TESTES (ADERÊNCIA, INDEPENDÊNCIA E HOMOGENEIDADE)

1. Teste de Aderência

1.1. Teste de *Kolmogorov-Smirnov*

Este teste é usado para comparar uma amostra com uma distribuição de probabilidade de referência ou para comparar duas amostras. O estatística de *Kolmogorov-Smirnov* qualifica a distância, ou seja, a máxima diferença absoluta entra a função de distribuição empírica da amostra e a função de distribuição acumulada da distribuição de referência, ou entre função de distribuição empírica de duas amostras. Como critério, comparamos esta diferença com um valor crítico, para um dado nível de significância. O teste é utilizado para avaliar as hipóteses:

$$\begin{cases} H_0: A\ amostra\ segue\ a\ distribuição\ F(x) \\ H_1: A\ amostra\ não\ segue\ a\ distribuição\ F(x) \end{cases}$$

1.2. Teste de *Anderson-Darling*

O teste foi proposto por *Theodore Wilbur Anderson* e *Donald A. Darling*, em 1952. É uma estatística que visa testar se uma dada amostra tenha vindo de função de distribuição acumulada $F(x)$, isto é, seja $x_1, \dots, x_n$ amostra aleatória e que $F(x)$ seja uma candidata a função de distribuição acumulada para a amostra, então o teste de hipóteses para verificar se a distribuição é adequada é:

$$\begin{cases} H_0: A\ amostra\ tem\ distribuição\ F(x) \\ H_1: A\ amostra\ não\ tem\ distribuição\ F(x) \end{cases}$$

1.3. Teste de *Shapiro-Wilk*

O teste *Shapiro-Wilk*, proposto em 1965, é baseado na estatística W, calculada como a seguir:

$$W = \frac{b^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

onde x_i são os valores da amostra ordenados (x_1 é o menor). Menores valores de W são evidências de que os dados são normais. A constante b é determinada da seguinte forma:

$$b = \sum_{i=1}^{n/2} a_{n-i+1} \times (x_{n-i+1} - x_i)$$

onde a_i são constantes geradas pelas médias, variâncias e covariâncias das estatísticas de ordem de uma amostra de tamanho n de uma distribuição Normal. Seus valores, tabelados, são dados abaixo.

Coeficientes α_{N-i+1} para o teste de normalidade W de SHAPIRO – WILK (Para $N = 2(1)50$)

$i \setminus N$	2	3	4	5	6	7	8	9	10	
1	0.7071	0.7071	0.6872	0.6646	0.6431	0.6233	0.6052	0.5888	0.5739	
2		0.0000	0.1677	0.2413	0.2806	0.3031	0.3164	0.3244	0.3291	
3				0.0000	0.0875	0.1401	0.1743	0.1976	0.2141	
4						0.0000	0.0561	0.0947	0.1224	
5								0.0000	0.0399	

$i \setminus N$	11	12	13	14	15	16	17	18	19	20
1	0.5601	0.5475	0.5359	0.5251	0.5150	0.5056	0.4968	0.4886	0.4808	0.4734
2	0.3315	0.3325	0.3325	0.3318	0.3306	0.3290	0.3273	0.3253	0.3232	0.3211
3	0.2260	0.2347	0.2412	0.2460	0.2495	0.2521	0.2540	0.2553	0.2561	0.2565
4	0.1429	0.1586	0.1707	0.1802	0.1878	0.1939	0.1988	0.2027	0.2059	0.2085
5	0.0695	0.0922	0.1099	0.1240	0.1353	0.1447	0.1524	0.1587	0.1641	0.1686
6	0.0000	0.0303	0.0539	0.0727	0.0880	0.1005	0.1109	0.1197	0.1271	0.1334
7			0.0000	0.0240	0.0433	0.0593	0.0725	0.0837	0.0932	0.1013
8					0.0000	0.0196	0.0359	0.0496	0.0612	0.0711
9							0.0000	0.0163	0.0303	0.0422
10									0.0000	0.0140

Para realizar o teste de *Shapiro-Wilk*, deve-se seguir os seguintes passos:

1. Formulação da Hipótese:

$$\begin{cases} H_0: A\ mostra\ provém\ de\ uma\ população\ Normal \\ H_1: A\ mostra\ não\ provém\ de\ uma\ população\ Normal \end{cases}$$

2. Estabelecer o Nível de significância do teste α, normalmente $0{,}05$;

3. Calcular a estatística de teste:

- ✓ Ordenar as n observações da amostra $x_1, \dots, x_n$;
- ✓ Calcular $\sum_{i=1}^{n}(x_i - \bar{x})^2$;
- ✓ Calcular b;
- ✓ Calcular W.

4. Tomar a decisão: Rejeitar H_0 ao nível de significância α se $W_{calculado} < W_{\alpha}$ (os valores críticos da estatística W de *Shapiro-Wilk* são dados na Tabela abaixo).

Valores críticos da estatística W de SHAPIRO-WILK

Tamanho da Amostra, N	Nível de significância α								
	0.01	0.02	0.05	0.10	0.50	0.90	0.95	0.98	0.99
3	0.753	0.756	0.767	0.789	0.959	0.998	0.999	1.000	1.000
4	0.687	0.707	0.748	0.792	0.935	0.987	0.992	0.996	0.997
5	0.686	0.715	0.762	0.806	0.927	0.979	0.986	0.991	0.993
6	0.713	0.743	0.788	0.826	0.927	0.974	0.981	0.986	0.989
7	0.730	0.760	0.803	0.838	0.928	0.972	0.979	0.985	0.988
8	0.749	0.778	0.818	0.851	0.932	0.972	0.978	0.984	0.987
9	0.764	0.791	0.829	0.859	0.935	0.972	0.978	0.984	0.986
10	0.781	0.806	0.842	0.869	0.938	0.972	0.978	0.983	0.986
11	0.792	0.817	0.850	0.876	0.940	0.973	0.979	0.984	0.986
12	0.805	0.828	0.859	0.883	0.943	0.973	0.979	0.984	0.986
13	0.814	0.837	0.866	0.889	0.945	0.974	0.979	0.984	0.986
14	0.825	0.846	0.874	0.895	0.947	0.975	0.980	0.984	0.986
15	0.835	0.855	0.881	0.901	0.950	0.975	0.980	0.984	0.987
16	0.844	0.863	0.887	0.906	0.952	0.976	0.981	0.985	0.987
17	0.851	0.869	0.892	0.910	0.954	0.977	0.981	0.985	0.987
18	0.858	0.874	0.897	0.914	0.956	0.978	0.982	0.986	0.988
19	0.863	0.879	0.901	0.917	0.957	0.978	0.982	0.986	0.988
20	0.868	0.884	0.905	0.920	0.959	0.979	0.983	0.986	0.988
21	0.873	0.888	0.908	0.923	0.960	0.980	0.983	0.987	0.989
22	0.878	0.892	0.911	0.926	0.961	0.980	0.984	0.987	0.989
23	0.881	0.895	0.914	0.928	0.962	0.981	0.984	0.987	0.989
24	0.884	0.898	0.916	0.930	0.963	0.981	0.984	0.987	0.989
25	0.888	0.901	0.918	0.931	0.964	0.981	0.985	0.988	0.989
26	0.891	0.904	0.920	0.933	0.965	0.982	0.985	0.988	0.989
27	0.894	0.906	0.923	0.935	0.965	0.982	0.985	0.988	0.990
28	0.896	0.908	0.924	0.936	0.966	0.982	0.985	0.988	0.990
29	0.898	0.910	0.926	0.937	0.966	0.982	0.985	0.988	0.990
30	0.900	0.912	0.927	0.939	0.967	0.983	0.985	0.988	0.900

1.4. Teste de *Jarque-Bera*

O teste objetiva verificar se os erros obedecem o pressuposto do modelo de regressão de seu valor esperado ser igual a zero.

O teste se baseia nos resíduos do método dos mínimos quadrados. Para sua realização o teste necessita dos cálculos da assimetria (*skewness*) e da curtose (*kurtosis*) da amostra. A estatística de teste é definida por:

$$JB = \frac{n}{6}\left(S^2 + \frac{1}{4}K^2\right)$$

onde n e o número de observações (ou graus de liberdade geral); S é a assimetria da amostra; e K é a curtose da amostra.

$$S = \frac{\hat{\mu}_3}{\hat{\sigma}^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{3/2}}$$

$$K = \frac{\hat{\mu}_4}{\hat{\sigma}^4} - 3 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{2}} - 3$$

onde $\hat{\mu}_3$ e $\hat{\mu}_4$ são as estimativas do terceiro e quarto momentos, respectivamente; $\bar{x}$ é a média da amostra, e $\hat{\sigma}^2$ é a estimativa do segundo momento, a variância.

Observação: O terceiro e o quarto momentos de uma distribuição são usados com freqüência para estudar o "formato" ou a "aparência" de uma distribuição de probabilidade, em particular, seu *grau de assimetria*, S (isto é, falta de simetria) ou *curtose*, K (isto é, grau de elevação ou de achatamento). Onde:

$$S = \frac{E[(X-\mu)^3]}{E[(X-\mu)^2]^{\frac{3}{2}}} = \frac{E[(X-\mu)^3]}{\sigma^3} = \frac{terceiro\ momento\ em\ torno\ da\ média}{desvio\ padrão\ ao\ cubo};$$

$$K = \frac{E[(X-\mu)^4]}{E[(X-\mu)^2]^2} = \frac{quarto\ momento\ em\ torno\ da\ média}{quadrado\ do\ segundo\ momento}.$$

Onde:

- ✓ $S = 0$ é simétrica;

- ✓ $S > 0$ é assimétrica à direita (cauda mais pesada à direita);
- ✓ $S < 0$ é assimétrica à esquerda (cauda mais pesada à esquerda);
- ✓ $K < 3$ é platicúrtica (Caudas curtas ou caudas leves);
- ✓ $K > 3$ é leptocúrtica (Caudas longas ou caudas pesadas);
- ✓ $K = 3$ é mesocúrtica (Distribuição Gaussiana).

A estatística JB tem uma distribuição chi-quadrado assintótica com dois graus de liberdade e pode ser usada para testar a hipótese nula que os dados são de uma distribuição Normal. A hipótese nula é uma hipótese conjunta de que a assimetria é igual a zero e a curtose ser também igual a zero. Amostras de uma distribuição Normal tem uma assimetria esperada de zero e uma curtose esperada de zero (que é o mesmo que uma curtose igual a 3).

A aproximação chi-quadrado, no entanto, é bastante sensível para amostras pequenas, rejeitando a hipótese nula quando na verdade ela é verdadeira. Além disso, a distribuição dos *p – valores* vem de uma distribuição uniforme e se torna uma distribuição uni-modal assimétrica para a direita, especialmente para *p – valores* pequenos. Isto leva para uma grande taxa de erro do Tipo I.

Exercício: Verifique se existe as variáveis X e Y seguem uma distribuição Gaussiana por meio dos testes de *Kolmogorov-Smirnov*, *Anderson Darling*, *Shapiro-Wilk* e *Jarque-Bera*.

$\boldsymbol{Variável\ X}$	9,4	10,5	12,4	5,6	12,2	8,6	6,3	8,5	8,9	10,3	9,3	12,7
$\boldsymbol{Variável\ Y}$	34,9	45,1	46,9	39,1	35,7	33,2	36,1	31,5	38,4	38,5	38,5	39,7

2. Testes Qui-Quadrado

Dentre outros testes que podem ser realizados sob a distribuição Qui-Quadrado, existem 3 importantes testes, são eles:

- ✓ Teste de Aderência;
- ✓ Teste de Independência;
- ✓ Teste de Homogeneidade.

2.1. Teste de Aderência

Este teste é utilizado quando deseja-se validar a hipótese que um conjunto de dados é gerado por uma determinada distribuição de probabilidade.

Seja X uma variável aleatória para a qual tem-se uma amostra de valores observados e deseja-se verificar a adequação ou não de um certo modelo probabilístico.

O principio básico do teste consiste em dividir a amostra em k categorias e avaliar a diferença entre a frequência observada em cada categoria e a frequência esperada de acordo com o modelo de probabilidade teório (em linha com a hipótese).

Se X for uma variável aleatória discreta, as categorias são os próprios valores observados. Se X for uma variável aleatória contínua, as categorias são definidas a partir dos valores observados.

Categoria	**1**	**2**	**3**	**…**	***K***
Freq. Observada	o_1	o_2	o_3	…	o_k
Freq. Esperada	e_1	e_2	e_3	…	e_k

Se X seguir o modelo probabilístico proposto, as frequências observadas e esperada não devem ser muito discrepantes, portanto, o Teste de Aderência cria, então, um critério para decidir se os dados amostrais aderem ao modelo ou não.

As hipóteses do teste são:

$$\begin{cases} H_o: X\ segue\ o\ modelo\ proposto \\ H_1: X\ não\ segue\ o\ modelo\ proposto \end{cases}$$

2.2. Teste de Independência

Esse teste é utilizado quando deseja-se validar a hipótese de independência entre duas variáveis aleatórias. Se existe a função de probabilidade conjunta das duas variáveis aleatórias, pode-se verificar se, para todos os possíveis valores das variáveis, o produto das probabilidades marginais é igual à probabilidade conjunta.

Caso não se tenha a função de probabilidade conjunta, o procedimento é avaliar a diferença entre a frequência observada conjunta das variáveis e a frequência esperada sob a hipótese de independência entre as variáveis.

As hipóteses do teste são:

$$\begin{cases} H_0: As\ variáveis\ aleatórias\ X\ e\ Y\ são\ independentes \\ H_1: As\ variáveis\ aleatórias\ X\ e\ Y\ não\ são\ independentes \end{cases}$$

2.3. Teste de Homogeneidade

Esse teste é utilizado quando deseja-se validar a hipótese de que uma variável aleatória apresenta comportamento similar, ou homogêneo, em relação às suas várias subpopulações.

Este teste apresenta a mesma mecânica do Teste de Independência, mas uma distinção importante se refere à forma como as amostras são coletadas. No Teste de homogeneidade fixa-se o tamanho da amostra em cada uma das subpopulações e, então, seleciona-se uma amostra de cada uma delas.

Na tabela abaixo tem-se a estrutura das observações, onde as linhas representam as subpopulações e, as colunas, os diferentes valores ou categorias da variável.

Subpopulações	**Valores da variável**			**Total de linha**
1	o_{11}	o_{12}	**…**	n_1
2	o_{21}	o_{22}	**…**	n_2
…	⋮	⋮	⋮	⋮
Total da coluna				**Total geral**

As hipóteses do teste são:

$$\begin{cases} H_0: As\ subpopulações\ da\ variáveis\ aleatórias\ X\ são\ homogêneas \\ H_1: As\ subpopulações\ da\ variáveis\ aleatórias\ X\ não\ são\ homogêneas \end{cases}$$

Defini-se a tabela de valores esperados calculando-se cada casela (i, j) da tabela por:

$$e_{i,j} = n_i \frac{total\ da\ coluna\ j}{total\ geral}.$$

Se houver alguma casela com frequência inferior a 5 é necessário agrupar as categorias para melhorar a aproximação para a utilização do teste Qui-Quadrado.

Caso haja homogeneidade de comportamento da variável, espera-se que a proporção de ocorrências do valor da variável correspondente à coluna j seja a mesma, em todas as subpopulações.

COEFICIENTES DE CORRELAÇÃO

Freqüentemente tem-se o interesse em verificar a existência de associação entre dois conjuntos de dados e também o seu grau desta associação. O coeficiente de correlação, por si só, representa o grau de associação. É necessário porém testar a significância estatística deste coeficiente. No caso paramétrico, a medida usual é o coeficiente de correlação r de *Pearson* que exige mensuração dos escores no mínimo ao nível intervalar. Ainda, se hover interesse e necessidade em comprovar a significância de um valor observado de r de *Pearson* deve-se supor que os escores provenham de uma distribuição Gaussiana. Quando estas suposições não são atendidas é necessário utilizar um dos coeficientes de correlação não-paramétricos (*Spearman* e *Kendall*) e suas respectivas provas de significância.

1.1. Coeficiente de Correlação de *Pearson*

O coeficiente de correlação r de *Pearson*, também chamado de coeficiente de correlação produto-momento, foi desenvolvido por *Karl Pearson.* Esta medida estabelecer o nível da relação linear entre duas variáveis. Em outras palavras, mede em grau e sentido (crescente/decrescente) a associação linear entre duas variáveis, como pode ser visto na figura abaixo. É definido por:

$$r_{x;y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}};\ -1 \leq r \leq +1.$$

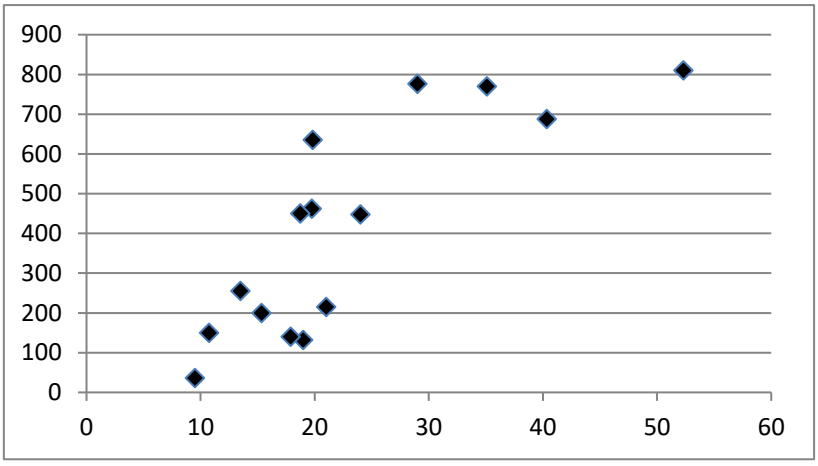
Deve-se lembrar que o coeficiente de correlação populacional é dado por:

$$\rho_{x;y} = \frac{Cov(x;y)}{\sqrt{Var(x)}\sqrt{Var(y)}};\quad -1 \leq \rho_{X;Y} \leq +1.$$

O coeficiente de correlação de *Pearson* esta sempre entre $-1{,}00$ e $+1{,}00$, onde o sinal indica a direção, se a correlação é positiva (direta) ou negativa (inversa). O valor do coeficiente de *Pearson* indica a força da correlação, onde nos intervalos $(+0{,}90;\ +1{,}00)$ ou $(-1{,}00;\ -0{,}90)$ indica muito forte correlação linear, $(+0{,}60;\ +0{,}90)$ ou $(-0{,}90;\ -0{,}60)$ indica uma forte correlação linear, entre $(+0{,}30;\ +0{,}60)$ ou $(-0{,}60;\ -0{,}30)$ indica uma moderada correlação linear e entre $(0{,}00;\ +0{,}30)$ ou $(-0{,}30;\ 0{,}00)$ indica uma fraca correlação linear.

Exemplo: Uma amostra com 15 observações do tempo de entrega (em minutos) de pizza de uma *Pizzaria Delivery* e a distância de entrega. Deseja-se verificar se as variáveis são correlacionadas. O gráfico de dispersão das variáveis, abaixo, sugere que há uma relação positiva e linear. Desta forma, utilizar-se-á o coeficiente de correlação de *Pearson* para checar ser as variáveis são correlacionadas.

Tempo	40	21	14	20	24	29	15	19	10	35	18	52	19	20	11
Distância	688	215	255	462	448	776	200	132	36	770	140	810	450	635	150



Cálculo do coeficiente r de *Pearson*:

$$r_{x;y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{2591{,}0190}{11{,}6775 \times 270{,}0534} = 0{,}8216.$$

Portanto, conclui-se que existe uma relação linear forte e positiva entre as variáveis.

O coeficiente de correlação r de *Pearson* é apenas uma estimativa do coeficiente de correlação populacional, pois é calculado com base em uma amostra aleatória de n pares de dados.

Ressalta-se que a amostra observada pode apresentar uma correlação e, no entanto a população não, neste caso, tem-se um problema de inferência, pois $r \neq 0$ não é garantia de que $\rho \neq 0$.

Pode-se resolver este problema por meio de um teste de hipóteses para verificar se o valor de r é é estatisticamente significativo, ao nível de significância α. Ou seja, se realmente existe correlação linear entre as variáveis. As hipóteses a serem formuladas são:

$$\begin{cases} H_o: \rho = 0 \\ H_1: \rho \neq 0 \end{cases}.$$

E tem-se que a estatística teste é:

$$t = r\sqrt{\frac{n-2}{1-r^2}};$$

Onde $t \sim t_{n-2;\alpha/2}$, ou seja, t segue uma distribuição $t - Student$ com $n-2$ graus de liberdade, ao nível α de significância.

Observação: Ressalta-se que esta estatística teste pode ser utilizada sob as suposições:

- ✓ A relação entre as variáveis aleatórias é linear;
- ✓ As variáveis aleatórias são medidas em escala intervalar ou razão;
- ✓ As duas variáveis aleatórias tem distribuição normal bivariada conjunta. Esta hipótese é impressindível quando a amostra é pequena.

Exemplo: Verifique se o coeficiente de correlação de *Pearson* do exemplo anterior é significativo, ao nível de 5%.

As hipóteses são:

$$\begin{cases} H_o: A\ correlação\ entre\ tempo\ de\ entrega\ e\ distância\ é\ zero\ (\rho = 0) \\ H_1: A\ correlação\ entre\ tempo\ de\ entrega\ e\ distância\ não\ é\ zero\ (\rho \neq 0) \end{cases}.$$

A estatística teste sob o dados é:

$$t = \rho\sqrt{\frac{n-2}{1-\rho^2}} = 0{,}8216\sqrt{\frac{15-2}{1-0{,}8216^2}} = 5{,}1965;$$

Dado que $t \sim t_{n-2;\alpha/2}$, tem-se, a partir da estatística $t - Student$ com 13 graus de liberdade, os pontos críticos $\pm 2{,}1604$. Portanto, rejeita-se H_o ao nível de significância de 5%. Ou seja, a correlação entre o tempo de entrega e a distância percorrida é diferente de zero, então, existe uma relação linear e positiva da ordem de $r = 0{,}8216$.

1.2. Coeficiente de Correlação de *Spearman*

O coeficiente de correlação de postos de *Spearman* foi desenvolvido por *Charles Edward Spearman*, e normalmente utiliza-se a letra grega ρ (rho) para denotá-la. É uma medida de correlação não-paramétrica, isto é, ele avalia uma função monótona arbitrária que pode ser a descrição da relação entre duas variáveis, sem fazer nenhumas suposições sobre a distribuição de frequências das variáveis.

Em outras palavras, o coeficiente ρ de *Spearman* mede a intensidade da relação entre variáveis no mímino ordinais. Usa, em vez do valor observado, apenas a ordem das observações.

Deste modo, este coeficiente não é sensível a assimetrias na distribuição, nem à presença de outliers, não exigindo portanto que os dados provenham de duas populações normais.

Aplica-se igualmente em variáveis intervalares/rácio como alternativa ao coeficiente de correlação de *Pearson*, quando neste último se viola a normalidade. Nos caso em que os dados não formam uma nuvem “bem

comportada", com alguns pontos muito afastados dos restantes, ou em que parece existir uma relação crescente ou decrescente em formato de curva, o coeficiente ρ de *Spearman* é mais apropriado.

Uma fórmula fácil para calcular o coeficiente ρ de *Spearman* é dada por:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)};$$

Onde n é o número de pares $(X_i - Y_i)$ e d_i é a diferença entre cada posto dos pares $(X_i - Y_i)$.
Se os postos de x são exatamente iguais aos pontos de y, então todos os d_i serão zero e ρ será 1.

O coeficiente ρ de *Spearman* varia entre -1 e $+1$. Quanto mais próximo estiver destes extremos, maior será a associação entre as variáveis. O sinal negativo da correlação significa que as variáveis variam em sentido contrário, isto é, as categorias mais elevadas de uma variável estão associadas a categorias mais baixas da outra variável.

Observação: Ocasionalmente podem ocorrer empates entre os escores de dois indivíduos na mesma variável. Quando isto ocorre, a cada um deles é atribuído a média dos postos que seriam atribuídos caso o empate não ocorresse, isto é, adota-se o procedimento usual. Se a proporção de empates não é grande seu efeito sobre o coeficiente de correlação é desprezível. Quando a proporção de empates é grande torna-se necessário a utilização de um fator de correção. O efeito de postos empatados na variável X, consiste em reduzir a soma dos quadrados. Portanto, quando houver empates em X é necessário corrigir a soma dos quadrados, onde o coeficiente é definido por:

$$\rho = \frac{\sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n} d_i^2}{2\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}};$$

Onde $\sum_{i=1}^{n} x_i^2 = \frac{n^3-n}{12} - \sum_{i=1}^{n} T_x$ e $\sum_{i=1}^{n} y_i^2 = \frac{n^3-n}{12} - \sum_{i=1}^{n} T_y$, e $T = \frac{t^3-t}{12}$, tal que t é o número de observações empatadas em determinado posto.

Para verificar a significância do valor observado do coeficiente ρ de *Spearman*, para n entre 4 e 30, deve-se consultar a tabela abaixo. Para n maior ou igual a 10, pode utilizar a estatística de teste:

$$t = \rho\sqrt{\frac{n-2}{1-\rho^2}};$$

Onde $t \sim t_{n-2;\alpha/2}$, ou seja, t segue uma distribuição $t - Student$ com $n-2$ graus de liberdade, ao nível α de significância. Ressalta-se que este teste é idêntico ao teste para o coeficiente de correlação de *Pearson*.

Tabela

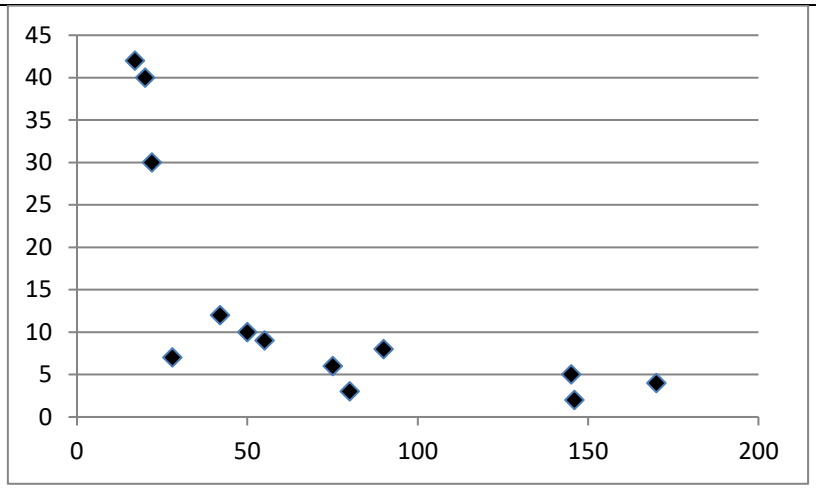
Valores Críticos de rS, coeficiente de correlação de Spearman

N	Nível de significância (unilateral)	
	0,05	0,01
4	1,000	
5	0,900	1,000
6	0,829	0,943
7	0,714	0,893
8	0,643	0,833
9	0,600	0,783
10	0,564	0,746
12	0,506	0,712
14	0,456	0,645
16	0,425	0,601
18	0,399	0,564
20	0,377	0,534
22	0,359	0,508
24	0,343	0,485
26	0,329	0,465
28	0,317	0,448
30	0,306	0,432

Exemplo: Uma amostra com 13 observações de duas variáveis X e Y foram coletadas e deseja-se verificar se as variáveis são correlacionadas. O gráfico de dispersão das variáveis, abaixo, sugere que há uma relação negativa, mas não linear entreas variáveis. Desta forma, utilizar-se-á o coeficiente de correlação de *Spearman* para checar ser as variáveis são correlacionadas.

Variável X	17	20	22	28	42	50	55	75	80	90	145	146	170

Variável Y	42	40	30	7	12	10	9	6	3	8	5	2	4



Cálculo do coeficiente ρ de *Spearman*:

Variável X	17	20	22	28	42	50	55	75	80	90	145	146	170
Variável Y	42	40	30	7	12	10	9	6	3	8	5	2	4
Postos de X	1	2	3	4	5	6	7	8	9	10	11	12	13
Postos de Y	13	12	11	6	10	9	8	5	2	7	4	1	3
$\boldsymbol{d_i}$	-12	-10	-8	-2	-5	-3	-1	3	7	3	7	11	10
$\boldsymbol{d_i^2}$	144	100	64	4	25	9	1	9	49	9	49	121	100

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} = 1 - \frac{6(144 + 100 + \cdots + 100)}{13(13^2-1)} = 1 - \frac{6 \times 717}{13(169-1)} = -0{,}9698;$$

As hipóteses são:

$$\begin{cases} H_o: A\ correlação\ entre\ X\ e\ Y\ é\ zero\ (\rho = 0) \\ H_1: A\ correlação\ entre\ \ X\ e\ Y\ não\ é\ zero\ (\rho \neq 0) \end{cases}.$$

A estatística teste sob o dados é:

$$t = \rho\sqrt{\frac{n-2}{1-\rho^2}} = -0{,}9698\sqrt{\frac{13-2}{1-0{,}9698^2}} = 13{,}5982;$$

Dado que $t \sim t_{n-2;\alpha/2}$, tem-se, a partir da estatística $t - Student$ com 11 graus de liberdade, os pontos críticos $\pm 2{,}2010$. Portanto, rejeita-se H_o ao nível de significância de 5%. Ou seja, a correlação entre as variáveis X e Y é diferente de zero, então, existe uma relação não-linear e negativa da ordem de $r = -0{,}9698$.

1.3. Coeficiente de Correlação de Kendall

O coeficiente de correlação τ por Postos de *Kendall* foi desenvolvido por *Maurice George Kendall*. É uma medida de correlação utilizada para dados ordinais, como no caso do coeficiente de correlação de *Spearman*. Ambas as variáveis devem ser medidas no mínimo em nível ordinal, de forma que seja possível atribuir postos a cada uma das variáveis.

O estimador do coeficiente de correlação τ por postos de *Kendall* é definido por:

$$\tau = \frac{S}{\frac{1}{2}n(n-1)};$$

Onde: n é o número de elementos aos quais atribui-se postos, S é a soma da variável Y à direita que são superiores menos o número de postos à direita que são inferiores.

Para o cálculo do coeficiente de correlação por postos de *Kendall* ordena-se inicialmente uma das variáveis em ordem crescente de postos e o S correspondente a cada elemento será obtido fazendo o número de elementos cujo posto é superior ao que se está calculando menos o número de elementos cujo posto é inferior ao mesmo.

Para verificar a significância do valor observado do coeficiente τ de *Kendall*, para $n \leq 10$ deve-se consultar a tabela abaixo. Para $n > 10$, pode utilizar a estatística de teste:

$$z = \frac{\tau - \mu_{H_0}}{\sigma_\tau};$$

Onde $\sigma_\tau = \sqrt{\frac{2(2n+5)}{9n(n-1)}}$ e $Z \sim N(0;1)$. Ou seja, z segue uma distribuição $Normal$ com média 0 e variância 1.

Observação: Pode-se fazer uma comparação entre coeficiente de correlação de *Spearman* e o coeficiente de correlação por postos de *Kendall*. Os valores numéricos não são iguais, quando calculados para os mesmos pares de postos, e não são comparáveis numericamente. Contudo, pelo fato de utilizarem a mesma quantidade de informação contida nos dados, ambos têm o

mesmo poder de detectar a existência de associação na população, e rejeitarão a hipótese nula para um mesmo nível de significância.

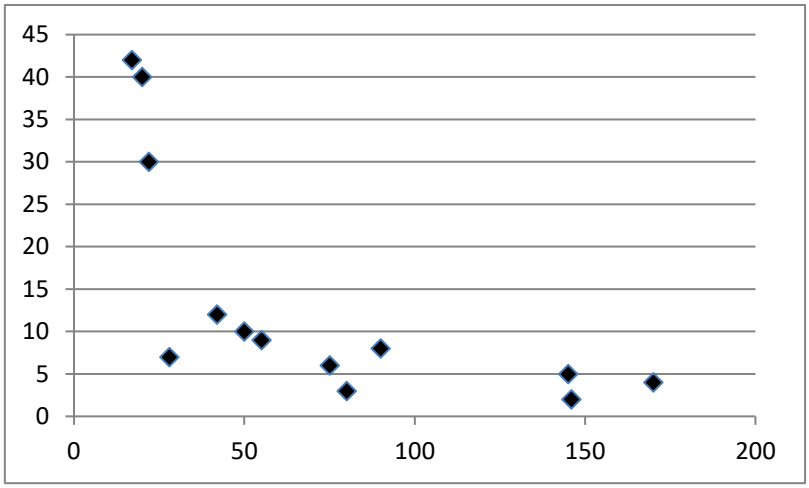
Tabela

Probabilidades associadas a valores tão grandes quanto os valores observados de S no coeficiente de correlação de Kendall

S	Valores de N				S	Valores de N		
	4	5	8	9		6	7	10
0	0,625	0,592	0,548	0,540	1	0,500	0,500	0,500
2	0,375	0,408	0,452	0,460	3	0,360	0,386	0,431
4	0,167	0,242	0,360	0,381	5	0,235	0,281	0,364
6	0,042	0,117	0,274	0,306	7	0,068	0,191	0,300
8		0,042	0,199	0,238	9	0,028	0,119	0,242
10		0,0083	0,138	0,179	11	0,0083	0,068	0,190
12			0,089	0,130	13	0,0014	0,035	0,146
14			0,054	0,090	15		0,015	0,108
16			0,031	0,060	17		0,0054	0,078
18			0,016	0,038	19		0,0014	0,054
20			0,0071	0,022	21		0,00020	0,036
22			0,0028	0,012	23			0,023
24			0,00087	0,0063	25			0,014
26			0,00019	0,0029	27			0,0083
28			0,000025	0,00012	29			0,0046
30				0,00043	31			0,0023
32				0,000012	33			0,0011
34				0,000025	35			0,00047
36				0,0000028	37			0,00018
					39			0,000058
					41			0,000015
					43			0,0000028
					45			0,00000028

Exemplo: Uma amostra com 13 observações de duas variáveis X e Y foram coletadas e deseja-se verificar se as variáveis são correlacionadas. O gráfico de dispersão das variáveis, abaixo, sugere que há uma relação negativa, mas não linear entreas variáveis. Desta forma, utilizar-se-á o coeficiente de correlação de *Kendall* para checar ser as variáveis são correlacionadas.

Variável X	17	20	22	28	42	50	55	75	80	90	145	146	170
Variável Y	42	40	30	7	12	10	9	6	3	8	5	2	4



Cálculo do coeficiente τ de *Kendall*:

Variável X	17	20	22	28	42	50	55	75	80	90	145	146	170
Variável Y	42	40	30	7	12	10	9	6	3	8	5	2	4
Postos de Y	13	12	11	6	10	9	8	5	2	7	4	1	3
Sup. Y à direita	0	0	0	4	0	0	0	1	3	0	0	1	0
Inf. Y à direita	12	11	10	5	8	7	6	4	1	3	2	0	0
S	−12	−11	−10	−1	−8	−7	−6	−3	2	−3	−2	1	0

$$\tau = \frac{S}{\frac{1}{2}n(n-1)} = \frac{-12-11-\cdots+1+0}{\frac{1}{2}\times 13(13-1)} = -0{,}7692$$

As hipóteses são:

$$\begin{cases} H_o: A\ correlação\ entre\ X\ e\ Y\ é\ zero\ (\tau = 0) \\ H_1: A\ correlação\ entre\ \ X\ e\ Y\ não\ é\ zero\ (\tau \neq 0) \end{cases}.$$

A estatística teste sob o dados é:

$$z = \frac{\tau - \mu_{H_0}}{\sigma_\tau} = \frac{\tau - \mu_{H_0}}{\sqrt{\frac{2(2n+5)}{9n(n-1)}}} = \frac{-0{,}7692 - 0}{\sqrt{\frac{2(2\times 13+5)}{9\times 13(13-1)}}} = -3{,}6604;$$

Dado que $Z \sim N(0;1)$, tem-se, a partir da estatística $Normal\ padrão$, os pontos críticos $\pm 1{,}96$. Portanto, rejeita-se H_o ao nível de significância de 5%. Ou seja, a correlação entre as variáveis X e Y é diferente de zero, então, existe uma relação não-linear e negativa da ordem de $\tau = -0{,}7692$.

Exercício: Verifique se existe associação, e qual o seu grau de associação, entre as notas obtidas por estudantes em Estatística e o seu quociente de inteligência – QI, por meio de uma amostra de 10 estudantes. Mensure os três coeficientes de correlação apresentados – *Pearson*, *Spearman* e *Kendall*.

Notas	8	14	18	10	6,5	9	14	5,2	10	13
Q I	70	190	304	100	42	80	169	27	105	159

Exercício: Verifique se existe associação, e qual o seu grau de associação, entre o tempo de experiência profissional e a remuneração destes profissionais, por meio de uma amostra de 12 profissionais. Mensure os três coeficientes de correlação apresentados – *Pearson*, *Spearman* e *Kendall* e teste a significância ao nível de 2%.

Tempo Exper.	12	20	13	4	19	12	17	12	13	7	6	9
Remuneração	4695	7525	5045	2580	7130	4695	6395	4720	5020	3295	3055	3830

ESTIMADORES DA CORRELAÇÃO

CORRELAÇÃO DE PEARSON

$$r_{x;y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}; \qquad t = \frac{r-\rho}{\sqrt{\frac{1-r^2}{n-2}}} \sim t_{n-2;\alpha/2}$$

CORRELAÇÃO DE SPEARMAN

$$\hat{\rho} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}; \qquad t = \frac{r-\rho}{\sqrt{\frac{1-r^2}{n-2}}} \sim t_{n-2;\alpha/2}$$

Onde n é o número de pares $(X_i - Y_i)$ e d_i é a diferença entre cada posto dos pares $(X_i - Y_i)$.

CORRELAÇÃO DE KENDALL

$$\tau = \frac{S}{\frac{1}{2}n(n-1)}; \quad z = \frac{\tau - \rho}{\sqrt{\frac{2(2n+5)}{9n(n-1)}}} \sim N(0;1)$$

Onde n é o número de pares $(X_i - Y_i)$ e S é a soma da variável Y à direita que são superiores menos o número de postos à direita que são inferiores.