
FIAP
MBA

FIAP MBA+

MBA em DATA SCIENCE

STATISTICS FUNDAMENTALS

Dra. Regina Tomie Ivata Bernal

Cientista de Dados na área da Saúde

Formação Acadêmica:

- **Estatístico - UFSCar**
- **Mestre em Saúde Pública – FSP/USP**
- **Doutor em Ciências – Epidemiologia - FSP/USP**

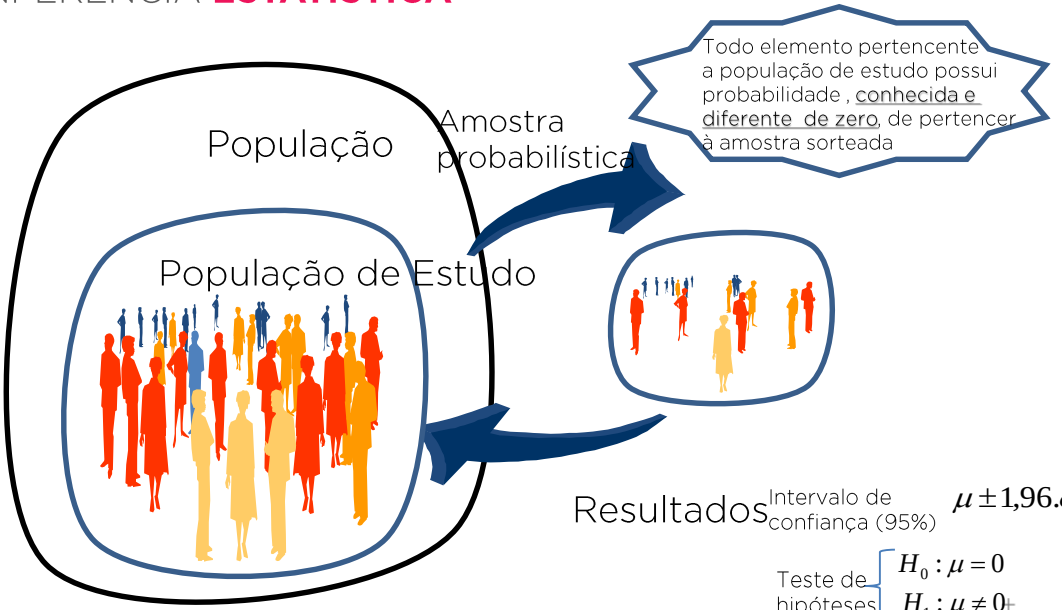
profregina.bernal@fiap.com.br
reginabernal@terra.com.br

Atividades Profissionais:

- **Professora de pós-graduação na FIAP**
- **Consultora externa da SVS/MS**
- **Cientista de Dados em Saúde**

Aula 4
Inferência Estatística
Teste de Hipóteses
Análise de Associação

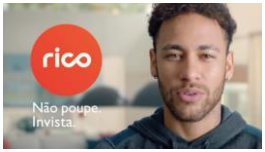
INFERÊNCIA **ESTATÍSTICA**

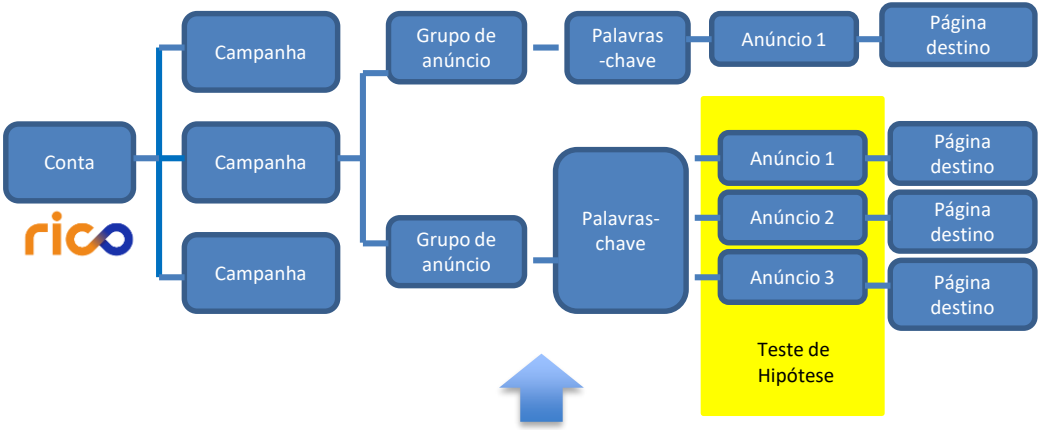


TESTE DE **HIPÓTESES**

Planejamento de campanhas

EXEMPLO

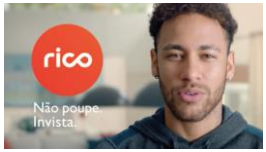


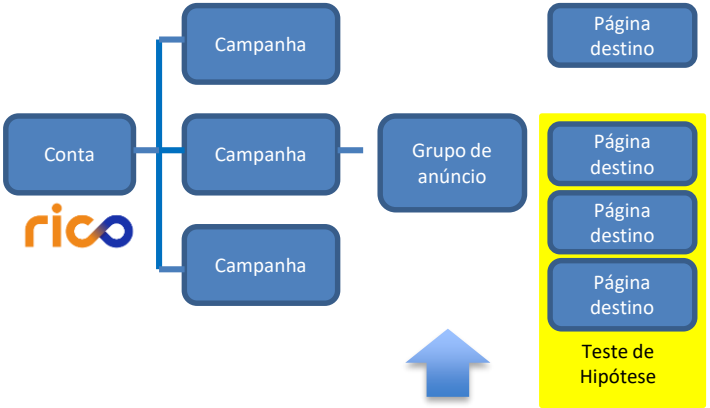


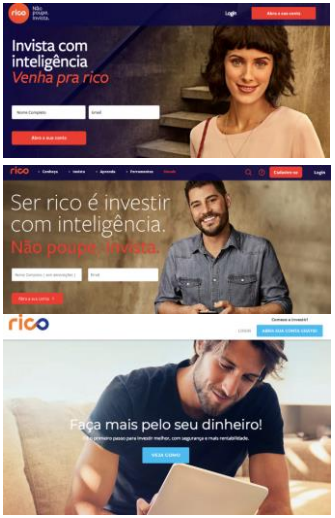
Não poupe. Invista. Abra a sua conta na Rico

Planejamento de campanhas

EXEMPLO







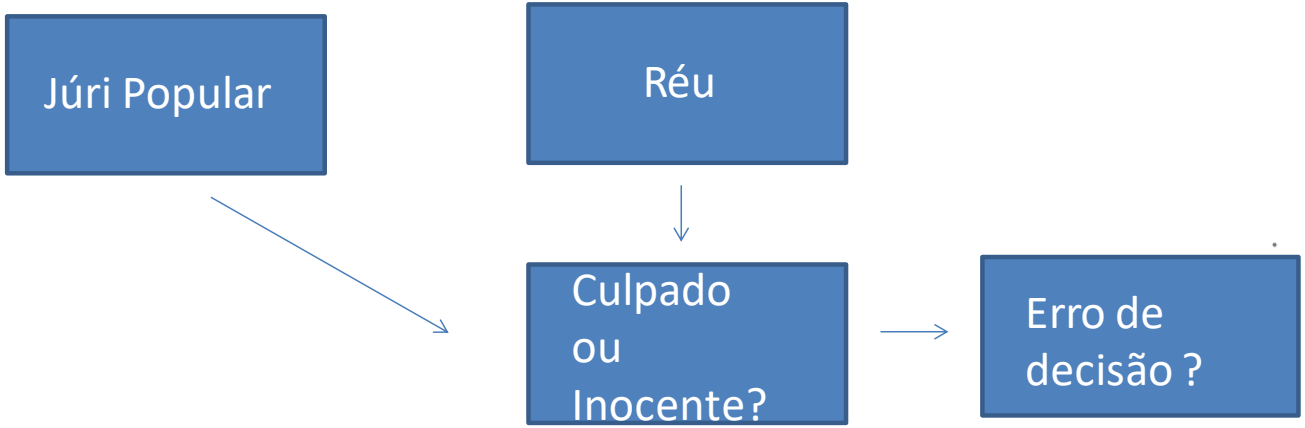
Não poupe. Invista.
Abra a sua conta na Rico

Teste de Hipóteses

H0: Inocente
H1: Culpado

Exemplo:

Erro=5%



Inferência Estatística

Estimação: Teste de hipóteses

- Tipos de erros

DECISÃO	REALIDADE DE H_0	
	VERDADE	FALSA
REJEITA	ERRO TIPO I α	Acerto
NÃO REJEITA	Acerto	ERRO TIPO II β

PLANEJAMENTO DE **CAMPANHAS**

Tipos de testes

1. Identificar a melhor forma de passar a mensagem para que a pessoa entenda.
2. Melhorar a acuracidade dos Leads depende do público-alvo.
3. Identificar o horário com maior conversão.
4. Uma celebridade/influenciador ajuda a aumentar a credibilidade da marca?
5. Comparar a conversão por dispositivo.
6. Entre outros.

PLANEJAMENTO DE **CAMPANHAS**

Teste de Hipótese

1. Problema.
2. Hipóteses estatísticas.
3. Fixa critério de decisão.....para decidir rejeição de Ho.
4. Calcular o tamanho da amostra.
5. Sorteia a amostra aleatória simples (n>30).
6. Estimativas da média e erro padrão ($\bar{x}$; $ep(\bar{x}) = \frac{s_x}{\sqrt{n}}$).
7. Decisão.....Rejeita/Não Rejeita H0, com risco de errar igual a α .

EXEMPLO

TESTE DE **HIPÓTESES**

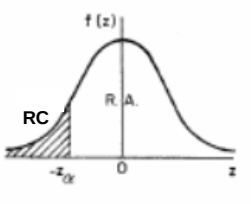
Exemplo: Desejando-se conhecer a média de gasto anual com medicamentos na cidade Y, selecionou-se uma amostra aleatória de 100 adultos maiores de 40 anos. Teste a hipótese de que o gasto anual dessa população é inferior ao gasto médio de R$ 120,00 a.a. com nível de significância de 5%?

Hipótese estatística: H1: $\mu < 120$

H1: $\mu < 120$

Critério de decisão: $\alpha = 0,05$

$z_{0,05} = 1,64$



RA: Região de aceitação de H0
RC: Região crítica de rejeição de H0

- Sortear uma amostra aleatória de 100 adultos

EXEMPLO

TESTE DE **HIPÓTESES**

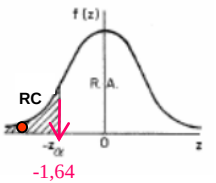
Exemplo:

Resultados da pesquisa:

	N	Mean	Std. Deviation	Std. Error Mean
gasto	100	95,10	63,333	6,333

	Test Value = 120					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
gasto	-3,932	99	,000	-24,900	-37,47	-12,33

Ao nível de 5% de significância, há evidências para rejeição de H_0. Portanto, o gasto médio anual de medicamentos
na população de adultos maiores de 40 anos residentes na cidade Y é inferior a R$ 120.

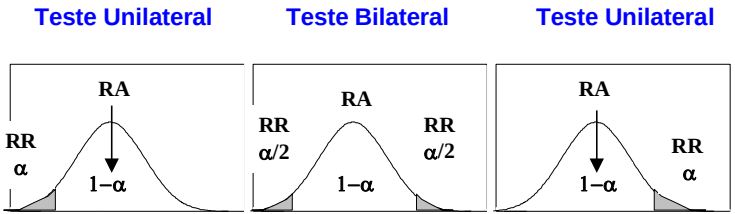


Conceito

Teste de Hipóteses

De acordo com a formulação das hipóteses, os testes podem ser monocaudal ou bicaudal.

- Monocaudal:
  - H_0: $\mu = \mu_0$
  - H_1: $\mu > \mu_0$
  - H_0: $\mu = \mu_0$
  - H_1: $\mu < \mu_0$
- Bicaudal
  - H_0: $\mu = \mu_0$
  - H_1: $\mu \neq \mu_0$



onde: RA = região de aceitação e RR = região de rejeição

Conceito

p-valor

É uma estimativa do nível de significância observada na amostra. Indica a probabilidade de ocorrer valores da estatística mais extremos do que o observado , sob a hipótese H_0 ser verdadeira.

Se p-valor $< \alpha$ então rejeito H_0, caso contrário não rejeito.

TESTE A/B

Origem: Wikipédia, a enciclopédia livre.

Teste A/B é um método de teste de design através do qual comparam-se elementos aleatórios com duas variantes, A e B, em que estes são o controle e o tratamento de uma experiência controlada, com o objetivo de melhorar a percentagem de aprovação. Estas experiências são muito utilizadas em desenvolvimento web e de marketing, e até mesmo em formas tradicionais de publicidade. Teste A/B também se designa por experiência aleatória controlada, experiência online controlada e teste de divisão. Em web design, o teste A/B é utilizado para identificar alterações nas páginas web que podem provocar mudanças positivas ou negativas no interesse dos utilizadores. Como o nome já diz, duas versões são comparadas, as quais são idênticas exceto por uma variante que pode impactar o comportamento do utilizador. A versão A pode ser a versão utilizada atualmente (controle), enquanto a Versão B é a modificada (tratamento). Podem ser vistas melhorias significativas através de testes de elementos como copiar o texto, layouts, imagens e cores, mas nem sempre. Os testes multivariados ou teste de balde são semelhantes ao teste A/B, mas estes testes abordam mais de duas versões diferentes ao mesmo tempo. [1]

Referências

1. ↑ «Split Testing Guide for Online Stores». *webics.com.au* (em inglês). Consultado em 17 de setembro de 2018

Fonte: https://pt.wikipedia.org/wiki/Teste_A/B

TESTE DE **HIPÓTESE**

Origem: Wikipédia, a enciclopédia livre.

Teste de hipóteses, **teste estatístico** ou **teste de significância**[1] é um procedimento estatístico que permite tomar uma decisão (aceitar ou rejeitar a hipótese nula H_0) entre duas ou mais hipóteses (hipótese nula H_0 ou hipótese alternativa H_1), utilizando os dados observados de um determinado experimento.[2] Há diversos métodos para realizar o teste de hipóteses, dos quais se destacam o método de Fisher (teste de significância),[3] o método de Neyman–Pearson[4] e o método de Bayes.[5]

Por meio da teoria da probabilidade, é possível inferir sobre quantidades de interesse de uma população a partir de uma amostra observada de um experimento científico. Por exemplo, estimar pontualmente e de forma intervalar um parâmetro de interesse, testar se uma determinada teoria científica deve ser descartada, verificar se um lote de remédios deve ser devolvido por falta de qualidade, entre outros. Por meio do rigor matemático, a inferência estatística pode ser utilizada para auxiliar a tomada de decisões nas mais variadas áreas.[6]

Os testes de hipóteses são utilizados para determinar quais resultados de um estudo científico podem levar à rejeição da hipótese nula H_0 a um nível de significância pré–estabelecido. O estudo da teoria das probabilidades e a determinação da estatística de teste correta são fundamentais para a coerência de um teste de hipótese. Se as hipóteses do teste de hipóteses não forem assumidas de maneira correta, o resultado será incorreto e a informação será incoerente com a questão do estudo científico. Os tipos conceituais de erro (erro do tipo I e erro do tipo II) e os limites paramétricos ajudam a distinguir entre a hipótese nula H_0 e a hipótese alternativa H_1.[7]

Fonte: https://pt.wikipedia.org/wiki/Testes_de_hipóteses

TESTE DE **HIPÓTESES**

Origem: Wikipédia, a enciclopédia livre.

São fundamentais os seguintes conceitos para um teste de hipóteses:[7]

- Hipótese nula (H_0): é a hipótese assumida como verdadeira para a construção do teste. É a teoria, o efeito ou a alternativa que se está interessado em testar.
- Hipótese alternativa (H_1): é considerada quando a hipótese nula não tem evidência estatística.
- Erro do tipo I (α): é a probabilidade de se rejeitar a hipótese nula quando ela é verdadeira.
- Erro do tipo II: é a probabilidade de se rejeitar a hipótese alternativa quando ela é verdadeira.

	Hipótese nula H_0 é verdadeira	**Hipótese nula H_0 é falsa**
Hipótese nula H_0 é rejeitada	Erro do tipo I	Não há erro
Hipótese nula H_0 não é rejeitada	Não há erro	Erro do tipo II

Fonte: https://pt.wikipedia.org/wiki/Testes_de_hipóteses

ANÁLISE **DE ASSOCIAÇÃO**

ANÁLISE DE **ASSOCIAÇÃO**

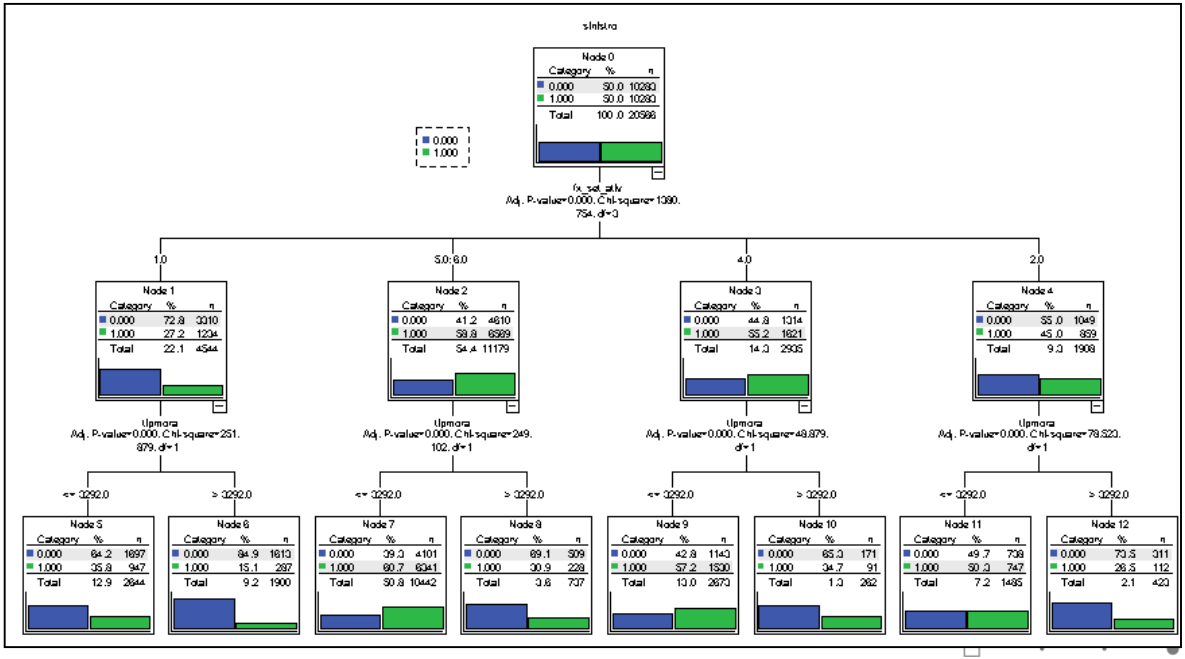
Analisa o comportamento conjunto de duas variáveis qualitativas apresentada em tabela bivariada.

- Teste Qui-Quadrado (Variáveis Qualitativas)
- Correlação de Pearson (Variáveis Quantitativas)

TÉCNICA DE CLASSIFICAÇÃO:

ÁRVORE DE DECISÃO

EXEMPLO

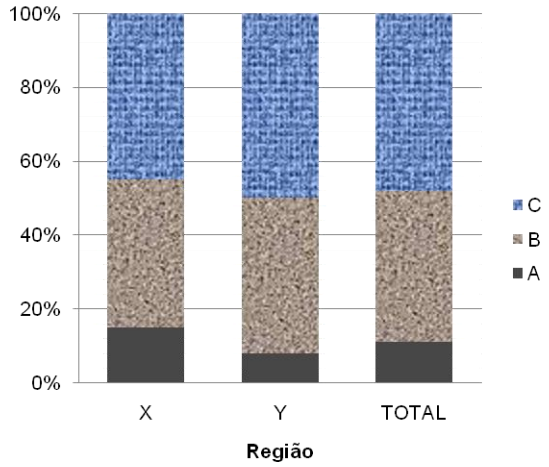


Existe associação entre as vendas de produto e região?

Exemplo:

Tabela 1 – Distribuição de vendas segundo produto e região. 2018

Produto	Região				Total	
	X		Y			
	N	%	N	%	N	%
A	300	15	200	8	500	11
B	800	40	1000	42	1800	41
C	900	45	1200	50	2100	48
Total	2000	100	2400	100	4400	100



Existe associação entre as vendas de produto e região?

Existe associação entre as vendas de produto e região?

		REGIAO		
		X	Y	TOTAL
Produto	A	300	200	500
	B	800	1000	1800
	C	900	1200	2100
		2000	2400	4400

Chi Square for R by C Table

Chi Square=	49.12
Degrees of Freedom=	2
p-value=	<0.0000001

Cochran recommends accepting the chi square if:
1. No more than 20% of cells have expected < 5.
2. No cell has an expected value < 1.

In this table:
None of 6 cells have expected values < 5.
No cells have expected values < 1.

Using these criteria, this chi square can be accepted.

Expected value = row total*column total/grand total

Rosner, B. Fundamentals of Biostatistics. 5th ed. Duxbury Thompson Learning. 2000; p. 395

Teste de independência qui-quadrado

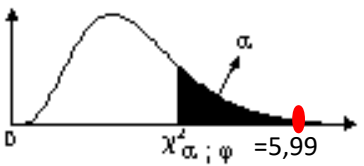
H_0: independentes

H_1: dependentes

α= 5%

Conclusão:
Rejeito H0, portanto há associação.

Graus de liberdade=



φ = graus de liberdade

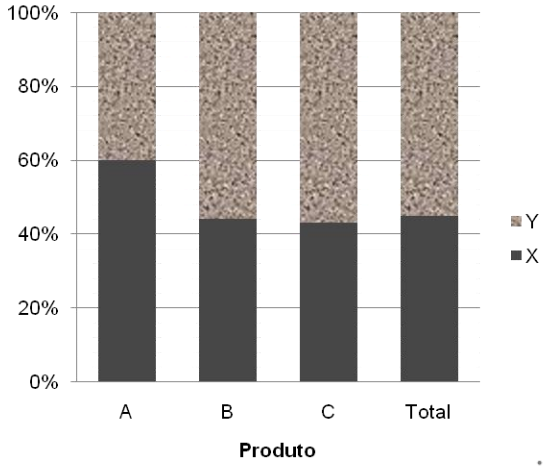
Existe associação entre as vendas de produto e região?

Exemplo:

Tabela 2 – Distribuição de vendas segundo região e produto. 2018

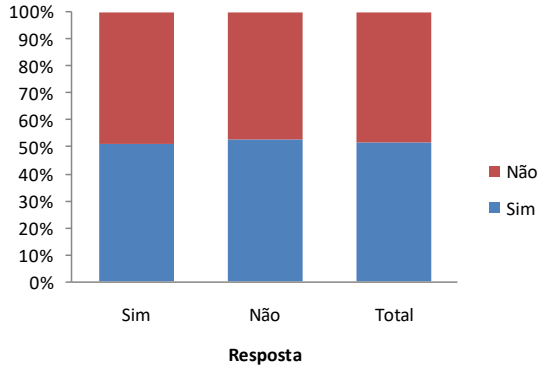
Produto	Região				Total	
	X		Y			
	N	%	N	%	N	%
A	300	60	200	40	500	100
B	800	44	1000	56	1800	100
C	900	43	1200	57	2100	100
Total	2000	45	2400	55	4400	100

Fonte:zzzz



Será que uma carta de pré-notificação afeta a taxa de resposta dos médicos participantes da pesquisa?

Resposta	Carta				Total	
	Sim		Não			
	N	%	N	%	N	%
Sim	2570	51.2	2645	52.6	5215	51.9
Não	2448	48.8	2384	47.4	4832	48.1
Total	5018	100.0	5029	100.0	10047	100.0



Será que uma carta de pré-notificação afeta a taxa de resposta dos médicos participantes da pesquisa?

2 x 2 Table Statistics

Single Table Analysis

		Carta: Sim	Carta: Nao	
Resposta	Sim	2570	2645	5215
	Nao	2448	2384	4832
		5018	5029	10047

Chi Square and Exact Measures of Association

Test	Value	p-value(1-tail)	p-value(2-tail)
Uncorrected chi square	1.914	0.08332	0.1666
Yates corrected chi square	1.859	0.08644	0.1729
Mantel-Haenszel chi square	1.914	0.08333	0.1667
Fisher exact		'?'(P)	'?'
Mid-P exact		'?'(P)	'?'

All expected values (row total*column total/grand total) are >=5
OK to use chi square.

Conclusão:
Não rejeito H0, portanto
Não há associação.

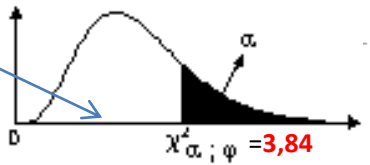
Teste de independência qui-quadrado

H_0: independentes

H_1: dependentes

α= 5%

Graus de liberdade= 1



φ = graus de liberdade

TESTE DE **HIPÓTESES NÃO PARAMÉTRICO**

Teste Qui-Quadrado (Independência)

Distribuições bivariadas de frequências para variáveis qualitativas são apresentadas em tabelas de contingência, que facilitam a análise estatística da possível relação entre duas características observadas em determinada população. A estatística chamada qui-quadrado sintetiza as diferenças entre as frequências observadas de uma tabela bivariada e as correspondentes frequências esperadas.

Definindo as hipóteses H_0 e H_1

H_0: as variáveis são independentes
H_1: as variáveis não são independentes, isto é, apresentam algum grau de associação entre si.

TESTE DE **HIPÓTESES NÃO PARAMÉTRICO**

VANTAGENS

Não é necessário fazer suposições sobre a distribuição da população da qual tenham sido extraídos os dados para a análise.

Aplicáveis a variáveis não contínuas (variáveis categóricas nominais e ordinais).

Simplicidade do ponto de vista de cálculo.

Aplicabilidade a pequenas amostras. Estudo Piloto (ex: amostras de pessoas portadoras de uma certa doença)

➔As técnicas não-paramétricas são em geral menos eficazes que as paramétricas quando aplicadas a dados onde é possível usar as técnicas paramétricas.

TESTE DE **HIPÓTESES NÃO PARAMÉTRICO**

A estatística Qui-Quadrado de Pearson é calculada pela expressão

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{k}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Sob H_0, tem distribuição Qui-Quadrado, com (r-1)(c-1) graus de liberdade, sendo r o número de linhas e c o número de colunas

Onde: O_{ij} = número de casos observados classificados na linha i da coluna j;
E_{ij} = número de casos esperados , sob Ho, na linha i da coluna j;

$\sum_{i=1}^{r}\sum_{j=1}^{k}$ Indica somatório sobre todas as células

Distribuição do Qui-Quadrado - χ^2_n

Os valores tabelados correspondem aos pontos x tais que: $P(\chi^2_n \leq x)$

	$P(\chi^2_n \leq x)$													
n	0,005	0,01	0,025	0,05	0,1	0,25	0,5	0,75	0,9	0,95	0,975	0,99	0,995	
1	3,93E-05	0,000157	0,000982	0,003932	0,016	0,102	0,455	1,323	2,706	3,841	5,024	6,635	7,879	1
2	0,010	0,020	0,051	0,103	0,211	0,575	1,386	2,773	4,605	5,991	7,378	9,210	10,597	2
3	0,072	0,115	0,216	0,352	0,584	1,213	2,366	4,108	6,251	7,815	9,348	11,345	12,838	3
4	0,207	0,297	0,484	0,711	1,064	1,923	3,357	5,385	7,779	9,488	11,143	13,277	14,860	4
5	0,412	0,554	0,831	1,145	1,610	2,675	4,351	6,626	9,236	11,070	12,832	15,086	16,750	5
6	0,676	0,872	1,237	1,635	2,204	3,455	5,348	7,841	10,645	12,592	14,449	16,812	18,548	6
7	0,989	1,239	1,690	2,167	2,833	4,255	6,346	9,037	12,017	14,067	16,013	18,475	20,278	7
8	1,344	1,647	2,180	2,733	3,490	5,071	7,344	10,219	13,362	15,507	17,535	20,090	21,955	8
9	1,735	2,088	2,700	3,325	4,168	5,899	8,343	11,389	14,684	16,919	19,023	21,666	23,589	9
10	2,156	2,558	3,247	3,940	4,865	6,737	9,342	12,549	15,987	18,307	20,483	23,209	25,188	10
11	2,603	3,053	3,816	4,575	5,578	7,584	10,341	13,701	17,275	19,675	21,920	24,725	26,757	11
12	3,074	3,571	4,404	5,226	6,304	8,438	11,340	14,845	18,549	21,026	23,337	26,217	28,300	12
13	3,565	4,107	5,009	5,892	7,041	9,299	12,340	15,984	19,812	22,362	24,736	27,688	29,819	13
14	4,075	4,660	5,629	6,571	7,790	10,165	13,339	17,117	21,064	23,685	26,119	29,141	31,319	14
15	4,601	5,229	6,262	7,261	8,547	11,037	14,339	18,245	22,307	24,996	27,488	30,578	32,801	15
16	5,142	5,812	6,908	7,962	9,312	11,912	15,338	19,369	23,542	26,296	28,845	32,000	34,267	16
17	5,697	6,408	7,564	8,672	10,085	12,792	16,338	20,489	24,769	27,587	30,191	33,409	35,718	17
18	6,265	7,015	8,231	9,390	10,865	13,675	17,338	21,605	25,989	28,869	31,526	34,805	37,156	18
19	6,844	7,633	8,907	10,117	11,651	14,562	18,338	22,718	27,204	30,144	32,852	36,191	38,582	19
20	7,434	8,260	9,591	10,851	12,443	15,452	19,337	23,828	28,412	31,410	34,170	37,566	39,997	20
21	8,034	8,897	10,283	11,591	13,240	16,344	20,337	24,935	29,615	32,671	35,479	38,932	41,401	21
22	8,643	9,542	10,982	12,338	14,041	17,240	21,337	26,039	30,813	33,924	36,781	40,289	42,796	22
23	9,260	10,196	11,689	13,091	14,848	18,137	22,337	27,141	32,007	35,172	38,076	41,638	44,181	23
24	9,886	10,856	12,401	13,848	15,659	19,037	23,337	28,241	33,196	36,415	39,364	42,980	45,558	24
25	10,520	11,524	13,120	14,611	16,473	19,939	24,337	29,339	34,382	37,652	40,646	44,314	46,928	25
26	11,160	12,198	13,844	15,379	17,292	20,843	25,336	30,435	35,563	38,885	41,923	45,642	48,290	26
27	11,808	12,878	14,573	16,151	18,114	21,749	26,336	31,528	36,741	40,113	43,195	46,963	49,645	27
28	12,461	13,565	15,308	16,928	18,939	22,657	27,336	32,620	37,916	41,337	44,461	48,278	50,994	28
29	13,121	14,256	16,047	17,708	19,768	23,567	28,336	33,711	39,087	42,557	45,722	49,588	52,335	29
30	13,787	14,953	16,791	18,493	20,599	24,478	29,336	34,800	40,256	43,773	46,979	50,892	53,672	30
40	20,707	22,164	24,433	26,509	29,051	33,660	39,335	45,616	51,805	55,758	59,342	63,691	66,766	40
50	27,991	29,707	32,357	34,764	37,689	42,942	49,335	56,334	63,167	67,505	71,420	76,154	79,490	50

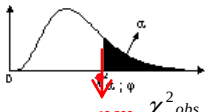
TESTE DE **HIPÓTESES NÃO PARAMÉTRICO**

Exemplo:

A Associação de Imprensa do Estado de São Paulo fez um levantamento com 1300 leitores, para verificar se a preferência por leitura de um determinado jornal é independente do nível de instrução do indivíduo. Os resultados obtidos foram:

Observado	A	B	C	D	Total
1. Grau	10 20%	8 16%	5 10%	27 54%	50
2. Grau	90 20%	162 36%	125 28%	73 16%	450
Universitário	200 25%	250 31%	220 28%	130 16%	800
Total	300 23%	420 32%	350 27%	230 18%	1300

Esperado	A	B	C	D	Total
1. Grau	12	16	13	9	50
2. Grau	104	145	121	80	450
Universitário	185	258	215	142	800
Total	300	420	350	230	1300



χ^2_{obs} =53,9

Gl= (linha -1)*(colunas-1) = (3-1)*(4-1) = 6

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{k}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Rejeitamos Ho ao nível de 5%, isto é os dados trazem evidência de uma forte dependência entre o fatores: Grau de escolaridade e preferência de jornal

TESTE DE **HIPÓTESES NÃO PARAMÉTRICO**

Exemplo:

Após uma pesquisa de satisfação estamos interessados em verificar se a preferência pela operadora(OP) estava associada com o fator regional.

Estado e Operadora	OP1	OP2	OP3	OP4	Total
SP	214 33%	237 37%	78 12%	119 18%	648 100%
Sul	51 17%	102 34%	126 42%	22 7%	301 100%
RJ	111 18%	304 51%	139 23%	48 8%	602 100%
Total	376 24%	643 42%	343 22%	189 12%	1551 100%

➔ Se tivesse independência todos os estados teriam a mesma distribuição

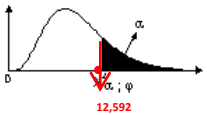
➔ Número esperado em SP: 648*0.24=157; Sul: 301*0.24= 73; RJ: 602*0.24=146 ;

H_0: **São independentes**

H_a: **Não são independentes**

α= **5%**

Graus de liberdade= **(l-1)(c-1) = 3*2 = 6**



TESTE DE HIPÓTESES NÃO PARAMÉTRICO

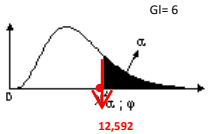
Exemplo:

Valores Observados

Estado e Operadora	OP1	OP2	OP3	OP4	Total
SP	214	237	78	119	648
Sul	51	102	126	22	301
RJ	111	304	139	48	602
Total	376	643	343	189	1.551

Valores Esperado

Estado e Operadora	OP1	OP2	OP3	OP4	Total
SP	157	269	143	79	648
Sul	73	125	67	37	301
RJ	146	250	133	73	602
Total	376	643	343	189	1.551



$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{k}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

χ^2_{obs} = (214-157)²/157.....(48-73)²/73=173.24

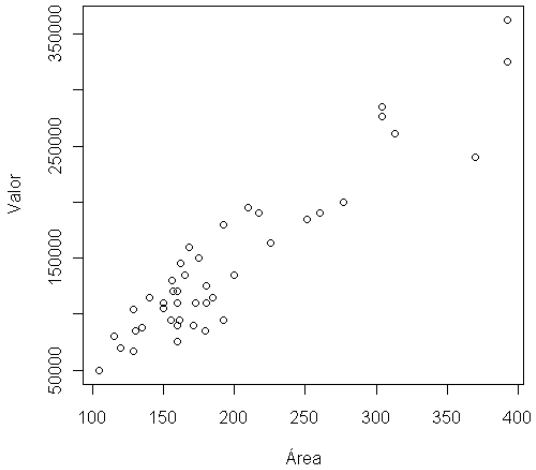
Rejeitamos Ho ao nível de 5%, isto é os dados trazem evidência de uma forte dependência entre o fatores: Operadora de Celular e Região

ANÁLISE DE **ASSOCIAÇÃO**

Analisa o comportamento conjunto de duas variáveis qualitativas apresentada em tabela bivariada.

- Teste Qui-Quadrado (Variáveis Qualitativas)
- Correlação de Pearson (Variáveis Quantitativas)

Existe correlação entre o valor do imóvel e a área?



Teste Correlação de Pearson

H_0: r = 0 (ausência de correlação)

H_1: r≠ 0 (presença de correlação)

Erro de decisão: 0,05 ou 5%

R Console

```
> corr_t<-cor.test(Valor,Área,method="pearson",alternative="two.sided")
> corr_t

        Pearson's product-moment correlation

data:  Valor and Área
t = 17.0563, df = 41, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8845899 0.9651600
sample estimates:
      cor
0.9362024
```

Conclusão:

Rejeito H0, portanto há associação.

CORRELAÇÃO DE **PEARSON**

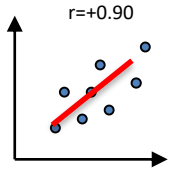
Correlação indica a força e a direção do relacionamento linear entre duas variáveis aleatórias. No uso estatístico geral, correlação se refere à medida da relação entre duas variáveis, embora correlação não implique causalidade. Nesse sentido geral, existem vários coeficientes medindo o grau de correlação, adaptados à natureza dos dados.

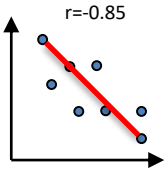
Vários coeficientes são utilizados para situações diferentes. O mais conhecido é o coeficiente de correlação de Pearson, o qual é obtido dividindo a covariância de duas variáveis pelo produto de seus desvios padrão. Apesar do nome, ela foi apresentada inicialmente por Francis Galton, em meados do século XVII.

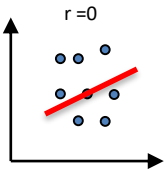
Coeficiente de correlação de Pearson, em geral é expresso por (R ou ρ).

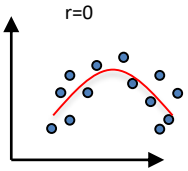
CORRELAÇÃO DE **PEARSON**

Análise de correlação









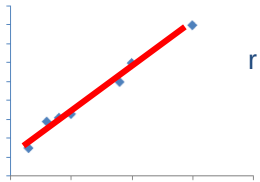
Correlação Linear Simples (r de Pearson)

$$\frac{\sum_{i=1}(X_i - \bar{X}) * (Y_i - \bar{Y})}{\sqrt{\sum_{i=1}(X_i - \bar{X})^2 * \sum_{i=1}(Y_i - \bar{Y})^2}}$$

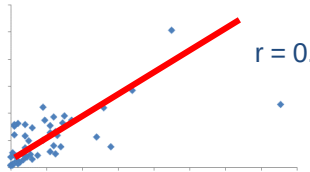
- Para avaliar-se a correlação entre variáveis, é importante conhecer a magnitude ou força tanto quanto a significância da correlação.

CORRELAÇÃO DE **PEARSON**

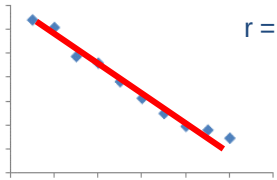
Análise de correlação



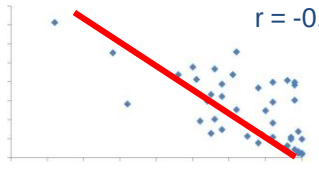
r = 1



r = 0.66

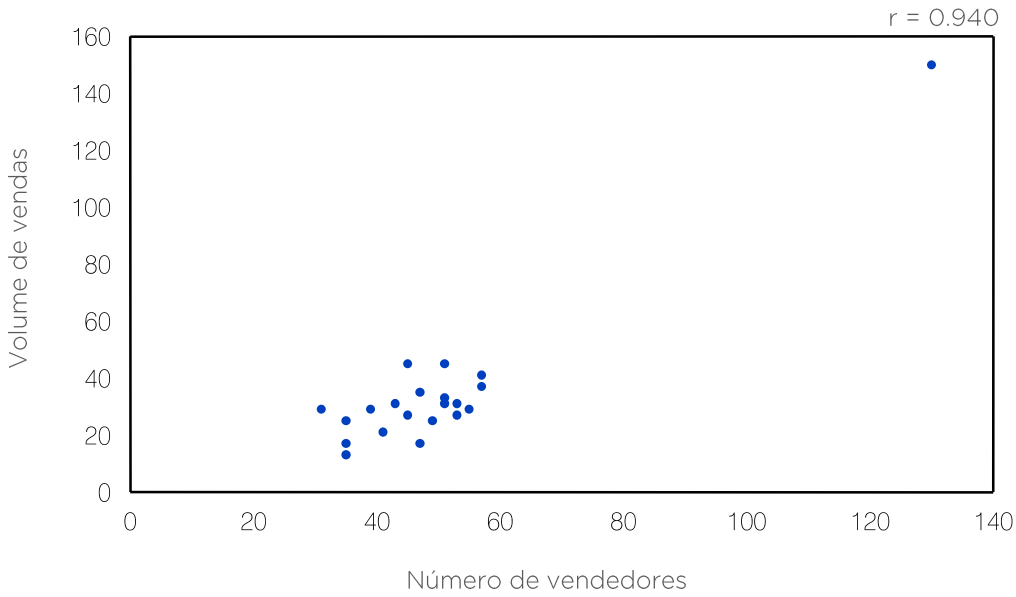


r = -1

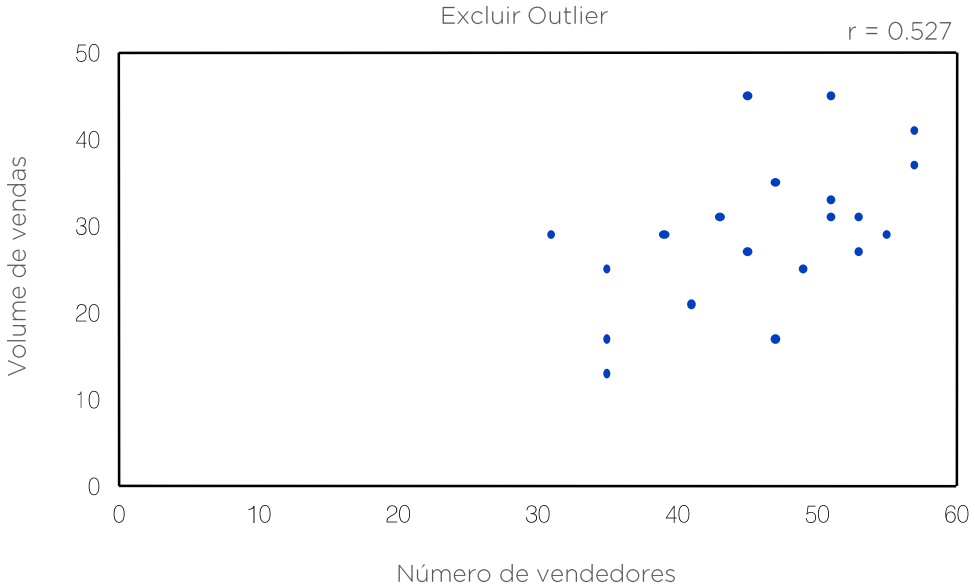


r = -0.7

Volume de vendas X Número de vendedores



Volume de vendas X Número de vendedores

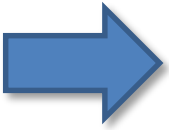


Associações Espúrias

- Associação entre dois fatores e quando queremos saber se um causa o outro ?

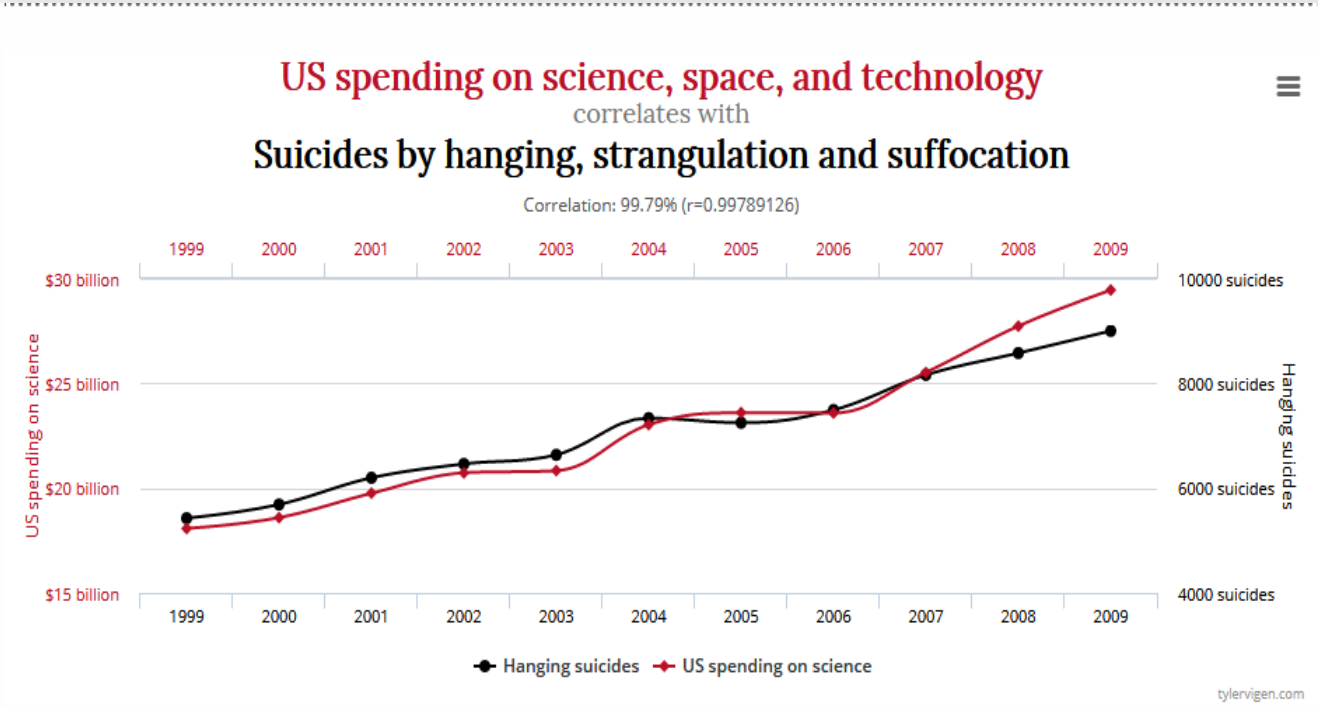
- Big data muitos resultados estatisticamente significativos que não fazem sentido causal

- variável de confusão quando há muitas variáveis na análise



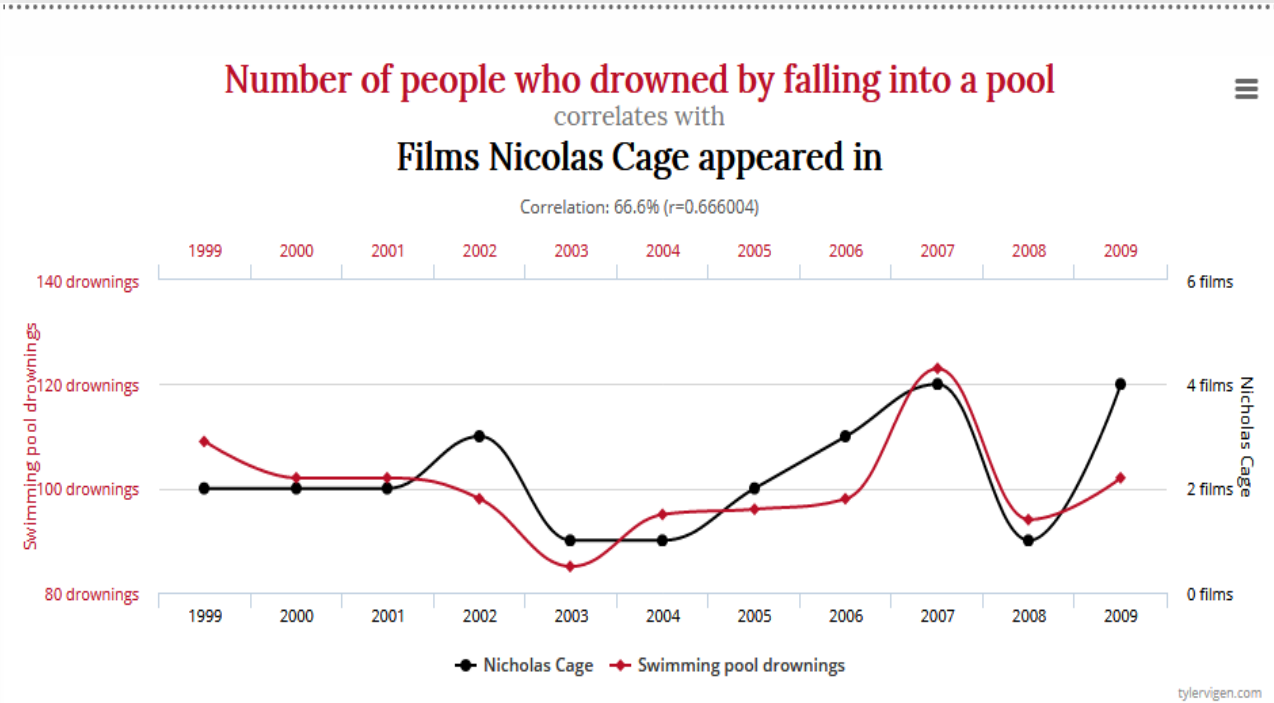
Uma relação estatística existente entre duas variáveis, mas onde não existe nenhuma relação causa-efeito entre elas. Essa relação estatística pode ocorrer por pura coincidência ou por causa de uma terceira variável.

Associações Espúrias



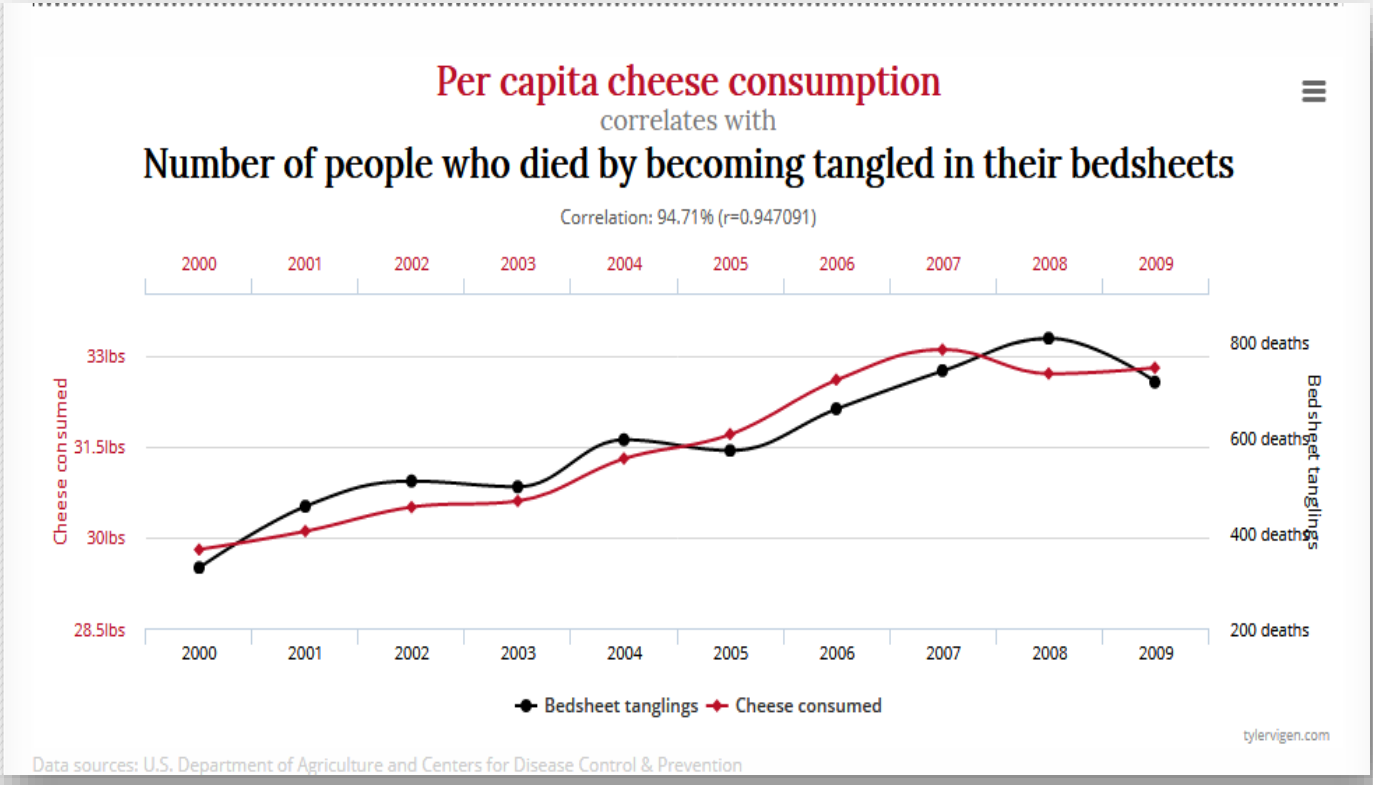
Fonte de dados: U.S. Office of Management and Budget e Centers for Disease Control & Prevention

Associações Espúrias



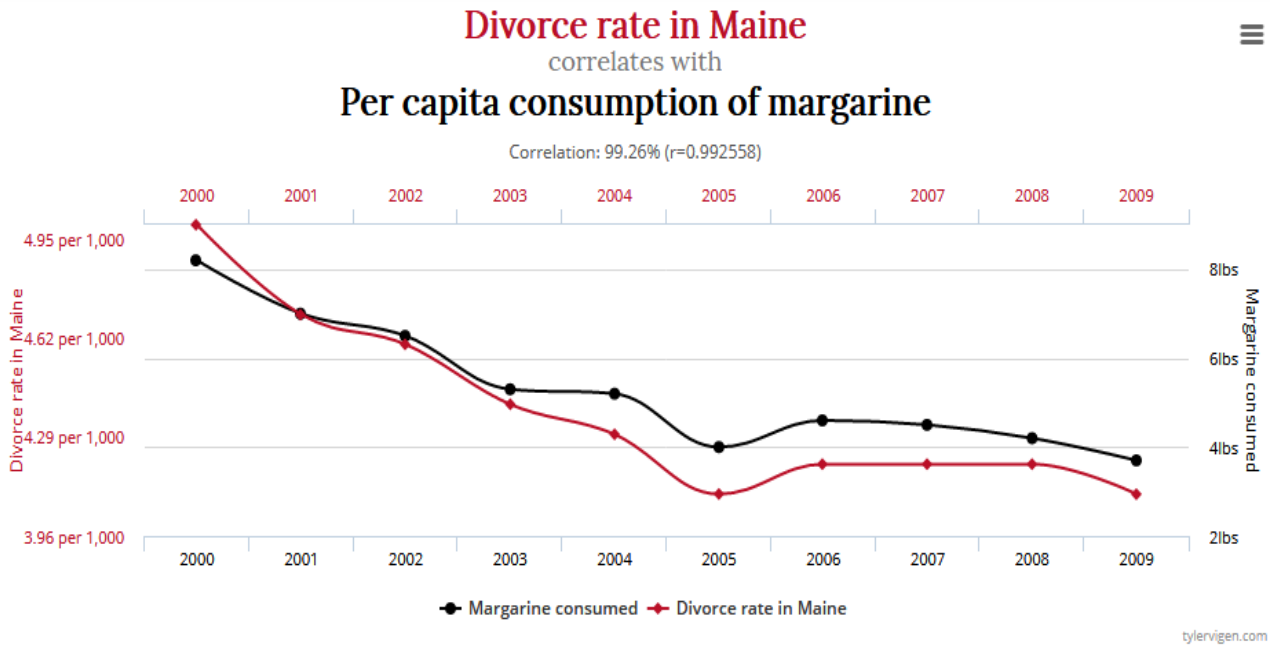
Fonte de dados: Centers for Disease Control & Prevention and Internet Movie Database

Associações Espúrias



Fonte de dados: U.S. Department of Agriculture e Centers for Disease Control & Prevention

Associações Espúrias



Fonte de dados: National Vital Statistics Report e U.S. Department of Agriculture

REGRESSÃO LINEAR

REGRESSÃO **LINEAR**

Objetivo

Unir de forma paramétrica os dados históricos, buscando sua relação de dependência entre períodos de tempo e na relação de causa e efeito entre variáveis.

Técnicas de Previsão - Técnicas Quantitativas

- O Modelo Causal permite:
  - Expressar as relações de Causa-Efeito entre variáveis;
  - Entender melhor os mecanismos geradores do fato em estudo;
  - Simular situações de forma a se avaliar o seu impacto na previsão;
  - Analisar situações independentes do tempo.

MODELO DE REGRESSÃO:

Esse modelo relaciona, funcionalmente, uma variável dependente às suas possíveis variáveis explicativas.

- Eficácia de propaganda sobre as vendas
- Número de acidentes pela velocidade desenvolvida
- Prever o tempo gasto no caixa de um supermercado em função do valor de compra
- Satisfação do Cliente em função do tempo de relacionamento e intensidade de uso

Conceito

MODELO PROBABILÍSTICO

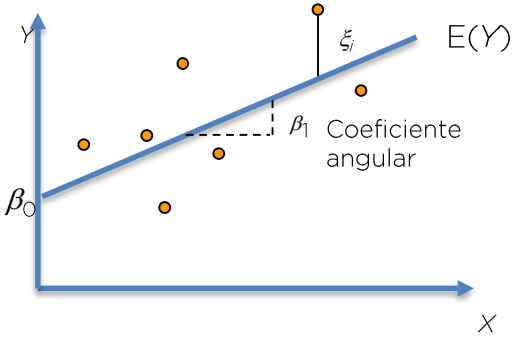
y = Componente Determinístico + Erro Aleatório

onde y é a variável dependente

Escrever a equação linear envolve dois parâmetros:

- ✓ O Intercepto de y
- ✓ A inclinação da reta

Modelo de Regressão Linear Simples



$$E(Y) = \beta_0 + \beta_1 X$$

Inclinação populacional

Intercepto populacional

$$Y_i = \beta_0 + \beta_1 X_i + \xi_i$$

Exemplo

Regressão Linear Simples

Faça a previsão das vendas (R$) mensal no período de 12 meses da empresa XYZ a partir dos dados disponíveis de Vendas (R$) e Budget Advertising (R$) da empresa (arquivo: Vendas_2016a2018.csv). Use o modelo de regressão linear simples.

```
Coefficients:
                          Estimate  Std. Error t value         Pr(>|t|)
(Intercept)           1060550.396  151771.308   6.988 0.0000000463123 ***
df$Budget_Advertising       4.964       0.524   9.473 0.0000000000458 ***
---
```

```
> somaxy = sum(df$xy); somaxy [1] 25202836280625
> somaxx = sum(df$x2); somaxx [1] 3020024869670
> ybarra = mean(df$Vendas); ybarra [1] 2388186
> xbarra = mean(df$Budget_Advertising); xbarra [1] 267445.2
> xbarra2 = xbarra*xbarra; xbarra2 [1] 71526917173
> xybarra = xbarra*ybarra; xybarra [1] 638708784704 >
```

$$\hat{\beta}_1 = \frac{25202836280625 - 36 * 267445.2 * 2388186}{3020024869670 - 36 * 71526917173}$$

$$\hat{\beta}_1 = 4.964141$$

$\hat{\beta}_0$ = 2388186 -4.964141*] 267445.2 = 1060550

Teste de Hipóteses

TESTANDO OS PARÂMETROS B'S

H_o: $B_i = 0$

H_1: $B_i \neq 0$

$$t = \frac{B_i}{erro\_padrao(B_i)}$$ com gl = n - p

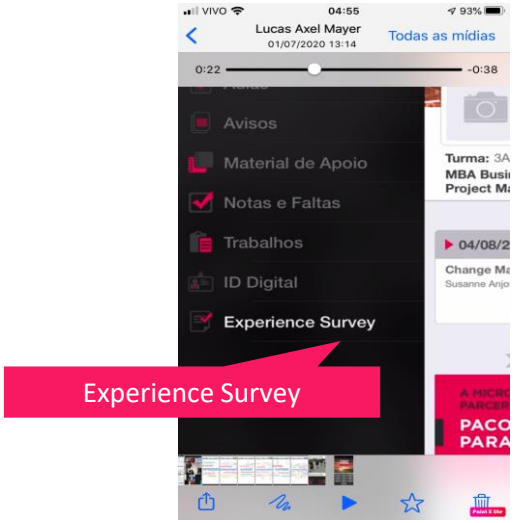
Quando $t > t_{\alpha/2} \Rightarrow$ região de rejeição

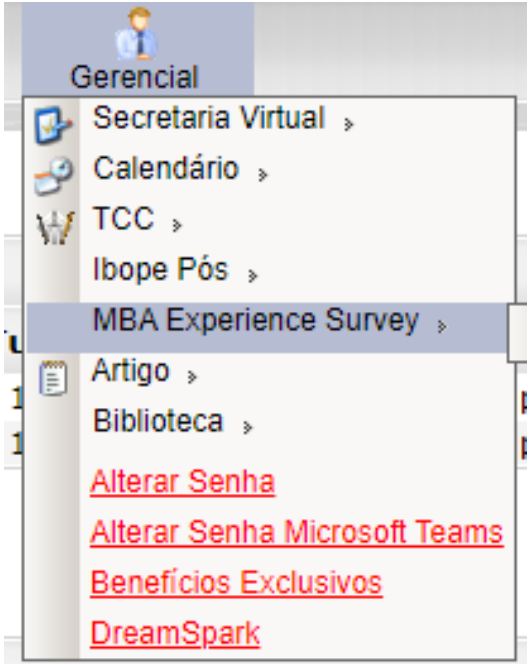
$$IC : \overline{b}_i + t_{\alpha/2} Sb_i$$

O que você achou da aula de hoje?

Pelo aplicativo da FIAP

(Entrar no FIAPP, e no menu clicar em Experience Survey)





A grande finalidade do conhecimento não é conhecer, mas agir.

T. Huxley

OBRIGADO

 / Regina T. I. Bernal

FIAP


FIAP