Gabarito - Lista 5

Distribuições e Probabilidade

Gabarito

Exercício 1 [2 pontos]. Calcule as probabilidades abaixo, baseadas nos dados na pesquisa *O Brasil, as Américas e o Mundo*. Utilize os conceitos de probabilidade apresentados no capítulo 4 de AF, e demonstre claramente como você chegou às suas respostas. Não serão aceitas respostas que apenas mostrarem resultados do Stata (embora você possa, se preferir, utilizá-lo para confirmar se suas respostas estão certas).

a Dentre os 1889 entrevistados, 80 simpatizantes do PSDB dizem que o Brasil deveria procurar ser um líder na América Latina. Qual é a probabilidade de um entrevistado escolhido aleatoriamente ser simpatizante do PSDB e considerar que o Brasil deve ser um líder na América Latina?

b Na lista 3, vimos que 38,64% dos entrevistados na pesquisa são simpatizantes do PT e que, entre os simpatizantes do PT, 51,10% consideram que o Brasil deveria procurar ser um líder na América Latina. Qual é a probabilidade de um entrevistado escolhido aleatoriamente ser simpatizante do PT e considerar que o Brasil deve ser um líder na América Latina?

a

$$\frac{80}{1889} = 0,0423 \Rightarrow 4,23\%$$

b Neste exercício deve ser aplicada a terceira regra básica de probabilidade apresentada por Agresti: $P(A \text{ e } B) = P(A) \times P(B|A)$

$$0,3864 \times 0,511 = 0,1974 \Rightarrow 19,74\%$$

Exercício 2 [2 pontos]. Considere que $P(x)$ é a probabilidade de um brasileiro escolhido aleatoriamente falar fluentemente x idiomas, e que a distribuição dessa probabilidade é $P(x = 0) = 0,001$, $P(x = 1) = 0,87$, $P(x = 2) = 0,114$ e $P(x \geqslant 3) = 0,015$.

a Você classificaria essa variável como contínua, categórica ordinal ou categórica nominal? Justifique.

b Qual é a probabilidade de escolhermos uma pessoa aleatoriamente e ela não ser poliglota?

c Não é possível encontrar a média exata dessa variável, pois a última categoria é aberta. Calcule o valor mínimo que a média pode assumir.

a A variável pode ser considerada categórica ordinal, por assumir poucos valores diferentes e possuir hierarquia clara entre as respostas. Outra abordagem seria considerá-la uma variável contínua (ou, na terminologia de Agresti, uma variável quantitativa discreta), já que a variável não separa as observações em grupos.

b A resposta vai depender da definição adotada para poliglota, mas assumindo um valor de $x \geqslant 3$, a probabilidade é de $0,001 + 0,87 + 0,114 = 0,985 \Rightarrow 98,5\%$

c Para calcular o valor mínimo que a média pode assumir, basta impor o menor valor possível na categoria aberta: 3. Assim, o valor mínimo da média é de $0 \times 0,001 + 1 \times 0,87 + 2 \times 0,114 + 3 \times 0,015 = 1,14$

Exercício 3 [1 ponto]. TOEFL e IELTS são dois dos testes de proficiências de inglês mais conhecidos e utilizados pelas universidades. Enquanto o TOEFL varia de uma escala de 0 a 120, o IELTS varia de uma escala de 1 a 9. No IRI, a nota mínima para admissão na pós-graduação é de 72 pontos para o TOEFL e de 6 pontos para o IELTS.

Suponha que a distribuição das notas obtidas nesses testes é normal, com os parâmetros TOEFL ($\mu = 82,5$, $\sigma = 19,3$) e IELTS ($\mu = 6,1$, $\sigma = 1,3$). A nota mínima exigida pelo IRI nas duas provas é semelhante, em termos relativos? Justifique.

Para que as notas do TOEFL e do IELTS sejam comparáveis, precisamos convertê-las a uma distribuição padronizada: z.

$$z_{TOEFL} = \frac{y - \mu}{\sigma} = \frac{72 - 82,5}{19,3} = -0,54$$

$$z_{IELTS} = \frac{y - \mu}{\sigma} = \frac{6 - 6,1}{1,3} = -0,08$$

A nota exigida pelo IRI para o IELTS está um pouco mais próxima da média do exame do que a nota exigida para o TOEFL. Em termos práticos, porém, ambas estão a menos de um desvio-padrão da média, de maneira que não podemos dizer que o critério em uma prova é substancialmente maior do que em outra.

Exercício 4 [2 pontos]. Marque as assertivas como verdadeiras (V) ou falsas (F). Para as frases falsas, aponte o erro, proponha uma correção e justifique.

V | F Uma consequência do Teorema do Limite Central é que quanto mais observações temos, mais a distribuição da nossa amostra se parece com uma Normal.

V | F A forma da distribuição amostral se assemelha à da distribuição populacional à medida que aumentamos o número de observações.

V | F O desvio-padrão é uma medida de variação da distribuição dos dados amostrais, enquanto o erro-padrão é uma medida de variação da distribuição amostral

V | F Podemos pensar no erro-padrão como o desvio-padrão das médias amostrais. Cada amostra tem um desvio em relação à média populacional devido a problemas como o viés de seleção, perguntas mal formuladas ou erros no desenho de pesquisa.

V | F Uma consequência do Teorema do Limite Central é que quanto mais observações temos, *mais a distribuição da nossa amostra se parece com uma Normal*.

Correção: O Teorema do Limite Central nos diz que, quanto mais amostras do mesmo tamanho temos, mais a distribuição das suas médias (ou outra estatística) se aproximam da Normal.

V | F A forma da *distribuição amostral* se assemelha à da distribuição populacional à medida que aumentamos o número de observações.

Correção: A distribuição amostral diz respeito à distribuição das estatísticas quando produzimos diferentes amostras. De acordo com o Teorema do Limite Central, independentemente da distribuição populacional, a distribuição amostral convergirá para a Normal à medida em que aumentamos o número de amostras.

V | F O desvio-padrão é uma medida de variação da distribuição dos dados amostrais, enquanto o erro-padrão é uma medida de variação da distribuição amostral

V | F Podemos pensar no erro-padrão como o desvio-padrão das médias amostrais. Cada amostra tem um desvio em relação à média populacional *devido a problemas como o viés de seleção, perguntas mal formuladas ou erros no desenho de pesquisa*.

Correção: O erro-padrão está associado ao fato de não observarmos o verdadeiro valor do parâmetro. Em outras palavras, mesmo que o desenho de pesquisa respeite todos os pressupostos da inferência estatística, esperamos que o erro-padrão seja maior do que 0 simplesmente pelo fato de não termos observado os valores reais de toda a população.

Exercício 5 [1 ponto]. A figura abaixo mostra as distribuições dos valores médios que encontramos quando jogamos 3, 5 e 20 dados 10000 vezes.

a Os gráficos se referem à distribuição da amostra ou à distribuição amostral? Justifique.

b Explique como essas figuras se relacionam com a Lei dos Grandes Números e com o Teorema do Limite Central.

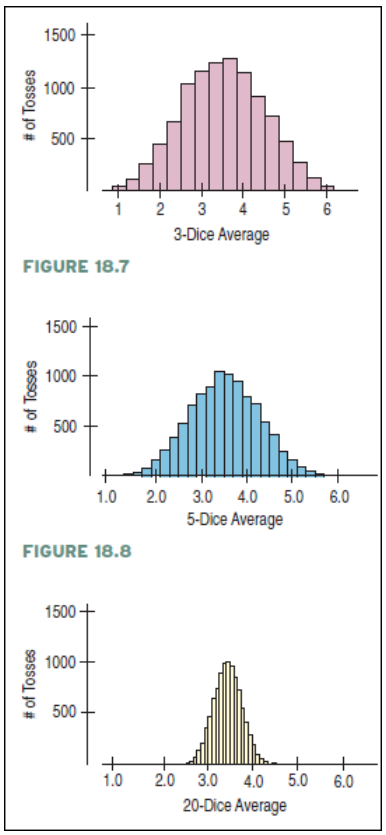

FIGURE 18.7

FIGURE 18.8

a Os gráficos apresentam a distribuição dos valores médios dos dados quando os lançamos 10000 vezes. Trata-se, portanto, de uma distribuição amostral, e não de uma distribuição das observações.

b O Teorema do Limite Central está ilustrado pela forma da distribuição das médias. Como repetimos as jogadas 10000 vezes, esperamos que as médias se distribuam em forma de sino *independentemente de jogarmos 3, 5 ou 20 dados*.

A Lei dos Grandes Números, por sua vez, está ilustrada pela convergência das médias em torno do Valor Esperado de 3,5. Quando jogamos 3 dados, a distribuição das médias é razoavelmente dispersa, assumindo todos os valores possíves (1 até 6). A medida em que aumentamos o número de dados que jogamos, *e mantendo fixa a repetição de 10000 vezes*, observamos que o erro-padrão diminui, e as médias estão mais concentradas em torno do Valor Esperado, com variação entre $\approx 2,5$ e $\approx 4,5$.

Exercício 6 [2 pontos]. As imagens abaixo mostram a média, o desvio-padrão e o erro-padrão da variável *literacy*, diisponível no banco de dados *world95.dta*. Na primeira imagem temos todos os países do banco, enquanto as demais mostram as mesmas estatísticas para sub-amostras de 49 e 9 países, respectivamente. Qual é a explicação das variações observadas na média, no desvio-padrão e no erro-padrão?

```
. summarize literacy

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    literacy |        100        77.8    23.40638        18        100

. mean literacy

Mean estimation                   Number of obs    =       100

--------------------------------------------------------------
             |       Mean   Std. Err.      [95% Conf. Interval]
-------------+------------------------------------------------
    literacy |       77.8   2.340638       73.15567    82.44433
--------------------------------------------------------------
```

```
. summarize literacy

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    literacy |         49     74.4898    24.57228        18        100

. mean literacy

Mean estimation                   Number of obs    =        49

--------------------------------------------------------------
             |       Mean   Std. Err.      [95% Conf. Interval]
-------------+------------------------------------------------
    literacy |    74.4898   3.510325       67.43181    81.54778
--------------------------------------------------------------
```

```
. summarize literacy

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    literacy |          9    68.55556    26.40128        18         99

. mean literacy

Mean estimation                   Number of obs    =         9

--------------------------------------------------------------
             |       Mean   Std. Err.      [95% Conf. Interval]
-------------+------------------------------------------------
    literacy |    68.55556   8.800428       48.26173    88.84938
--------------------------------------------------------------
```

Podemos observar que, a medida em que cai o número de observações na amostra, o erro-padrão aumenta sistematicamente. Essa é uma consequência direta da Lei dos Grandes Números: quanto mais observações na nossa amostra, maior será a convergência da estimativa pontual ao valor real do parâmetro. A média e o desvio-padrão, por sua vez, variam de maneira não-sistemática em torno do valor real do parâmetro.