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 / 2016

Prof. Glauco Peres da Silva

LISTA DE EXERCÍCIOS 04

Data de entrega: 04/04/2016 (noturno) e 06/04/2016 (vespertino).

Exercício 1(3 pontos)

Indique se cada uma das afirmativas abaixo é "Verdadeira" (V) ou "Falsa" (F) e justifique suas escolhas em no máximo 5 linhas quando a opção escolhida for "Falso".

(   ) Uma variável aleatória somente pode assumir um único resultado e apenas é considerada aleatória se for de tipo contínua.

Falso. Segundo Agresti e Finlay (2012), uma variável aleatória assume ao menos dois resultados que variam de observação a observação, podendo ser atribuída uma probabilidade possível para cada um deles. Denomina-la "variável aleatória" reforça a ideia de que existe variação para os valores assumidos pela variável e que esta variação ocorre de forma aleatória. Tanto variáveis discretas como contínuas podem ser variáveis aleatórias, desde que se encaixem na definição acima.

(   ) A média de uma distribuição de probabilidade ($\bar{x}$) é diferente de seu valor esperado (E(x)), já que a média é igual a soma do total de observações dividida pelo tamanho da amostra; e o valor esperado é dado pela soma de todos os valores possíveis da variável.

Falso. Como afirmam Agresti e Finlay (2012), a média da distribuição de probabilidade de uma variável é também seu valor esperado, assumindo que se espera o valor médio encontrado após uma longa série de repetições das observações ou extrações. A média é obtida pela soma do total de valores dividida pelo numero de observações, mesmo procedimento para obter o valor esperado (soma de todos os valores possíveis da variável multiplicados pela probabilidade de ocorrerem).

( ) A distribuição normal é uma distribuição simétrica com forma de sino, média μ e desvio padrão σ.

Verdadeira.

( ) A distribuição amostral de uma estatística especifica a probabilidade de uma variável assumir determinado valor em cada amostra coletada.

Falso. A distribuição amostral de uma estatística é a distribuição da probabilidade dos possíveis valores que aquela estatística (uma proporção, a média, o desvio padrão, por exemplo) pode assumir. Esta distribuição não especifica ou depende das probabilidades da estatística para observações individuais, mas traça a distribuição dos valores possíveis que a estatística pode (ou não vir a assumir).

( ) O Teorema Central do Limite afirma que a média amostral tende a aproximar-se da média da população na medida em que *n* aumenta.

Falso. O Teorema Central do Limite afirma que a distribuição amostral da média amostral é aproximadamente normal. A afirmação acima define a Lei dos Grandes números.

Exercício 2 (3 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 numero 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 $rae\_ele = \sum_{i=1}^{m} v_i^2$, onde *v* representa o númeo 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.

Considerando a fórmula $\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$ para a média, $S^2 = \sqrt{\frac{\sum_i^n (Y_i - \bar{Y})^2}{n-1}}$ para a variância e $S = \sqrt{\frac{\sum_i^n (Y_i - \bar{Y})^2}{n-1}}$, para o desvio padrão, 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:

$\bar{Y}$ =4,273785833

S^2 = 2,785030822

e S= 1,668841161

Lembrando que o aluno deve indicar que a variável contém missing.

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

É mais correto tratar os dados acima como dados 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 além de a variável apenas apresentar 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.

Ano	País	effpar_ele
1980	Switzerland	5,534402
1964	Italy	4,163215
2006	Bulgaria	5,830768

1968	Greece	
2007	Czech Republic	3,912011
1981	Belgium	9,040284
1964	Denmark	3,759059
2009	Denmark	5,402485
1960	Norway	3,350813
1961	Luxembourg	3,26143

Considerando a fórmula $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$ para a média, $s^2 = \sqrt{\frac{\sum_i^n (y_i - \bar{y})^2}{n-1}}$ para a variância e $s = \sqrt{\frac{\sum_i^n (y_i - \bar{y})^2}{n-1}}$, para o desvio padrão, 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*= 9) temos:

$\bar{y} = 4{,}917162936$

$s^2 = 3{,}320774285$

e $s = 1{,}822299176$

Novamente há a presença de missing.

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 2

Ano	País	effpar_ele
2005	Bulgaria	5,830767796
1972	Germany	2,856057555
1992	Estonia	8,893473969
2000	Switzerland	5,911528071
1998	Switzerland	6,918356475
2000	Greece	2,6388636
1977	USA	2,018660498
1987	Iceland	5,78760642
2003	Austria	3,019515126
1998	Denmark	4,740369938

Amostra 3

Ano	País	effpar_ele
1978	Sweden	3,579469594
1970	Netherlands	6,226727606
1974	Iceland	3,469620007
1966	New Zealand	2,614262896
1978	Norway	3,75949743
2003	Luxembourg	4,714357103
1988	Finland	6,157673385
2001	Lithuania	5,589652435
1981	Italy	3,920231137
1977	Austria	2,267712536

Amostra 4

Amostra 5

Ano	País	effpar_ele
2013	Croatia	4,351591159
2007	Cyprus	4,305593828
2008	Latvia	7,638310711
2001	Czech Republic	4,736283722
2008	Belgium	9,04273597
1981	Belgium	9,040283503
1983	United Kingdom	3,120436362
1995	Lithuania	3,832842091
1974	Belgium	6,107951943
1993	Belgium	9,842325938

Ano	País	effpar_ele
1981	United Kingdom	2,874174393
1981	Luxembourg	4,165018014
1998	Canada	4,111436371
1992	Spain	4,159699171
1983	Ireland	2,718181645
2004	Sweden	4,507286028
1994	Japan	5,29590891
2009	New Zealand	3,073934267
1991	Belgium	9,842325938
2010	Iceland	4,55302913

Amostra 2: $\bar{y} = 4{,}861519945$ $s^2 = 4{,}87093488$ e $s = 2{,}207019456$	Amostra 3: $\bar{y} = 4{,}229920413$ $s^2 = 1{,}951775304$ e $s = 1{,}397059521$
Amostra 4: $\bar{y} = 6{,}201835523$ $s^2 = 6{,}191897485$ e $s = 2{,}488352363$	Amostra 5: $\bar{y} = 4{,}530099387$ $s^2 = 4{,}155396797$ e $s = 2{,}03847904$

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.

Amostras	**Médias**

Amostra 1	4,917163
Amostra 2	4,86152
Amostra 3	4,22992
Amostra 4	6,201836
Amostra 5	4,530099
Média	**4,948108**
Variância	**0,567703**
Desvio Padrão	**0,753461**

A média das médias ainda está 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%). Como cada amostra foi selecionada de forma aleatória, suas médias variam entre si. A amostra 4, por exemplo, tem uma média muito distante da média calculada para as 1159 observações e isso provavelmente influencia o valor da média das 5 médias, levando-o a se distanciar de 4,273. 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 esta 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 99 observações (já que há um missing) foi 4,039234677 e o desvio padrão, 1,401915284. Observa-se que especialmente o desvio padrão está mais próximo daquele para todas as observações. A média também teve seu valor reduzido 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)

Como visto no artigo de Brambor e Ceneviva (2012) o argumento central é o de que determinadas qualidades e recursos, como por exemplo, exposição na mídia; facilidades para obter recursos para o financiamento de campanhas; disponibilidade de recursos governamentais que podem ser utilizados para mobilizar e angariar o apoio do eleitorado e, finalmente, capacidade para dissuadir desafiantes competitivos de concorrer, dariam ao incumbente alguma vantagem sobre os seus opositores.

Tendo esse argumento em mente, vamos analisar o que ocorreu nas últimas eleições presidenciais de 2014.

A última pesquisa[1] realizada pelo instituto Ibope (Instituto brasileiro de opinião pública e estatística) para o segundo turno das eleições presidenciais do ano de 2014 foi realizada entre os dias 24 e 25 de outubro de 2014. A pesquisa ouviu 3.010 eleitores em 206 municípios brasileiros. A margem de erro divulgada pelo instituto foi de 2 pontos para mais ou para menos com um intervalo de confiança de 95%. Abaixo segue a tabela 1 com os resultados divulgados pela pesquisa:

Tabela 1: Resultado Ibope

Candidato	Intenções de Voto
Dilma (PT)	49%
Aécio Neves (PSDB)	43%
Branco/Nulo	5%
Não Sabem/ Não Opinaram	3%

a-) O que podemos inferir dessa tabela? É possível mobilizar os argumentos sobre a vantagem do incumbente para tentar explicar os resultados acima? Em quais outros fatores você consegue pensar que possam explicar esses resultados? (Máximo de 12 linhas).

A tabela mostra, de acordo com a pesquisa realizada às vésperas do segundo turno pelo instituto de pesquisa Ibope, que a candidata Dilma Rousseff ganharia as eleições com 49% dos votos. Tal valor demonstra que ela teria uma vantagem de 6 pontos sobre o segundo candidato Aécio Neves. Mobilizando os argumentos a respeito da vantagem do incumbente é possível tentar explicar a possível vitória da candidata Dilma Rousseff pelas variáveis apresentadas por Brambor e Ceneviva (2012). Além disso, fatores

[1] A pesquisa está registrada do TSE sob o protocolo nº BR-01195/2014

relacionados a conjuntura social e econômica do país, como por exemplo desempenho da economia ou taxa de desemprego, também poderiam explicar a possível vitória da candidata.

b-) Supondo que estamos interessados em medir a probabilidade de sucesso da candidata Dilma Rouseff. Calcule a média e o desvio padrão para a amostra dessa pesquisa. Lembre-se de que estamos interessados no sucesso de uma candidata, apenas. Demonstre seus cálculos e interprete os resultados.

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 resulto 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	**Intenções de Voto (Frequência Absoluta)**
Dilma (PT)	1475
Aécio Neves (PSDB)	1294
Branco/Nulo	151
Não Sabem/ Não Opinaram	90

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

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} = \frac{(1475*1)+(1294*0)+(151*0)+(90*0)}{3010} = 0,49$$

Ou

$$\mu = \sum yP(y) = (1*0,53)+(0*0,47) = 0,49$$

3. Cálculo do Desvio Padrão:

$$s = \sqrt{\frac{\sum_{i}^{n}(y_i-\bar{y})^2}{n-1}} = \sqrt{\frac{1475(1-0,53)^2+1294(0-0,53)^2+151(0-0,53)^2+90(0-0,53)}{3010-1}} = 0,5$$

Interpretando os resultados: levando em consideração somente o interesse em medir o sucesso eleitoral da candidata Dilma Rouseff tem-se que, em média, a probabilidade do sucesso eleitoral da candidata será igual a 53%. Já o desvio – padrão nos indica que a distância típica dos dados em relação à média é de 0,5, 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?

1. Erro- Padrão

$$\sigma_{\overline{y}} = \frac{\sigma}{\sqrt{n}}$$

$$\sigma_{\overline{y}} = \frac{0{,}5}{\sqrt{3010}} = 0{,}009$$

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 3010 entrevistados irá variar 0,009 de uma amostra para outra.

2. Intervalo de Confiança $= \bar{Y} \pm 1 \; ou \; 2 \; ou \; 3 * \sigma_y$

68%

$\bar{Y} = 0{,}49 \pm$ 1*0,009 = O sucesso eleitoral da candidata Dilma Rousseff está entre 48,1% e 49,9%

95%

$\bar{Y} = 0{,}49 \pm$ 2*0,009 = O sucesso eleitoral da candidata Dilma Rousseff está entre 47,2% e 50,8%

99%

$\bar{Y} = 0{,}49 \pm$ 3*0,009 = O sucesso eleitoral da candidata Dilma Rousseff está entre 46,3% e 51,7%.

As pesquisas de intenção de votos, como a elucidada acima, têm como objetivo a opinião da população como um todo e não somente a opinião dos indivíduos presentes na amostra. A partir da pesquisa realizada pelo instituto de pesquisa Ibope, sabemos que a média do sucesso da candidata Dilma Rousseff, contudo não temos certeza se essa é a média para a população.

A tabela 2 abaixo traz a mesma pesquisa realizada para o Ibope, contudo somente considerando os votos válidos (desconsideram-se os votos em Branco/ Nulo e as "Não respostas". Abaixo, a figura 1 traz os resultados oficiais das eleições presidenciais de 2014 para o segundo turno:

Tabela 2. Pesquisa Ibope

Candidato	Intenções de Voto
Dilma (PT)	53%
Aécio Neves (PSDB)	47%

Figura 1: Resultados Oficiais

d-) Como podemos explicar a diferença entre os resultados da pesquisa do Ibope (tabela 2) e o resultado oficial das eleições? Essa diferença era prevista pelo Ibope? Justifique sua resposta. (Máximo de 8 linhas)

Aqui o aluno deveria se atentar para o enunciado da pesquisa do Ibope 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 53+2 = 55% e 53-2=51%. Sendo assim, os resultados do

Ibope poderiam variar entre 51% e 55% e o resultado oficial de 53% encontra-se dentro desse intervalo com 95% de confiança.

e-) Considerando que os dados apresentados na figura 1 referem-se aos dados da população e pensando a respeito do que aprendemos a respeito do Teorema do Limite Central, como podemos aproximar os resultados da pesquisa do Ibope dos resultados oficiais das eleições? (máximo de 8 linhas)

À luz do Teorema do Limite Central seria possível aproximar os resultados encontrados pelo Ibope 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)