

Universidade de São Paulo / Instituto de Relações Internacionais

Análise Quantitativa e Métodos Empíricos com Aplicações em Política Comparada e Relações Internacionais

PRI-5003

1º semestre / 2016

Avaliação Intermediária
(20 de Maio de 2016)

GABARITO

Atenção: esta avaliação é individual e sem consulta. Todas as fórmulas necessárias foram fornecidas. Todas as tabelas necessárias também foram fornecidas.

Peso: 40% da Nota Final

Instruções:

Esta avaliação consiste em 8 questões para serem resolvidas em um período de 4 horas.

Está autorizado o uso de aparelho celular para a utilização de calculadora. Mas está proibido o uso do celular para comunicação entre os alunos, bem como para consulta à internet ou a outros materiais de referência. Para evitar inconvenientes e/ou suspeitas, bem como distúrbios externos, recomendamos colocar o celular em “modo avião”.

Utilize as duas últimas folhas para rascunho. Caso necessite de mais folhas, peça aos monitores.

Responda às perguntas conceituais de maneira clara, completa e concisa. Para os cálculos, demonstre o seu raciocínio, não forneça apenas os resultados das contas.

Fórmulas

Estatística Descritiva	Amostral	Populacional
Média	$\bar{Y} = \sum_{i=1}^n \frac{y_i}{n}$	$\mu = \sum_{i=1}^N \frac{y_i}{N}$
Variância	$s^2 = \sum_{i=1}^n \frac{(y_i - \bar{Y})^2}{n - 1}$	$\sigma^2 = \sum_{i=1}^N \frac{(y_i - \mu)^2}{N}$
Desvio-Padrão	$s = \sqrt{s^2} = \sqrt{\sum_{i=1}^n \frac{(y_i - \bar{Y})^2}{n - 1}}$	$\sigma = \sqrt{\sigma^2} = \sqrt{\sum_{i=1}^N \frac{(y_i - \mu)^2}{N}}$

Estatística Inferencial	Médias	Proporções
Estimativa Pontual	\bar{Y}	\hat{p} ou $\hat{\pi}$
Erro-Padrão	$ep = \sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}$	$ep = \sigma_{\bar{Y}} = \sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n}}$

Tamanho Amostral	Médias	Proporções
n	$n = \sigma^2 \times \left(\frac{Z}{M}\right)^2$	$n = \pi(1 - \pi) \times \left(\frac{Z}{M}\right)^2$

Distribuições	Normal	t-Student
Notação	$\mathcal{N}(\mu, \sigma^2)$	t_{α}
Graus de Liberdade	-----	$gl = n - 1$

Teste de Hipóteses	Z-teste	t-teste
Estatística-Teste	$Z = \frac{(\bar{Y} - \mu)}{ep}$	$t = \frac{(\bar{Y} - \mu)}{ep}$

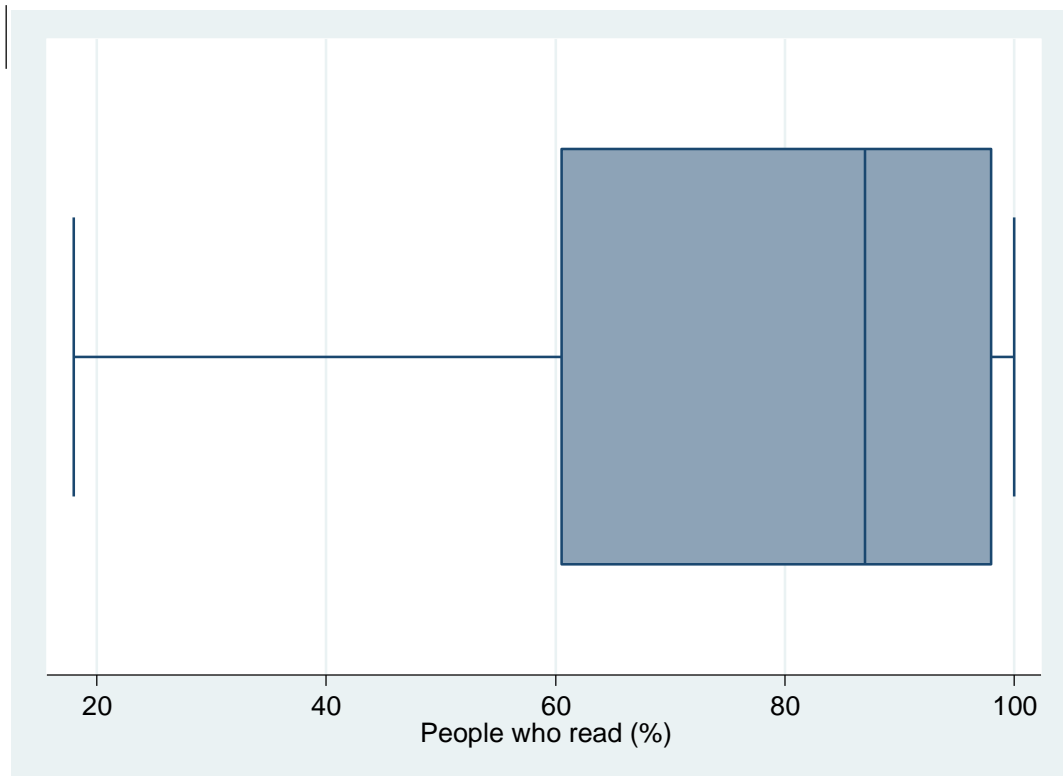
Exercício 1 [0,5 ponto] Uma variável independente (VI) e uma variável interveniente (VInt) estabelecem relações distintas com a variável dependente (VD) na construção da hipótese de pesquisa. Explique:

Cf. Comentários na prova

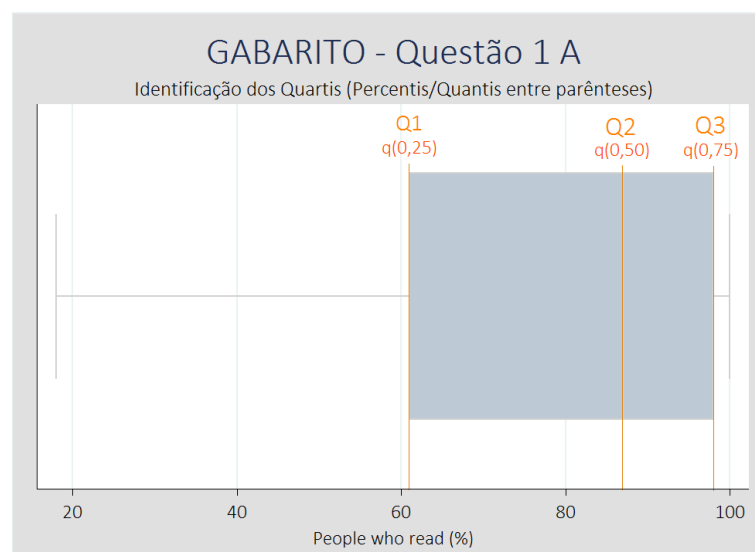
Exercício 2 [0,5 ponto] O fator explicativo-chave distingue os dois desenhos clássicos de pesquisa: MSSD e MDSD. Enquanto para um desenho o fator explicativo-chave funciona como elemento de unificação da variável dependente, para o outro desenho este fator funciona como elemento de discriminação. Explique:

Cf. Comentários na prova

Exercício 3 [1,0 ponto] A imagem abaixo mostra o *boxplot* correspondente às taxas de alfabetização de 102 países, de acordo com os dados disponíveis no banco *world95.dta*. Com base nele, responda as questões abaixo:



- a) Identifique as regiões do gráfico que representam os valores do 1º quartil, 2º quartil e 3º quartil.



- b) A mediana dessa variável é 87%, e um pesquisador te diz que o gráfico demonstra que a maioria dos países têm taxas de alfabetização menores do que essa. Você concorda com essa afirmação? Justifique.

A afirmação é incorreta. A mediana é a medida de tendência central que divide a quantidade de observações exatamente ao meio. Isso significa que 50% dos países estão abaixo dos 87% e os demais 50% dos países estão acima dos 87%. O objetivo do exercício é verificar se o aluno compreende e interpreta corretamente um gráfico de *box-plot*. A área do *box-plot* não indica a quantidade de observações, mas apenas como se distribuem.

- c) Julgando a distribuição dos dados, você diria que essa variável é simétrica, assimétrica à direita ou assimétrica à esquerda? Justifique.

Essa variável é assimétrica à esquerda. Pelo *box-plot* é possível perceber pela posição da mediana (Q1 ou $q[0,50]$), assim como dos demais quartis e limites (superior e inferior), a maneira como se distribuem os dados dessa variável. Caso fosse uma variável com distribuição simétrica, a mediana estaria bem ao centro da “caixa” (*box*). Mas, como a posição da mediana encontra-se mais próxima ao 3º quartil (e limite superior), isso significa que os dados acima da mediana são mais concentrados. Vale dizer, os 50% dos países acima da mediana encontram-se muito mais dispersos do que os 50% dos países acima da mediana.

O exercício também avaliou se o aluno compreendeu os conceitos de simetria da distribuição, assimetria à direita e assimetria à esquerda.

Exercício 4 [1,0 ponto] A tabela abaixo mostra o número de pessoas que migraram de países sul-americanos para os Estados Unidos em 2014:

Table 1. Distribution of South American Immigrants by Country of Origin, 2014		
Region and Country	Number of Immigrants	Percent (%)
South America	2,856,000	100.0
Colombia	707,000	24.8
Peru	449,000	15.7
Ecuador	424,000	14.8
Brazil	336,000	11.8
Guyana	273,000	9.6
Venezuela	216,000	7.6
Argentina	185,000	6.5
Chile	94,000	3.3
Bolivia	81,000	2.8
Uruguay	53,000	1.8

Tabela 1. Fonte: Migration Policy Institute (MPI), 2014.

Calcule as seguintes estatísticas com base nos dados fornecidos:

a) Média

$$\begin{aligned}
 \bar{Y} &= \sum_{i=1}^n \frac{y_i}{n} = \left(\frac{1}{n}\right) \times \left(\sum_{i=1}^n y_i\right) = \\
 &= \left(\frac{1}{n}\right) \times (707.000 + 449.000 + 424.000 + 336.000 + 273.000 + 216.000 + 185.000 + 94.000 + 81.000 + 53.000) = \\
 &= \left(\frac{1}{10}\right) \times (2.818.000) = \frac{2.818.000}{10} = 281.800
 \end{aligned}$$

b) Mediana

Como os dados já estão em ordem, basta achar a posição central. Como n é par, deve-se calcular a média aritmética dos valores das posições do meio, isto é, a média entre o valor da Guiana e o valor da Venezuela:

$$Med(Y) = \frac{273.000 + 216.000}{2} = 244.500$$

Para melhor compreender todos os quartis, sugerimos o seguinte esquema:

País	Quartil	IIQ
Colombia (707.000)		
Peru (449.000)		
Ecuador (424.000)	← Q3	← ← ← ←
Brazil (336.000)		↑
Guyana (273.000)		↑
	← Q2 (Mediana)	IIQ = Q3 - Q1
Venezuela (216.000)		↓
Argentina (185.000)		↓
Chile (94.000)	← Q1	← ← ← ←
Bolivia (81.000)		
Uruguay (53.000)		

c) Primeiro quartil

Para o primeiro quartil, deve-se encontrar a posição central entre a mediana e a primeira posição.

$$Q1(Y) = 94.000$$

d) Terceiro quartil

Para o primeiro quartil, deve-se encontrar a posição central entre a mediana e a última posição.

$$Q3(Y) = 424.000$$

e) Intervalo interquartil

O Intervalo Interquartil (IIQ) é calculado como a diferença entre o Q3 e o Q1:

$$IIQ = Q3 - Q1 = 424.000 - 94.000 = 330.000$$

f) Desvio-padrão

Para facilitar os cálculos de desvio-padrão, sugerimos a seguinte tabela:

	Y_i	$(Y_i - \bar{Y})$	$(Y_i - \bar{Y})^2$	$\left(\frac{1}{n-1}\right) \times \sum_{i=1}^n (Y_i - \bar{Y})^2$	$\sqrt{\left(\frac{1}{n-1}\right) \times \sum_{i=1}^n (Y_i - \bar{Y})^2}$
Colombia	707.000	425.200	180.795.040.000	$\left(\frac{1}{10-1}\right) \times$ (373.625.600.000)	$\sqrt{41.513.955.556}$
Peru	449.000	167.200	27.955.840.000		
Ecuador	424.000	142.200	20.220.840.000		
Brazil	336.000	54.200	2.937.640.000		
Guyana	273.000	-8.800	77.440.000		
Venezuela	216.000	-65.800	4.329.640.000		
Argentina	185.000	-96.800	9.370.240.000		
Chile	94.000	-187.800	35.268.840.000		
Bolívia	81.000	-200.800	40.320.640.000		
Uruguay	53.000	-228.800	52.349.440.000		
TOTAL	2.818.000	-----	373.625.600.000	41.513.955.556	203.749,74
n	10	-----	-----	-----	-----
Média (\bar{Y})	281.800	-----	-----	-----	-----

g) Intervalo de confiança de 95% para a média

Para o cálculo do Intervalo de Confiança, necessitamos das seguintes informações:

- Estimativa Pontual: Média de $\bar{Y} = 281.800$
- Margem de Erro:
 - Escore-t (*t-score*)
 - Nível de Confiança de 95% = Nível de Significância de 5%
 - Distribuição *t-Student* bicaudal ($t_{0,025}$)
 - Graus de Liberdade: $g.l. = n - 1 = 10 - 1 = 9$
 - Pela tabela: *t-score* = 2,2621572
 - Erro Padrão
 - $\sigma_{\bar{Y}} = ep = \frac{s}{\sqrt{n}} = \frac{203.749,74}{\sqrt{10}} = 64.431,32$
 - M = Margem de Erro = $t_{\alpha/2} \times ep = 2,2621572 \times 64.431,32$

Limite Inferior	Estimativa Pontual	Limite Superior
$\bar{Y} - t_{\alpha/2} \times ep$	$\bar{Y} = 281.800$	$\bar{Y} + t_{\alpha/2} \times ep$
$281.800 - 2,2621572 \times 64.431,32$		$281.800 + 2,2621572 \times 64.431,32$
136.046,22		427.553,78

Intervalo de Confiança de 95%: $IC = [136.046,22 ; 427.553,78]$

Exercício 5 [2 pontos] Julgue cada afirmativa abaixo como verdadeira ou falsa. Para as alternativas que julgar como falsa, aponte o erro, proponha uma correção e justifique.

- a) [V | F] A estatística descritiva é uma forma de resumir os dados que observamos em uma amostra ou população. Como não a utilizamos para tirar conclusões a respeito de dados que não observamos, podemos falar dela sem comunicar incerteza.

[V] A estatística descritiva é uma forma de resumir os dados que observamos em uma amostra ou população. Como não a utilizamos para tirar conclusões a respeito de dados que não observamos, podemos falar dela sem comunicar incerteza.

- b) [V | F] Quando fazemos inferência estatística, é preciso comunicar nosso grau de incerteza, já que estamos tirando conclusões sobre parâmetros que não observamos. Comunicamos essa incerteza por meio de uma estimativa intervalar e de um nível de confiança, que indica a probabilidade de a média populacional estar dentro do intervalo de confiança.

[F] Quando fazemos inferência estatística, é preciso comunicar nosso grau de incerteza, já que estamos tirando conclusões sobre parâmetros que não observamos. Comunicamos essa incerteza por meio de uma estimativa intervalar e de um nível de confiança, que indica a probabilidade de a média populacional estar dentro do intervalo de confiança.

Correção: comunicamos essa incerteza por meio de uma estimativa intervalar e de um nível de confiança, que indica a proporção das vezes que se espera que o parâmetro populacional esteja contido dentro dos intervalos em uma sequência bastante longa de repetições.

Justificativa: a interpretação do nível de confiança é a interpretação clássica, frequentista, ou seja, a da esperança de conter o parâmetro em uma longa sequência de repetições do intervalo.

- c) [V | F] Quando aumentamos o número de observações de uma amostra, diminuimos a margem de erro e, conseqüentemente, aumentamos o nível de confiança de uma estimativa intervalar.

[F] Quando aumentamos o número de observações de uma amostra, diminuimos a margem de erro e, conseqüentemente, aumentamos o nível de confiança de uma estimativa intervalar.

Correção: diminuimos o erro-padrão e, conseqüentemente, a margem de erro de uma estimativa intervalar.

Justificativa: ao se aumentar o n amostral, o nível de confiança permanece inalterado. O impacto desse aumento é sobre o erro-padrão, não sobre o nível de confiança. Pela fórmula, $IC = (\bar{Y} \text{ ou } \hat{p}) \pm M = (\bar{Y} \text{ ou } \hat{p}) \pm (Z_{\alpha/2} \text{ ou } t_{\alpha/2}) \times ep$, ou seja, $IC = (\bar{Y} \text{ ou } \hat{p}) \pm (Z_{\alpha/2} \text{ ou } t_{\alpha/2}) \times \left(\frac{\sigma}{\sqrt{n}} \text{ ou } \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$. O $Z_{\alpha/2}$ ou $t_{\alpha/2}$ permanecem inalterados

com variações no n amostral. Mas o aumento do n reduz o erro-padrão e, consequentemente, reduz a margem de erro.

- d) [V | F] Quando falamos em “distribuição dos dados amostrais”, estamos falando do comportamento dos dados da nossa amostra, ou seja, de estatística descritiva. Quando falamos de “distribuição amostral”, estamos falando da distribuição de estimativas, ou seja, de estatística inferencial.

[V] Quando falamos em “distribuição dos dados amostrais”, estamos falando do comportamento dos dados da nossa amostra, ou seja, de estatística descritiva. Quando falamos de “distribuição amostral”, estamos falando da distribuição de estimativas, ou seja, de estatística inferencial.

- e) [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.

[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: cada amostra tem um desvio em relação à média populacional devido à própria aleatoriedade das amostras e devido ao fato de que a média populacional não é observada.

Justificativa: a variabilidade das amostras em relação ao parâmetro populacional (que caracteriza o erro-padrão) é decorrência da aleatoriedade dos dados amostrais. Viés de seleção, perguntas mal formuladas, erros no desenho de pesquisa ou erros de mensuração impactam no processo de inferência, dado que a estimativa muito provavelmente apresentará problemas de precisão e acurácia. Sob essas condições, o erro-padrão será calculado para médias (ou proporções) enviesadas em relação ao verdadeiro parâmetro populacional. Mas o ponto importante é que o desvio calculado pelo erro-padrão decorre da variabilidade dos próprios dados, não dos problemas citados. Esses problemas citados são relativos à inferência, não são relativos à variabilidade dos dados.

- f) [V | F] O Teorema do Limite Central diz que a distribuição das estatísticas amostrais segue uma distribuição normal. Esse resultado nos permite fazer inferências para quaisquer variáveis, independentemente de sua distribuição original.

[V] O Teorema do Limite Central diz que a distribuição das estatísticas amostrais segue uma distribuição normal. Esse resultado nos permite fazer inferências para quaisquer variáveis, independentemente de sua distribuição original.

Observação: foram consideradas corretas as respostas que classificaram a alternativa como falsa, mas que justificaram que é necessário um aumento do n para que a distribuição de uma estatística amostral se aproxime a uma distribuição normal. O motivo é que, pelo TLC, o n deve ser suficientemente grande para que se caracterize o processo de convergência em distribuição.

- g) [V | F] Em um teste de hipóteses, um p-valor alto (por exemplo, $p = 0,25$) indica evidência forte a favor da hipótese nula, de maneira que podemos aceitar que ela é verdadeira.

[F] Em um teste de hipóteses, um p-valor alto (por exemplo, $p = 0,25$) indica evidência forte a favor da hipótese nula, de maneira que podemos aceitar que ela é verdadeira.

Correção: de maneira que não podemos rejeitar que ela seja verdadeira.

Justificativa: em um teste de hipóteses não se deve “aceitar” a hipótese nula. Por se tratar de um teste estatístico, não se pode ter certeza absoluta sobre parâmetros populacionais a partir de dados amostrais.

- h) [V | F] Quando diminuimos a probabilidade de cometer erro do Tipo I, aumentamos a probabilidade de cometer erros do tipo II.

[V] Quando diminuimos a probabilidade de cometer erro do Tipo I, aumentamos a probabilidade de cometer erros do tipo II.

Exercício 6 [2 pontos] A imagem abaixo mostra as estatísticas descritivas da variável PIB *per capita*, disponível no banco *world95.dta*.

. summarize gdp_cap					
Variable	Obs	Mean	Std. Dev.	Min	Max
gdp_cap	102	5883.333	6603.27	122	23474

- a) Calcule o erro-padrão para a média

A fórmula do erro-padrão para a média é $ep = \sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$. Assim:

$$ep = \sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}} = \frac{\hat{s}}{\sqrt{n}} = \frac{6603,27}{\sqrt{102}} = 653,8212$$

- b) Calcule a margem de erro com nível de confiança de 95%

A fórmula da margem de erro para a média é $M = (Z_{\alpha/2} \text{ ou } t_{\alpha/2}) \times ep$

Neste caso, o aluno deveria escolher entre utilizar $Z_{\alpha/2}$ ou $t_{\alpha/2}$. Pelo fato de n ser suficientemente grande, o aluno poderia utilizar o $Z_{\alpha/2}$, desde que justificasse pelo Teorema do Limite Central. De qualquer maneira, demonstramos os cálculos com os dois, para demonstrar que a diferença é razoavelmente pequena:

Pela Tabela-t	Pela Tabela-Z
$M = (t_{\alpha/2}) \times ep$	$M = (Z_{\alpha/2}) \times ep$
$g.l. = n - 1 = 102 - 1 = 101$ $\alpha = 0,05$ Pela tabela-t, $t_{\alpha/2} = 1,983731$	$\alpha = 0,05$ Pela tabela-Z, $Z_{\alpha/2} = 1,959964$
$M = (t_{\alpha/2}) \times ep$ $M = 1,983731 \times 653,8212$ $M = 1.297,0054$	$M = (Z_{\alpha/2}) \times ep$ $M = 1,959964 \times 653,8212$ $M = 1.281,4660$

- c) Calcule a margem de erro com nível de confiança de 97%. Quando aumentamos o nível de confiança, a margem de erro aumenta ou diminui?

Para a mudança no nível de significância, basta alterar o $t_{\alpha/2}$ ou o $Z_{\alpha/2}$.

Pela Tabela-t	Pela Tabela-Z
$M = (t_{\alpha/2}) \times ep$	$M = (Z_{\alpha/2}) \times ep$
$g.l. = n - 1 = 102 - 1 = 101$ $\alpha = 0,03$ Pela tabela-t, $t_{\alpha/2} = 2,2011817$	$\alpha = 0,03$ Pela tabela-Z, $Z_{\alpha/2} = 2,1700904$
$M = (t_{\alpha/2}) \times ep$ $M = 2,2011817 \times 653,8212$ $M = 1.439,1791$	$M = (Z_{\alpha/2}) \times ep$ $M = 2,1700904 \times 653,8212$ $M = 1.418,8510$

Em comparação com as margens de erros obtidas no item b, ao aumentarmos o nível de confiança de 95% para 97%, observa-se um aumento na margem de erro. Isso ocorre, pois, na abordagem frequentista, deseja-se que mais intervalos contenham o parâmetro de interesse. Isto é, deseja-se que os intervalos que contenham o parâmetro passem, na média (esperança), de 95% (item b) a 97% das repetições.

Trata-se de uma escolha mais conservadora do pesquisador ao aumentar o nível de confiança. Mas, ao mesmo tempo, perde-se em precisão, pois a estimativa intervalar passa a conter limites (inferior e superior) mais amplos. Assim, a escolha passa por um *trade-off* entre precisão, pois deseja-se fazer estimativas cada vez mais precisas, e acerto (conservadorismo), pois se deseja que os intervalos contenham de fato o parâmetro, mas, que não seja tão conservadora a ponto de a estimativa ser pouco informativa, isto é, intervalos extremamente amplos.

- d) Indique o intervalo de confiança de 95% para a média. Qual é a interpretação correta desse intervalo?

O intervalo de confiança de 95% para a média, considerando-se a Tabela-Z:

Limite Inferior	Estimativa Pontual	Limite Superior
$\bar{Y} - Z_{\alpha/2} \times ep$	$\bar{Y} = 5.883,333$	$\bar{Y} + Z_{\alpha/2} \times ep$
$\bar{Y} - M$		$\bar{Y} + M$
$5.883,333 - 1.281,4659$		$5.883,333 + 1.281,4659$
$4.601,8674$		$7.164,7993$

O intervalo de confiança de 95% para a média, considerando-se a Tabela-t:

Limite Inferior	Estimativa Pontual	Limite Superior
$\bar{Y} - t_{\alpha/2} \times ep$	$\bar{Y} = 5.883,333$	$\bar{Y} + t_{\alpha/2} \times ep$
$\bar{Y} - M$		$\bar{Y} + M$
$5.883,333 - 1.297,0053$		$5.883,333 + 1.297,0053$
$4.586,3280$		$7.180,3386$

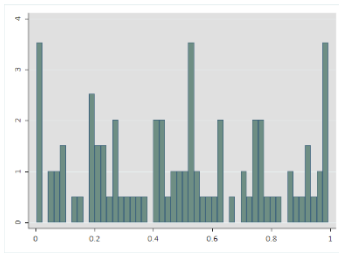
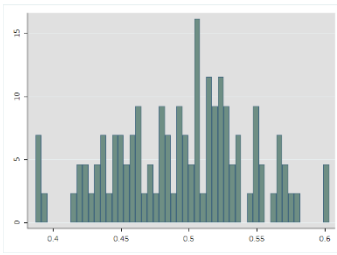
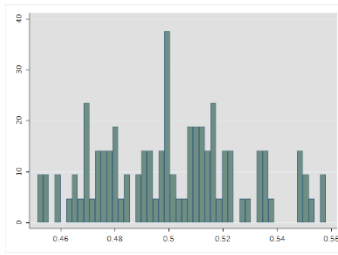
Para a interpretação correta do intervalo de confiança de 95% para a média, deve-se atentar ao fato de que, para uma longa sequência de repetições, espera-se (em média) que 95% dos intervalos contenham a média populacional. Isto é, após uma longa sequência amostragens, em média, 95 de cada 100 intervalos conterão a média populacional.

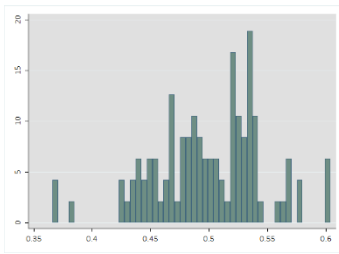
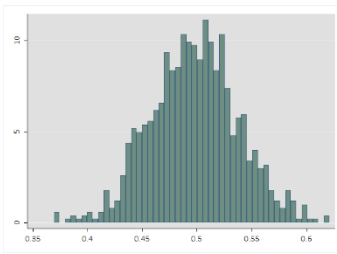
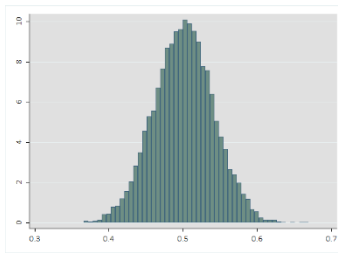
Assim, para o intervalo de confiança $[4.601,8674 ; 7.164,7993]$ (utilizando-se a *Tabela-Z*), espera-se que, após uma longa sequência de repetições, em 95% dessas repetições, a média populacional esteja contida nos intervalos.

Da mesma maneira, utilizando-se a *Tabela-t*, espera-se que, após uma longa sequência de repetições, em 95% dos casos, a média populacional esteja contida nos intervalos sequenciais de $[4.586,3280 ; 7.180,3386]$.

A interpretação incorreta e comum em muitas das respostas foi a de que o intervalo de confiança de 95% seria a probabilidade de 95% de que o parâmetro esteja dentro do intervalo, ou a probabilidade de 95% de que o intervalo cubra (ou inclua) o parâmetro. Não é essa a interpretação clássica. Vale ressaltar que essas interpretações existem, mas não é a interpretação clássica. Caso o aluno tenha interesse em conhecer mais, recomenda-se o estudo dos intervalos de credibilidade da estatística bayesiana.

Exercício 7 [1,5 ponto] Observe os gráficos a seguir de uma dada distribuição qualquer para a qual são apresentadas variações no tamanho das amostras e na quantidade de amostras.

Caso 1		
		
n de cada amostra = 1 Quantidade de amostras = 100	n de cada amostra = 30 Quantidade de amostras = 100	n de cada amostra = 200 Quantidade de amostras = 100

Caso 2		
		
n de cada amostra = 50 Quantidade de amostras = 100	n de cada amostra = 50 Quantidade de amostras = 1000	n de cada amostra = 50 Quantidade de amostras = 10000

Responda:

- a) Por que alterar o n de cada amostra (mantendo-se constante a quantidade de amostras) é diferente de alterar a quantidade de amostras (mantendo-se constante o n de cada amostra)? Utilize o que se observa nos gráficos para auxiliar em sua resposta.

Os efeitos são diferentes ao se alterar o n de cada amostra (mantendo-se constante a quantidade de amostras) em contraste com a alteração da quantidade de amostras (mantendo-se constante o n de cada amostra).

Conforme se observa no Caso 1, ao se aumentar o n de cada amostra de 1 para 30, a amplitude de x_i diminui de (0,0–1,0) para (0,36–0,65) (aproximadamente). Da mesma maneira, quando se aumenta o n de cada amostra de 30 para 200, a amplitude de x_i

diminui ainda mais, passando a (0,45–0,56) (aproximadamente). O que se verifica nos gráficos do Caso 1 é que a distribuição se apresenta cada vez mais concentrada em torno de um valor específico (em um processo de convergência pontual).

De modo distinto, no Caso 2, observa-se que, mantendo-se constante o n de cada amostra, mas aumentando-se a quantidade de amostras, altera-se o formato da distribuição. A amplitude nas três alterações do Caso 2 parece não sofrer alteração. À medida que se aumenta a quantidade de amostragens de 100 para 1.000 e 10.000, observa-se que a distribuição da variável aleatória x_i adquire cada vez mais um formato de sino (aproxima-se cada vez mais a uma distribuição normal).

- b) Quais são os nomes técnicos dos fenômenos observados para cada caso? Quais são as definições técnicas de cada um deles? Como se distinguem?

Os nomes técnicos dos fenômenos são:

- Caso 1: Lei dos Grandes Números (LGN);
- Caso 2: Teorema do Limite Central (TLC).

Para as definições técnicas:

- LGN: ao aumentar o n de cada amostra, obtém-se uma convergência pontual ao parâmetro de interesse;
- TLC: para um dado tamanho amostral suficientemente grande (e.g. $n > 30$), ao aumentar a quantidade de amostragens, obtém-se uma convergência em distribuição (da distribuição da estatística amostral) a uma distribuição normal, independente de como seja a distribuição original ou a distribuição populacional.

Para as distinções:

- LGN: convergência pontual com aumento do n de cada amostra e consequente redução do erro-padrão e melhoria na precisão do estimador (“consistência”);
- TLC: convergência em distribuição com o aumento da quantidade de amostras para uma distribuição normal e consequente mais facilidade nos cálculos do estimador (“normalidade assintótica”).

- c) Caso se queira utilizar as repetidas amostras para se inferir sobre parâmetros populacionais, pode-se utilizar o desvio-padrão ou o erro-amostral como margem para a estimativa pontual? Justifique com base nos conceitos de estimativa intervalar e margem de erro.

Não se deve utilizar o desvio-padrão ou o erro-padrão (apenas) como margem na estimativa intervalar.

Conceitualmente, para a estimativa intervalar necessita-se de uma estimativa pontual (uma média ou uma proporção) e uma margem de erro: $IC = (\bar{Y} \text{ ou } \hat{p}) \pm M$. Já a margem de erro é a multiplicação do Z -score ou do t -score (dependendo do caso) pelo erro-padrão:

$M = (Z_{\alpha/2} \text{ ou } t_{\alpha/2}) \times ep$. Por fim, o erro-padrão pode ser definido como $ep = \frac{s}{\sqrt{n}}$ para

médias ou $ep = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ para proporções.

Portanto, indiretamente tanto o desvio-padrão quanto o erro-padrão são utilizados nos cálculos de intervalos, mas não podem ser utilizados por si sós como margem.

Exercício 8 [1,5 ponto] Um acordo bilateral ambiental entre o país A e o país B determina um nível máximo de 500 galões de água contaminada por hora a ser despejada por cada parte em um lago fronteiriço comum. Algumas ONGs e alguns ativistas ambientais alegam que esse limite está sendo excedido. Mas, como o monitoramento permanente é custoso e nenhum dos países está disposto a arcar com esses gastos, uma equipe de pesquisadores foi enviada ao local para coletar algumas amostras para verificar as alegações. As amostras foram coletadas em diferentes horas, diferentes dias e sob diferentes condições climáticas.

Os dados da equipe de pesquisa mostram que para um $n = 121$, a média das amostras foi de 503 galões de água contaminada por hora. Supondo-se que as distribuições sejam normais e que a variância das amostras sejam todas iguais a 225, siga as etapas propostas nos itens abaixo, considerando um nível de confiança de 95%.

- a) Formalize as hipóteses e indique qual o nível de significância do teste.

Formalização do teste:

$$\begin{aligned}H_0: \mu &= 500 \\H_a \text{ ou } H_1: \mu &> 500\end{aligned}$$

Obs: é importante observar que este é um teste monocaudal (ou unicaudal ou unilateral) à direita. Estamos testando se os limites estão sendo ultrapassados.

Conforme foi reiteradamente explicado nas monitorias, o teste é sobre parâmetros populacionais. Portanto, estão incorretas as respostas que testaram estatísticas amostrais, como em $H_0: \bar{Y} = 500$ (menos grave), ou que simplesmente nem indicaram o que está se testando, como em $H_0 = 500$ (muito grave).

Nível de significância do teste:

$$\alpha = 1 - 0,95 = 0,05$$

- b) Calcule a estatística-teste. Trata-se um Z -teste ou t -teste?

Neste caso, como as distribuições são todas normais, podemos considerar a estatística-teste como um Z -teste para médias:

$$Z = \frac{\bar{Y} - \mu}{ep} = \frac{503 - 500}{\frac{\sqrt{225}}{\sqrt{121}}} = \frac{3}{\frac{15}{11}} = 2,2 \text{ (estatística-teste)}$$

Utilizamos as seguintes informações:

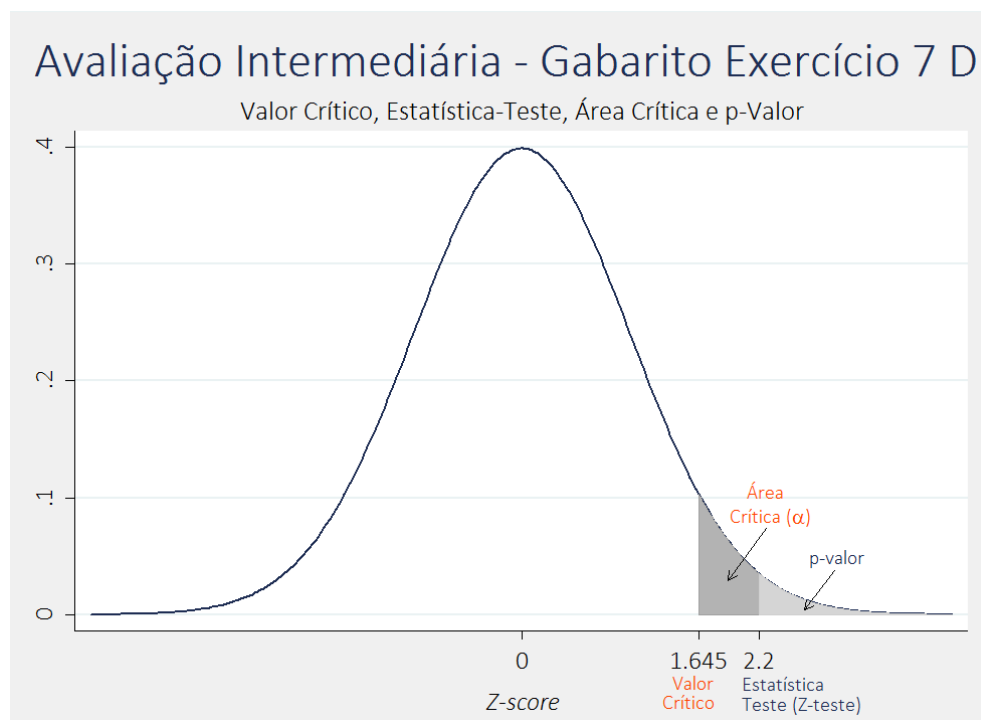
- Média amostral: $\bar{Y} = 503$
- Média populacional (sob H_0): $\mu = 500$
- Erro-padrão para médias: $ep = \frac{s}{\sqrt{n}} = \sqrt{\frac{s^2}{n}}$. Obs: o valor de 225 é o da variância.

- c) Encontre o p-valor com base na tabela apropriada. Encontre também o valor crítico.

Para o p-valor, necessitamos consultar a *Tabela-Z* para um *Z-score* de 2,2 (linha do 2.2 e coluna do zero): *p - valor* = 0,01390345

Para o valor crítico, devemos fazer o procedimento contrário, ou seja, buscar o valor mais próximo da probabilidade de 0,05 (α como área crítica) e retornar o *Z* correspondente: $Z = 1,6448536$ (que é o valor crítico).

- d) Apresente um desenho com a estatística-teste, o valor crítico, o p-valor e a região crítica.



- e) Decida se rejeita ou não rejeita a hipótese nula. Faça as comparações adequadas para isso.

Como regra de decisão, deve-se comparar a estatística-teste vs. o valor crítico ou o p-valor vs. a área crítica (α):

Comparação para Decisão	
Por Pontos	Estatística-Teste > Valor Crítico
Por Áreas	p-valor < Área Crítica (α)

Decisão: rejeita-se H_0

- f) Apresente uma conclusão formal adequada (em palavras) a ser enviada aos países.

Como conclusão, deve-se enfatizar que, com base nas amostras coletadas, com 95% de confiança, existem indícios estatísticos de que o limite está sendo excedido.

Serão descontados pontos às respostas que fizerem afirmações contundentes. Em um teste de hipóteses não se pode concluir com afirmações imperativas, dado que não se conhece toda a população (que se deseja estudar) e os cálculos somente se baseiam em amostras (aleatórias). Portanto, as respostas devem sempre considerar que os dados disponíveis para os cálculos fornecem apenas indícios estatísticos favoráveis (ou não favoráveis) a H_0 , para dado nível de confiança.

Utilize esse espaço para rascunho

Utilize esse espaço para rascunho
