Gabarito - Lista 6

Inferência: Estimação Pontual e Intervalar

Gabarito

Exercício 1 [0,5 ponto]. Encontre os valores de z associados aos seguintes níveis de confiança:

(a) 50%

(b) 93%

(c) 97%

(d) 98%

(e) 99,73%

(a) $z = 0,675$

(b) $z = 1,81$

(c) $z = 2,17$

(d) $z = 2,33$

(e) $z = 3$

Exercício 2 [0,5 ponto]. Encontre os seguintes valores de t:

(a) Nível de confiança de 80%, com 14 graus de liberdade

(b) Nível de confiança de 95%, com 5 graus de liberdade

(c) Nível de confiança de 99,8%, com 9 observações

(d) Nível de confiança de 99%, com 36 observações

(e) Nível de confiança de 90%, com 250 observações

(a) $t = 1,345$

(b) $t = 2,571$

(c) $t = 4,501$

(d) $t = 2,727$

(e) $t = 1,645$

Exercício 3 [2 pontos]. Descreva brevemente a diferença entre os conceitos abaixo:

(a) **Estimativa pontual** vs **estimativa intervalar**

(b) **Erro-padrão** vs **margem de erro**

(c) **Intervalo de confiança** vs **nível de confiança**

(d) **Distribuição z** vs **distribuição t**

(a) A **estimativa pontual** é um valor único que representa o melhor julgamento que podemos fazer sobre o valor verdadeiro do parâmetro, enquanto a **estimativa intervalar** é um conjunto de valores que acreditamos conter o valor verdadeiro do parâmetro.

(b) O **erro-padrão** é o desvio médio da distribuição amostral em torno do valor verdadeiro do parâmetro, enquanto a **margem de erro** é um múltiplo do erro-padrão utilizado para padronizá-lo de acordo com o nível de confiança definido pelo pesquisador.

(c) O **intervalo de confiança** é um conjunto de valores que acreditamos conter o valor verdadeiro do parâmetro com determinada probabilidade. Essa probabilidade, definida pelo pesquisador antes da análise dos dados, é denominada **nível de confiança**.

(d) Tanto a **distribuição z** quanto a **distribuição t** são distribuições padronizadas utilizadas para calcularmos determinadas probabilidades. Ambas têm formato muito parecido, e a diferença mais importante é que a distribuição t tem, em geral, caudas mais pesadas do que a distribuição z. À medida em que aumentamos os graus de liberdade, porém, a distribuição t se aproxima da z.

Exercício 4 [2 pontos]. Uma pesquisadora interessada em investigar as condições de vida das mulheres em diversos países produziu a tabela abaixo com uma amostra de 102 países, extraída a partir da variável *Average female life expectancy* no banco de dados *world95.dta*

```
. mean lifeexpf

Mean estimation                   Number of obs    =       102
```

	Mean	Std. Err.	[95% Conf.	Interval]
lifeexpf	69.92157	1.070663	67.79766	72.04548

(a) Para calcular o intervalo de confiança dessa variável, você utilizaria a tabela z ou t? Justifique.

(b) Demonstre, **passo a passo**, como utilizar as informações da tabela para reproduzir o intervalo de confiança obtido (67.79766;72.04548). Para facilitar o cálculo, arredonde todos os valores para duas casas decimais (seu intervalo não será exatamente igual ao da tabela, mas será próximo).

(c) Se tivéssemos dados apenas para 50 países, você espera que o intervalo de confiança aumentaria, diminuiria, ou permaneceria o mesmo? E o nível de confiança?

(a) Geralmente utilizamos a tabela t para calcular o intervalo de confiança para médias. Porém, quanto mais graus de liberdade temos na t, mais nos aproximamos dos valores da tabela z. Com um n de 102 observações, os intervalos de confiança produzidos pelas tabelas z e t são bastante semelhantes, de modo que a diferença entre eles é desprezível.

(b) Para calcular o intervalo de confiança, vamos utilizar a tabela z (respostas utilizando a t também serão aceitas). Portanto,

$$IC = \bar{y} \pm z \overbrace{\frac{s}{\sqrt{n}}}^{\text{e.p.}}$$

$$IC = 69,92 \pm 1,96 \times 1,07$$

$$IC = 69,92 \pm 2,1$$

$$(67,82; 72,02)$$

(c) Com um numero mais baixo de observações, o intervalo de confiança irá aumentar para refletir o fato de que temos menos informações. O nível de confiança, por sua vez, é definido pelo pesquisador, e não sofre influência direta da queda no n.

Exercício 5 [2 pontos]. Ainda se referindo à tabela do exercício 4, a pesquisadora redigiu algumas interpretações para os resultados, e pediu sua ajuda para avaliá-las. Indique a(s) interpretação(ões) corretas, e aponte os erros das demais.

(a) “Com esses resultados, concluímos que a expectativa de vida feminina média em 1995 era de 69,9 anos.”

(b) “Os resultados não nos permitem concluir com certeza qual é a expectativa de vida média das mulheres em 1995, mas podemos dizer que ela não é menor do que 67,8 anos nem maior do que 72 anos.”

(c) “A partir dessa amostra, podemos inferir que um país sorteado aleatoriamente teria, em 1995, 95% de probabilidade de ter expectativa de vida feminina entre 67,8 e 72 anos.”

(d) “Os resultados indicam que, em 1995, 95% dos países tinham expectativa de vida feminina entre 67,8 e 72 anos.”

(e) “A partir desses resultados, podemos dizer com 95% de confiança que, em 1995, a expectativa de vida média das mulheres em cada país estava entre 67,8 e 72 anos.”

(a) **Incorreto**. Como não coletamos dados para toda a população, precisamos comunicar a incerteza que temos sobre nossa estimativa. 69,9 é a estimativa pontual para a expectativa média de vida das mulheres, não o valor verdadeiro do parâmetro.

(b) **Incorreto**. Assim como na alternativa anterior, essa frase apenas comunica nossa estimativa (desta vez intervalar) para a expectativa média de vida das mulheres, sem comunicar a incerteza.

(c) **Incorreto**. Ao calcularmos um intervalo de confiança, não estamos falando da probabilidade de o parâmetro estar dentro do intervalo encontrado. A interpretação probabilística pode ser feita no âmbito da inferência bayesiana, um ramo da estatística que está fora do escopo deste curso.

(d) **Incorreto**. A interpretação desta alternativa só faria sentido em um contexto de análise descritiva, na qual teríamos as informações de expectativa média de vida das mulheres para todos os países

(e) **Correto, porém...** Esta é a alternativa mais próxima da interpretação correta, e as respostas que assim a julgarem serão aceitas. Porém, vale fazer uma observação sobre um aspecto sutil de interpretação que não está claramente exposto em Agresti e Finlay. Quando especificamos os valores do intervalo, condicionamos nosso resultado a eles. Tecnicamente, a leitura mais correta seria “A partir desses resultados, podemos dizer com 95% de confiança que, em 1995, a expectativa de vida média das mulheres em cada país *se encontra entre os limites inferior e superior do intervalo de confiança*.” Trata-se de uma diferença aparentemente pequena, mas as consequências são consideráveis. Como o livro não cobre bem esse aspecto, não vamos insistir nesse detalhe, mas agora você sabe que ele existe. Se você quiser ler um debate mais aprofundado sobre as nuances de interpretação de intervalos de confiança (é claro que quer), confira o comentário do estatístico e cientista político Andrew Gelman neste link.

Exercício 6 [2 pontos]. Na última pesquisa do Datafolha para as eleições presidenciais de 2018, Lula lidera um dos cenários com 30% das intenções de voto. Você pode acessar o relatório completo da pesquisa neste link.

(a) Na página 8 do relatório, o Datafolha afirma que "foram realizadas 4194 entrevistas presenciais em 227 municípios, com margem de erro máxima 2 pontos percentuais para mais ou para menos considerando um nível de confiança de 95%. Isto significa que se fossem realizados 100 levantamentos com a mesma metodologia, em 95 os resultados estariam dentro da margem de erro prevista." Você concorda com essa afirmação? Justifique.

(b) Calcule qual seria o n mínimo de uma pesquisa com as mesmas características apresentadas pelo Datafolha.

(c) Aponte ao menos um motivo que explique a divergência entre o n calculado e o n realizado pelo instituto.

(d) Se o Datafolha, por motivos orçamentários, pudesse entrevistar apenas 1200 pessoas, qual seria a margem de erro da pesquisa a um nível de confiança de 95%?

(e) Imagine que você não tivesse informação alguma sobre a proporção de pessoas que pretendem votar no Lula, e como consequência que você não tenha um palpite razoável para π (essa é a situação mais comum em pesquisa). Identifique qual é a abordagem "segura" proposta por Agresti nesses casos, e recalcule o n da pesquisa usando essa abordagem.

(a) A leitura apresentada no relatório está correta: na inferência clássica, as conclusões que podemos tirar a partir do intervalo de confiança dizem respeito às diferentes amostras que poderíamos ter feito e não fizemos. O instituto acerta, inclusive, no detalhe sutil abordado na resposta do exercício 5*e*.

(b) Para calcular o número de observações necessárias para estimar uma proporção, utilizamos a seguinte fórmula:

$$n = \pi(1-\pi)(\frac{z}{M})^2$$

Como temos a informação da proporção de pessoas que declaram voto no Lula, definiremos $\pi = 0,3$

$$n = 0,3(0,7)(\frac{1,96}{0,02})^2$$

$$n = 0,21 \times 9604$$

$$n = 2017$$

(c) Um dos motivos pelos quais os institutos de pesquisa geralmente entrevistam mais pessoas do que o mínimo necessário é a demanda por dados desagregados em subgrupos socioeconômicos. Vale notar também que os institutos de pesquisa não fazem amostras puramente aleatórias, de modo que o cálculo do erro-padrão pode mudar de acordo com a técnica de amostragem utilizada.

(d) Sabemos que

$$M = z\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$

Então basta substituir os valores

$$M = 1,96\sqrt{\frac{0,3(0,7)}{1200}}$$

$$M = 1,96\sqrt{0,000175}$$

$$M = 1,96 \times 0,0132$$

$$M = 0,026$$

Com 1200 observações e nível de confiança de 95%, a margem de erro seria de ± 2,6 pontos percentuais

(e) Quando não temos uma avaliação bem informada do valor verdadeiro da proporção, Agresti indica fixar $\pi = 0,5$. Assim, o novo cálculo do n seria

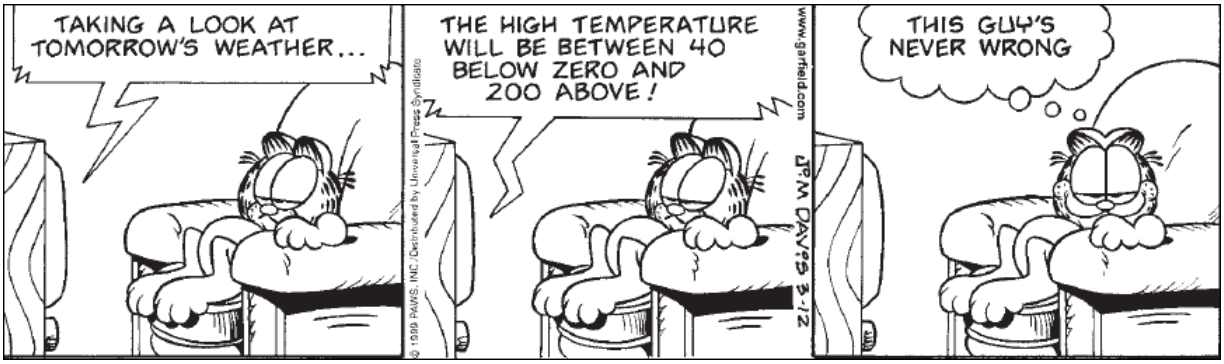
$$n = \pi(1-\pi)(\frac{z}{M})^2$$

$$n = 0,5(0,5)(\frac{1,96}{0,02})^2$$

$$n = 0,25 \times 9604$$

$$n = 2401$$

Exercício 7 [1 ponto]. "Todo cálculo de intervalo de confiança depende de um equilíbrio entre incerteza e precisão". Aponte quais são os elementos do intervalo de confiança que exprimem a tensão entre incerteza e precisão, e explique como essa tensão é ilustrada no quadrinho abaixo.



Ao calcularmos o intervalo de confiança, a incerteza é definida por meio do nível de confiança, enquanto a precisão é definida pelo tamanho da nossa margem de erro. Idealmente, gostaríamos de maximizar nosso nível de confiança ao mesmo tempo em que minimizamos nossa margem de erro, de maneira a termos muita certeza de que o valor verdadeiro do parâmetro está contido em um intervalo pequeno de valores.

Na prática, porém, quanto mais aumentamos nosso nível de confiança, mais aumentamos nossa margem de erro. O quadrinho ilustra uma situação extrema dessa ideia, na medida em que mostra uma previsão que acerta 100% das vezes ("this guy is never wrong"), mas com um intervalo muito pouco preciso e, por isso, absolutamente inútil. Remetendo ao exercício 6, poderíamos dizer com 100% de confiança que Lula tem entre 0 e 100% das intenções de votos, mas isso não nos ajudaria muito. O desafio na construção de um intervalo de confiança não é ter o maior nível de confiança possível, e sim chegar a um intervalo que seja substantivamente útil.