
```

```
[1] 0.01134144
```

```r
# comparando com a distr. amostral
p
```

```
[1] 0.35
```

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

```
[1] 0.011375
```

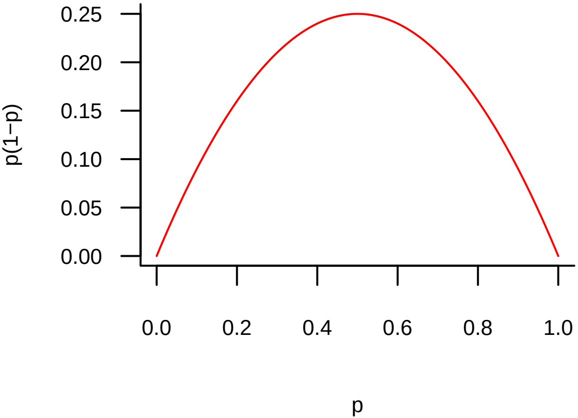
Intervalo de Confiança como Estimativa de p

- Veremos como construir intervalos de confiança para a proporção p, utilizando a aproximação Normal. Vimos que,
$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right).$$
- Seja $1 - \alpha$ o grau de confiança do intervalo.
- Geralmente usamos $\alpha = 0.05$, então o grau de confiança é 95%.
- Queremos encontrar um intervalo tal que a probabilidade do intervalo conter o verdadeiro valor de p seja $(1 - \alpha) \times 100\%$.

$$
\begin{aligned}
0.95 &= P\left(-1.96 \le Z \le 1.96\right) \\
&= P\left(-1.96 \le \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \le 1.96\right) \\
&= P\left(-1.96\sqrt{p(1-p)/n} \le \hat{p} - p \le 1.96\sqrt{p(1-p)/n}\right) \\
&= P\left(\hat{p} - 1.96\sqrt{\frac{p(1-p)}{n}} \le p \le \hat{p} + 1.96\sqrt{\frac{p(1-p)}{n}}\right)
\end{aligned}
$$

- Note que p é desconhecido, mas a variância depende da função de $p(1-p)$, dada no seguinte gráfico:

- Note que p é desconhecido, mas a variância depende da função de $p(1-p)$, dada no seguinte gráfico:



- A função $p(1-p)$ atinge o valor máximo quando $p = 1/2$, ou seja, $p(1-p) \leq \frac{1}{4}$.

- Vimos que $p(1-p) \leq \frac{1}{4}$, então erro-padrão é maximizado por:

$$\sqrt{\frac{p(1-p)}{n}} \leq \sqrt{\frac{1}{4n}} \quad \Longleftrightarrow \quad -\sqrt{\frac{p(1-p)}{n}} \geq -\sqrt{\frac{1}{4n}}$$

- Portanto, $0.95 \leq P\left(\hat{p} - 1.96\sqrt{\frac{1}{4n}} \leq p \leq \hat{p} + 1.96\sqrt{\frac{1}{4n}}\right)$.

- Caso geral (conservador): Um IC de $100(1-\alpha)\%$ para p é dado por

$$IC(p, 1-\alpha) = \left[\hat{p} - z_{\alpha/2}\sqrt{\frac{1}{4n}}; \hat{p} + z_{\alpha/2}\sqrt{\frac{1}{4n}}\right].$$

- Podemos também usar a estimativa $\hat{p}$ para estimar o erro-padrão $\sqrt{\frac{p(1-p)}{n}}$.

- Portanto, é possivel construir o seguinte IC de $100(1-\alpha)\%$

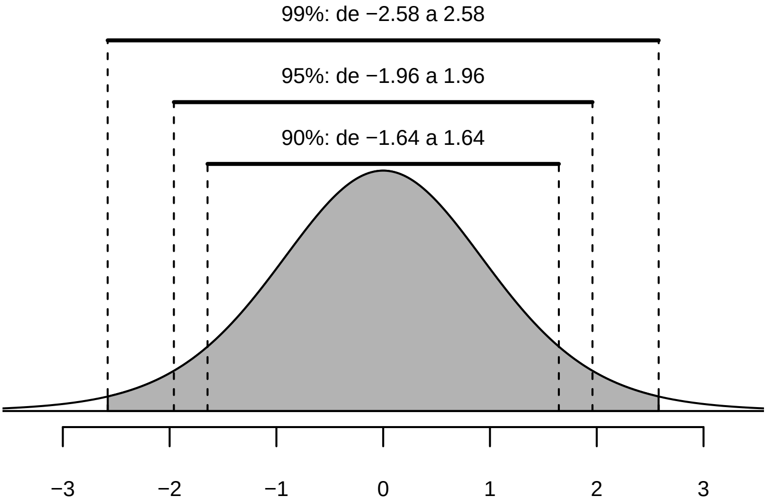
$$IC(p, 1-\alpha) = \left[\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}; \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right].$$

- Veja que nos dois casos temos que escolher as quantidades $z_{\alpha/2}$ tal que:

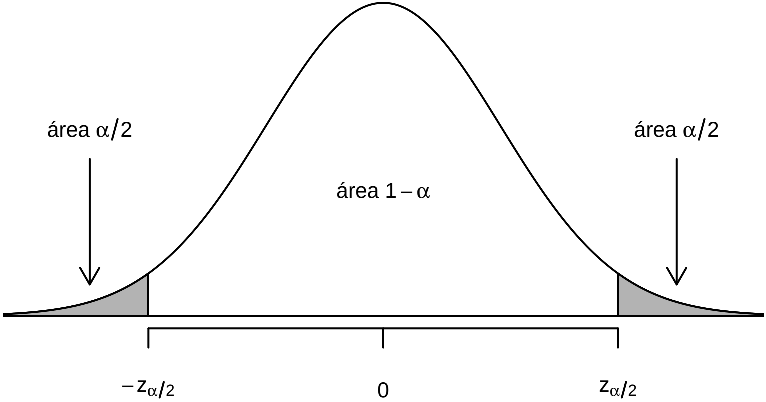
$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1-\alpha.$$

Como encontrar $z_{\alpha/2}$

$$P(|Z| \leq z_{\alpha/2}) = P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$$



$$P(|Z| \leq z_{\alpha/2}) = P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$$



- Seja $Z \sim N(0,1)$. O percentil $z_{\alpha/2}$ é tal que

$$1 - \alpha = P\left(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}\right).$$

- Como determinar $z_{\alpha/2}$?

$$\begin{aligned}
1 - \alpha = P\left(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}\right) &= P(Z \leq z_{\alpha/2}) - P(Z \leq -z_{\alpha/2}) \\
&= P(Z \leq z_{\alpha/2}) - P(Z \geq z_{\alpha/2}) \\
&= P(Z \leq z_{\alpha/2}) - \left[1 - P(Z \leq z_{\alpha/2})\right] \\
&= 2P(Z \leq z_{\alpha/2}) - 1 \\
&= 2\Phi(z_{\alpha/2}) - 1.
\end{aligned}$$

- Portanto, $1 - \frac{\alpha}{2} = \Phi(z_{\alpha/2}) \quad \Rightarrow \quad \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) = z_{\alpha/2}$.

- Procure na tabela o valor de z tal que a probabilidade acumulada até o valor de z, isto é $P(Z \leq z) = \Phi(z)$, seja $1 - \alpha/2$.

```r
qnorm(0.995)
```

```
[1] 2.575829
```

```r
qnorm(0.975)
```

```
[1] 1.959964
```

```r
qnorm(0.95)
```

```
[1] 1.644854
```

Interpretação do Intervalo de Confiança

- Se várias amostras forem retiradas da população e calcularmos um IC de 95% para cada amostra, cerca de 95% desses intervalos irão conter a verdadeira proporção na população, p.

- INCORRETO: Dizer que "a probabilidade de que p esteja dentro do intervalo é 95%".

- Por que incorreto? p é uma constante, não é variável aleatória. Ou p está no intervalo ou não está. O intervalo é que é aleatório.

Exercício

- Numa amostra aleatória de 800 crianças menores de 4 anos da cidade de Campinas foram encontradas 300 crianças que sofrem com a bronquite. Ache um intervalo de confiança de nível 95% para a proporção p crianças com bronquite.

Exercício

- Numa amostra aleatória de 800 crianças menores de 4 anos da cidade de Campinas foram encontradas 300 crianças que sofrem com a bronquite. Ache um intervalo de confiança de nível 95% para a proporção p crianças com bronquite.

```
xobs=300
n=800
p.hat=xobs/n
p.hat
```

```
[1] 0.375
```

```r
zalpha=qnorm(0.975)
zalpha
```

```
[1] 1.959964
```

```r
limInf=p.hat-zalpha*sqrt(p.hat*(1-p.hat)/n)
limInf
```

```
[1] 0.3414526
```

```r
limSup=p.hat+zalpha*sqrt(p.hat*(1-p.hat)/n)
limSup
```

```
[1] 0.4085474
```

- Então, o intervalo de confiança é dado por

$$IC(p, 95\%) = (0,341; 0,409)\,.$$

- Interpretação: Com um grau de confiança de 95%, estimamos que a proporção de crianças com bronquite está entre 34,1% e 40,9%.

- Uma outra maneira de obtermos um intervalo de confiança para proporção é através da aproximação normal com correção de continuidade. Considerando o processo anterior, a única diferença é que aqui não consideraremos simplesmente a proporção amostral $\hat{p}$, mas sim uma correção dela. Assim, para determinar o intervalo de confiança consideramos uma modificação da proporção $\hat{p}$, dada por:

$$\hat{p}_c = \begin{cases} \hat{p} + \dfrac{1}{2n} \text{ se } \hat{p} < 0,5; \\ \\ \hat{p} - \dfrac{1}{2n} \text{ se } \hat{p} > 0,5. \end{cases}$$

- Assim, o intervalo de confiança para proporção p com correção de continuidade, é dado por

$$IC(p, 1-\alpha) = \left(\hat{p}_c - Z_{\alpha/2}\sqrt{\frac{\hat{p}_c(1-\hat{p}_c)}{n}}, \hat{p}_c + Z_{\alpha/2}\sqrt{\frac{\hat{p}_c(1-\hat{p}_c)}{n}}\right).$$

- Repetindo o exercício anterior aplicando a correção de continuidade.

```
xobs=300
n=800
p.hat=ifelse(xobs/n > 0.5, xobs/n-(1/(2*n)),xobs/n+(1/(2*n)))
p.hat
```

```
[1] 0.375625
```

```
zalpha=qnorm(0.975)
limInf=p.hat-zalpha*sqrt(p.hat*(1-p.hat)/n)
limInf
```

```
[1] 0.3420665
```

```
limSup=p.hat+zalpha*sqrt(p.hat*(1-p.hat)/n)
limSup
```

```
[1] 0.4091835
```

- Os intervalos de $100(1-\alpha)\%$ de confiança para p podem então ser de duas formas:
  1. Método Conservador
  $$IC_1(p, 1-\alpha) = \left[\hat{p} - z_{\alpha/2}\sqrt{\frac{1}{4n}}; \hat{p} + z_{\alpha/2}\sqrt{\frac{1}{4n}}\right].$$
  2. Usando $\hat{p}$ para estimar o erro-padrão
  $$IC_2(p, 1-\alpha) = \left[\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}; \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right].$$

- Coletamos uma amostra aleatória $X_1, X_2, \ldots, X_n$ de uma população com distribuição de Bernoulli com probabilidade de sucesso igual a p, portanto com média p e a variância $p(1-p)$ e usamos $\bar{X} = \hat{p}$ para estimar p.

- Pelo TCL: $\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$. Pela propriedade da Normal:

$$Z = \frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \sim N(0,1), \quad P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1-\alpha.$$

- Então, um intervalo de $100(1-\alpha)\%$ de confiança para p:

$$IC(p, 1-\alpha) = \left[\hat{p} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\,;\, \hat{p} + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right].$$

- Problema: não conhecemos p. Portanto, usamos:

$$IC(p, 1-\alpha) = \left[\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\,;\, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right],$$

ou, pelo método conservador,

$$IC(p, 1-\alpha) = \left[\hat{p} - z_{\alpha/2}\sqrt{\frac{1}{4n}}\,;\, \hat{p} + z_{\alpha/2}\sqrt{\frac{1}{4n}}\right].$$

Teste de Hipóteses para proporção

- Objetivo em muitos estudos: checar se os dados apóiam certas afirmações que são feitas para uma população.
- Afirmações a serem testadas: **hipóteses**.
- Expressamos as hipóteses em termos dos parâmetros da população.
- Por exemplo: o parâmetro pode ser uma proporção populacional, ou seja, voltando ao nosso exercicio: A proporção de crianças menores que 4 anos em campinas com bronquite é 0,35?
  - $p = 0,35$?

- Nesse exercício, para saber se proporção de crianças menores que 4 anos em campinas com bronquite é 0,35, usaríamos as seguintes **hipóteses**:

$$\begin{cases} H_0 : p = 0,35 \quad \text{(hipótese nula)} \\ H_1 : p \neq 0,35 \quad \text{(hipótese alternativa)} \end{cases}$$

- No experimento com as 800 crianças, observar uma proporção alta ou muito baixa de crianças com bronquite pode ser uma evidência contra a hipótese de que $p = 0,35$.

Passos de um teste de hipótese

- **Passo 1: suposições.**
  O teste é válido sob algumas suposições. A mais importante assume que os dados do experimento foram produzidos através de um processo de aleatorização.

- **Passo 2: hipóteses.**
  O teste de hipótese tem sempre duas hipóteses sobre o parâmetro populacional de interesse. As hipóteses devem ser definidas **antes** de se realizar o experimento e coletar dados.

  - **Hipótese Nula - H_0:** afirma que o parâmetro populacional assume um dado valor.
  - **Hipótese Alternativa - H_1:** afirma que o parâmetro populacional assume outros valores, diferente do valor descrito na H_0.
  - **Em teste de hipóteses, mantém-se a favor de H_0 a menos que os dados tragam grande evidência contra.** A hipótese nula é conservadora: "o réu é inocente até que se prove o contrário".

- **Passo 3: estatística do teste.**
  Vimos que podemos usar uma estatística para estimar um parâmetro populacional. A **estatística do teste** descreve quão longe do parâmetro populacional usado na H_0 a estimativa está.

  - Por exemplo, se $H_0 : p = 0,35$, e se $\hat{p} = 300/800 = 0,375$, queremos uma estatística que quantifique quão longe está $\hat{p} = 0,375$ de $p = 0,35$.

- **Passo 4: valor-de-p.**
  Para interpretar uma estatística do teste, vamos usar uma probabilidade para resumir a evidência contra H_0. Esta probabilidade é o que chamamos de **valor-de-p**.

- **Passo 5: conclusão.**
  - Baseado no valor-de-p, decidir se rejeita ou não a hipótese nula. Note que a conclusão é sempre em termos da hipótese nula: rejeitar ou não H_0.
  - Mas quão pequeno deve ser o valor-de-p para ser considerado forte evidência contra H_0?
  - Geralmente, fixamos o **nível de significância** do teste (α), e usamos a seguinte regra. É comum usarmos $\alpha = 0.05$.
  - Se valor-de-p $\leq \alpha$: rejeitamos H_0, ou seja, os dados trazem forte evidência contra a hipóstese nula.
  - Se valor-de-p $> \alpha$: não rejeitamos H_0, ou seja, não temos evidência nos dados contra a hipótese nula.

- Resumindo:
  - Assumimos primeiro que H_0 é verdadeira.
  - Consideramos então todos os valores possíveis para a estatística do teste, de acordo com sua distribuição amostral.
  - Calculamos a estatística do teste observada para o experimento realizado e verificamos onde, na distribuição amostral, ela se posiciona.
  - Calculamos a probabilidade de um valor igual ou mais extremo ao da estatística do teste observada (valor-de-p). Mais extremo: mais evidência contra H_0.
  - Se o valor-de-p obtido é bem pequeno, por exemplo, 0.01, isto quer dizer que se H_0 é verdadeira, então seria incomum obter uma amostra com os resultados como o observado. Um valor-de-p muito baixo traz fortes evidências contra H_0.

- **Estatística do teste:** Da amostra temos que $\hat{p} = 300/800 = 0.375$

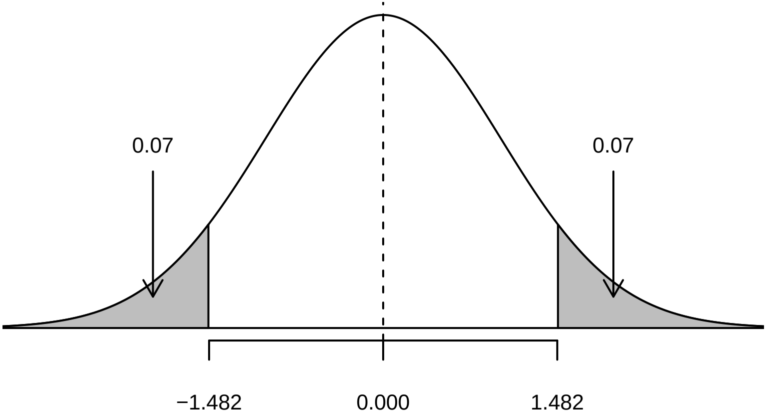
$$z_{obs} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.375 - 0.35}{\sqrt{\frac{0.35(1-0.35)}{800}}} = 1.483$$

- Então,

$$\begin{aligned} \text{valor-de-p} &= P(|Z| \geq 1.483) = P(Z \leq -1.5) + P(Z \geq 1.483) \\ &= 2P(Z \leq -1.483) = 2 \times 0.07 = 0.14 \end{aligned}$$

```
pnorm(-1.483)
```

```
[1] 0.06903721
```


0.07
0.07
−1.482
0.000
1.482

```r
binom.test(300, 800, p = 0.35, alternative = "two.sided")
```

```

	Exact binomial test

data:  300 and 800
number of successes = 300, number of trials = 800, p-value =
0.1384
alternative hypothesis: true probability of success is not equal to 0.35
95 percent confidence interval:
 0.3413412 0.4095837
sample estimates:
probability of success 
                 0.375 
```

```r
prop.test(300, 800, p = 0.35, alternative = "two.sided",
          correct = F)
```

```
	1-sample proportions test without continuity correction

data:  300 out of 800, null probability 0.35
X-squared = 2.1978, df = 1, p-value = 0.1382
alternative hypothesis: true p is not equal to 0.35
95 percent confidence interval:
 0.3421249 0.4090698
sample estimates:
    p
0.375
```

Exercício

- Alice faz um exame com 100 questões de multipla escolha, em que cada questão tem 5 alternativas. O objetivo do estudo é investigar se ela tem um desempenho melhor do que ela faria por adivinhação. Assim, as hipóteses são

$$\begin{cases} H_0 : p = 0,2 & \text{(hipótese nula)} \\ H_1 : p > 0,2 & \text{(hipótese alternativa)} \end{cases}$$

- Suponha que Alice acerta 28 questões em seu exame.

Referências

- Wardrop, R. L. (1995). *Statistics: Learning in the presence of variation.*
- *Notas de aula da Profa. Samara F. Kiihl*
- *Bussab, W. O. & Morettin, P. A. (1987). Estatística Básica. Atual Editora Ltda., São Paulo.*
- *Magalhães, Marcos N.; Lima, Antonio Carlos P. (2010). Noções de probabilidade e estatística. São Paulo: Edusp, 2010.*