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 06

Data de entrega: 25/04/2016 (noturno) e 27/04/2016 (vespertino)

Para essa lista, o termo "**calcule**" sempre implicará na apresentação no corpo do texto de todos os cálculos necessários para a obtenção dos valores especificados nas fórmulas correspondentes. Isso pode ser feito sem desconto de nota:

a) à mão, ou seja, sem o uso de ferramentas ou *softwares* estatísticos;
b) em ferramentas ou *softwares* de cálculo ou estatísticos, desde que todas as etapas estejam claras no desenvolvimento do exercício, podendo caso a(o) aluna(o) julgue necessário ser complementadas textualmente para esse fim.

Exercício 1 (3 pontos)

Marque "V" para verdadeiro e "F" para falso sobre as afirmações, indicando as justificativas para a escolha.

(V) A amplitude dos intervalos de confiança aumenta quando adotam-se níveis de confiança maiores e diminui com o aumento do tamanho das amostras.

Conforme Agresti & Finlay (2012, p. 139) a "amplitude de um intervalo de confiança (1) aumenta à medida que o nível de confiança aumenta e (2) diminui à medida que o tamanho da amostra aumenta" constituem duas propriedades que se aplicam a todos os intervalos de confiança.

(V) Amostras aleatórias grandes permitem a utilização da distribuição normal padrão em substituição à distribuição-t para intervalos de confiança de médias.

A utilização da distribuição normal padrão é utilizada para amostras aleatórias grandes quando observada a estimativa intervalar $\hat{\pi} \pm 1{,}96\sigma_{\hat{\pi}}$ para proporções, sendo necessárias ao menos 15 observações tanto na categoria de interesse quanto nas demais (AGRESTI & FINLAY, 2012, p. 140). Para médias, no entanto, a utilização do erro padrão estimado introduz um erro adicional que pode ser considerável em amostras pequenas, o que leva à utilização da distribuição-t para construção de intervalos de confiança. No entanto, a distribuição-t se aproxima da normal padrão a partir de 30 ou mais observações, apesar de suas caudas mais grossas e sua maior dispersão. (AGRESTI & FINLAY, 2012, p. 142)

(F) A violação do pressuposto da distribuição da população ser normal afeta a robustez da construção dos intervalos de confiança para uma média.

A construção de intervalos de confiança para médias utilizando a distribuição-t é robusta contra violações de normalidade da população, uma vez que, mesmo em casos em que a distribuição da população não seja normal, os intervalos de confiança baseados na distribuição-t funcionam bem. Quanto maior a amostra, menor a importância da suposição de normalidade, seguindo o Teorema do Limite Central (AGRESTI & FINLAY, 2012, p. 146)

(F) Por ser muito difícil satisfazer todos os pressupostos para construção de intervalos de confiança, pode-se dizer a utilização da distribuição-t é suficiente para conferir robustez aos intervalos de confiança construídos.

Como dito no gabarito da afirmação anterior, a distribuição-t é robusta para a suposição de normalidade da população. No entanto, a construção de intervalos de confiança baseados na distribuição-t não é robusta a violações de aleatorização, produzindo, dessa forma, resultados de validade questionável quando utilizados dados não aleatórios (AGRESTI & FINLAY, 2012, p. 146)

(F) É possível definir o tamanho da amostra para estimação de proporções e médias com critérios objetivos, desde que utilizadas grandes amostras.

Apesar da existência de fórmulas que auxiliam na definição do tamanho da amostra para estimação de proporções e de médias, esse processo também deve considerar outras questões como a complexidade da análise planejada ou limitações de tempo, dinheiro e recursos. A utilização de amostras pequenas não é inviabilizada, mas requer maiores

cuidados com violações de normalidade e valores atípicos. (AGRESTI & FINLAY, 2012, p. 152-153).

Exercício 2 (4 pontos)

Suponha que em uma pesquisa recente de opinião realizada semana passada, uma pesquisadora tenha descoberto que, de uma amostra aleatória de 10000 brasileiros e brasileiras, 6660 se mostraram contra o impeachment do vice-presidente Michel Temer e todas as restantes se mostraram a favor, não havendo não-respostas, nem pessoas que não quiseram opinar. Suponha ainda que não tenha havido erro de mensuração.

a) Imagine que essa pesquisadora quer saber a quantidade de pessoas que são a favor do impedimento (considerando essa categoria como "sucesso" e a outra como "falha") e (i) **classifique** essa variável (a quantidade de "sucessos"), de acordo com os valores que ela pode assumir; e (ii) desconsiderando totalmente o número total de indivíduos da amostra (n), e olhando apenas para suas outras características, **indique** o tipo de distribuição de valores ao qual ela converge dentre os estudados (ex.: Poisson, Binomial, Bernoulli, Normal, dentre outros), **justificando** a razão por trás de seu raciocínio.
(Máximo de 9 linhas)

A referida **quantidade** trata-se e variável quantitativa discreta, que pode convergir numa distribuição Binomial, já que parte de uma série de *n* provas de Bernoulli (*Bernoulli trials*) com sucessos e falhas, de probabilidades complementares cuja soma é igual a 1.

b) Considerando agora que a pesquisadora está interessada na porcentagem de indivíduos favoráveis ao impeachment, (i) **identifique** e (ii) **calcule** a estimativa por ponto mais apropriada para os fins da pesquisadora (entre média e proporção), justificando sua escolha e (iii) **considerando** e **comentando** sobre sua eficiência e tendenciosidade.
(Cálculo livre; máximo de 10 linhas para os comentários e justificativas)

O estimador de ponto apropriado é uma **proporção**, pois a porcentagem reporta-se a um resumo de uma categoria, e não ao centro de uma distribuição de variável quantitativa.
Cálculo:

$$\hat{\pi} = \frac{n_S}{n} \ , onde$$

n_S é a quantidade de indivíduos na categoria de "sucesso" (S) (favoráveis ao impeachment);

$\hat{\pi}$ é a proporção estimada de indivíduos favoráveis e

n é a quantidade total de indivíduos na amostra.

$$\hat{\pi} = \frac{10000 - 6660}{10000} = \frac{3340}{10000} = 33{,}4\%$$

Esse estimador **não é tendencioso**, pois sua distribuição amostral está centrada em torno do parâmetro populacional. Em outras palavras, considerando dados outros pressupostos do método de estimação de verossimilhança, a média das proporções estimadas não subestimaria ou superestimaria o parâmetro π. Considerando o tamanho grande da amostra, reafirmamos o argumento da não tendenciosidade e acrescentamos que o estimador é também **eficiente**, pois sua distribuição amostral se aproxima do parâmetro em questão, tendo pequena variabilidade da distribuição amostral.

c) Considere a seguinte fórmula, referente ao desvio padrão de uma distribuição de probabilidade discreta (desconsiderando aqui também o número de indivíduos da amostra) para uma variável aleatória y: $s = \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 P(y)}$

A partir dela, (i) **chegue na fórmula simplificada** do cálculo do desvio padrão de proporções, disponível nas leituras e no Laboratório dessa semana.

Dica: para um guia prático mais detalhado para a execução dessa etapa de derivação, ver o exercício 4.55 (especialmente item b), no capítulo 4 do livro de Agresti & Finlay, disponibilizado no Moodle. Perceba que a derivação parte pela demonstração de que $\mu = \bar{y}$.

Agora, (ii) **calcule** o desvio padrão referente à variável.

A variável aleatória y pode ser considerada a posição favorável ou contrária ao impeachment, atribuindo ao valor 1 ("sucesso") a probabilidade π e ao valor 0 ("fracasso") a probabilidade complementar $1 - \pi$.

Além disso, sabemos que $\bar{y}$ é igual à probabilidade de sucesso π. Senão, vejamos:

$$E(Y) = \bar{y} = 1 \cdot \pi + 0 \cdot (1 - \pi) = \pi$$

Dessa forma, partindo dessas informações e da fórmula do enunciado, temos:

$$s = \sqrt{(1 - (\pi))^2 \cdot \pi + (0 - \pi)^2 \cdot (1 - \pi)}$$

$$s = \sqrt{\pi(1-\pi)^2 + \pi^2(1-\pi)}$$

Utilizando a regra distributiva, temos:

$$s = \sqrt{\pi(1-\pi)((1-\pi)+\pi)}$$

$$s = \sqrt{\pi(1-\pi)}$$

Olhando agora para nossa variável, que mede a proporção de pessoas favoráveis ao impeachment de Temer, podemos utilizar a fórmula derivada ara calcular seu desvio padrão:

$$s = \sqrt{0{,}334(0{,}666)}$$

$s = 0{,}47$, aproximadamente.

Exercício 3 (3 pontos)

a) Ainda utilizando os mesmos dados do exercício anterior, **calcule** e **construa** dois intervalos de confiança para a variável: um com nível de confiança de 95% e outro com probabilidade de erro de 2%. **Interprete** os resultados.
(Cálculo livre; máximo de 8 linhas para a interpretação).

Para construir os intervalos de confiança, precisamos primeiro calcular o erro-padrão.
Para uma amostra de 10000 indivíduos, temos:

$$s_{\bar{y}} = \frac{s}{\sqrt{n}}$$

$$s_{\bar{y}} = \frac{0{,}47}{\sqrt{10000}}$$

$$s_{\bar{y}} = 0{,}0047$$

Nível de confiança de 95%:
Probabilidade de erro: 5%
z-crítico (encontrado a partir da tabela z): 1,96

$$IC_{95\%} = \hat{\pi} \pm z s_{\bar{y}}$$
$$IC_{95\%} = 0{,}334 \pm 1{,}96 \cdot 0{,}0047$$
$$IC_{95\%} = 0{,}334 \pm 0{,}009212$$
$$IC_{95\%} = [32{,}48\%, 34{,}32\%]$$

Probabilidade de erro de 2%:

Nível de significância: $\alpha = 0{,}02$.

Nível de confiança de 98%.

z-crítico (encontrado a partir da tabela z): aproximadamente 2,33.

$$IC_{98\%} = 0{,}334 \pm 0{,}010951$$
$$IC_{98\%} = [32{,}3\%, 34{,}49\%]$$

Ambos os intervalos de confiança foram construídos arredondando os valores para duas casas decimais.

Interpretação [(uma dentre muitas possíveis)]:

Dois intervalos de confiança foram construídos, um com probabilidade de erro (de tipo I, ou nível de significância) de 5% e outro de 2%. O valor do erro-padrão é bem baixo, visto que temos uma amostra com muitas observações (*n* grande). Conforme o nível de confiança aumenta e o nível de significância consequentemente diminui, como esperado o z-crítico e os intervalos tornam-se maiores, mas esse aumento não é tão substantivo (de menos de 0,2% para mais e para menos), pelo tamanho pequeno do erro-padrão. De fato, se arredondássemos ainda mais os valores dos percentuais, sequer vislumbraríamos essa diferença, ficando com o intervalo de [32%, 34%].

b) Imagine agora que, dada a nova conjuntura política dessa semana, a pesquisadora queira empreender uma nova pesquisa para verificar as mesmas informações, mas tenha orçamento mais limitado. Qual o número mínimo de indivíduos para os quais ela deve aplicar o questionário, supondo que ela não sabe nenhuma especificação sobre a nova distribuição de valores para a população e quer construir dois intervalos como os calculados no item 3.a, com margem de erro de dois pontos percentuais, para mais ou para menos? **Calcule** e **justifique** seu raciocínio.

Supomos novamente uma amostra aleatória sem erro de mensuração – o que, vale ressaltar, costuma ser mais uma suposição teórica que uma viabilidade prática. Tratando-se de uma proporção, podemos calcular o número de observações necessárias para o empreendimento da pesquisa a partir do **escore z**, da **margem de erro** e, de forma complementar e restritiva, **de informações acerca da proporção de indivíduos suposta para a população**, através da seguinte formula:

$$n = \pi(1-\pi) \cdot \left(\frac{Z}{M}\right)^2$$

Como uma pesquisa anterior fora empreendida no mesmo sentido, a princípio a informação sobre a proporção suposta poderia ser utilizada, mas devemos nos atentar para o fato exposto no enunciado de que ela **supõe não saber nenhuma especificação sobre a nova distribuição de valores para a população**. De fato, esse é um cuidado relevante, porque dadas as mudanças na conjuntura política (votação do impeachment da presidenta Dilma na Câmara dos Deputados), é muito razoável supor que a nova proporção seja bem diferente da primeira calculada. Quando não temos informações sobre a proporção que esperamos encontrar na população, podemos alternativamente utilizar a proporção conservadora que maximiza o valor de *n*, que é 50%.

$$\max_{0 \leq \pi \leq 1} n = \max_{0 \leq \pi \leq 1} \pi(1-\pi) \cdot \left(\frac{Z}{M}\right)^2 \mid \left(\frac{Z}{M}\right)^2$$

$$\max_{0 \leq \pi \leq 1} \pi(1-\pi) = 0{,}5(1-0{,}5)$$

$$\max_{0 \leq \pi \leq 1} \pi(1-\pi) = 0{,}25$$

A fórmula para o cálculo de *n* torna-se:

$$n = 0{,}25 \left(\frac{Z}{M}\right)^2$$

$M = 0{,}02$

Nível de confiança de 95%:

$$n_{95\%} = 0{,}25 \cdot \left(\frac{1{,}96}{0{,}02}\right)^2$$

$$n_{95\%} = 2401$$

Nível de confiança de 98%:

$$n_{98\%} = 0{,}25 \cdot \left(\frac{2{,}33}{0{,}02}\right)^2$$

$$n_{98\%} \approx 3393$$

Exercício 04 – Pós Graduação (5 pontos)

Segundo Brambor e Ceneviva (2012), os incumbentes possuem vantagens sobre os opositores em eleições em decorrência de vantagens relacionadas ao cargo ocupado, como exposição na mídia, recursos governamentais e maior facilidade para obtenção de financiamento eleitora, além da capacidade de levar a desafiantes competitivos à dissuasão de sua candidatura. Esse é um argumento que já foi discutido em listas anteriores e será retomado agora.

Em lista passada, foi apresentada uma tabela sobre a última pesquisa de opinião realizada pelo Ibope (Instituto Brasileiro de Opinião Pública e Estatística) para o segundo turno das eleições presidenciais do ano de 2014 realizada entre 24 e 25 de outubro de 2014. Esta pesquisa ouviu 3.010 eleitores em 206 municípios brasileiros, com uma margem de erro divulgada de 2 pontos percentuais para um nível de confiança de 95%. Reproduzimos a tabela a seguir.

Candidato	Intenção de voto
Dilma (PT)	49%
Aécio Neves (PSDB)	43%
Branco/Nulo	5%

Não sabem/Não opinaram	3%

Considerando apenas os votos válidos, a distribuição de votos seria de 53% para a candidata Dilma (PT) e 47% para Aécio Neves (PSDB). Os resultados eleitorais apontaram a vitória de Dilma com 51,64% contra 48,36% de Aécio Neves.

a) Utilizando apenas os votos válidos, é possível inferir vantagem do incumbente a partir da pesquisa realizada pelo Ibope a um nível de 95% de confiança? Justifique e represente graficamente (Dica: para o número de observações, desconsidere as observações de votos brancos/nulos e não sabem/não opinaram). **(2,0 pontos)**

Inicialmente, calculamos o número de observações a ser utilizado para o cálculo, ou seja, como as indicações de branco/nulo e não sabem/não opinaram combinam para 8%, consideramos 92% das 3.010 observações, sendo:

$$n = 3010 * 0{,}92 \cong 2769$$

O próximo passo é observar a proporção de votos a favor de algum dos candidatos. Como o interesse é pelo incumbente, tomaremos a proporção dos votos para a candidata Dilma (PT). Ela já é fornecida pelos 53% de intenção e voto na candidata e pode ser expressa como:

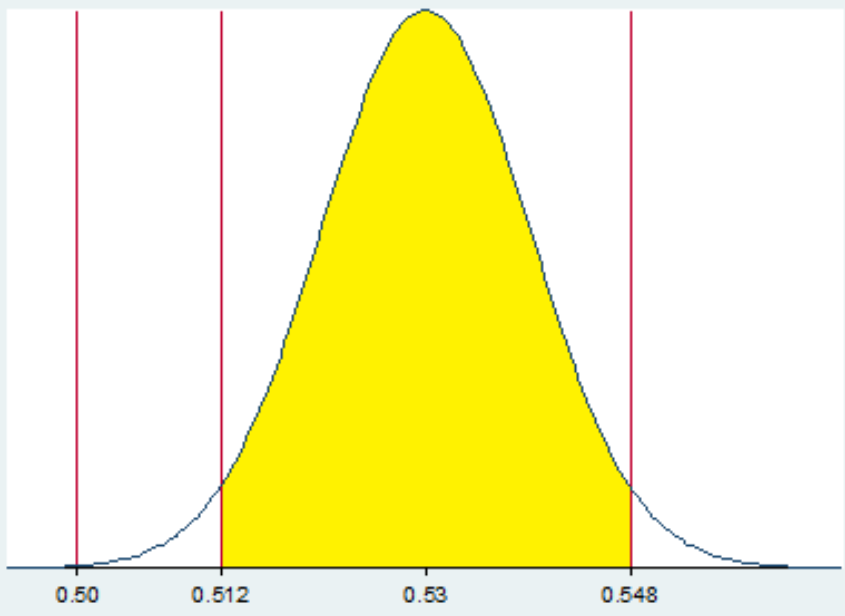
$$\hat{\pi} = 0{,}53$$

Com a proporção e o número de observações, podemos calcular o erro-padrão para construção do intervalo de confiança:

$$ep = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} = \sqrt{\frac{0{,}53(1-0{,}53)}{2769}} \cong 0{,}009$$

Considerando a proporção estimada e o erro padrão estimado, podemos construir os intervalos de confiança para 95%, utilizando o escore-z de 1,96.

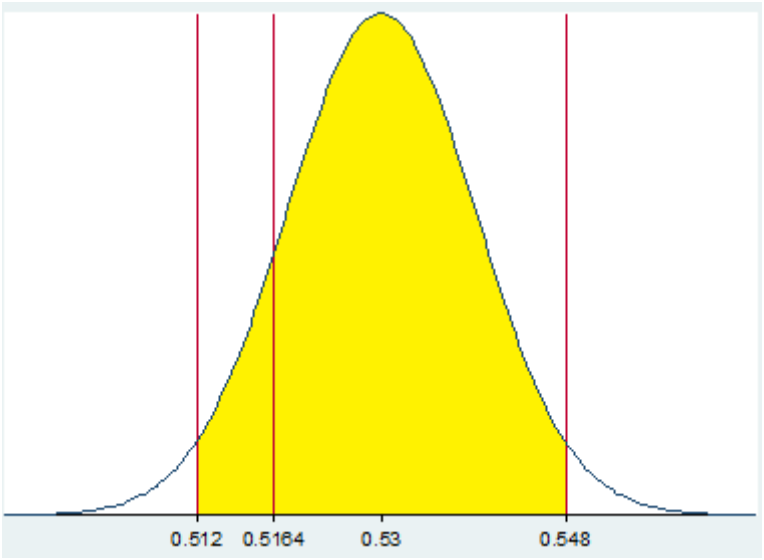
$$IC = 0{,}53 \pm 1{,}96 * 0{,}009$$
$$IC = 0{,}53 \mp 0{,}018$$
$$\therefore 0{,}512 \leq \hat{\pi} \leq 0{,}548$$



O intervalo de confiança indica que, a um nível de 95% de confiança, o incumbente sairia vitorioso. Dessa forma, a pesquisa vai ao encontro da ideia de vantagem do incumbente.

b) Pode-se dizer que a pesquisa realizada pelo Ibope foi capaz de prever a proporção de votos válidos recebida pela candidata Dilma? Justifique e represente grafiamente. **(1,0 ponto)**

A observação do intervalo de confiança da pesquisa realizada pelo Ibope é utilizada também para responder essa questão. Para isso, deve-se observar que a proporção de votos recebidos por Dilma é $\pi = 0{,}5164$, são os 51,64% dos votos recebidos. Como o intervalo de confiança vai de 0,512 a 0,548, o valor de 0,5164 está contido no intervalo de confiança a um nível de 95% de confiança.



c) Uma das críticas realizadas a muitas pesquisas eleitorais é a utilização de pontos de fluxo para coleta dos dados. Caso a pesquisa tenha utilizado essa forma de coleta, quais as consequências para a construção do intervalo de confiança? **(2,0 pontos)**

Como colocam Agresti e Finlay (2012) e Kellstedt e Whitten (2015), a construção dos intervalos de confiança utilizando amostras selecionadas aleatoriamente é vital para a construção de intervalos de confiança para médias em uma distribuição amostral. Isso ocorre porque a seleção de uma amostra não aleatória impede a utilização do Teorema do Limite Central. Dessa forma, os resultados obtidos com a construção dos intervalos de confiança não robustos à violação da aleatorização dos dados e, consequentemente, não trazem informações confiáveis para a análise. Como apontam Kellstedt e Whitten (2015, p. 162), por uma amostra claramente não aleatória não ser uma seleção aleatória da população, ela não diz "nada".

Boa Lista!