Gabarito - Lista 3

Stata - Exercícios de Fixação

Gabarito

Para fazer os exercícios desta lista, você precisará baixar o banco de dados `OpiniaoPublica.dta`, disponível no Moodle. Trata-se do resultado da pesquisa "Brasil, as Américas e o Mundo", conduzida em parceria pelo IRI e pelo CIDE, do México. A pesquisa busca compreender o posicionamento da opinião pública a respeito de diversos temas que dizem respeito às relações internacionais e à política externa, e faz parte de um projeto mais amplo envolvendo diversos países da América Latina. Para ter mais informações, acesse `http://lasamericasyelmundo.cide.edu/`

Importante: nas perguntas que exigirem o uso do Stata é necessário apresentar o comando utilizado e, quando relevante, o *output* obtido. Respostas incompletas terão pontuação parcial.

Exercício 1 [1 ponto]. Base de dados: descrição e noções básicas.

a Você classificaria essa base de dados como *cross-section*, painel ou série de tempo? Justifique em no máximo 50 palavras.

b Faça uma tabela que contenha as estatísticas descritivas básicas de todas as variáveis do banco (para este exercício, não precisa colar a tabela, que seria muito grande. Apenas mostre o comando).

c De maneira geral, o banco é formado por uma maioria de variáveis contínuas ou categóricas? Justifique e dê um exemplo do tipo de variável que você respondeu (máx. 50 palavras).

d Identifique a pergunta na qual os respondentes atribuem uma nota de 0 a 100 a líderes políticos do continente, e renomeie a variável que diz respeito a Cristina Kirchner para "kirchner" (sem aspas).

a Trata-se de uma base *cross-section*, pois temos variação apenas entre observações. Como não há variação no tempo, o banco de dados é um retrato da realidade no momento em que a pesquisa foi feita, e não nos permite observar tendências.

b O comando para gerar estatísticas descritivas básicas de todas as variáveis é o `summarize`.

c A grande maioria das variáveis do banco é categórica. Um tipo comum de variáveis categóricas são escalas Likert, em que o entrevistado responde se concorda muito, concorda, discorda ou discorda muito de uma afirmação. Escalas Likert podem ter três ou mais pontos, e um exemplo de escala de 4 pontos é a pergunta 16_1 (variável p16_1 no banco de dados).

d O nome original da variável sobre Cristina Kirchner é p37a, e o comando para renomeá-la é `rename p37a kirchner`

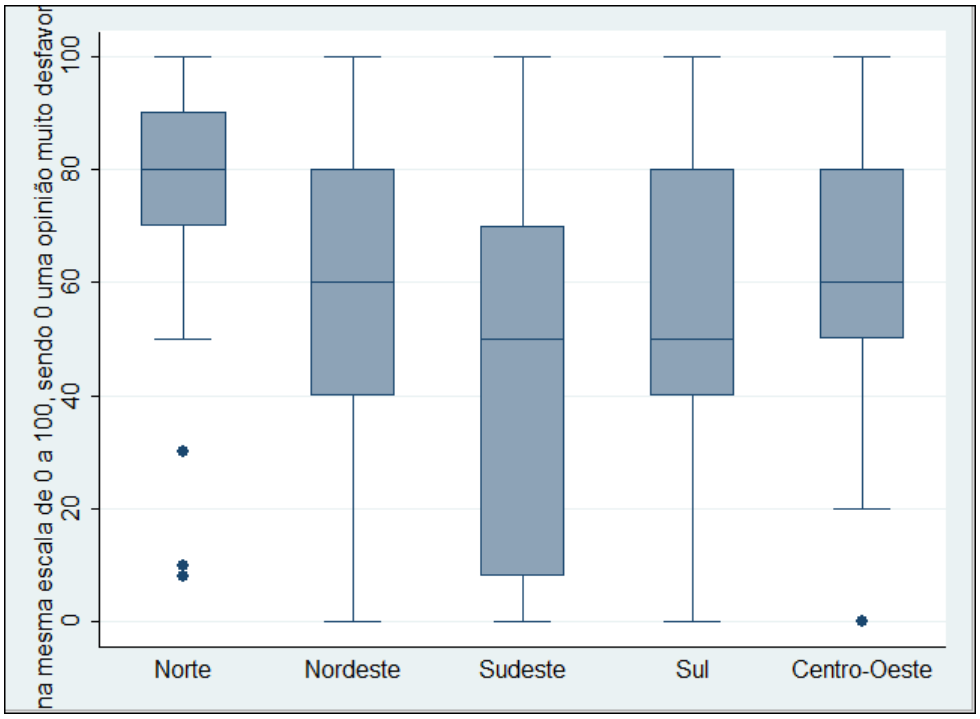
Exercício 2 [0.5 ponto]. Qual é o percentual de apoiadores do PT que consideram que o Brasil deveria procurar ser um líder na América Latina?

Existe mais de uma maneira de conseguir essa resposta. Uma delas é fazer uma tabulação cruzada entre a variável de identificação partidária (ps4) e a variável sobre a posição do Brasil na América Latina (p29), com o comando `tab ps4 p29, column row`. Na tabela abaixo, podemos observar que entre os respondentes que simpatizam com o PT, 51,1% responderam que o Brasil deve ser um líder da região

S4) Independentemente do partido que você tenha votado, normalmente você se cons	P29) Qual das seguintes informações se aproxima mais do que você pensa sobre o p					
	O Brasil	O Brasil	O Brasil	Não sabe	Não respo	Total
dem	5	3	1	0	0	9
	55.56	33.33	11.11	0.00	0.00	100.00
	0.53	0.49	0.51	0.00	0.00	0.48
psdb	80	56	15	10	0	161
	49.69	34.78	9.32	6.21	0.00	100.00
	8.48	9.20	7.61	7.52	0.00	8.52
pt	373	243	67	47	0	730
	51.10	33.29	9.18	6.44	0.00	100.00
	39.55	39.90	34.01	35.34	0.00	38.64
pv	21	19	5	1	0	46
	45.65	41.30	10.87	2.17	0.00	100.00
	2.23	3.12	2.54	0.75	0.00	2.44
Outro partido	42	12	0	4	0	58
	72.41	20.69	0.00	6.90	0.00	100.00
	4.45	1.97	0.00	3.01	0.00	3.07
Não se identifica com	422	276	109	71	7	885
	47.68	31.19	12.32	8.02	0.79	100.00
	44.75	45.32	55.33	53.38	100.00	46.85
Total	943	609	197	133	7	1,889
	49.92	32.24	10.43	7.04	0.37	100.00
	100.00	100.00	100.00	100.00	100.00	100.00

Exercício 3 [0.5 ponto]. Faça um *boxplot* comparando a avaliação dos entrevistados (de 0 a 100) sobre o Mercosul nas cinco regiões do país.

```
graph box p44g, over(regiao)
```



Exercício 4 [1 ponto]. Faça uma tabela comparando a avaliação sobre a Organização das Nações Unidas (de 0 a 100) entre as pessoas que acertaram e aquelas que erraram o significado da sigla ONU. O que os resultados indicam?

O comando para gerar a tabela é `tab p12b, sum(p44a)`. Os resultados (cf. tabela na próxima página) mostram que os entrevistados que não reconheceram a sigla da ONU têm uma opinião mais positiva sobre a instituição do que aqueles que acertaram. Não é possível dizer por que isso acontece, mas deve-se notar que ainda não podemos dizer que essa diferença é estatisticamente significativa. Em outras palavras, trata-se meramente de uma diferença na amostra, e ainda não podemos dizer que essa diferença pode ser projetada na população. Esse segundo passo, que é o mais interessante em termos substantivos, será dado a partir das aulas de inferência estatística

P12) Por favor, diga-me o significado das siglas que você vê no seguinte cartão:	Summary of P44) E na mesma escala de 0 a 100, sendo 0 uma opinião muito desfavorável, 100 u		
	Mean	Std. Dev.	Freq.
Correto	57.614773	32.672547	880
Incorreto	61.336245	32.603755	229
Não sabe	60.294011	28.943593	551
Não respo	61.666667	40.70217	6
Total	59.027011	31.514635	1666

Exercício 5 [1 ponto]. Quantos entrevistados que votaram no PSDB ou DEM na eleição passada acham que o Brasil deve reduzir barreiras comerciais?

170 pessoas se encaixam no perfil pedido pelo enunciado. Vamos mostrar aqui duas estratégias (mas não as únicas) para encontrar essa resposta.

Resposta 1: `count if (ps5 == 1 | ps5 == 2) & (p24 == 1 | p24 == 2)`
A primeira estratégia é utilizar um código sucinto, mas um pouco complexo. A vantagem é que encontramos a resposta em apenas uma linha e sem criar nenhuma variável adicional. A desvantagem é que ela exige a manipulação de 7 operadores lógicos e o uso correto do ordenamento de operações.

Resposta 2: A segunda alternativa é cognitivamente mais simples, mas dá um pouco mais de trabalho. Primeiramente, será necessário criar uma nova variável que diferencia quem votou no PSDB ou no DEM dos demais. Vamos chamar essa nova variável de `psdb_dem`:

```
recode ps5 (1/2 = 1) (else = 0), gen(psdb_dem)
```

Na sequência, vamos criar uma nova variável juntando todas as pessoas que concordam um pouco ou concordam muito com a redução de barreiras tarifárias. Chamaremos essa nova variável de `protecionismo`:

```
recode p24 (1/2 = 1) (else = 0), gen(protecionismo)
```

Por fim, vamos criar uma tabela cruzada entre essas duas novas variáveis. A intersecção entre as pessoas que votaram no PSDB/DEM e as pessoas que defendem menos protecionismo comercial é de 170 pessoas (cf. tabela na próxima página):

```
tab psdb_dem protecionismo
```

RECODE of ps5 (S5) Para qual partido você votou na eleição presidenci al passada?	RECODE of p24 (P24) Você concordaria ou não se o Brasil reduzisse as barreiras d		
	0	1	Total
0	629	1,103	1,732
1	98	170	268
Total	727	1,273	2,000

Exercício 6 [2 pontos]. Obtenha as estatísticas descritivas básicas da variável "kirchner", renomeada no item *d* do exercício 1. O que te faz desconfiar desses dados? Identifique o problema e faça a correção apropriada.

Se rodarmos o comando `summarize kirchner`, teremos o seguinte resultado:

Variable	Obs	Mean	Std. Dev.	Min	Max
kirchner	2000	688.2625	447.3636	0	999

Se nos lembrarmos que a pergunta pede para o entrevistado avaliar cada líder político de 0 a 100, duas coisas soltam imediatamente aos olhos: tanto a média quanto o valor máximo da variável são maiores do que 100.

Se olharmos o questionário original, veremos que as respostas `Não sabe`, `Não respondeu` e `Nunca ouviu falar` são codificadas como 998, 999 e 997, respectivamente. Esses valores puxarão a média para cima a não ser que deixemos claro que elas devem contar como *missing data*. Um jeito de fazer isso é utilizar o seguinte comando:

```
recode kirchner 997 998 999 = .
```

Agora, se pedirmos a estatística descritiva novamente, os resultados serão coerentes:

Variable	Obs	Mean	Std. Dev.	Min	Max
kirchner	647	41.91963	28.53253	0	100

Exercício 7 [2 pontos]. Um pesquisador trabalha com a hipótese de que pessoas que se identificam no campo da direita são mais extremistas do que aquelas que se identificam no campo da esquerda. Você ainda não tem as ferramentas para testar formalmente essa hipótese, mas o banco de dados pode ajudar a dar os primeiros passos. Construa uma tabela com a qual seja possível observar se, entre os respondentes da pesquisa, a hipótese tem mérito. Os resultados indicam que o pesquisador está na direção certa?

O banco de dados não tem variáveis de extremismo político ou de classificação ideológica binária, então o desafio é perceber que ambas podem ser criadas a partir da variável `ps12`, que é uma escala de 0 a 10 em que 0 é extrema esquerda e 10 é extrema direita.

Assim, a criação da variável esquerda/direita é simples: basta codificar todas as respostas de 0 a 5 como esquerda e todas as respostas de 6 a 10 como de direita.

```
recode ps12 (0/5 = 0 Esquerda) (6/10 = 1 Direita), gen(DirEsq)
```

Note que, neste caso, seria errado codificar a nova variável usando o argumento `(0/5 = 0 Esquerda) (else = 1 Direita)`, pois essa variável tem um número razoável de *missing data*, e todos eles seriam convertidos para "Direita". É sempre importante tomar cuidado com o que você está pedindo no `recode`

A criação da variável de extremismo é operacionalmente tão simples quanto a anterior, mas precisamos escolher um ponto de corte relativamente arbitrário. Para este exemplo, vamos classificar como extremistas os valores de 0 a 2 e 8 a 10, e como centristas os valores de 3 a 7. Os pontos de corte da sua resposta podem ser diferentes, desde que coerentes com a ideia de extremistas vs centristas.

```
recode ps12 (0/2 8/10 = 0 Extremista) (3/7 = 1 Centrista), gen(CentrExtr)
```

Finalmente, fazemos uma tabela cruzada para comparar as duas variáveis:

```
tab DirEsq CentrExtr, column row
```

RECODE of ps12 (S12) Em uma escala de 0 a 10, em que 0 significa politicamente a	RECODE of ps12 (S12) Em uma escala de 0 a 10, em que 0 significa politicamente a		
nte a	Extremist	Centrista	Total
Esquerda	110	563	673
	16.34	83.66	100.00
	20.45	56.30	43.76
Direita	428	437	865
	49.48	50.52	100.00
	79.55	43.70	56.24
Total	538	1,000	1,538
	34.98	65.02	100.00
	100.00	100.00	100.00

Nessa tabela, fica claro que os dados *desta amostra* confirmam a suspeita do pesquisador. De fato, entre os entrevistados que classificamos de extremistas, 80% são de direita, enquanto 20% são de esquerda. Olhando de outra maneira, entre os esquerdistas temos 16% de extremistas e 84% de centristas, enquanto entre os direitistas essa divisão é 50%–50%.

Por fim, um comentário colateral sobre testes feitos com base em cortes arbitrários, como nos casos acima. Em um trabalho acadêmico, seu ponto de corte geralmente tem que ser justificado com base na literatura, e não é bom que seus resultados dependam demais do ponto de corte que você escolheu. Sempre se faça a seguinte pergunta: se eu mudar meu corte um ponto para cima ou para baixo, meus resultados mudam substancialmente? Se sim, sua análise não é robusta. Na literatura empírica, chamamos de *teste de sensibilidade* a prática de verificar se a análise é robusta a diferentes pontos de corte.

Exercício 8 [2 pontos]. Crie uma variável *dummy* cujo valor é 1 se o respondente tem ensino superior completo, e 0 caso contrário. Em seguida, crie rótulos de valores para essa nova variável, de maneira que valores 1 tenham o rótulo "Com Ensino Superior" e os valores 0 tenham o rótulo "Sem Ensino Superior".

Dica: No laboratório de introdução ao Stata, mostramos como incluir rótulos para as variáveis, mas não mostramos como incluir rótulos para valores específicos das variáveis. Neste exercício, você terá que pesquisar como fazer isso. Utilize os comandos de ajuda para consultar a documentação.

O comando para criar a *dummy* é `recode ps3 (1/8 = 0) (9 = 1), gen(SupComp)`.

Para criar os rótulos de valor, um exemplo de busca na documentação é `findit variable value label`. A partir do comando `label define` podemos incluir os dois rótulos de valor:

```
label define SUPERIOR 1 "Com Ensino Superior", add
```

```
label define SUPERIOR 0 "Sem Ensino Superior", add
label values SupComp SUPERIOR
```

Ou, como alternativa mais fácil, simplesmente defina os rótulos na hora de criar a *dummy*: `recode ps3 (1/8 = 0 "Sem Ensino Superior") (9 = 1 "Com Ensino Superior"), gen(SupComp)`