Desvendando os Mistérios do Coeficiente de Correlação de Pearson (r)*

Dalson Britto Figueiredo Filho

Universidade Federal de Pernambuco (UFPE)

*José Alexandre da Silva Júnior***

Universidade Federal de Pernambuco (UFPE)

Resumo: Existe relação entre X e Y? Essa é uma pergunta recorrente no cotidiano de qualquer pesquisador. O objetivo desse trabalho é discutir o conceito de correlação de Pearson (r) a partir de uma lógica intuitiva. Isso porque muitas vezes os livros de Estatística e/ou Econometria apresentam esse conceito adotando uma abordagem mais técnica, o que dificulta a compreensão. O texto apresenta as principais propriedades do coeficiente de correlação de Pearson (r), suas respectivas aplicações e limites a partir de uma abordagem descritiva. Em termos substantivos, espera-se facilitar a compreensão desse conceito nas ciências sociais em geral e na ciência política em particular.

* Esse artigo é o primeiro resultado do Projeto "*Political Science Quotation Database*" desenvolvido conjuntamente pelos autores. Além disso, esse trabalho se beneficiou dos comentários dos participantes do V Seminário de Ciência Política e Relações Internacionais da UFPE. Em especial, gostaríamos de agradecer a Giuseppe Lobo (UFMG) pelo apoio logístico, a Natalia Leitão pela leitura atenta de diferentes versões anteriores e ao parecerista anônimo da Revista Política Hoje por importantes sugestões. Assumimos total responsabilidade pelos erros remanescentes. Esse trabalho é financiado por duas principais fontes: CAPES e CNPQ.

** Ambos os autores são doutorandos em Ciência Política pela Universidade Federal de Pernambuco (UFPE).

1. Introdução

Existe relação entre X e Y? Essa é uma pergunta recorrente na vida de qualquer pesquisador. Por exemplo, ao afirmar que a taxa de suicídio entre protestantes é maior do que entre católicos, Durkheim sugere uma correlação entre denominação religiosa e propensão ao autocídio. Da mesma forma, ao postular que o sistema eleitoral majoritário tende a produzir sistemas bipartidários, a Lei de Duverger sugere a existência de uma correlação entre o tipo de regra eleitoral (majoritária ou proporcional) e a quantidade de partidos. Mas o que significa dizer que duas variáveis estão correlacionadas?[1] Essa é a questão de pesquisa que norteia esse trabalho.

Uma motivação adicional que orienta esse artigo é a "hostilidade em relação aos métodos quantitativos e à estatística [na ciência social brasileira]" (Soares, 2005: 27). Um rápido passeio nos textos de Werneck Vianna *et al* (1988), Valle e Silva (1999) e Santos e Coutinho (2000) corrobora esse diagnóstico. Isso porque os dados levantados por esse autores apontam para uma mesma direção: a utilização de técnicas básicas de estatística descritiva e inferencial ainda é bastante limitada na Ciências Sociais brasileira. De forma mais preocupante, essa análise se mantém consistente independente do tipo de produção (artigo, dissertações ou teses). O resultado prático disso é o enfraquecimento metodológico generalizado, o que por sua vez, influencia negativamente a capacidade das ciências sociais explicarem os fenômenos que elas se propõem.

[1] Esse é um debate polêmico na Estatística. Para o leitor interessado em aprofundar seus conhecimentos na área ver Aldrich (1995), Andres, Tejedor e Mato (1995), Blyht (1994), Carroll (1961), Devlin, Gnanadesikan e Kettering (1975), Kronmal (1993), Muddapur (1988), Niles (1921), O'Brien (1979), Pearson, Fisher e Inman (1994), Rodgers e Nicewander (1988), Schield (1995) e Stigler (1989). Para uma aplicação prática utilizando o SPSS ver Pallant (2007). Para uma aplicação prática utilizando o STATA ver Pollock (2006).

Consideramos que o método funciona como a "lente" que o pesquisador utiliza para auxiliar a teoria no sentido de interpretar e explicar os fenômenos de seu interesse[2]. Para King, Keohane e Verba (1994), "a substância da ciência é primordialmente os métodos e as técnicas" (King, Keohane e Verba, 1994: 09). Collier, Seawright e Munck (2004) defendem que "a credibilidade dos métodos empregados deve ser um critério central para avaliar os resultados de pesquisa" (Collier, Seawright e Munck, 2004: 23). Dessa forma, partindo do pressuposto de que o método é um componente central do conhecimento científico, esse artigo tem dois principais objetivos: (1) discutir o conceito de correlação de Pearson (r) a partir de uma lógica intuitiva. Isso porque muitas vezes os livros de Estatística e/ou Econometria apresentam esse conceito adotando uma abordagem mais técnica, o que dificulta a compreensão (Field, 2005); (2) chamar a atenção dos pesquisadores para as aplicações e os limites dessa medida na formulação dos seus desenhos de pesquisa.

Para tanto, o artigo está dividido em cinco seções. A primeira define o conceito e apresenta as principais propriedades do coeficiente de correlação de Pearson (r)[3]. A segunda seção demonstra, passo a passo, como essa medida é calculada. O objetivo é oferecer ao leitor a lógica intuitiva do processo. A terceira parte apresenta alguns cuidados básicos que os pesquisadores devem tomar durante a utilização dessa estatística na análise de seus dados. A quarta seção oferece um exemplo prático da aplicação e dos limites dessa medida. Por fim, a quinta parte sumariza as conclusões desse artigo.

---

[2] Para se aprofundar nesse debate sugerimos o seguinte: Almond (1990), Collier, Seawright e Munck (2004), Geddes (2003), Gerring (2001), King, Keohane e Verba (1994), Marsh e Stoker (2002) e Van Evera (1997).

[3] Como nosso principal objetivo é pedagógico, procuramos minimizar a formalização algébrica dos conceitos. Para o leitor interessado em um maior grau de detalhamento técnico sugerimos conferir a bibliografia citada.

1.1 Definição e Propriedades

O coeficiente de correlação de Pearson não tem esse nome por acaso. É comum atribuir exclusivamente a Karl Pearson o desenvolvimento dessa estatística, no entanto, como bem lembrou Stanton (2001), a origem desse coeficiente remonta o trabalho conjunto de Karl Pearson e Francis Galton (Stanton, 2001: 01). Garson (2009) afirma que correlação "é uma medida de associação bivariada (força) do grau de relacionamento entre duas variáveis". Para Moore (2007), "A correlação mensura a direção e o grau da relação linear entre duas variáveis quantitativas" (Moore, 2007: 100/101). Em uma frase: o coeficiente de correlação de Pearson (r) é uma medida de associação linear entre variáveis. Sua fórmula é a seguinte:

$$r = \frac{1}{n-1}\sum\left(\frac{xi - \bar{X}}{sx}\right)\left(\frac{yi - \bar{Y}}{sy}\right)$$

Dois conceitos são chaves para entendê-la: "associação" e "linearidade". Afinal, o que significa dizer que duas variáveis estão associadas? Em termos estatísticos, duas variáveis se associam quando elas guardam semelhanças na distribuição dos seus escores. Mais precisamente, elas podem se associar a partir da distribuição das freqüências ou pelo compartilhamento de variância. No caso da correlação de Pearson (r) vale esse último parâmetro, ou seja, ele é uma medida da variância compartilhada entre duas variáveis. Por outro lado, o modelo linear supõe que o aumento ou decremento de uma unidade na

variável X gera o mesmo impacto em Y[4]. Em termos gráficos, por relação linear entende-se que a melhor forma de ilustrar o padrão de relacionamento entre duas variáveis é através de uma linha reta. Portanto, a correlação de Pearson (r) exige um compartilhamento de variância e que essa variação seja distribuída linearmente[5].

1.2 Como interpretar?

O coeficiente de correlação Pearson (r) varia de -1 a 1. O sinal indica direção positiva ou negativa do relacionamento e o valor sugere a força da relação entre as variáveis. Uma correlação perfeita (-1 ou 1) indica que o escore de uma variável pode ser determinado exatamente ao se saber o escore da outra. No outro oposto, uma correlação de valor zero indica que não há relação linear entre as variáveis[6].

Todavia, como valores extremos (0 ou 1) dificilmente são encontrados na prática é importante discutir como os pesquisadores podem interpretar a magnitude dos coeficientes. Para Cohen (1988), valores entre 0,10 e 0,29 podem ser considerados pequenos; escores entre 0,30 e 0,49 podem ser considerados como médios; e valores entre 0,50 e 1 podem ser interpretados como grandes. Dancey e Reidy (2005) apontam para uma classificação ligeiramente diferente: r = 0,10 até 0,30 (fraco); r = 0,40 até 0,6 (moderado); r = 0,70 até 1 (forte). Seja como for,

[4] O modelo linear esta baseado na função linear, um caso particular da função afim, que tem domínio de ¡ (f: ¡ − ¡) definida por f(x) = ax para todo x ∈ a R onde a ≠ 0.

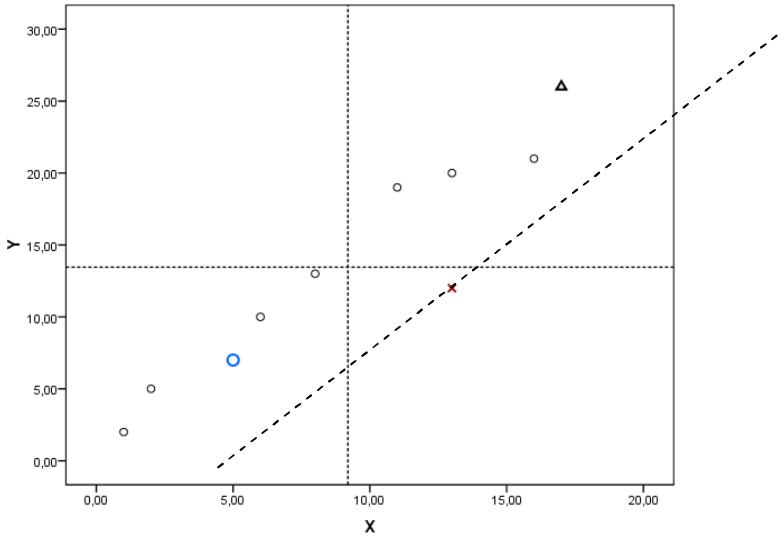
[5] Uma associação entre duas variáveis pode ser descrito por outros modelos, como por exemplo, o quadrático: f: ¡ − ¡ definida f(x) = ax²+b+c, onde existem números reais,

a, b, c com a ≠ 0 para todo x ∈ R.

[6] Para acompanhar o debate ver Aldrich (1995), Haig (2007) e Kozak (2009).

o certo é que quanto mais perto de 1 (independente do sinal) maior é o grau de dependência estatística linear entre as variáveis. No outro oposto, quanto mais próximo de zero, menor é a força dessa relação. O gráfico de dispersão abaixo apresenta um exemplo de uma relação linear entre duas variáveis hipotéticas X e Y.

Gráfico 01 - Correlação linear entre X e Y



Como pode ser observado, há uma correlação linear positiva entre X e Y. Detalhadamente, isso implica que quando um escore está acima da média de X espera-se que ele também esteja acima da média de Y (as linhas pontilhadas representam as médias das respectivas variáveis, sendo 9,20 para X e 13,5 para Y). Por exemplo, ao se considerar o triângulo preto, observa-se que ele está acima da média em ambas as variáveis (17; 26). No outro oposto, ao saber que o círculo azul está abaixo da média de X, observa-se que ele também está abaixo da média de Y (5; 7). Em quase todas as oportunidades que X assumiu um valor acima da média Y também o fez. Da mesma forma, quase todas as vezes que X ficou abaixo da média Y também ficou. A única exceção fica por conta da cruz

vermelha já que essa observação está acima da média em X, mas ficou abaixo do termo médio em Y (13;12). Se ela fosse excluída da análise chegaríamos a um coeficiente de valor igual a 1, ou seja, haveria uma compartilhamento de 100% da variância entre X e Y.

1.3 Propriedades: efeitos e desvios

Uma vez definido o conceito e fornecida a sua interpretação é importante entender algumas de suas propriedades. Baseado em Moore e McCabe (2004), destacamos as propriedades do coeficiente e as condições que precisam ser satisfeitas para realizar a análise de correlação de Pearson (r). Portanto, as observações são as seguintes:

1) *O coeficiente de correlação de Pearson não diferencia entre variáveis independentes e variáveis dependentes*. Dessa forma, o valor da correlação entre X e Y é o mesmo entre Y e X. Schield (1995) lembra que a correlação não se aplica a distinção de causalidades simples ou recursiva. Ou seja, por ela dificilmente pode-se afirmar quem varia em função de quem. Simplesmente pode-se dizer que há semelhanças entre a distribuição dos escores das variáveis[7].
2) *O valor da correlação não muda ao se alterar a unidade de mensuração das variáveis.* Por ser tratar de uma medida padronizada, o valor da correlação entre quilos e litros será o mesmo caso o pesquisador utilize toneladas e mililitros[8]. Padronização torna possível a comparação entre diferentes

[7] Correlação não deve ser confundida com relação de causa e efeito (causalidade). Para uma análise mais detalhada ver Asher (1983), Blalock (1971), Holland (1986) e Rubin (1974).
[8] Para uma discussão mais detalhada ver Carroll (1961).

variáveis no que diz respeito a sua magnitude e dispersão. Para tanto, deve-se subtrair cada observação (X) pela média (μ) e dividir o resultado pelo desvio padrão (σ)[9]. A média será zero com desvio padrão assumindo valor 1. Algebricamente,

$$z = \frac{x - \mu}{\sigma}$$

3) *O coeficiente tem um caráter adimensional, ou seja, ele é desprovido de unidade física que o defina.* Não faz sentido interpretar uma correlação de 0,3 como sendo 30%, por exemplo. Além disso, ele não se refere à proporção. Logo, uma correlação de 0,4 não pode ser interpretada como representando o dobro de uma correlação de 0,2 (Chen e Popovic, 2002: 09);

Para além das propriedades do coeficiente, algumas condições precisam ser satisfeitas:

4) *A correlação exige que as variáveis sejam quantitativas (contínuas ou discretas).* Não faz sentido utilizar a correlação

[9] O desvio padrão é uma medida de dispersão dos valores em torno da média. Quanto maior o seu valor, maior é o grau de heterogeneidade dos casos vis-à-vis o valor da média. Quanto menor, mais homogênea é a distribuição dos casos em torno do termo médio.

de Pearson (r) para dados categóricos já que é impossível calcular o desvio padrão da variável sexo, por exemplo[10].

5) Os *valores observados precisam estar normalmente distribuídos*[11]. Dessa forma, assume-se que:

$$N\ (\mu, \sigma)$$

Esse pressuposto é especialmente importante em amostras pequenas (N<40). Isso porque, a partir do Teorema do Limite Central, sabe-se que na medida em que o número de observações aumenta, a distribuição das médias amostrais se aproxima da curva normal, independente do formato da distribuição dos dados na população.

6) *Faz-se necessário uma análise de outliers, o coeficiente de correlação é fortemente afetado pela presença deles.* A presença de *outliers* pode comprometer fortemente as estimativas dos pesquisadores, levando inclusive a cometer erros do tipo I ou do tipo II.

7) *Faz-se necessária a independência das observações, ou seja, a ocorrência de uma observação X_1 não influencia a*

[10] Para dados categóricos deve-se utilizar a correlação de Kendall's tau-b ou Spearman. Para uma abordagem prática ver Pallant (2007). Para uma discussão mais aprofundada ver Tabachnick e Fidell (2007).

[11] Existem diferentes testes para estimar a normalidade da distribuição dos dados. Por exemplo, no teste de Kolmogorov-Smirnov um resultado não significativo (p>0,05) indica normalidade. Outros testes de normalidade incluem Anderson-Darling, Cramer-von Mises e Shapiro-Wilk. Graficamente, a normalidade pode ser observada a partir de histogramas e Q-Q plots. Agradecemos ao parecerista anônimo por nos lembrar desse detalhe.

> *ocorrência de outra observação* X_2. Segundo Schield (1995), a violação desta orientação implica risco de assumir correlações espúrias. Em termos mais técnicos, o pesquisador pode enfrentar o problema de *lurking* ou *counfouding variables.*

Para Osborne e Waters (2002), a violação desses pressupostos pode comprometer os resultados, levando o pesquisador a cometer os erros do tipo I ou tipo II (Osborne e Waters, 2002: 01). O erro do tipo I consiste em concluir que a hipótese nula é falsa quando ela é verdadeira. Logo, não existe relação entre as variáveis (H_o é verdadeira), mas o pesquisador argumenta que X e Y são estatisticamente dependentes. Ou seja, ele não poderia ter rejeitado a hipótese nula. O erro do tipo II consiste em concluir que a hipótese nula é verdadeira quando ela é falsa. Logo, existe relação entre X e Y (H_o é falsa), mas o pesquisador defende que as variáveis são estatisticamente independentes. Ou seja, ele deveria ter rejeitado a hipótese nula[12].

1.4 Calculando o coeficiente de correlação de Pearson (r)

Uma vez apresentada a sua definição e compreendida as suas propriedades o próximo passo é entender como o coeficiente de correlação é calculado. Suponha que um pesquisador esteja interessado

[12] Em estatística a hipótese nula (H_o) descreve o comportamento esperado de um determinado conjunto de dados. No teste de hipótese, o pesquisador procura estimar em que medida as evidências coletadas permitem rejeitar a hipótese nula em função da hipótese alternativa H_a (em geral a hipótese de pesquisa) ou não. Por exemplo, suponha que a Ho: μ=10. A hipótese alternativa (H_a) pode assumir que: $H_a > 10$; $H_a < 10$ (teste unicaldal) ou $H_a \neq 10$ (teste bicaudal).

em analisar a relação entre duas variáveis X e Y. A tabela abaixo ilustra esses dados[13].

Tabela 01 - Variáveis X e Y

ID	X	Y
1	29	0,49
2	40	1,59
3	54	1,69
4	55	1,82
5	72	3,10
Média	**50**	**1,738**

A primeira coluna (ID) registra a identificação de cada observação. O primeiro passo para estimar o coeficiente de correlação de Pearson é padronizar as observações, ou seja,

$$Z_x = \frac{X_1 - \bar{X}}{S_x}$$

onde X_1 representa o valor da observação 01 (29), $\bar{X}$ representa a média (50) e S_x indica o valor do desvio padrão (16,32). O mesmo deve ser feito para Y. Depois disso, o pesquisador deve somar o produto cruzado dos valores padronizados de X e Y (Z_x * Z_y). A tabela abaixo ilustra esse procedimento.

[13] Esses dados foram retirados de Moore (2007).

Tabela 02 - Variáveis padronizadas (Z_x e Z_y)

ID	Z_x	Z_y	Z_x *
1	-1,286	-1,345	1,730
2	-0,613	-0,160	0,098
3	0,245	-0,052	-0,013
4	0,306	0,008	0,027
5	1,348	1,46	1,978

A terceira coluna ilustra os produtos de Z_x * Z_y. A soma dos produtos (1,730 + 0,098 + -0,013 + 0,027 + 1,978) resulta em 3,821. Para finalizar o cálculo deve-se aplicar a fórmula[14]:

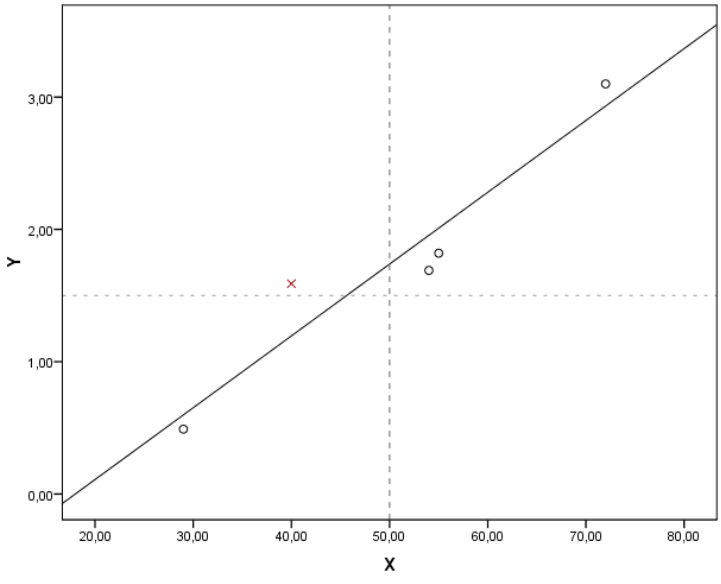
$$r = \frac{1}{n-1}\sum\left(\frac{xi - \bar{X}}{sx}\right)\left(\frac{yi - \bar{Y}}{sy}\right)$$

O resultado encontrado é de = 0,955. Ou seja, existe uma correlação forte e positiva entre X e Y[15]. O gráfico abaixo ilustra esses dados.

[14] Rodgers e Nicewander (1988) apresentam 13 diferentes fórmulas para estimar o coeficiente de correlação.

[15] Na prática, o pesquisador não precisa se preocupar em calcular essa medida já que os diferentes pacotes estatísticos fazem isso de forma rápida e eficiente. No entanto, consideramos importante entender a *rationale* do procedimento. Para uma excelente introdução ao coeficiente de correlação ver Chen e Popovic (2002). Para um site bastante informativo ver http://faculty.chass.ncsu.edu/garson/PA765/statnote.htm.

Gráfico 02 – Correlação entre X e Y



Quando X está acima da média espera-se que Y também esteja (as linhas pontilhadas representam as médias das respectivas variáveis, sendo 50 para X e 1,738 para Y). A única exceção fica por conta da observação 02 (cruz vermelha) na medida em que ela está acima da média de Y (1,59), mas abaixo da média em X (40). Se ela fosse excluída da análise chegaríamos a um coeficiente de valor igual a 1. E o que aconteceria se uma observação assumisse um valor muito distante da média? O que acontece com o coeficiente de correlação de Pearson (r) quando existe um *outlier* na amostra? Essas questões serão abordadas na próxima seção.

1.5 Cuidados básicos: *outliers* e *lurking variables*

O coeficiente de correlação de Pearson (r) é fortemente influenciado pela média da distribuição. Por esse motivo, um dos pressupostos centrais para que essa medida seja adequadamente utilizada é de que as observações obedeçam a uma distribuição normal. Existem

testes disponíveis para averiguar em que medida as observações estão normalmente distribuídas, sendo o teste de Kolmogorov-Smirnov e a observação gráfica dos dados um dos procedimentos mais comumente utilizados. No caso do teste, um resultado não significativo ($p>0,05$) indica normalidade. Caso o p valor assuma valores abaixo desse patamar ($p<0,05$), isso é um indicativo de que o pressuposto da normalidade foi violado. Em relação à análise gráfica, é comum a utilização de histogramas e Q-Q *plots* para analisar o formato da distribuição. Em relação ao histograma, o pesquisador deve observar em que medida a distribuição dos seus dados se aproxima da curva normal.

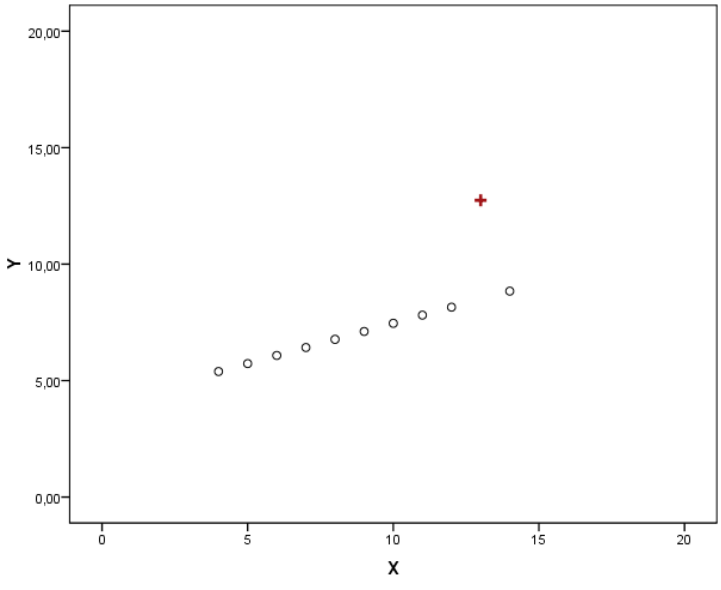
A presença de *outliers* tende a distorcer o valor da média e, por conseqüência, o valor do coeficiente de correlação. Dessa forma, a presença de *outliers* pode comprometer fortemente as estimativas dos pesquisadores, levando inclusive a cometer erros do tipo I ou do tipo II. Para ilustrar esse efeito esse trabalho replicará os dados apresentados por Anscombe (1973). A tabela abaixo ilustra essas informações.

Tabela 03 - Dados de Anscombe (1973)

Observação	X_{1-3}	Y_1	Y_2	Y_3	X_4	Y_4
1	10	8,04	9,14	7,46	8	6,58
2	8	6,95	8,14	6,77	8	5,76
3	13	7,58	8,74	12,74	8	7,71
4	9	8,81	8,77	7,11	8	8,84
5	11	8,33	9,26	7,81	8	8,47
6	14	9,96	8,10	8,84	8	7,04
7	6	7,24	6,13	6,08	8	5,25
8	4	4,26	3,10	5,39	19	12,50
9	12	10,84	9,13	8,15	8	5,56
10	7	4,82	7,26	6,42	8	7,91
11	5	5,68	4,74	5,73	8	6,89

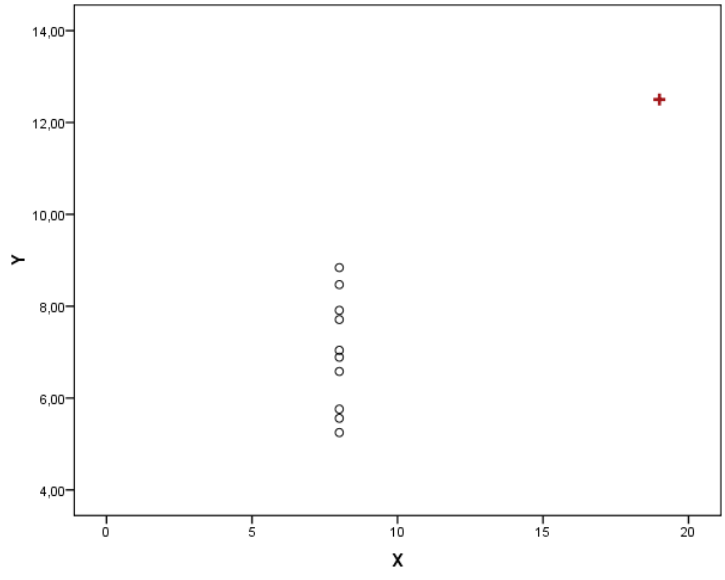
As variáveis (X e Y) estão agrupadas em quatro pares. São eles: (X_1 e Y_1), (X_2 e Y_2), (X_3 e Y_3) e (X_4 e Y_4). O coeficiente de correlação de Pearson (r) para cada par de variáveis é o mesmo: 0,816, sugerindo uma relação positiva e forte entre os respectivos pares de variáveis. Os gráficos abaixo replicam parte desses dados (X_3 e Y_3) e (X_4 e Y_4).

Gráfico 03 - Correlação X_3 e Y_3 - Anscombe (1973)



No caso acima, a cruz vermelha representa um *outlier* em Y. Caso essa observação fosse eliminada da análise, no entanto, continuaria existindo uma correlação positiva e linear entre as variáveis, sendo a diferença a sua magnitude. Logo, a presença desse *outlier* subestimou a verdadeira relação entre X_3 e Y_3.

Gráfico 04 - Correlação entre X_4 e Y_4 - Anscombe (1973)



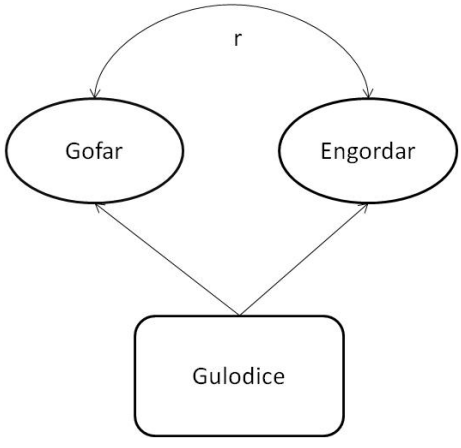
No caso acima, o coeficiente de correlação de Pearson (r) também é de 0,816. No entanto, o padrão de relacionamento entre as variáveis é bastante diferente do observado no gráfico 03. Logo, ao se considerar apenas o valor dessa estatística o pesquisador pode chegar à conclusão de que existe uma relação positiva entre X_4 e Y_4 quando na verdade tudo isso não passa de ilusão. Para a maior parte das observações, não há variação no valor da variável X_4. No entanto, em uma observação houve grande variação o que distorce o padrão de associação entre X_4 e Y_4. Em termos mais técnicos, a presença do *outlier* distorceu o padrão encontrado nos dados, qual seja: independência estatística das observações. A lição deixada por Anscombe (1973) é bastante clara: nem sempre os resultados obtidos através das tabelas de correlação são informativos a respeito do padrão de relacionamento entre as variáveis de

interesse do pesquisador[16]. Ou seja, uma inspeção gráfica mais detalhada dos dados na fase inicial da análise pode evitar muita "dor de cabeça" na hora de realizar inferências[17].

O último tópico concernente ao coeficiente de correlação de Pearson (r) diz respeito ao problema das *lurking* ou *counfouding variables* já que elas podem produzir correlações espúrias[18]. Por exemplo, ao se estimar a relação entre o número leitos hospitalares e a taxa de mortalidade de um determinado estado, o pesquisador pode chegar à conclusão de que quanto mais camas, maior é a taxa de mortalidade. A variável omitida, nesse caso, é o número de pessoas internadas. Para os propósitos desse artigo, será utilizado como exemplo a relação entre gofar e engordar. Isso porque existe a crença de que bebê que gofa muito, ganha peso mais rápido[19]. A figura abaixo ilustra essa relação.

[16] Para outros exemplos nesse sentido ver Magnusson e Mourão (2003). Agradecemos ao parecerista anônimo por essa observação pontual.
[17] É importante lembrar que muitas vezes um *outlier* pode representar simplesmente um erro de digitação.
[18] Ver Tufte (1976). O livro pode ser eletronicamente acessado a partir do seguinte endereço: http://www.edwardtufte.com/tufte/
[19] Essa relação tem sido tradicionalmente utilizada pelo professor Jorge Alexandre no curso intensivo de Metodologia Quantitativa (MQ) em Ciências Sociais da UFMG. Por considerá-lo um excelente exemplo optamos por utilizá-lo. De acordo com o Houaiss, o termo gulodice é proveniente da alteração da palavra gulosice e remonta ao século XV. No nordeste brasileiro a palavra gulodice é usualmente utilizada para designar pessoas que comem de forma excessiva.

Figura 01 - correlação entre gofar e engordar



A correlação observada entre gofar e engordar pode ser explicada na medida em que elas têm a mesma causa: gulodice. Ou seja, essa última variável estava agindo como *lurking variable*. Ao se controlar pelo efeito da gulodice, a correlação entre as variáveis desaparece. Dessa forma, os pesquisadores, antes de apresentar suas conclusões, devem analisar cuidadosamente os seus dados e investigar em que medida uma correlação entre suas variáveis de interesse pode estar sendo afetada pela presença de *lurking variables*.

Exemplo prático: Reapresentação e Conservação na Câmara dos Deputados

Para ilustrar a aplicação do coeficiente de correlação de Pearson (r) com um exemplo mais próximo da Ciência Política, optamos por utilizar alguns dados eleitorais (LEEX, 2009). Em particular, estamos interessados em duas principais variáveis: (1) Taxa de Reapresentação (Divisão do número de candidatos que se reapresentaram pelo total); (2) Taxa de Conservação (Divisão do número de reeleitos pelo total de

candidatos que se reapresentaram, ou seja, reeleitos + derrotados). A tabela abaixo ilustra a correlação entre essas duas variáveis no período 1945-2006.

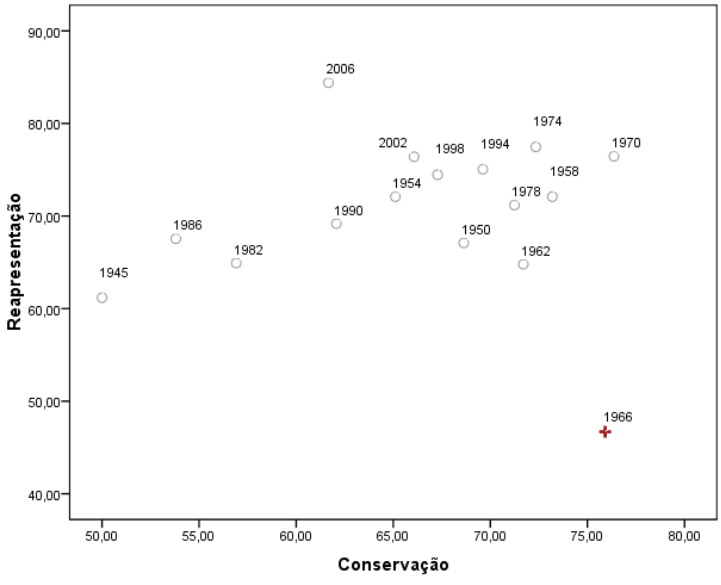
Tabela 04 - Correlação entre Reapresentação e Conservação

Período	R	p valor	n
1945-2006	0,047	0,861	16

Como pode ser observado, o coeficiente de correlação de Pearson (r) entre a taxa de reapresentação e a taxa de conservação é de 0,047 (p=0,861). Logo, o pesquisador chegaria à conclusão de que não existe relação linear entre as variáveis, ou seja, as variáveis são estatisticamente independentes. Além disso, o escore do p valor não permite inferir que os valores encontrados para a amostra podem ser generalizados para a população[20]. O gráfico abaixo ilustra esses dados.

[20] Um dos objetivos centrais da estatística é fazer inferências válidas para a população a partir de dados amostrais. É nesse sentido que a significância estatística, assim como o intervalo de confiança, é uma medida de incerteza a respeito de uma determinada estimação. Para Moore (2007), "a probabilidade, estimada assumindo que H_o é verdadeira, de que a estatística assumiria um valor extremo ou maior do que foi de fato observado é chamado de p valor" (Moore, 2007: 368). O p valor apresenta a probabilidade dos valores encontrados a partir de dados amostrais serem representativos dos parâmetros populacionais, dado que a hipótese nula é verdadeira. Quanto menor o seu valor, maior é a confiança do pesquisador em rejeitar a hipótese nula. No outro oposto, valores altos do p indicam que a hipótese nula não pode ser rejeitada. Em ciências sociais, é comum adotar três diferentes patamares para analisar o p valor: 0,1 (significativo no nível de 10%); 0,05 (significativo no nível de 5%) e 0,01 (significativo no nível de 1%). Para uma discussão sobre o assunto ver Blalock (1967), Carver (1978, 1993), Daniel (1998), McLean e Ernest (1998) e Sawilowsky (2003).

Gráfico 05 – Reapresentação e Conservação (1945-2006)

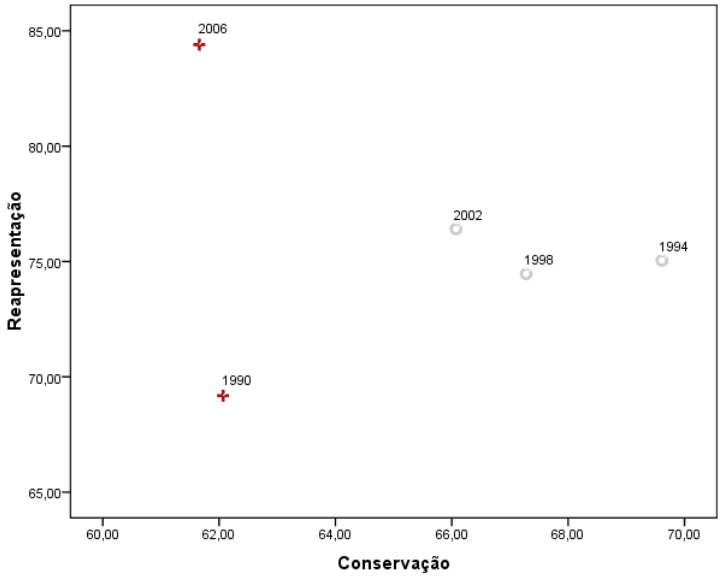


Ao se analisar o gráfico de dispersão, no entanto, observa-se que a eleição de 1966 se comporta como um *outlier* (cruz vermelha), distorcendo o padrão de associação entre as variáveis. Uma forma de testar essa afirmação é remover esse caso da análise e realizar um novo teste de correlação. Com efeito, caso essa observação seja excluída da análise, o nível de associação entre as variáveis assume o valor de 0,447 (p=0,095). Logo, o pesquisador chegaria à conclusão de que existe relação linear positiva entre as variáveis (valor moderado), ou seja, elas são estatisticamente dependentes.

Ao se considerar apenas o período pós-abertura, ou seja, somente as eleições ocorridas entre 1990 e 2006, observa-se uma correlação negativa entre as taxas de reapresentação e conservação (-0,205; p=0,741). Logo, o pesquisador seria levado a acreditar que algo aconteceu e, por esse motivo, o padrão de relacionamento entre as variáveis não só mudou de

magnitude, mas também mudou de direção. O gráfico abaixo ilustra esses dados.

Gráfico 06 – Reapresentação e Conservação (1990-2006)



Aqui emergem dois padrões interessantes. Em primeiro lugar, ao se excluir o ano de 1990 de amostra, o coeficiente de correlação de Pearson (r) passa de -0,205 (p=0,741) para -0,926 (p=0,074) (mesma direção mas diferente magnitude). Logo, o pesquisador chegaria à conclusão de que existe relação linear negativa entre as variáveis (valor alto), ou seja, elas são estatisticamente dependentes. Mas como explicar esse fenômeno? Como explicar que cada vez mais os *incumbents*[21] se recandidatam e não levam? Teoricamente orientado, o pesquisador argumenta que uma possível resposta pode ser especulada via competição eleitoral. Tecnicamente, uma forma de testar essa hipótese é através de uma correlação parcial entre as taxas de reapresentação e conservação, tendo como variável de controle a competição eleitoral. De acordo com

[21] *Incumbent* é um termo utilizado para designar os políticos que já ocupam um cargo na Câmara dos Deputados.

Pallant (2007), a correlação parcial permite controlar em um teste bivariado por uma variável adicional. Através desse controle, é possível estimar mais acuradamente o grau de associação entre as variáveis de interesse[22]. A tabela abaixo sintetiza esses dados.

Tabela 05 - Correlação entre Reapresentação e Conservação controlada pelo Índice de Competitividade Eleitoral

	Sem controle			Com controle		
Período	R	p valor	n	R	p valor	Gl
1990-2006	-0,205	0,741	5	0,668	0,332	2

Os dados indicam que, controlando pela competição política, o grau de associação entre as variáveis aumenta e muda de direção, passando de -0,205 para 0,668 (p=0,332). Ou seja, a associação negativa que antes se observava não se sustenta se a competição eleitoral for considerada. Isso porque a reapresentação influencia positivamente a concorrência, ao permitir a disputa entre *incumbents*. Na correlação parcial (com controle), esse efeito é incluído no modelo, fazendo com que as taxas de reapresentação e conservação se correlacionem positivamente. Diante desses resultados, o pesquisador postula os seus achados: a taxa de representação está positivamente correlacionada com a taxa de reeleição e com o índice de competitividade eleitoral. Este último, por sua vez, está negativamente correlacionado com a taxa de reeleição.

O segundo padrão emerge ao se excluir o ano de 2006 da amostra já que o coeficiente de correlação de Pearson (r) passa de -0,205 para 0,790 (diferente magnitude e diferente direção). Ou seja, ao excluir uma observação o pesquisador chegaria à mesma conclusão caso tivesse

[22] Ver apêndice para mais detalhes.

optado por inserir o controle em sua análise. Em outras palavras, seja por um motivo teoricamente orientado (inclusão do controle), seja pela exclusão de uma das observações 1990 ou 2006 (motivo aleatório – falta de dados, por exemplo), o pesquisador poderia chegar ao mesmo resultado ou a uma conclusão diametralmente oposta. Para Geddes (2003), "relações que parecem existir entre causas e efeitos em amostras pequenas selecionadas a partir da variável dependente podem desaparecer e mesmo mudar de direção quando mais casos que contemplam a amplitude da variação na variável dependente são examinados" (Geddes, 2003:129). A tabela a seguir apresenta as diferentes conclusões possíveis a partir da análise desses dados:

Tabela 06 - Síntese dos resultados encontrados

Período de análise	Valor do coeficiente (r)	Observação metodológica	Conclusão
1945-2006	0,047	Foram analisados todos os anos (sem controle)	As variáveis são estatisticamente independentes.
1945-2006	0,447	Por ser um *outlier*, o ano de 1966 foi excluído da amostra	As variáveis estão positivamente correlacionadas (valor moderado).
1990-2006	-0,205	Foram analisadas apenas as eleições mais recentes (pós-abertura)	As variáveis estão negativamente correlacionadas (valor fraco).
1990-2006	0,668	Inclusão do índice de competitividade eleitoral como controle	As variáveis estão positivamente correlacionadas (valor moderado).
1994-2006	-0,926	Exclusão do ano de 1990 (*outlier*) ou falta de dados	As variáveis estão negativamente relacionadas (valor alto).
1990-2002	0,790	Exclusão do ano de 2006 (*outlier*) ou falta de dados	As variáveis estão positivamente relacionadas (valor alto).

Afinal, qual é a verdadeiro padrão de correlação entre as taxas de reapresentação e de conservação? Dado o número reduzido de casos,

qualquer resposta tende a ser tentativa e as inferências devem ser interpretadas com bastante cautela. Isso porque estatísticas extraídas de amostras pequenas tendem a ser não representativas dos parâmetros populacionais[23]. Em termos metodológicos, uma possível saída para responder essa questão seria aumentar o número de casos (King, Keohane e Verba, 1994). Para tanto o pesquisador poderia sugerir analisar em que medida o padrão de correlação entre essas variáveis para a Câmara dos Deputados se mantém constante para as assembléias estaduais. Reportam-se os seguintes resultados:

Tabela 07 – Correlação entre Reapresentação e Conservação (Assembléias estaduais)

	Sem controle		
Período	R	p valor	n
1990-2006	-0,199	0,021	134

Ou seja, os dados sugerem que também no nível estadual existe uma correlação negativa entre as taxas de reapresentação e conservação (-0,199; p =0,021). Além disso, o pesquisador reporta que esse padrão é consistente ao se desagregar os dados por ano: -0,028 em 1990; -0,552 em 1994; -0,184 em 1998; -0,096 em 2002 e -0,294 em 2006. A primeira vista, esses resultados poderiam ser utilizados para corroborar a teoria do pesquisador de que na ausência de controle pelo índice de competição eleitoral, as taxas de reapresentação e reeleição estão negativamente associadas. Todavia, esses resultados não podem ser comparados já que foram extraídos de amostras diferentes. Isso porque a diferença da

[23] Ver o trabalho seminal de Fisher (1921). Para um trabalho sobre as diferentes contribuições de Fisher ver Anderson (1996).

magnitude entre correlações extraídas de amostras diferentes podem variar simplesmente porque a variância é diferente e não porque o padrão de relacionamento entre as variáveis é mais ou menos consistente (Achen, 1977: 807).

Em síntese, esses dados sugerem alguns cuidados ao se utilizar o coeficiente de correlação de Pearson (r):

(1) O coeficiente de correlação de Pearson (r) deve ser acompanhado por análises gráficas (gráficos de dispersão). Apenas depois disso o pesquisador deve utilizar o coeficiente de correlação de Pearson (r) para medir o grau e a direção da associação entre as suas variáveis de interesse;

(2) Além disso, o pesquisador deve se certificar de que os pressupostos estão sendo respeitados (nível de mensuração das variáveis, linearidade da relação, normalidade da distribuição, etc.);

(3) Inferências realizadas a partir de uma quantidade reduzida de observações devem ser interpretadas com bastante cautela. Isso porque amostras pequenas não fornecem estimativas confiáveis dos parâmetros populacionais;

(4) A presença de *outliers* e/ou de variáveis omitidas compromete fortemente a confiabilidade dos resultados encontrados. Dessa forma, o pesquisador deve verificar em que medida o seu banco de dados foi devidamente construído e "*cleaning*";

(5) Correlações não podem ser comparadas entre diferentes amostras já que elas podem diferir porque apresentam variâncias diferentes, mesmo que o padrão de relacionamento entre as variáveis seja consistente (Achen, 1977: 807). Dessa forma, não se deve utilizar coeficientes de amostras diferentes como um indicativo de

existência de uma relação mais geral entre as variáveis (para isso devem ser utilizados os coeficientes não padronizados).

2. Conclusão

Por um lado, estima-se que o coeficiente de correlação de Pearson e suas derivações são escolhidos em 95% dos casos para descrever o padrão de relacionamento entre variáveis ou para fazer inferências válidas para a população a partir de dados amostrais (Chen e Popovic, 2002: 09). Por outro, Carroll (1961) afirma que o coeficiente de correlação é geralmente utilizado de forma inapropriada (Carroll, 1961: 01) Mas o que significa dizer que duas variáveis estão correlacionadas? O principal objetivo desse artigo é pedagógico. Procuramos apresentar as principais propriedades do coeficiente de correlação de Pearson (r), suas respectivas aplicações e limites a partir de uma abordagem descritiva.

Além disso, queremos chamar a atenção dos pesquisadores para as aplicações e os limites dessa medida na formulação dos seus desenhos de pesquisa. Concordamos fortemente com a afirmação de que não é a estatística que determina se relações causais podem ser ou não alcançadas (Chen e Popovic, 2002: 07). No entanto, acreditamos também que o que distingue o conhecimento científico de outras formas de conhecimento é exatamente a utilização sistemática e rigorosa do método. Nesse sentido, compreender melhor o significado do coeficiente de correlação de Pearson (r) é um passo fundamental para lidar com os problemas enfrentados pelos cientistas sociais em geral e pelos cientistas políticos, em particular. Dessa forma, independente do que será servido, "filé" (explicação) ou "picanha" (interpretação), é preciso que o *chef* tenha habilidade suficiente para preparar esses pratos, caso contrário, corre-se o risco de oferecer um guizado de filé ou ensopado de picanha.

Esperamos contribuir para tornar essa receita menos tortuosa. Afinal, existe uma correlação positiva entre interpretação e explicação.

Apêndice

Desvio padrão

$$s = \sqrt{\frac{1}{n-1}\sum(xi - \bar{X})^2}$$

Média

$$\bar{X} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum xi$$

Padronização

$$z = \frac{x - \mu}{\sigma}$$

Correlação

$$r = \frac{1}{n-1}\sum\left(\frac{xi - \bar{X}}{sx}\right)\left(\frac{yi - \bar{Y}}{sy}\right)$$

Correlação parcial

$$\hat{\rho}_{XY\cdot \mathbf{Z}} = \frac{N\sum_{i=1}^{N} r_{X,i}r_{Y,i} - \sum_{i=1}^{N} r_{X,i}\sum r_{Y,i}}{\sqrt{N\sum_{i=1}^{N} r_{X,i}^2 - \left(\sum_{i=1}^{N} r_{X,i}\right)^2}\sqrt{N\sum_{i=1}^{N} r_{Y,i}^2 - \left(\sum_{i=1}^{N} r_{Y,i}\right)^2}}.$$

Referências Bibliográficas

ACHEN, Christopher H. (1977), "Measuring Representation: Perils of the Correlation Coefficient". *American Journal of Political Science*, 21, 4: 805-815.

ALDRICH, John H. (1995), "Correlations Genuine and Spurious in Pearson and Yule". *Statistical Science*, 10, 4: 364-376.

ALMOND, Gabriel. (1990), *A Discipline Divided: Schools and Sects in Political Science*. Newbury Park, Calif.: Sage Publications.

ANDERSON, Theodore W. (1996), "R. A. Fisher and Multivariate Analysis". *Statistical Science*, 11, 1: 20-34.

ANDRES, Martin I.; TEJEDOR, Herranz & MATO, A. Silva. (1995), "The Wilcoxon, Spearman, Fisher, χ^2-, Student and Pearson Tests and 2×2 Tables". *Journal of the Royal Statistical Society*, 44, 4: 441-450.

ANSCOMBE, Frank J. (1973), "Graphs in Statistical Analysis". *The American Statistician*, 27: 17-21.

ASHER, Hebert. (1983), *Causal Modeling*. London, Sage.

BLALOCK, Hubert. (1967), "Causal Inferences, Closed Populations, and Measures of Association". *The American Political Science Review*, 61, 1: 130-136.

BLALOCK, Hubert. (1971), Causal Models in the Social Sciences, Chicago: Aldine-Atherton.

BLYTH, Stephen. (1994), "Karl Pearson and the Correlation Curve". *International Statistical Review*, 62, 3: 393-403.

CARROLL, John B. (1961), "The Nature of the Data, or How to Choose a Correlation Coefficient". *Psychometrika*, 26: 347-372.

CARVER, Ronald P. (1978), "The case against statistical significance

testing". *Harvard Educational Review, 48*, 378- 399.
CARVER, Ronald P. (1993), "The case against statistical significance testing, revisited". *Journal of Experimental Education, 61,* 287-292.
CHEN, Peter Y. & POPOVIC, Paula M. (2002), *Correlation*. London, Sage.
COHEN, Jacob. (1988), *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ, Erlbaum.
COLLIER, David; SEAWRIGHT, Jason & MUNCK, Gerardo L. (2004), "Sources of Leverage in Causal Inference: Toward an Alternative View of Methodology", in H. Brady & D. Collier (eds), *Rethinking Social Inquiry: Diverse Tools, Shared Standards*, Lanham, MD, Rowman and Littlefield.
DANCEY, Christine & REIDY, John. (2006), *Estatística Sem Matemática para Psicologia: Usando SPSS para Windows.* Porto Alegre, Artmed.
DANIEL, Larry G. (1998), "Statistical Significance Testing: A Historical Overview of Misuse and Misinterpretation with Implications for the Editorial Policies of Educational Journals". *Research in the Schools*, 5, 2: 23-32.
DEVLIN, Susan J.; GNANADESIKAN, Ramanathan & KETTENRING, Jon R. (1975), "Robust Estimation and Outlier Detection with Correlation Coefficients". *Biometrika.* 62, 3: 531-545.
FIELD, Andy (2005). *Discovering Statistics Using SPSS*. London, Sage.
FISHER, Ronald A. (1921), "On the "probable error" of a coefficient of correlation deduced from a small sample". *Metron*, 1: 3-32.
FRIEDRICH, Robert. (1982), "In Defense of Multiplicative Terms in Multiple Regression Equations". *American Journal of Political Science*, 26: 797-833.
GARSON, G. David. (2009), *Statnotes: Topics in Multivariate Analysis*. Disponível em: http://faculty.chass.ncsu.edu/garson/PA765/statnote.htm
GEDDES, Barbara. (2003), *Paradigms and Sand Castles: Theory Building and Research Design in Comparative Politics.* Ann Arbor: University of Michigan Press.
GERRING, John. (2001), *Social Science Methodology: A Criterial Framework*. Cambridge: Cambridge University Press.
HAIG, Brian D. (2007), "Spurious correlation", in N. J. Salkind (ed.), *Encyclopedia of Measurement and Statistics*, Thousand Oaks, Sage.

HAIR Jr., Joseph F.; ANDERSON, Ralph E.; TATHAM, Ronald T. & BLACK, Willian C. (2005), *Análise Multivariada de Dados*. Porto Alegre, Bookman.

HOLLAND, Paul. W. (1986), "Statistics and Causal Inference". *Journal of American Statistical Association*, 81, 396: 945-960.

KENNEDY, Peter. (2009), *A Guide to Econometrics*. Boston: MIT Press.

KING, Garry. (2001), *How not to lie with statistics: avoiding common mistakes in quantitative political science.* Disponível em: http://gking.harvard.edu/#

KING, Garry.; KEOHANE, Robert. & VERBA, Sidney. (1994), *Designing social inquiry: scientific inference in qualitative research*. Princeton: Princeton University Press.

KLECKA, William R. (1980), *Discriminant Analysis*. Beverly Hills, Sage.

KOZAK, Marcin. (2009), "What is strong correlation?". *Teaching Statistics*, 31: 85-86.

KRONMAL, Richard A. (1993), "Spurious Correlation and the Fallacy of the Ratio Standard Revisited". *Journal of the Royal Statistical Society*, 156, 3: 379-392.

LEEX (2009). *Almanaque de dados eleitoriais*. Disponível em: http://www.ucam.edu.br/leex/

MAGNUSSON, William E. & MOURÃO, Guilherme. (2003), *Estatística sem Matemática*. Londrina, Editora Planta.

McLEAN, James E. & ERNEST, James M. (1998), "The Role of Statistical Significance Testing In Educational Research". *Research in the Schools*, 5, 2: 15-22.

MOORE, David S. & McCABE, George. (2004), *Introduction to the practice of statistics*. New York, Freeman.

MOORE, David S. (2007), *The Basic Practice of Statistics*. New York, Freeman.

MUDDAPUR, M. V. (1988), "A Simple Test for Correlation Coefficient in a Bivariate Normal Distribution". *The Indian Journal of Statistics*, 50, 1: 60-68.

NILES, Henry E. (1921), "Correlation, Causation and Wright's theory of "Path Coefficients". *Genetics*, 7: 258.

O'BRIEN, Robert M. (1979), "The Use of Pearson's with Ordinal Data". *American Sociological Review*, 44, 5: 851-857.

OSBORNE, Jason & WATERS, Elaine. (2002), "Four assumptions of multiple regression that researchers should always test". *Practical Assessment, Research & Evaluation*, 8, 2. Disponível em: http://PAREonline.net/getvn.asp?v=8&n=2

PALLANT, Julie. (2007), *SPSS Survival Manual*. Open University Press.
PEARSON, Karl. (1892), *The grammar of science*. London, J. M. Dent and Company.
PEARSON, Karl; FISHER, Ronald & INMAN, Henry F. (1994), "Karl Pearson and R. A. Fisher on Statistical Tests: A 1935 Exchange from Nature". *The American Statistician*, 48,1: 2-11.
PEDHAZUR, Elazar J. (1997), *Multiple Regression in Behavioral Research*. Orlando, Harcourt Brace.
POLLOCK III, Philip H. (2006), *A Stata Companion to Political Analysis*. Washington, DC: CQ Press.
RODGERS, Joseph Lee & NICEWANDER, W. Alan. (1988), "Thirteen Ways to Look at the Correlation Coefficient". *The American Statistician*, 42,1: 59-66.
RUBIN, D.B. (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies", Journal of Educational Psychology, 66, 688-701.
RUMMEL, Rudolph J. (1976), *Understanding Correlation*. Disponível em: http://www.hawaii.edu/powerkills/UC.HTM
SANTOS, Maria Helena & COUTINHO, Marcelo. (2000), "Política comparada: estado das artes e perspectivas no Brasil". *BIB*, 54: 3-146.
SAWILOWSKY, Shlomo S. (2003), "Deconstructing Arguments From The Case Against Hypothesis Testing". *Journal of Modern Applied Statistical Methods*, 2, 2: 467-474.
SCHIELD, Milo. (1995), "Correlation, Determination And Causality In Introductory Statistics**".** *American Statistical Association, Section on Statistical Education*.
SOARES, Gláucio (2005), "O calcanhar metodológico da ciência política no Brasil". *Sociologia*, 48: 27-52.
STANTON, Jeffrey M. (2001), "Galton, Pearson, and the peas: A brief history of linear regression for statistics instructors". *Journal of Statistical Education*, 9,3. Disponível em: http://www.amstat.org/publications/JSE/v9n3/stanton.html
STIGLER, Stephen. (1989), "Francis Galton's Account of the Invention of Correlation". *Statistical Science*, 4, 2: 73-79.
STOKER, Getry & MARSH, David. (2002), "Introduction", in D. Marsh & G. Stoker (eds.), *Theory and Methods in Political Science*, Palgrave, Macmillan.
TABACHNICK, Barbara & FIDELL, Linda. (2007), *Using multivariate analysis*. Needham Heights, Allyn & Bacon.
TUFTE, Edward. (1976), *Data Analysis for Politics and Policy*.

Englewood Cliffs, Prentice-Hall.
VALLE E SILVA, Nelson (1999), *Relatório de Consultoria sobre Melhoria do Treinamento em Ciência Social Quantitativa e Aplicada no Brasil*. Rio de Janeiro, Laboratório Nacional de Computação Científica.
VAN EVERA, Stephen. (1997), *Guide to Methods for Students of Political Science*. Ithaca, Cornell University Press.
WERNECK VIANNA, Luiz *et al* (1998). "Doutores e teses em ciências sociais". *Dados*, 41, 3: 453-515.