Métodos Estatísticos – L.EIC

Semana 12

Aula 10

6 de junho de 2022

Métodos Estatísticos – L.EIC

Aula 10

Inferência Estatística

Intervalo de Confiança para Proporções

Métodos de Wald e de Agresti-Coull

Escolha do Tamanho da Amostra

Intervalo de Confiança Unilateral

Intervalo de Confiança para Diferença de Proporções

Distribuição por Amostragem

Testes de Hipóteses

Exemplos

7

INFERÊNCIA ESTATÍSTICA

– Proporções –

Inferência Estatística
Proporções

Nos capítulos anteriores foram abordados problemas envolvendo variáveis quantitativas e estudados processos para inferir sobre as médias dessas variáveis.
Iremos agora abordar variáveis categóricas e vamos estudar processos de inferência sobre proporções.

Há imensas **questões que se relacionam com proporções**, por exemplo:

- a proporção de pacientes recuperados com um dado tratamento;
- a proporção de votos num dado candidato;
- a proporção de pessoas imunes a uma dada doença.

Inferência Estatística
Proporções

O parâmetro em que estamos agora interessados é então a proporção populacional p, isto é, a proporção de indivíduos da população com uma determinada característica.

Comecemos por considerar uma variável categórica dicotómica, isto é, uma variável que tem apenas dois valores possíveis.

Seja, por exemplo, a variável aleatória X associada à

> **"presença de uma doença nos indivíduos de uma certa população".**

Pretende-se saber qual a proporção da população que tem a doença.

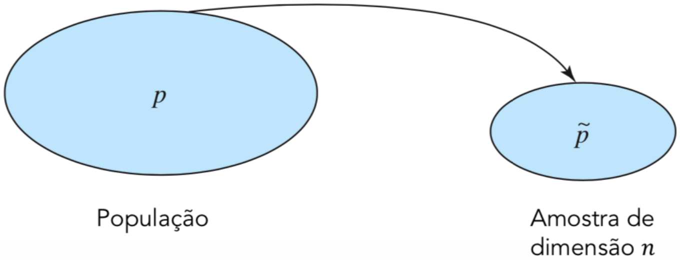
Inferência Estatística
Proporções

Os nossos objetivos são:

- Construir estimadores para uma dada proporção numa população
- Construir e interpretar Intervalos de Confiança para uma proporção
- Fornecer um método para encontrar o tamanho ótimo da amostra para estimar uma proporção

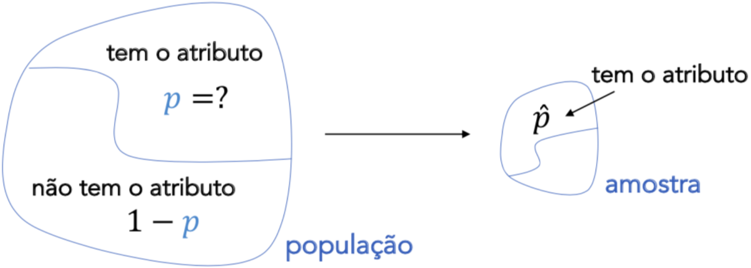
Inferência Estatística
Proporções

Pretendemos estimar a proporção p da população



Inferência Estatística

Proporções – Estimador $\widehat{P}$



- população dicotómica (0,1)
- amostra aleatória (dimensão n)
- atributo de interesse (p.e. pessoa imune a uma doença)

Inferência Estatística

Proporções – Estimador $\widehat{P}$

Consideremos então:

- a proporção p de indivíduos (numa população dicotómica) com uma determinada caraterística.
- a variável aleatória X – representa o número de indivíduos, numa amostra aleatória $X_1, X_2, \ldots, X_n$, de tamanho n, retirada dessa população dicotómica, que têm essa característica.

Um estimador pontual (natural) de p, é:

$$\widehat{P} = \frac{X}{n}$$

Inferência Estatística

Proporções – Estimador $\widehat{P}$

A partir de uma amostra 'concreta', a estimativa 'natural' para a proporção populacional p, é então a proporção nessa amostra, ou proporção amostral,

$$\widehat{p} = \frac{x}{n}$$

onde n é o tamanho da amostra, e x o número de observações na amostra com o atributo de interesse.

Inferência Estatística

Proporções – Estimador $\widehat{P}$

$$\widehat{P} = \frac{X}{n}; \qquad \Big(X \sim Bi(n,p)\Big)$$

Valor esperado de $\widehat{P}$

$$\mu_{\widehat{P}} = \frac{1}{n}E(X) = \frac{1}{n}\mu_X = \frac{1}{n}np = p$$

Variância de $\widehat{P}$

$$\sigma^2_{\widehat{P}} = \frac{1}{n^2}\sigma^2_X = np(1-p)/n^2 = p(1-p)/n$$

Para n **grande**, e $\{np > 5; \quad n(1-p) > 5\}$ podemos aproximar a distribuição de $\widehat{P}$ à distribuição normal.

Inferência Estatística

Proporções – Estimador $\widehat{P}$

O **erro padrão** da estimativa $\hat{p}$ obtida a partir de uma amostra aleatória de dimensão n é dado por:

$$s_{\hat{P}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

O erro padrão $s_{\hat{P}}$ descreve a fiabilidade que se tem na proporção $\hat{p}$ da amostra como estimativa da proporção populacional p.

Exemplo

Para estimar a proporção p de máquinas de sumos contaminadas com uma dada bactéria, foram selecionadas aleatoriamente 30 máquinas e encontradas 5 contaminadas.

A estimativa 'natural' para p, com base na amostra, é

$$\hat{p} = \frac{5}{30} \approx 0.167$$

O erro padrão desta estimativa é:

$$s_{\hat{P}} = \sqrt{\frac{5}{30}\left(1 - \frac{5}{30}\right)/30} \approx 0.068$$

Inferência Estatística

Proporções – I.C. – Estimador $\widehat{P}$

Convém lembrar que para grandes valores de n (pelo Teorema do Limite Central), um intervalo de confiança a aproximadamente 95% para a média μ, de uma população, tem a forma:

$$\overline{x} \pm 1.96 \cdot \frac{s}{\sqrt{n}}$$

ou seja

$$\overline{x} \pm 1.96 \cdot \mathrm{se}_{\overline{X}}$$

Pode construir-se de modo análogo, um intervalo de confiança para p a aproximadamente 95%.

Inferência Estatística

Proporções – I.C. – Estimador $\widehat{P}$

Pelo Teorema do Limite Central tem-se:

$$\widehat{P} = \frac{X}{n} \dot{\sim} N\left(p, \frac{p(1-p)}{n}\right)$$

ou seja

$$\frac{\widehat{P} - p}{\sqrt{\frac{p(1-p)}{n}}} \dot{\sim} N(0,1)$$

Para grandes valores de n, tem-se então

$$\frac{\widehat{P} - p}{se_{\widehat{P}}} \dot{\sim} N(0,1) \quad \text{onde} \quad se_{\widehat{P}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Método de Wald – Estimador $\widehat{P}$

Pergunta: Como construir um intervalo de confiança para a proporção populacional p?

Há diversos métodos!

De acordo com o designado Método de Wald,

$$\hat{p} \pm z_{\alpha/2} \underbrace{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}_{se_{\widehat{P}}}$$

é um I.C., com grau de confiança aproximadamente $1-\alpha$, para a proporção populacional p, baseado numa amostra de tamanho n, e na estimativa $\hat{p}$.

Inferência Estatística
Proporções – Intervalo de Confiança

Método de Wald – Estimador $\widehat{P}$

Na verdade, seja

$$Z = \frac{X - np}{\sqrt{np(1-p)}} = \frac{\hat{P} - p}{\sqrt{p(1-p)/n}}$$

Se n é grande, temos $Z \dot{\sim} N(0,1)$, e para construir o intervalo aleatório consideramos que,

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

i.e.:

$$P\left(-z_{\alpha/2} \leq \frac{\hat{P} - p}{\sqrt{p(1-p)/n}} \leq z_{\alpha/2}\right) \approx 1 - \alpha$$

Inferência Estatística
Proporções – Intervalo de Confiança

Método de Wald – Estimador $\widehat{P}$

Para o IC, obtemos:

$$P\left(\hat{p} - z_{\alpha/2}\sqrt{p(1-p)/n} \leq p \leq \hat{p} + z_{\alpha/2}\sqrt{p(1-p)/n}\right) \approx 1-\alpha$$

Como o erro padrão de $\hat{p}$ é $\sqrt{p(1-p)/n}$, é necessário o valor de p para o calcular... **e não conhecemos** p ...

Mas se n é grande podemos substituir p por $\hat{p}$.

Assim, o intervalo de confiança aproximado, a $100(1-\alpha)\%$, para a proporção populacional p será dado por

$$\hat{p} - z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} \leq p \leq \hat{p} + z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$$

Método de Wald – Estimador $\widehat{P}$

Como acabamos de ver, a justificação teórica é baseada no facto da v.a. $\hat{P} = X/n$, representativa da proporção de observações com o atributo de interesse numa qualquer amostra aleatória de tamanho n, seguir aproximadamente uma distribuição normal para valores grandes de n:

$$\widehat{P} = \frac{X}{n} \dot{\sim} N\left(p, \frac{p(1-p)}{n}\right)$$

Por isso, o método de Wald para a construção de um intervalo de confiança para uma proporção, é adequado quando as amostras são muito grandes e p não é muito próximo de 0 nem de 1.

Inferência Estatística
Proporções – Intervalo de Confiança

Método de Wald – Estimador $\widehat{P}$

Usando o estimador $\widehat{P}$ o intervalo de confiança a aproximadamente 95% para a proporção p de uma população, é:

$$\hat{p} \pm 1.96 \cdot \mathrm{se}_{\hat{P}}$$

Para outros graus de confiança o intervalo de confiança é construído de modo análogo, bastando substituir o valor do quantil da distribuição normal.

Por exemplo, para uma confiança aproximada de 99%, temos

$$\hat{p} \pm 2.576 \cdot \mathrm{se}_{\hat{P}}$$

Inferência Estatística
Proporções – Intervalo de Confiança

Método de Wald – **Margem de Erro**

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Se usarmos $\hat{p}$ para estimar p, então podemos dizer com grau de confiança $1-\alpha$ que o erro que cometemos não excede:

$$z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Este valor designa-se por **margem de erro**.

Inferência Estatística
Proporções – Intervalo de Confiança

Método de Wald – **Exemplo**

Numa amostra aleatória de 85 plantas, 10 sofrem de uma dada mutação genética.

Construir o intervalo de confiança a 95% para a proporção p, de plantas que sofrem da mutação genética.

Uma estimativa pontual da proporção de plantas na população, que sofrem dessa mutação, é

$$\hat{p} = \frac{x}{n} = \frac{10}{85} \cong 0.12$$

Inferência Estatística
Proporções – Intervalo de Confiança

Método de Wald – **Exemplo**

Um intervalo de confiança a 95% para p, é determinado então por:

$$\hat{p} - z_{0.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{0.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

t

$$0.12 - 1.96\sqrt{0,12 \cdot \frac{0.88}{85}} \leq p \leq 0.12 + 1.96\sqrt{0.12 \cdot 0.88/85}$$

$$0.05 \leq p \leq 0.19$$

Inferência Estatística
Proporções – Intervalo de Confiança

Método de Agresti-Coull – Estimador $\widetilde{P}$

Podemos construir um intervalo de confiança a aproximadamente 95% para a proporção populacional p, usando uma amostra de dimensão n, e o estimador de Wilson ajustado $\widetilde{P}$.

O intervalo de confiança é determinado por:

$$\tilde{p} \pm 1.96 \cdot \mathrm{se}_{\tilde{P}}$$

onde 1.96 é o valor de $z_{0.025}$, e $\mathrm{se}_{\tilde{P}}$ é o erro padrão:

$$\mathrm{se}_{\tilde{P}} = \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}}$$

Inferência Estatística
Proporções – Intervalo de Confiança

Método de Agresti-Coull – Proporção de Wilson Ajustada

A estimativa da proporção p, obtida a partir de uma amostra de tamanho n, em que se observaram x 'sucessos', é dada neste caso por:

$$\tilde{p} = \frac{x+2}{n+4}$$

- Para amostras pequenas, este método é melhor que o de Wald.
- Esta estimativa da proporção fica mais próxima de $1/2$
- Intervalos de confiança com base nesta estimativa $\tilde{p}$, na verdade, são mais fiáveis do que aqueles baseados em $\hat{p}$.

Inferência Estatística
Proporções – Intervalo de Confiança

Método de Agresti-Coull – **Exemplo 1**

Num estudo sobre crescimento da população de uma certa região, de uma amostra de 496 mulheres entre 20 e 24 anos que deram à luz, 78 fumaram durante a gravidez $(15,7\% = 78/496)$.

$$\tilde{p} = \frac{78+2}{496+4} = 0.16 \quad \rightarrow \quad se_{\tilde{P}} = \sqrt{\frac{0.16(1-0.16)}{500}} = 0.016$$

Espera-se que a proporção p, de mulheres fumadoras durante a gravidez entre os 20 e 24 anos, esteja entre $\pm$ 2 desvios padrão, i.e. $p = 0.16 \pm 2 \times 0.016$.

Uma afirmação mais precisa é estabelecida através do I.C..

Inferência Estatística
Proporções – Intervalo de Confiança

Método de Agresti-Coull – **Exemplo 2**

Pretende-se construir um intervalo de confiança a 95% para a proporção p de máquinas de sumos contaminadas com uma bactéria.

Foram selecionadas aleatoriamente 30 dessas máquinas e encontradas 5 máquinas contaminadas.

$$\tilde{p} = \frac{7}{34} \approx 0.21; \qquad s_{\tilde{P}} = \sqrt{\frac{\frac{7}{34}\left(1 - \frac{7}{34}\right)}{34}} \approx 0.07$$

I.C. é dado por:

$$\tilde{p} \pm 1.96\, s_{\tilde{P}} \approx 0.21 \pm 0.138$$

Método de Agresti-Coull – **Exemplo 3**

A oxigenação por membrana extracorpórea (ECMO) é um procedimento médico usado para salvar recém-nascidos que sofrem de problemas respiratórios.

Num estudo envolvendo 11 recém-nascidos tratados com ECMO todos sobreviveram.

Pretendemos determinar um intervalo de confiança para a proporção de mortes de recém-nascidos tratados com ECMO.

Inferência Estatística
Proporções – Intervalo de Confiança

Método de Agresti-Coull – **Exemplo 3**

Uma vez que a amostra é pequena e que indicia que a proporção a estimar é próxima de zero, usaremos o método de Agresti-Coull e construiremos um IC a 95%.

$$\tilde{p} = \frac{0+2}{11+4} \approx 0.133; \qquad s_{\tilde{P}} = \sqrt{\frac{\frac{2}{15}\left(1-\frac{2}{15}\right)}{15}} \approx 0.088$$

I.C. é dado por:

$$\tilde{p} \pm 1.96\, s_{\tilde{P}} \approx 0.21 \pm 0.138$$

$$]-0.04, 0.306[$$

Inferência Estatística
Proporções – Intervalo de Confiança

Método de Agresti-Coull – **Exemplo 3**

$$]-0.04, 0.306[$$

Uma vez que $0 \leq p \leq 1$, tomaremos o I.C. (0, 0.306).

Podemos assim afirmar com uma confiança de 95% que a proporção de mortes de recém-nascidos tratados com ECMO se situa no intervalo]0, 0.306[.

Nota: Não faria sentido usar aqui o método de Wald. De facto, como $\hat{p} = 0$, também $s_P = 0$, o que conduziria ao I.C.]0, 0[...

Inferência Estatística
Proporções – Intervalo de Confiança

Método de Agresti-Coull – **Exemplo 4**

Em determinadas eleições entrevistaram-se 1000 eleitores para saber se eram a favor, ou contra, o candidato A.

Verificou-se que 350 eram a favor e 650 contra.

Encontrar um intervalo com 95% de confiança para a percentagem de eleitores na população favoráveis ao candidato A.

Inferência Estatística
Proporções – Intervalo de Confiança

Método de Agresti-Coull – **Exemplo 4**

Método de Agresti-Coull:

O intervalo de confiança é,

$$]\tilde{p} - z_{\alpha/2} \cdot \mathrm{se}_{\tilde{P}}, \tilde{p} + z_{\alpha/2} \cdot \mathrm{se}_{\tilde{P}}[$$

$$n = 1000; \quad 1 - \alpha = 0.95; \quad 1 - \alpha/2 = 0.975$$

$$z_{0.025} \approx 1.96; \quad \tilde{p} = \frac{350 + 2}{1000 + 4} = 0.3506$$

$$se_{\tilde{P}} = \sqrt{\frac{0.35(1 - 0.35)}{n + 4}} \quad \Rightarrow \quad I.C._{(95\%_6)} \approx]0.3211, 0.3801[$$

Inferência Estatística
Proporções – Intervalo de Confiança

Método de Agresti-Coull – **Exemplo 5**

Consideremos o 'Exemplo 1' anterior, de uma amostra de 496 mulheres entre 20 e 24 anos, em que 78 fumaram durante a gravidez, isto é, 15,7% $= 78/496$.

Pretende-se construir um intervalo de confiança para a proporção p, dessas mulheres fumadoras, a aproximadamente 95%.

Vamos utilizar os 2 métodos:

1. Método de Wald – estimador $\widehat{P}$
2. Método de Agresti-Coull – estimador $\widetilde{P}$

Inferência Estatística
Proporções – Intervalo de Confiança

Método de Agresti-Coull – **Exemplo 5**

1. Método de Wald – estimador $\widehat{P}$

$$\hat{p} = \frac{78}{496} \approx 0.157$$

$$se_{\hat{P}} = \sqrt{\frac{0.157(1-0.157)}{496}} \approx 0.016$$

Intervalo de confiança para p, a aproximadamente 95% é então determinado por $0.157 \pm 1.96 \times 0.016$, isto é

$$\text{I.C.}_{(95\%)} =]0.126, 0.189[$$

Inferência Estatística
Proporções – Intervalo de Confiança

Método de Agresti-Coull – **Exemplo 5**

2. Método de Agresti-Coull – estimador $\widetilde{P}$

$$\tilde{p} = \frac{78+2}{496+4} = 0.16$$

$$se_{\tilde{P}} = \sqrt{\frac{0.16(1-0.16)}{500}} = 0.016$$

Intervalo de confiança para p, a aproximadamente 95% é agora determinado por $0.16 \pm 1.96 \times 0.016$, isto é

$$\text{I.C.}_{(95\%)} =]0.129, 0.191[$$

Inferência Estatística
Proporções – Intervalo de Confiança

Alguns comentários

- O método de Wald **não deve ser usado** nas situações em que a proporção p poderá ser próxima de zero ou de um;
- O I.C. de Wald **só deve ser usado** para amostras que tenham um tamanho muito elevado;
- Note-se que, dada a elevada dimensão da amostra no exemplo anterior, os I.C. obtidos são próximos. Na verdade, para amostras de dimensão elevada os I.C. obtidos pelos **dois métodos** são muito próximos.
- Se nada for dito em contrário, na construção de I.C. a 95% deve ser usado o método de Agresti-Coull.

Inferência Estatística
Proporções – Intervalo de Confiança

Intervalos de confiança a 95% – **Comparação**

n	Wald	Agresti-Coull
10	]0.096, 0.704[	] 0.169,0.688[
30	]0.165, 0.502[	] 0.192, 0.514[
100	]0.229, 0.411[	]0.237, 0.417[
1000	]0.270, 0.326[	]0.270, 0.327[

Nota: valores arredondados, obtidos por simulação de amostras aleatórias ($p = 0.3$)

Inferência Estatística
Proporções – Planeamento

Escolha do Tamanho da Amostra

Supor que se pretende construir um I.C. para uma proporção p, que obedeça a determinadas características.

Qual deverá ser o tamanho da amostra a considerar?

Uma vez **especificados** o grau de confiança $1-\alpha$, e o erro máximo ε, é possível determinar o tamanho da amostra n, para se construir o I.C. com as características pretendidas.

O procedimento é o mesmo quer se use o método de Wald, ou o método de Agresti-Coull.

Inferência Estatística
Proporções – Planeamento

Escolha do Tamanho da Amostra

Queremos que o intervalo de confiança, ou estimativa, tenha a precisão necessária para o objetivo pretendido, sendo que o intervalo é centrado na estimativa.

(No que se segue usaremos sempre o método de Agresti-Coull.)

Sabemos que o I.C. é determinado por:

$$\tilde{p} \pm z_{\alpha/2} \cdot \mathrm{se}_{\tilde{P}}$$

A dimensão da amostra n, está relacionada com a margem de erro, através de $z_{\alpha/2}$ e do erro padrão $se_{\tilde{P}}$,

$$se_{\tilde{P}} = \sqrt{\frac{p(1-p)}{n+4}}$$

Inferência Estatística
Proporções – Planeamento

Escolha do Tamanho da Amostra

Podemos definir o I.C. desejado, a $100(1-\alpha)\%$, bem como o erro máximo pretendido ε, isto é,

$$|p - \tilde{p}| \leq \varepsilon$$

ou

$$z_{\alpha/2} \cdot se_{\tilde{P}} = z_{\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n+4}} \leq \varepsilon$$

E assim,

$$n \geq \left(\frac{z_{\alpha/2}}{\varepsilon}\right)^2 p(1-p) - 4$$

Inferência Estatística
Proporções – Planeamento

Escolha do Tamanho da Amostra

Ficamos então a conhecer o **limite inferior** para n, para construir o I.C. com as características pretendidas:

$$n \geq \left(\frac{z_{\alpha/2}}{\varepsilon}\right)^2 p(1-p) - 4$$

mas o valor de p **é desconhecido**...

Uma solução será utilizar um valor previamente estimado de p (ou uma estimativa subjetiva).

Em alternativa podemos considerar uma amostra preliminar para obter uma primeira estimativa de p.

Escolha do Tamanho da Amostra – **Exemplo 1**

Numa população de plantas de determinada espécie, há algumas que sofrem de uma dada mutação genética.

O método de Wald aplicado a uma amostra preliminar, permitiu fazer uma primeira estimativa da proporção p, de plantas que sofreram a mutação nesta população.

A estimativa feita foi $\hat{p} = 0.12$.

Determinar qual deve ser o tamanho da amostra, para se ter um erro máximo de 0.05, com um grau de confiança de 95%.

Inferência Estatística
Proporções – Planeamento

Escolha do Tamanho da Amostra – **Exemplo 1**

$$n \geq \left(\frac{z_{\alpha/2}}{\varepsilon}\right)^2 p(1-p)$$

Utilizando a estimativa preliminar para p, i.e., $\hat{p} = 0.12$, temos:

$$n \geq \left(\frac{z_{\alpha/2}}{\varepsilon}\right)^2 \hat{p}\,(1-\hat{p}) = \left(\frac{1.96}{0.05}\right)^2 \times 0.12 \times 0.88 \cong 163$$

Se não dispusermos de qualquer estimativa para p, podemos ainda assim calcular um limite inferior para n, como veremos mais à frente.

Inferência Estatística
Proporções – Planeamento

Escolha do Tamanho da Amostra – **Exemplo 2**

Numa amostra aleatória de 400 alunos do sexo masculino de uma determinada escola, constatou-se que 40 eram esquerdinos.

Queremos determinar o tamanho da amostra a utilizar, para estimar a proporção de alunos esquerdinos do sexo masculino, tal que $se_{\tilde{P}} \leq 1\%$.

A estimativa da amostra para p, é:

$$\tilde{p} = \frac{40 + 2}{400 + 4} = 0.104$$

$$se_{\tilde{P}} = \sqrt{\frac{0.104(0.896)}{n + 4}} \leq 0.01$$

$$n + 4 \geq 931.8 \quad \Rightarrow \quad n \geq 928 \text{ estudantes}$$

Inferência Estatística
Proporções – Planeamento

Escolha do Tamanho da Amostra – **Exemplo 3**

Suponha-se que, tal como no exemplo anterior, se pretende determinar o tamanho da amostra n, para estimar a proporção p de esquerdinos com erro máximo $\varepsilon < 0.01$.

Admitamos ainda que está fixado o grau de confiança, isto é, que conhecemos $z_{\alpha/2}$.

No entanto, agora não dispomos de qualquer estudo prévio.

Neste caso, que tamanho de amostra escolher?

Mesmo sem estudo prévio, é possível sugerir um valor para n!

Inferência Estatística
Proporções – Planeamento

Escolha do Tamanho da Amostra

$$n \geq \left(\frac{z_{\alpha/2}}{\varepsilon}\right)^2 p(1-p) - 4$$

Quando não temos qualquer estudo prévio para estimar p, podemos ainda assim calcular um limite superior para o fator $p(1-p)$. Na verdade,

$$p(1-p) = p - p^2 \leq 0.25$$

pois a função $f(p) = p - p^2$ tem um máximo para $p = 0.5$.

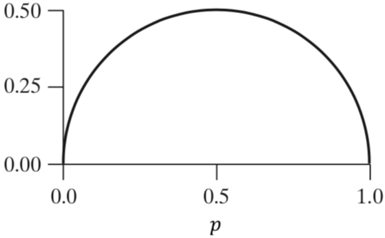
Podemos então utilizar esta relação para obter o limite para n:

$$n \geq \left(\frac{z_{\alpha/2}}{\mathcal{E}}\right)^2 \cdot 0.25 - 4$$

Inferência Estatística
Proporções – Planeamento

Escolha do Tamanho da Amostra

$$p(1-p) \leq 0.25 \quad \text{ou} \quad \sqrt{p(1-p)} \leq 0.5$$



Com $p = 0.5$ obtemos um valor de n “conservativo”, i.e., em muitas situações não é necessário um valor tão elevado.

Inferência Estatística
Proporções – Planeamento

Escolha do Tamanho da Amostra – **Exemplo 3**

Retomemos agora o exercício.

Pretende-se determinar o tamanho da amostra n, para estimar a proporção p de esquerdinos (pelo método de Agresti-Coull), com uma margem de erro (ou erro máximo) inferior a $\varepsilon = 1\%$. Considerar um grau de confiança de 95%.

$$\underbrace{1.96\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}}}_{\text{margem de erro}} \leq 1.96\,\frac{0.5}{\sqrt{n+4}} \leq 0.01$$

$$n+4 \geq \left(\frac{1.96 \cdot 0.5}{0.01}\right)^2 \Rightarrow n \geq 9600$$

Inferência Estatística
Proporções – Planeamento

Escolha do Tamanho da Amostra – **Exemplo 4**

Considerar ainda o caso do exemplo anterior, e supor que se pretende determinar o tamanho da amostra a considerar n, para estimar a proporção p de esquerdinos, pelo método de Agresti-Coull, mas agora com erro padrão $se_{\tilde{P}} < 0.01$.

Neste caso, sem estudo prévio, o menor valor para n pode ser então estimado a partir de

$$se_{\tilde{P}} = \sqrt{\frac{p(1-p)}{n+4}} = \sqrt{\frac{0.5 \times 0.5}{n+4}} \leq 0.01$$

ou seja

$$n + 4 \geq 2500 \quad \Rightarrow \quad n \geq 2496 \text{ estudantes}$$

Proporções

Intervalo de Confiança Unilateral

Inferência Estatística
Proporções – I.C. Unilateral

Os intervalos de confiança que vimos até agora, são da forma:

]estimativa − margem de erro, estimativa + margem de erro[

e designam-se por intervalos **bilaterais**.

Para um grau de confiança (aproximado), $1-\alpha$:

método	estimativa	margem de erro
Wald	$\hat{p}$	$z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
Agresti-Coull (α=0.05)	$\tilde{p}$	$z_{\alpha/2}\sqrt{\frac{\tilde{p}\,(1-\tilde{p})}{n+4}}$

Inferência Estatística
Proporções – I.C. Unilateral

No entanto, é possível construir um intervalo de confiança **unilateral**, apropriado quando nos interessa apenas um limite inferior/superior.

$$]\text{estimativa} - \text{margem de erro}, +\infty[$$

ou

$$]-\infty, \text{estimativa} + \text{margem de erro}[$$

método	estimativa	margem de erro
Wald	$\hat{p}$	$z_{\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
Agresti-Coull (α=0.05)	$\tilde{p}$	$z_{\alpha}\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}}$

Inferência Estatística
Proporções – I.C. Unilateral

Exemplo

Pretende-se estimar a probabilidade de morte p, de um recém nascido submetido a ECMO (ver um exemplo anterior); considerar $\alpha = 0.05$.

Aqui o interesse recai em obter um limite superior para p.

Como $z_{0.05} = 1.65$ e sabendo que $\tilde{p} \approx 0.133$, temos que $s_{\tilde{P}} \approx 0.088$.

Usando o método de Agresti-Coull, o I.C. unilateral, é:

$$]-\infty, \tilde{p} + 1.65\, s_{\tilde{P}}\,[=]-\infty, 0.2782\,[$$

Como $p \geq 0$ considera-se o intervalo $]\,0, 0.2782\,[$

Podemos assim afirmar com confiança de 95% que a probabilidade de morte é, no máximo, 0.2782.

Nota

Como $0 \leq p \leq 1$, considera-se

$$\left]\,\text{estimativa} - \text{margem de erro},\ 1\,\right[$$

e

$$\left]\,0,\ \text{estimativa} + \text{margem de erro}\,\right[$$

em vez de

$$\left]\,\text{estimativa-margem de erro},\ +\infty\,\right[$$

e

$$\left]\,-\infty,\ \text{estimativa} + \text{margem de erro}\,\right[$$

Proporções

I.C. para Diferença de Proporções

Inferência Estatística
Proporções – Intervalo de Confiança

Diferença de Proporções

Por vezes há interesse em construir um intervalo de confiança para a diferença de duas proporções.

Como construir um I.C. para $p_1 - p_2$?

	População 1 (p_1)	População 2 (p_2)
tamanho da amostra	n_1	n_2
nº de elementos da amostra com o atributo de interesse	x_1	x_2

$$\hat{p}_1 = \frac{x_1}{n_1}; \quad \hat{p}_2 = \frac{x_2}{n_2}$$

Inferência Estatística
Proporções – Intervalo de Confiança

Diferença de Proporções

Para o método de Wald, o erro padrão é:

$$se = \sqrt{\frac{\hat{p}_1\left(1-\hat{p}_1\right)}{n_1}+\frac{\hat{p}_2\left(1-\hat{p}_2\right)}{n_2}}$$

e então

$$\hat{p}_1-\hat{p}_2 \pm z_{\alpha/2}\, se$$

é um intervalo de confiança com grau de confiança aproximadamente $1-\alpha$ para a diferença entre duas proporções p_1-p_2, baseado em amostras independentes de tamanhos n_1 e n_2, e proporções amostrais respetivamente $\hat{p}_1$ e $\hat{p}_2$.

Diferença de Proporções

Se pretendermos a proporção de Wilson ajustada, procedemos de forma análoga, isto é:

X_1 é uma amostra de dimensão n_1 a partir da qual se pode calcular uma estimativa para a proporção p_1, que será

$$\tilde{p}_1 = \frac{x_1 + 1}{n_1 + 2}$$

X_2 é uma amostra de dimensão n_2, que para p_2 fornece a estimativa

$$\tilde{p}_2 = \frac{x_2 + 1}{n_2 + 2}$$

Diferença de Proporções

Para o método de Agresti-Coull, o erro padrão será:

$$se_{(\tilde{P}_1-\tilde{P}_2)} = \sqrt{\frac{\tilde{p}_1\left(1-\tilde{p}_1\right)}{n_1+2}+\frac{\tilde{p}_2\left(1-\tilde{p}_2\right)}{n_2+2}}$$

e então o intervalo de confiança para $p_1 - p_2$, com confiança de 95%, será:

$$\left]\,(\tilde{p}_1-\tilde{p}_2)-1.96\cdot se_{(\tilde{P}_1-\tilde{P}_2)},(\tilde{p}_1-\tilde{p}_2)+1.96\cdot se_{(\tilde{P}_1-\tilde{P}_2)}\,\right[$$

Inferência Estatística
Proporções – Intervalo de Confiança

Diferença de Proporções – **Exemplo 1**

Pacientes que sofrem de cefaleia, participaram num ensaio clínico para avaliar uma nova cirurgia.

Um grupo de 75 pacientes foi aleatoriamente designado para receber a nova cirurgia (Grupo 1; $n_1 = 49$), ou uma cirurgia simulada (Grupo 2; $n_2 = 26$).

Sejam x_1 e x_2 o número de pacientes que tiveram melhorias no Grupo 1 e no Grupo 2, respetivamente.

Os dados registados foram $x_1 = 41$ e $x_2 = 15$. Então,

$$\tilde{p}_1 = \frac{41+1}{49+2} = 0.824 \quad \text{e} \quad \tilde{p}_2 = \frac{15+1}{26+2} = 0.571$$

Inferência Estatística
Proporções – Intervalo de Confiança

Diferença de Proporções – **Exemplo 1**

Erro padrão:

$$se_{(\tilde{P}_1 - \tilde{P}_2)} = \sqrt{\frac{0.824 \times 0.176}{49+2} + \frac{0.571 \times 0.429}{26+2}} = 0.1077$$

Assim, o intervalo de confiança a 95% para $p_1 - p_2$, é

$$](0.824 - 0.571) - 1.96 \times 0.1077, (0.824 - 0.571) - 1.96 \times 0.1077[$$

$$\text{I.C.}_{95\%} =]0.042, 0.464[$$

A probabilidade de redução da cefaleia com a nova cirurgia é, com 95% de confiança, entre 0.042 e 0.464 superior à probabilidade de redução da cefaleia com a cirurgia simulada.

Inferência Estatística
Proporções – Intervalo de Confiança

Diferença de Proporções – **Exemplo 2**

Um medicamento foi administrado a 71 pacientes e 30 deles melhoraram.

Um grupo independente de 70 pacientes recebeu um placebo e 20 deles melhoraram.

Sejam p_1 e p_2 as probabilidades de melhora, usando respetivamente o medicamento e o placebo.

Pretende-se contruir um I.C. a 99% para $p_1 - p_2$.

Inferência Estatística
Proporções – Intervalo de Confiança

Diferença de Proporções – **Exemplo 2**

$$\hat{p}_1 = \frac{30}{71}; \quad \hat{p}_2 = \frac{20}{70}$$

$$se = \sqrt{\frac{\hat{p}_1\left(1-\hat{p}_1\right)}{n_1} + \frac{\hat{p}_2\left(1-\hat{p}_2\right)}{n_2}}$$

$$\hat{p}_1 - \hat{p}_2 \approx 0.14; \quad se \approx 0.080$$

$$\text{I.C.}_{(99\%)} \quad \rightarrow \quad \hat{p}_1 - \hat{p}_2 \pm 2.58\, se \approx 0.14 \pm 0.207$$

$$\text{I.C.}_{(99\%)} \quad \rightarrow \quad]-0.067, 0.347[$$

Inferência Estatística

Proporções – Intervalo de Confiança

Utilização do R

(prop.test)

```
# Exemplo1

# i.c. a 99% para proporção a partir de amostra de tamanho 632
com 114 sucessos:

prop.test(114,632,conf.level = 0.99)


# i.c. a 95% para diferença de proporções:

prop.test(c(30,20),c(71,70),conf.level = 0.95)
```

Proporções

Distribuição por Amostragem

Inferência Estatística
Proporções

Retomemos o exemplo das máquinas dispensadoras de refrigerantes contaminadas.

Para estimar a proporção p de máquinas contaminadas, foi selecionada aleatoriamente uma amostra de dimensão 30, e foram detetadas 5 máquinas contaminadas.

Assim, a proporção amostral de máquinas contaminadas é:

$$\hat{p} = \frac{5}{30} = 0.167$$

e este é o valor utilizado como estimativa de p, a proporção populacional de máquinas contaminadas.

Inferência Estatística
Proporções

Dimensão da amostra: $n = 30$

Número de máquinas contaminadas na amostra: $x = 5$

$$\hat{p} = \frac{x}{n} = \frac{5}{30} = 0.167$$

Notar que com outra amostra de tamanho 30 poderíamos observar um número diferente de máquinas contaminadas, e portanto obter uma estimativa também diferente.

Podemos pensar na probabilidade associada ao número de máquinas contaminadas numa amostra de dimensão 30, ou na probabilidade associada a $\hat{P}$, isto é, $P(\hat{P} = \hat{p})$.

Distribuição por Amostragem

Suponhamos que numa determinada região, 17% de todas as máquinas estão contaminadas.

Analisada uma amostra aleatória de 2 máquinas desta população, teremos então 0, 1 ou 2 máquinas contaminadas.

Nº de Máquinas Contaminadas	Probabilidade	Proporção $\tilde{P}$	Probabilidade
0	$(1\text{-}0.17)^2 = 0.6889$	$\tilde{p} = \frac{0+2}{2+4} = 0.33$	$P(\tilde{P} = 0.33) = 0.6689$
1	$2(0.17)(1\text{-}0.17) = 0.2822$	$\tilde{p} = \frac{1+2}{2+4} = 0.50$	$P(\tilde{P} = 0.50) = 0.2822$
2	$(0.17)^2 = 0.0289$	$\tilde{p} = \frac{2+2}{2+4} = 0.67$	$P(\tilde{P} = 0.67) = 0.0289$

Distribuição por Amostragem

Suponhamos agora que foi examinada uma amostra de **20 máquinas**, de uma população em que 17% **estão contaminadas**.

Quantas máquinas contaminadas se podem encontrar na amostra?

Y – representa o número de máquinas contaminadas (em 20)

$Y \sim Bi(20, 0.17)$

Calculemos, por exemplo, P(5 contaminadas e 15 não):

$$P(5 \text{ contaminadas e } 15 \text{ não }) = P(Y = 5)$$

Distribuição por Amostragem

$$P(5 \text{ contaminadas e } 15 \text{ não}) = C_5^{20}(0.17)^5(0.83)^{15}$$

$$P(5 \text{ contaminadas e } 15 \text{ não}) = P(Y = 5) = 0.1345$$

$$\tilde{p} = \frac{5+2}{20+4} = 0.2917$$

$$\therefore \; P(\tilde{P} = 0.2917) = 0.1345$$

A distribuição binomial pode então ser usada para determinar a distribuição por amostragem de $\tilde{P}$

Distribuição por Amostragem

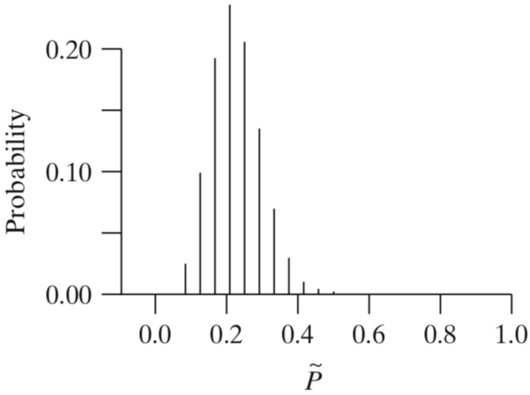
Distribuição amostral de Y (número de sucessos), e de $\widetilde{P}$ (proporção de sucessos pelo método de Wilson ajustado), para $n = 20$ e $p = 0.17$

Y	$\widetilde{P}$	Probability	Y	$\widetilde{P}$	Probability
0	0.0833	0.0241	11	0.5417	0.0001
1	0.1250	0.0986	12	0.5833	0.0000
2	0.1667	0.1919	13	0.6250	0.0000
3	0.2083	0.2358	14	0.6667	0.0000
4	0.2500	0.2053	15	0.7083	0.0000
5	0.2917	0.1345	16	0.7500	0.0000
6	0.3333	0.0689	17	0.7917	0.0000
7	0.3750	0.0282	18	0.8333	0.0000
8	0.4167	0.0094	19	0.8750	0.0000
9	0.4583	0.0026	20	0.9167	0.0000
10	0.5000	0.0006			

Inferência Estatística
Proporções

Distribuição por Amostragem

Distribuição amostral de $\widetilde{P}$ para $n = 20$ e $p = 0.17$



Inferência Estatística
Proporções

Distribuição por Amostragem – **Intervalo de Confiança**

Na inferência estatística de uma amostra para a população, é aceitável utilizar $\widetilde{p}$ como estimativa de p.

Para além disso podemos usar a distribuição por amostragem para avaliar o erro de amostragem na estimativa.

Voltemos ao exemplo anterior: amostra de 20 máquinas, de uma população com 17% contaminadas.

Inferência Estatística
Proporções

Distribuição por Amostragem – **Intervalo de Confiança**

Usando a tabela anterior, da probabilidade de amostragem, podemos escrever:

$$P(0.12 \leq \tilde{P} \leq 0.22) = 0.0986 + 0.1919 + 0.2358 \approx 0.53$$

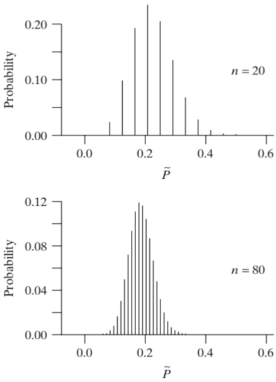
Sendo $p = 0.17$, para a amostra de tamanho 20 temos então que

$$P(\tilde{P} \in [p - 0.05, p + 0.05]) \approx 0.53$$

Inferência Estatística
Proporções

Distribuição por Amostragem – **Intervalo de Confiança**

A distribuição amostral de Y e consequentemente a de $\widetilde{P}$ dependem da dimensão da amostra. Exemplo das máquinas de refrigerantes.



A distribuição por amostragem de $\tilde{P}$, para $n = 20, 40$ e 80

Inferência Estatística
Proporções

Distribuição por Amostragem – **Intervalo de Confiança**

A distribuição amostral de $\widetilde{P}$ depende da dimensão da amostra.

n	$\Pr\{0.12 \leq \widetilde{P} \leq 0.22\}$
20	0.53
40	0.56
80	0.75
400	0.99

Assim, para uma amostra de tamanho 80, por exemplo, temos:

$$P\left(\widetilde{P} \in [p - 0.05, p + 0.05]\right) = 0.75$$

Proporções

Teste de Hipóteses

Testes para a Proporção de uma População

Definimos

$$H_0 : p = p_0$$
$$H_1 : p \neq p_0$$

Com o estimador $\widehat{P} = \frac{X}{n}$ para a proporção p, utilizamos a estatística de teste seguinte:

$$Z_0 = \frac{\hat{P} - p_0}{\sqrt{p_0 (1 - p_0) / n}} = \frac{X - np_0}{\sqrt{np_0 (1 - p_0)}}$$

e aceitamos H_0 se z_0 está na região de aceitação, definida por:

$$\left[-z_{\alpha/2}, z_{\alpha/2}\right]$$