Universidade de São Paulo Faculdade de Filosofia, Letras e Ciências Humanas
Departamento de Ciência Política

FLS 5028 Métodos Quantitativos e Técnicas de Pesquisa em Ciência Política
FLP0406 Métodos e Técnicas de Pesquisa em Ciência Política

1º semestre / 2018
Prof. Glauco Peres da Silva

LISTA DE EXERCÍCIOS 04

Data de entrega: 09/04/2018 (noturno) e 11/04/2018 (vespertino).

Exercício 01 (2 pontos)

Marque "Verdadeiro" (V) ou "Falso" (F) para as sentenças que serão apresentadas a seguir e **justifique** cada uma das suas escolhas em no máximo 5 linhas.

(**F**) Chamamos de escore-z a distância entre y (média da distribuição da probabilidade) e μ (valor da variável observada) medida em desvios padrão.
Falso. De acordo com Agresti e Finlay (2012, p. 103)o escore-z representa a distância entre *y* (valor observado) e μ (média do parâmetro dada uma grande quantidade de repetições) medida em desvios padrão.

(**F**) As distribuições de probabilidade das variáveis contínuas atribuem probabilidades a intervalos de números. Assim, a probabilidade de que o valor de uma variável esteja em algum dos intervalos é 1, enquanto que a probabilidade de o intervalo conter todos os valores possíveis que a variável assume varia de 0 a 1.
Falso. É justamente o contrário. A probabilidade de que o valor de uma variável esteja em algum dos intervalos varia de 0 a 1, enquanto que a probabilidade do intervalo conter todos os valores possíveis que uma variável assume é 1 (AGRESTI; FINLAY, 2012, p. 97).

(**V**) O erro padrão fica menor quanto maior é o tamanho da amostra dado por *n*.

Verdadeiro. O erro padrão fica menor à medida que o tamanho da amostra *n* fica maior. Isso acontece, porque amostras maiores fornecem estimativas mais precisas das características da população. E como o erro padrão descreve como $\bar{y}$ varia de amostra para amostra, quanto maior for o nosso *n*, menores ficam essas variações (AGRESTI; FINLAY, 2012, p. 111).

(**F**) Somente as variáveis discretas apresentam uma distribuição de probabilidade para os valores que elas assumem.

Falso. De acordo com Agresti e Finlay (AGRESTI; FINLAY, 2012, p. 96 e 97) a distribuição de probabilidade de uma variável discreta atribui uma probabilidade a cada valor possível de que a variável venha a assumir. Mas as variáveis contínuas também apresentam uma distribuição de probabilidade que é dada e estabelecida por intervalos de números.

(**V**) O Teorema do Limite Central afirma que para uma amostra aleatória com *n* grande, a distribuição amostral da média da amostra dada por $\bar{y}$ é aproximadamente uma distribuição normal com média $\mu = 0$ e desvio padrão $\sigma = 1$.

Verdadeiro. Para amostras aleatórias com *n* grande a distribuição amostral das médias das amostras dadas por $\bar{y}$ tende a ser muito próxima da média normal dada por μ=0 e desvio padrão $\sigma = 1$. Isso acontece porque a repetição das amostras faz com que nossos dados se aproximem dos dados da população.

Exercício 2 (4 pontos)

Para este exercício, utilize o banco de dados "CPDS_1960-2013" (e codebook), disponível no Moodle. Vamos trabalhar com a variável "effpar_ele", discutida em sala de aula.

a) Descreva brevemente o conceito que a variável está mensurando e como foi operacionalizada.

Conforme o codebook, a variável "effpar_ele" operacionaliza o **conceito de número efetivo de partidos** de acordo com nível de votos recebidos. Para operacionalizar este conceito, o índice utiliza os valores do "índice de fragmentação eleitoral" (rae_ele), de modo a ser calculado pela seguinte fórmula: effpar_ele = 1/(1-rae_ele), enquanto $rae\_ele = \Sigma vi2mi = 1$, onde *v* representa o número de votos recebido por partido *i* e *m* o número de partidos. A variável "effpar_ele" é uma variável quantitativa e contínua mensurada por ano e país entre 1960 e 2013 para 36 países.

b) Calcule a média, variância e desvio padrão desta variável. Reporte seus cálculos e resultados.

Passo 1: identificar os missings e exclui-los do banco

Passo 2: identificar as fórmulas

Média: Considerando a fórmula $\bar{Y} = \frac{\sum_{i=1}^{n} Yi}{n}$

Variância $S^2 = \sqrt{\frac{\sum_i^n (Y_i - \bar{Y})^2}{n-1}}$

Desvio padrão $S = \sqrt{\frac{\sum_i^n (Y_i - \bar{Y})^2}{n-1}}$

Onde y é o valor assumido pela variável "effpar_ele" por observação *i* (lembrando que existe mais de uma observação por país), e *n* é igual ao total de observações para esta variável (*n*= 1521, pois há missing na variável) temos:

Passo 3: fazer os cálculos

$\bar{Y}$=4,273785833

*S*2= 2,785030822

S= 1,668841161

c) É possível afirmar que os parâmetros obtidos acima correspondem aos parâmetros da população? Estamos trabalhando com uma população ou uma amostra?

O aluno deve olhar o codebook e reparar que os dados acima são amostrais, ainda que correspondam a percentuais gerais para 36 países da OCDE e União Europeia. Esta escolha se justificaria principalmente porque não estão sendo analisados dados para todos os países do globo. E ainda, a variável apenas apresenta observações para o período 1996 e 2013.

Exercício 3 (4 pontos)

Para este exercício vamos utilizar o mesmo banco de dados do exercício acima e a mesma variável discutida acima.

a) Selecionamos uma amostra aleatória de 10 observações da variável "effpar_ele". Calcule a média e o desvio padrão desta amostra.

2013	USA	2,192516
2013	Slovakia	4,387273
2012	Luxembourg	4,257221
1996	Luxembourg	4,728691
1970	Japan	3,392061
1981	Germany	3,097414
2011	New Zealand	3,164247
1975	Italy	4,073303
2002	United Kingdom	3,337483
1979	Denmark	4,995479

Passo 1: identificar as fórmulas

Média: Considerando a fórmula $\overline{Y} = \frac{\sum_{i=1}^{n} Yi}{n}$

Variância $S^2 = \sqrt{\frac{\sum_i^n (Y_i - \bar{Y})^2}{n-1}}$

Desvio padrão $S = \sqrt{\frac{\sum_i^n (Y_i - \bar{Y})^2}{n-1}}$

Média= 3,76

Variância: 0,752046

Desvio Padrão= 0,867

b) Agora selecionamos mais quatro amostras aleatórias da mesma variável, com o mesmo número de 10 observações. Calcule a média, variância e desvio padrão para cada uma delas. Os valores encontrados são os mesmos que aqueles do item "a"? Por que?

Amostra 1			Amostra 2		
1972	New Zealand	2,430902	1994	Switzerland	7,423629
2007	Poland	3,326558	1984	Japan	3,669132
1999	Netherlands	5,154134	1990	Slovakia	5,816253
1978	Austria	2,267713	1965	Germany	3,149706
1992	Denmark	4,859559	2007	Australia	3,013673
1979	Spain	4,299522	1977	Norway	3,759497
1986	Denmark	5,252846	2001	Austria	3,819841
2007	Bulgaria	5,830768	1997	Japan	4,090682
1969	Norway	3,51985	1989	Sweden	3,906433
1994	Denmark	4,766649	1986	Netherlands	3,779261

Amostra 3	Amostra 4

2011	Portugal	3,67381	1992	Germany	3,735902
2008	France	4,09311	1993	Switzerland	7,423629
1999	Austria	3,819841	2005	Austria	3,019515
2003	Bulgaria	3,950789	1995	Sweden	3,65473
1965	Norway	3,830537	1998	Romania	6,121562
1969	Ireland	2,824619	1979	Australia	3,117081
2012	Estonia	4,779292	1996	Ireland	3,950258
1981	Spain	4,299522	2009	Croatia	4,103743
1978	Ireland	2,758491	1995	Germany	3,747915
1995	Denmark	4,766649	2003	Slovenia	5,16177

Amostra 1	Amostra 2
Média: 4,17085 Variância: 1,502747136 Desvio Padrão: 1,225865872	Média: 4,242810794 Variância: 1,821306646 Desvio Padrão: 1,349557945
Amostra 3	**Amostra 4**
Média: 3,879666047 Variância: 0,471276342 Desvio Padrão: 0,686495697	Média: 4,403611 Variância: 1,99439978 Desvio Padrão: 1,412232198

Os valos encontrados não são os mesmos que em “a”. As novas amostras aleatórias trazem novas observações e isso altera os valores encontrados para as médias, variância e desvio padrão.

Espera-se que a cada nova amostra do total de 1159 observações, as estatísticas selecionadas também variem, se aproximando mais ou menos dos valores encontrados para o total de 1159.

O tamanho da amostra, menos de 1% do total de observações, também tem impacto sobre a diferença entre os valores das estatísticas encontradas e os parâmetros da amostra total.

c) Anote em uma planilha diferente as cinco médias que você encontrou a partir das cinco amostras dos itens “a” e “b”. Calcule a média deste conjunto de médias e compare-a com a média encontrada no exercício 2. À luz do Teorema Central do Limite, explique seus achados.

Amostra1	4,17
Amostra 2	4,24
Amostra 3	3,87
Amostra 4	4,4
Amostra 5	3,76
Média	4,088

A média das médias ainda está muito distante da média das 1159 observações, encontrada no exercício 2 (4,273). Isso porque, tal como no item “b” deste exercício, mesmo trabalhando com 5 amostras (e 50 observações), ainda estamos lidando com um percentual pequeno do número de observações (menos de 5%).

O teorema central do limite indica que a distribuição amostral da média amostral assume formato normal, centrado na média da população. Assumindo aqui que as 1159 observações são nossa “população” (ver a resolução do exercício 2b para esclarecimentos), o que tornaria a distribuição da média amostral normal e centrada em 4,273, teríamos alguma probabilidade α de encontrar cada uma das médias amostrais obtidas nos itens acima, probabilidade está dada pela área sobre a curva normal.

d) Agora selecionamos novamente uma segunda amostra da mesma variável, desta vez com 100 observações (veja a seleção na planilha 2 do banco de dados, intitulada AMOSTRA). Calcule a média e o desvio padrão desta nova amostra. Existe diferença em relação a primeira? Estamos diante de uma demonstração do Teorema Central do Limite?

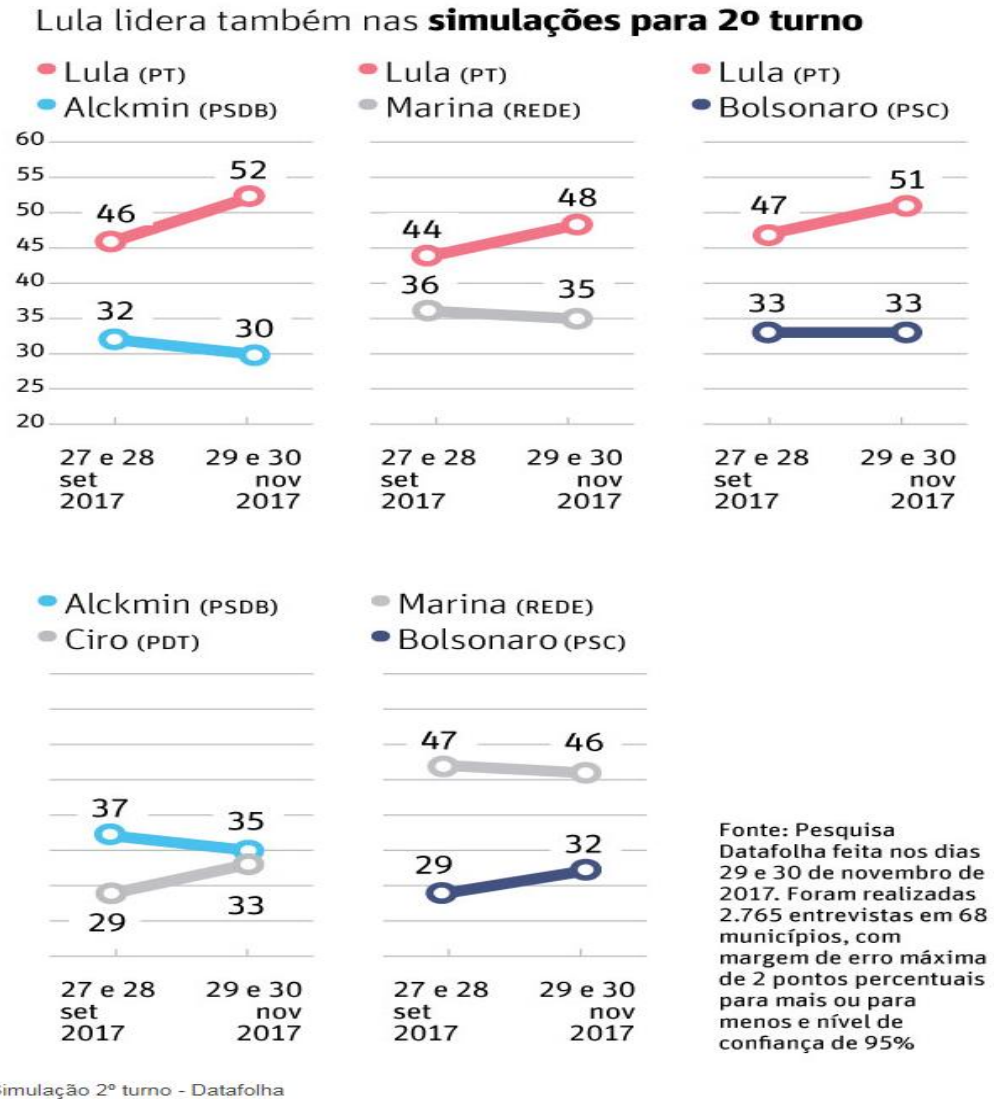
A média obtida para a amostra de 100 observações é de 4,13 e o desvio padrão, 1,60. Observa-se que especialmente o desvio padrão está mais próximo daquele para todas as observações (1,66).

A média também teve seu valor aumentado e mais próximo de 4,273, se comparada com o exercício “c”. Esse resultado, contudo, não é uma demonstração do Teorema Central do Limite, mas da Lei dos grandes números, que afirma que na medida em que

aumentamos o tamanho da amostra, os parâmetros dela se aproximam daqueles da população.

Exercício 4 Pós-Graduação (5 pontos)

Na lista 01, foi apresentado a você uma pesquisa que o Data Folha publicou no dia 02 de dezembro de 2017 sobre o possível resultado das eleições presidenciais de 2018. Para relembrar, segue novamente a divulgação dos resultados da pesquisa criada a partir de disputas hipotéticas para o segundo turno.



Simulação 2º turno - Datafolha

> Lula ganha em todos os cenários de segundo turno. Ele ampliou em quatro pontos percentuais sua vantagem, em relação à pesquisa feita no fim de setembro, no confronto com Alckmin (52% a 30%), Marina (48% a 35%) e Bolsonaro (51% a 33%).
>
> O tucano empata tecnicamente com Ciro (35% a 33%) e Marina ganharia de Bolsonaro (46% a 32%).

(Fonte: http://www1.folha.uol.com.br/poder/2017/12/1940171-lula-lidera-e-bolsonaro-se-consolida-em-2-aponta-datafolha.shtml)

a-) Repare que para as simulações do 2º turno, o candidato Luiz Inácio Lula da Silva lidera as pesquisas em todos os cenários. Você, enquanto pesquisador, está interessado em entender melhor a intenção de votos dada a esse candidato. Quais variáveis você mobilizaria para tentar explicar esse fenômeno de interesse? Qual seria sua variável dependente e suas variáveis explicativas? Como as variáveis selecionadas ajudam a explicar o fenômeno de interesse? Qual teoria você mobilizaria para justificar essas escolhas? Justifique todas as suas escolhas. (Máximo de 20 linhas)

O fenômeno de interesse que estamos interessados em explicar é a quantidade de votos, ou a intenção de votos, no candidato Luiz Inácio Lula da Silva para as eleições presidenciais de 2018. Uma teoria que poderia ser mobilizada para explicar os votos no candidato seria a teoria que lida com a vantagem dos incumbentes (teoria já citada em aula e pelos autores da bibliografia). Como Lula já foi presidente, determinados atributos – como, por exemplo, acesso a recursos públicos, cargos e outros incentivos seletivos; acesso a financiadores e à máquina pública; influência sobre políticas públicas; maior exposição na mídia; maior visibilidade junto ao eleitorado; etc – poderiam fornecer ao candidato maiores vantagens sobre seus concorrentes/ rivais em disputa. Para explicar, então, as intenções de voto no candidato Luiz Inácio Lula da Silva (variável dependente) poderiam ser mobilizadas variáveis como: escolaridade; acesso as políticas públicas, financiamento de campanha, tempo de propaganda eleitoral gratuita, etc. Aqui o aluno deve justificar o que ele espera que cada uma das variáveis impacte sobre a intenção de votos.

b-) Olhando apenas para o cenário colocado pelo Data Folha em que o candidato Lula (PT) e o candidato Bolsonaro (PSC) disputam o 2º turno das eleições temos o seguinte resultado:

Candidato	Intenção de Votos
Lula	51%
Bolsonaro	33%
Brancos e Nulos	16%

Você, enquanto pesquisador, continua interessado em medir a probabilidade de sucesso do candidato Luiz Inácio Lula da Silva. Tendo em vista o cenário descrito acima, calcule a média e o desvio padrão para a amostra dessa pesquisa. Lembre-se de que estamos interessados no sucesso de um candidato, apenas. Demonstre seus cálculos e interprete os resultados (Máximo de 20 linhas)

1. O aluno deve atentar-se para o fato de que ele precisará medir sucesso da candidata Dilma como sendo igual a 1 e os demais como sendo igual a 0 (Ver Guy e Whitten capítulo 6).
2. Cálculo da Média: o aluno pode chegar ao valor da média de duas formas: Nessa primeira ele deve calcular o quanto cada resultado em % representa do total da amostra, ou seja, ele deve calcular a frequência absoluta para cada opção. Realizando o cálculo o aluno deve chegar aos seguintes valores para cada uma das opções:

Candidato	Intenção de Votos
Lula	1410
Bolsonaro	913
Brancos e Nulos	442

Além disso o aluno deve se atentar para o fato de que ele precisa atribuir um valor para as opções de resposta.

Média

$$\bar{y} = \frac{(1*1410)+(0*913)+(0*442)}{2765} = 0{,}51$$

ou

$$\mu = \sum yP(y) = (1 * 0{,}51) + (0 * 0{,}33) + (0 * 0{,}16) = 0{,}51$$

Desvio padrão

$$s = \sqrt{\frac{1410(1-0{,}51)^2 + 913(0-0{,}33)^2 + 442(0-0{,}16)^2}{2765-1}} = 0{,}49$$

Interpretando os resultados: levando em consideração somente o interesse em medir o sucesso eleitoral do candidato Lula tem-se que, em média, a probabilidade do sucesso eleitoral do candidato será igual a 51%. Já o desvio – padrão nos indica que a distância típica dos dados em relação à média é de 0,16, não sendo considerado um desvio muito elevado, devido à pequena variabilidade das respostas.

c-) Calcule agora o erro-padrão e em seguida calcule o intervalo de confiança para 68%, 95% e 99% de confiança. O que o erro-padrão significa? O que é possível interpretar a partir dos intervalos de confiança calculados? (Máximo de 15 linhas)

$$\sigma_y = \frac{\sigma}{\sqrt{n}}$$

$$\sigma_y = \frac{0{,}49}{\sqrt{2765}} = 0{,}009$$

O erro padrão é o desvio padrão da distribuição amostral. Nesse caso o erro padrão de 0,009 indica que a proporção amostral da pesquisa com 2765 entrevistados irá variar 0,009 de uma amostra para outra.

Intervalo de Confiança = $\bar{y} \pm 1\ ou\ 2\ ou\ 3 * \sigma y$

68% $\bar{y}$= 0,51 ± 1*0,007 = O sucesso eleitoral do candidato Lula está entre 50,0% e 51,9%

95% $\bar{Y}$ = 0,51 ± 2*0,007 = O sucesso eleitoral do candidato Lula está entre 49,1% e 52,8%

99% $\bar{Y}$ = 0,51 ± 3*0,007 = O sucesso eleitoral do candidato Lula está entre 48,2% e 53,7%

d-) Em um cenário hipotético, suponha que nos resultados oficiais das eleições o candidato Lula tenha ganhado com 52,7% dos votos. Como podemos explicar a diferença entre os resultados da pesquisa do Data Folha e o resultado oficial das eleições? Essa

diferença era prevista pelo Data Folha? Justifique sua resposta e aponte sugestões para a melhora na precisão das pesquisas de intenção de voto. (Máximo de 15 linhas)

Aqui o aluno deveria se atentar para o enunciado da pesquisa do Data Folha quando eles declaram que a margem de erro é de dois pontos para mais ou para menos com 95% de confiança. Considerando essa margem de erro apresentada pelo instituto de pesquisa é possível afirmar que o resultado oficial estava previsto dentro do resultado que foi encontrado na amostra, pois 51+2 = 53% e 51-2=49%. Sendo assim, os resultados do Ibope poderiam variar entre 49% e 53% e o resultado oficial de 52,7% encontra-se dentro desse intervalo com 95% de confiança. Uma sugestão para melhorar a pesquisa, à luz do Teorema do Limite Central explica, seria possível aproximar os resultados encontrados pelo Data Folha do resultado oficial se realizássemos essa mesma pesquisa várias vezes. A média encontrada para todas essas amostras – umas acima e outras abaixo – geraria uma distribuição normal e se tirarmos a média das médias encontradas para todas as amostras, essa média final (ou média amostral) seria igual a média da população. (Para mais informações ver Guy e Whitten capítulo 6)