Cadernos Geográficos

Segredos da Estatística para Geografia

Jesué Graciliano da Silva

UNIVERSIDADE FEDERAL DE SANTA CATARINA

Reitor:	Luiz Carlos Cancellier de Olivo
Vice-Reitor:	Alacoque Lorenzini Erdmann

CENTRO DE FILOSOFIA E CIÊNCIAS HUMANAS

Diretor:	Paulo Pinheiro Machado
Vice-Diretor:	Sônia Weidner Maluf

DEPARTAMENTO DE GEOCIÊNCIAS

Chefe:	Antônio Fernado H. Fetter Filho
Sub-Chefe:	Nazareno José de Campos

PROGRAMA DE PÓS-GRADUAÇÃO EM GEOGRAFIA

Coordenador:	Aloysio M. De Araújo Junior
Sub-Coordenador:	Elson Manoel Pereira

UNIVERSIDADE FEDERAL DE SANTA CATARINA

Cadernos Geográficos

GCN / CFH / UFSC

ISSN 1519–4639
ISSNe 5448-265X

Cadernos Geográficos	Florianópolis	Nº35	128p.	Julho 2016

Cadernos Geográficos é uma publicação editada pelo Departamento de Geociências da Universidade Federal de Santa Catarina.

(Catalogação na fonte por Daurecy Camilo – CRB 14/416)

Cadernos Geográficos / Universidade Federal de Santa Catarina. Centro de Filosofia e Ciências Humanas. Departamento de Geociências. – n° 1 (maio de 1999) – Florianópolis: Imprense Universitária, 1999 – v.; 23.
ISSNe2448-265X ISSN 1519-4639
1. Geografia 2.Anais I. Universidade Federal de Santa Catarina.

Endereço para correspondência

E-mail: cadernosgeograficos@contato.ufsc.br
Endereço eletrônico: www.cadernosgeograficos.ufsc.br

Nota Editorial

Dando continuidade à política de publicação do departamento de Geociências da UFSC, o Caderno Geográfico de número 35 apresenta a obra de Jesué Graciliano da Silva, intitulada “Segredos da Estatística para Geografia”.

Professor do Instituto Federal de Santa Catarina desde 1993, Jesué Graciliano da Silva elabora uma obra que vem ao encontro com o anseio dos alunos de Graduação e, em grande medida, da pós-graduação em Geografia, pois se trata de um momento em que a discussão sobre a inserção da disciplina de Estatística no currículo do curso de Graduação em Geografia desta Universidade está presente.

A obra, ora apresentada, tem caráter extremamente didático, partindo de um breve contexto histórico sobre Estatística, passa em seguida para os principais indicadores estatísticos costumeiramente utilizados na ciência geográfica (IDH, PIB, IPCA, IDEB, etc.). Nas páginas seguintes a obra apresenta as possibilidades de se construir histogramas, gráficos e indicadores com o apoio de ferramentas estatísticas. Por fim, apresenta uma introdução aos principais bancos de dados estatísticos oficiais (RAIS, CAGED, IBGE, etc.). Além disto, traz diversos exercícios para fixação e compreensão prática dos conceitos apresentados.

Neste sentido, temos certeza de que a presente obra será uma importante contribuição para o aprendizado e interpretação das informações estatísticas que estão presentes em nosso cotidiano da geografia.

Comissão Editorial

Editorial Note

Continuing the publication policy of the geosciences department from UFSC, the Geographic Book number 30 presents the work of, Jesué Graciliano da Silva, entitled “Estatistic Secrets for Geography”.

Professor of the Federal Institute of Santa Catarina since 1993, Jesué Graciliano da Silva elaborates a work that meets the yearning of Graduation students and, to a great extent, of the postgraduate in Geography, because it is a moment in which the discussion about the insertion of the discipline of Statistics in the curriculum of the course of Graduation in Geography of this University is present.

The work presented here has an extremely didactic character, starting with a brief historical context on Statistics, then goes on to the main statistical indicators commonly used in geographic science (HDI, GDP, IPCA, IDEB, etc.). In the following pages the work presents the possibilities of constructing histograms, graphs and indicators with the support of statistical tools. Finally, it presents an introduction to the main official statistical databases (RAIS, CAGED, IBGE, etc.). In addition, it brings several exercises for fixing and practical understanding of the concepts presented.

In this sense, we are sure that the present work will be an important contribution to the learning and interpretation of the statistical information that is present in our daily life of geography.

Editorial Comission

Sumário

Summary

Introdução

Jesué Graciliano da Silva[1]

Segundo os historiadores, uma das primeiras aplicações da estatística, mesmo que ainda assim não se chamasse se deu a partir da necessidade de se quantificar os estoques de comida das primeiras civilizações e para aperfeiçoar a cobrança de impostos. Era comum a realização de censos populacionais desde os babilônicos, chineses, egípcios, gregos e romanos. Em torno de 1066, após a conquista da Bretanha pelos invasores normandos liderados por Willian, "o conquistador", foi realizado um censo e também uma listagem de todos os itens de propriedade no território, registrado no livro que ficou conhecido como *Domesday Book (dia do juízo final).* A palavra "estatística", conforme utilizamos na atualidade, parece ter sido introduzida pelo economista alemão Gottfried Achenwall (1719-1772) em 1748. Achenwall estudou a regularidade de fenômenos de caráter econômico e social. Mas antes dele, no século XVII John Graunt (1620 – 1674) já havia introduzido relatórios sobre mortalidade e natalidade à procura de regularidades. A estatística confundia-se, praticamente, com a demografia à qual fornecia métodos sistemáticos de enumeração e organização. Somente após o desenvolvimento da teoria das probabilidades por Blaise Pascal[2] (1623-1662) e por Pierre S. Laplace (1749 – 1827), acabou se tornando uma disciplina. A curva chamada de normal, fundamental para a compreensão dos fenômenos estatísticos, foi observada pela primeira vez por Abraham de Moivre (1667-1754) no ano 1733. O sociólogo e matemático belga Adolphe Jacques Quételet (1796-1874) usou a curva normal para realização de estudos sociais, mas somente anos mais tarde o matemático alemão Carl F. Gauss3 (1777-1855) determinou sua equação descritiva. O desenvolvimento da estatística moderna se deu a partir dos estudos de F. Galton (1822-1911), K. Pearson (1857-1936), R.A.Fischer (1890-1962) e W.S.Gosset (1876-1936). A história do desenvolvimento da Estatística como ciência é cheia de grandes personagens e passagens interessantes e permite compreender melhor o que levou à descoberta da curva normal e ao desenvolvimento dos Testes de Hipóteses[4].

[1] Ver perfil do autor página 139

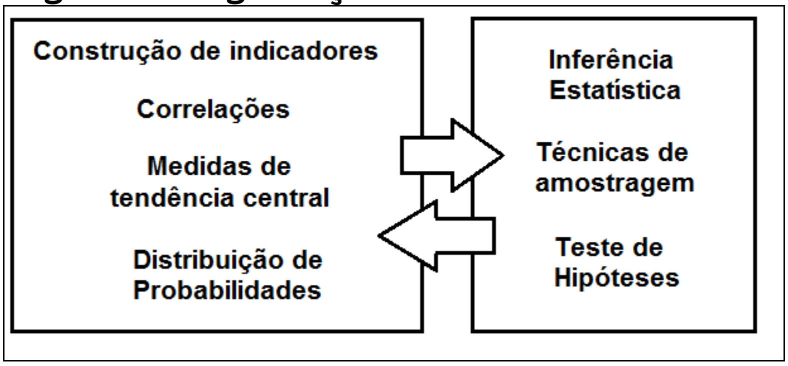
[2] Filme sobre Blaise Pascal: https://www.youtube.com/watch?v=C3fhX3q0-SQ

3 Livro recomendado: As 17 equações que mudaram o mundo, do autor Ian Stewart.

[4] Livro recomendado: Uma senhora toma chá do autor David Salsburg

Há diversas definições para Estatística. Podemos simplificar dizendo que “**estatística é o estudo da coleta, organização, análise, interpretação e apresentação de dados”.** Dados são valores coletados da variável em estudo. Para facilitar o aprendizado, organizamos os capítulos em dois grandes grupos conforme ilustrado na Figura 1.

Figura 1 – Organização do estudo da estatística.



Fonte: Elaborado pelo autor

No primeiro grupo, tem-se a Estatística Descritiva e no segundo grupo a Estatística Inferencial. A Estatística Descritiva utiliza um conjunto de técnicas tais como: medidas de posição e dispersão, tabelas e gráficos para resumir as características dos dados coletados. Já a Estatística Inferencial possibilita que uma população inteira seja conhecida a partir do estudo das características de uma amostra aleatória representativa do todo.

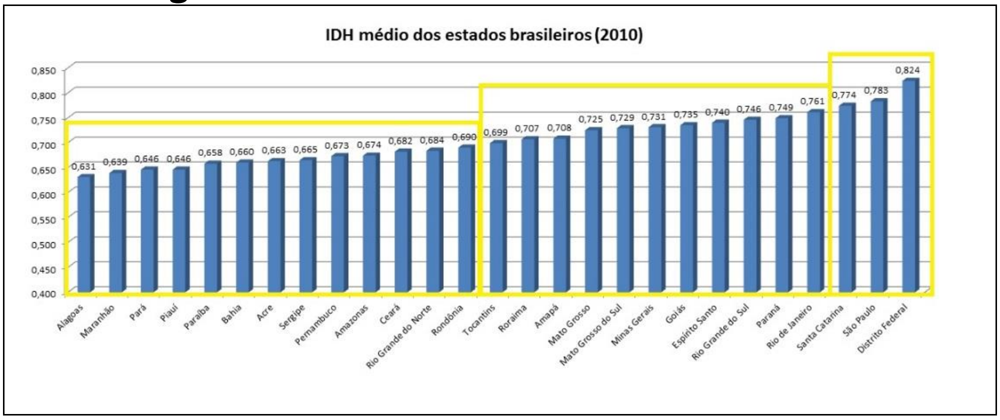
1 - Principais Indicadores Estatísticos

Para iniciar nosso estudo vamos analisar a seguir alguns indicadores muito utilizados no nosso dia a dia. Nem sempre paramos para refletir sobre como eles foram construídos e como eles podem nos auxiliar na interpretação e compreensão da realidade. Neste capítulo, vamos destacar alguns deles como IDH, PIB, PISA, IPCA PIB per capita e IDEB.

a) IDH – Índice de desenvolvimento humano

O Índice de desenvolvimento humano é um índice que serve de comparação entre os países, com objetivo de medir o grau de desenvolvimento econômico e a qualidade de vida oferecida à população. O relatório anual de IDH é elaborado pelo Programa das Nações Unidas para o Desenvolvimento (PNUD), órgão da ONU. Quanto mais próximo de 1, mais desenvolvido é o país. Este índice também é usado para apurar o desenvolvimento de cidades, estados e regiões. Na Figura 2 tem-se o IDH médio dos estados brasileiros. Normalmente, os países com IDH menor que 0,5 são considerados com baixo desenvolvimento humano. Os países com IDH entre 0,5 e 0,8 são considerados de médio desenvolvimento humano e os que possuem IDH superior a 0,8 apresentam desenvolvimento humano alto.

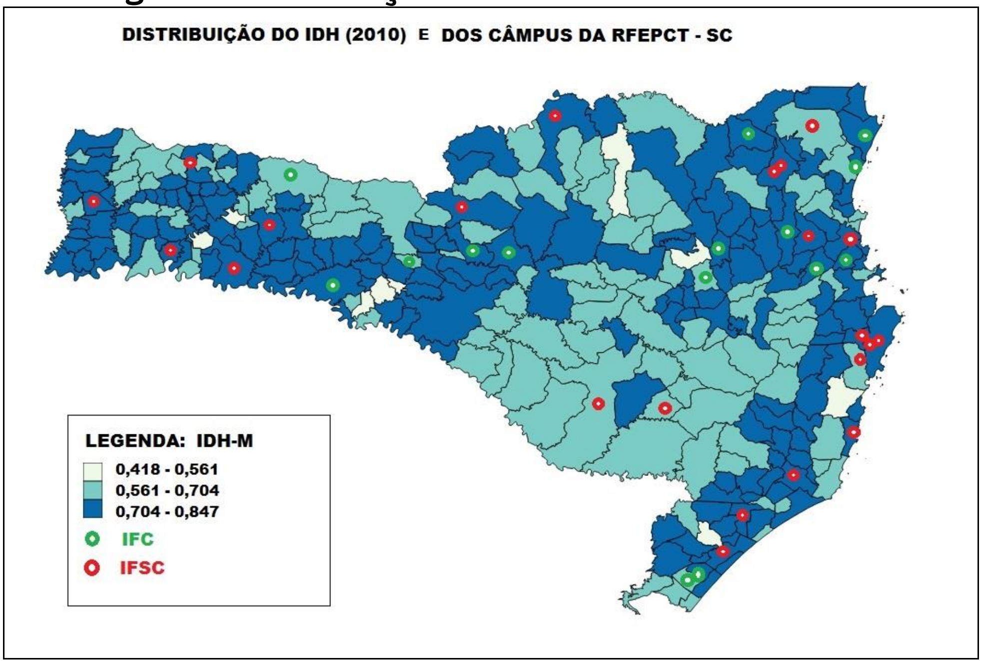
Figura 2 – IDH médio dos estados brasileiros.



Fonte: Elaborado pelo autor

O IDH é um índice que agrega três dimensões: educação, longevidade e renda, que são combinados. Seu valor médio não mostra as desigualdades existentes em um município, estado ou país. Mesmo estados como Santa Catarina, que apresenta um dos melhores IDHs do país tem grandes diferenças regionais, conforme Figura 3 (IBGE, 2010).

Figura 3- Distribuição do IDH-M – Santa Catarina



Fonte: Elaborado pelo autor

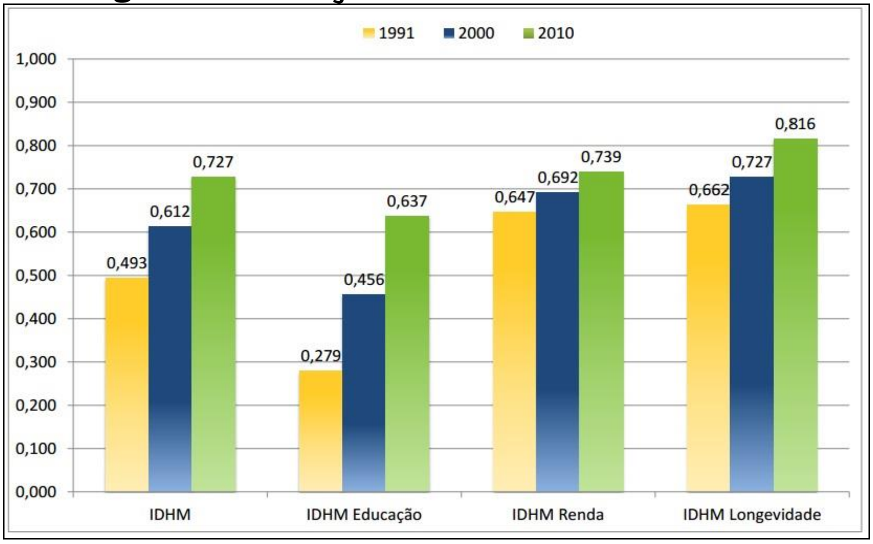
Assim como o Estado de Santa Catarina tem grandes desigualdades regionais, fruto das diferentes formações socioespaciais[5], dentro de uma cidade também há grandes diferenças na organização espacial. O município de Palhoça, por exemplo, apresenta diversos bairros organizados convivendo lado a lado com comunidades segregadas e pobres. Por esse motivo o indicador IDH médio da

[5] Formação socioespacial em SC: https://periodicos.ufsc.br/index.php/geosul/article/viewFile/13604/12471

Palhoça não é capaz de mostrar essas diferenças, que se escondem na média. Uma pessoa que esteja com a cabeça em uma temperatura de 40ºC graus e os pés a uma temperatura de 10ºC estará sujeita a uma temperatura média de 25 ºC. Em média a pessoa estará confortável, mas isso não reflete a realidade.

Esse é um dos cuidados que temos ao analisar os indicadores sociais tais como IDH6. As avaliações têm possibilitado fazer comparações ao longo do tempo. Na Figura 4 é possível verificar que o IDH médio brasileiro vem evoluindo nos últimos 20 anos nas suas três dimensões.

Figura 4- Evolução do IDH médio brasileiro.



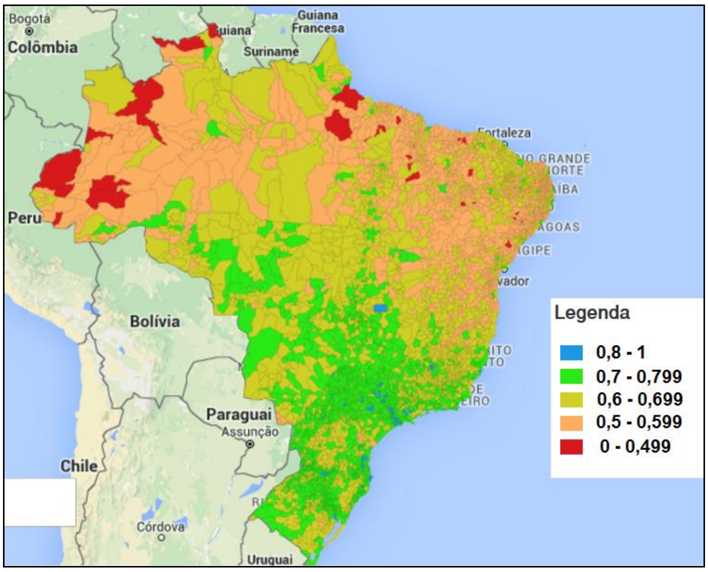
Fonte: Elaborado pelo autor

Como é possível perceber a dimensão IDH-M Educação é a que tem apresentado a maior evolução ao longo dos últimos 20 anos. O IDH-M tem variado de maneira diferente ao longo do Brasil. As regiões Nordeste e Norte são as que apresentaram a maior evolução entre os anos de 2000 e 2010 com variação média de IDH-M 2,5% e 2,4%, acima da variação média brasileira que foi de 1,7%.

No site: http://www.atlasbrasil.org.br/2013/ é possível visualizar graficamente como o IDH-M vem evoluindo ao longo dos anos de todas as regiões brasileiras, bem como construir diversos tipos de gráficos sobre o assunto. Como exemplo, na Figura 5 tem-se a distribuição do IDH-M brasileiro para o ano de 2010. Há 1399 municípios com IDH-M inferior a 0,6. Há 2223 municípios com IDH-M entre 0,6 e 0,69. Há 1890 municípios com IDH-M entre 0,7 e 0,79. Finalmente, há apenas 44 municípios brasileiros com IDH-M superior a 0,8.

6 Entrevista Canal Futura avaliando IDHM - https://www.youtube.com/watch?v=3QE4URPdoiA

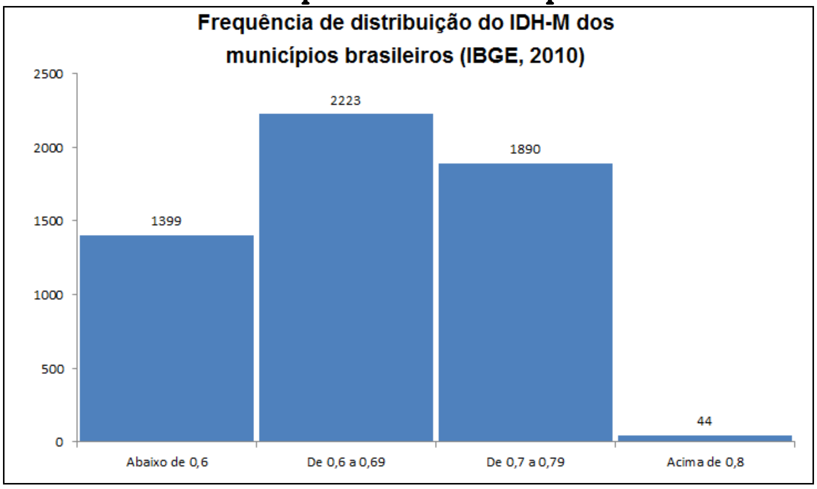
Figura 5- Distribuição do IDH-M do Brasil em 2010.



Fonte: Elaborado pelo autor

No histograma representado na Figura 6, tem-se a frequência de distribuição do IDH-M dos municípios brasileiros para o ano de 2010.

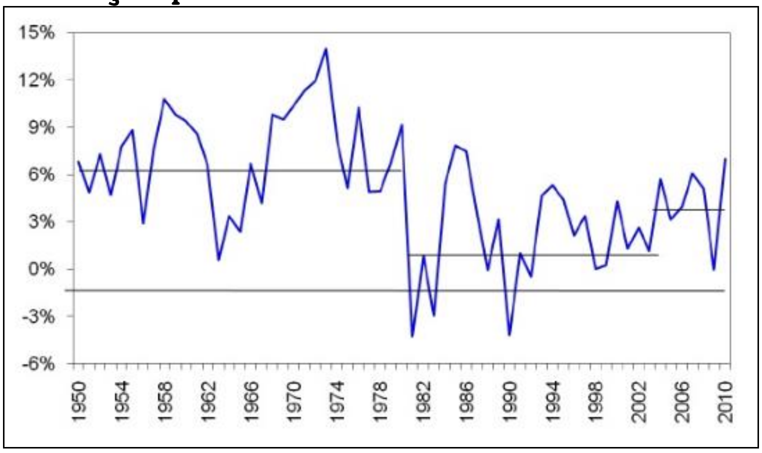
Figura 6- Frequência de distribuição
do IDH-M dos municípios brasileiros para o ano de 2010.



b) PIB – Produto Interno Bruto

O Produto Interno Bruto (PIB) é normalmente usado para medir o nível de atividade econômica de um país. É comum se dizer que o PIB é um bom indicador de crescimento, mas não de desenvolvimento, que envolve uma transformação qualitativa da estrutura econômica, social e cultural do país. Na Figura 7 é possível visualizar o comportamento percentual do PIB entre os anos de 1950 a 2010 representado em um gráfico de linha.

Figura 7- Evolução percentual do PIB entre os anos 1950 a 2010[7]



Fonte: Elaborado pelo autor

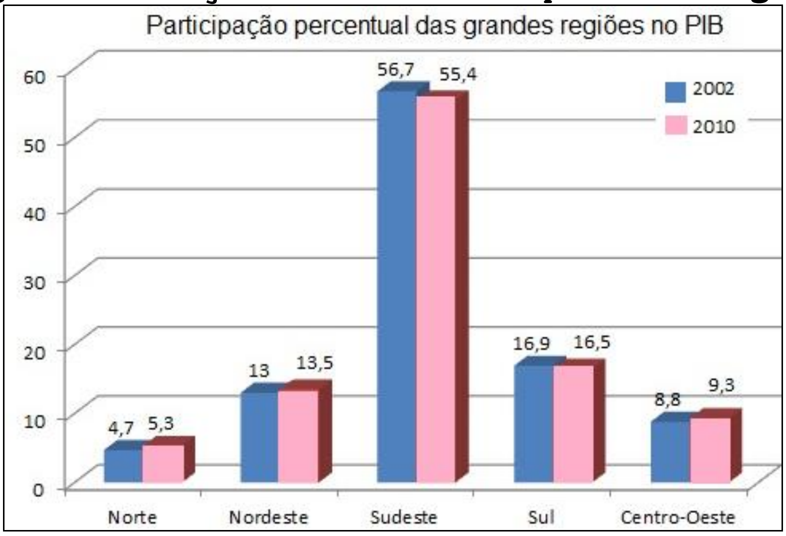
O PIB é calculado de diversas maneiras: uma delas é pela soma das riquezas produzidas dentro do país, incluindo nesse cálculo empresas nacionais e estrangeiras. Nesse cálculo entram os resultados da indústria, serviços e agropecuária. Entra no cálculo apenas o produto final vendido. Suponha que um marceneiro venda um armário de fabricação própria por R$500,00, e seu gasto com matéria-prima foi de R$200,00. Nesse caso a riqueza gerada por ele será de R$ 300,00.

Outra maneira de medir o PIB é por meio da avaliação da demanda. Nesse caso, são considerados o consumo das famílias, o consumo do governo, os investimentos do governo e de empresas privadas e a soma das exportações e das importações.

Na Figura 8, tem-se a evolução percentual do PIB das macrorregiões brasileiras entre os anos de 2002 e 2010.

7 http://pt.slideshare.net/feers/apresentacao-seminario-9576319

Figura 8- Evolução do PIB brasileiro por macrorregião.



Fonte: Elaborado pelo autor

Na Tabela 1, tem-se a projeção do PIB (trilhão de dólares) das maiores economias do mundo no ano de 2030. O maior avanço no período deverá ocorrer em países como a China e Índia, que terão seus PIB ampliados em quase 300% no período.

Tabela 1– Estimativa da distribuição do PIB de alguns países selecionados em 2030

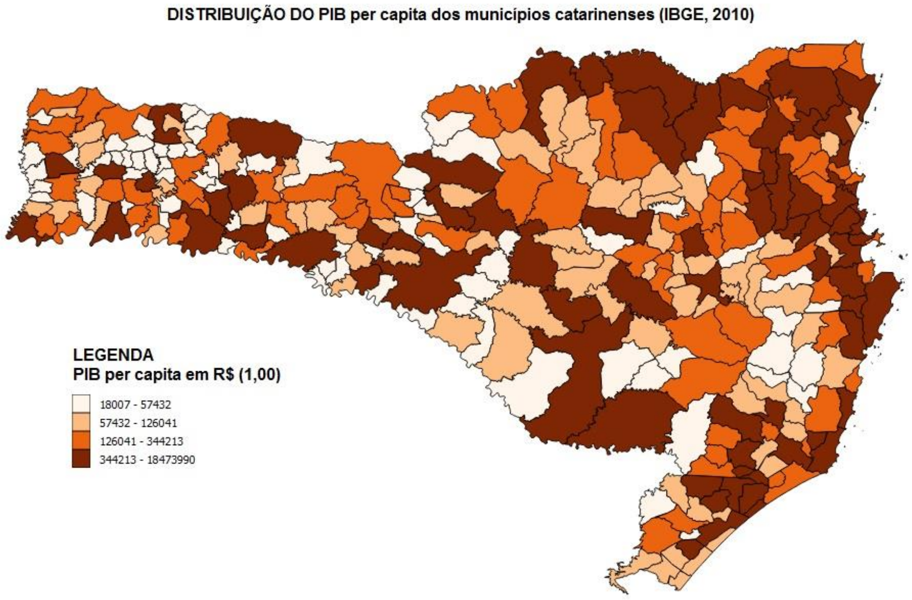
País	PIB 2015 (trilhão de U$)	PIB estimado para 2030 (trilhão de U$)
Estados Unidos	16,8	24,8
China	8,5	22,1
Índia	2,2	6,5
Japão	5,6	6,4
Alemanha	3,5	4,5
Brasil	2,2	3,9
Reino Unido	2,5	3,6

Fonte: BLOOMBERG NEWS

c) PIB per capita

O PIB per capita, calculado a partir da divisão do PIB total pelo número de habitantes da região, indica quanto cada habitante produziu em determinado período. No entanto, o PIB per capita é um indicador que precisa ser avaliado com atenção. A presença de uma grande empresa, um porto ou uma refinaria em uma cidade com baixa densidade populacional é suficiente para produzir um PIB per capita elevado. Na Figura 9 tem-se a distribuição do PIB per capita das cidades catarinenses.

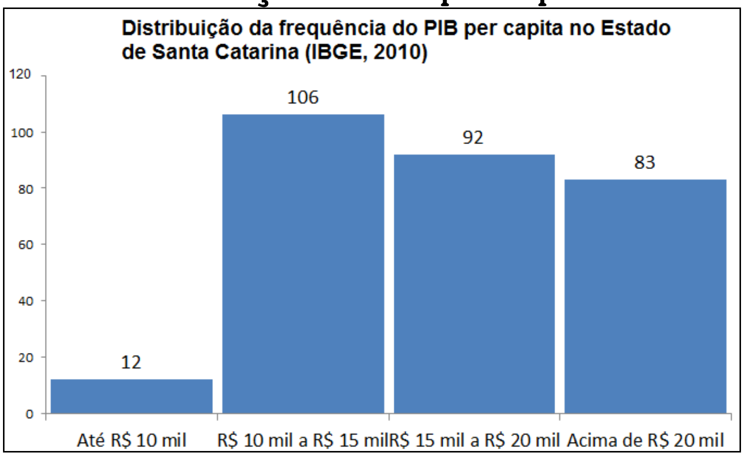
Figura 9- Distribuição do PIB per capita catarinense.



Fonte: Elaborada pelo autor

Já na Figura 10 tem-se a distribuição da frequência do PIB per capita das cidades catarinenses para o ano de 2010.

Figura 10- Distribuição do PIB per capita catarinense.

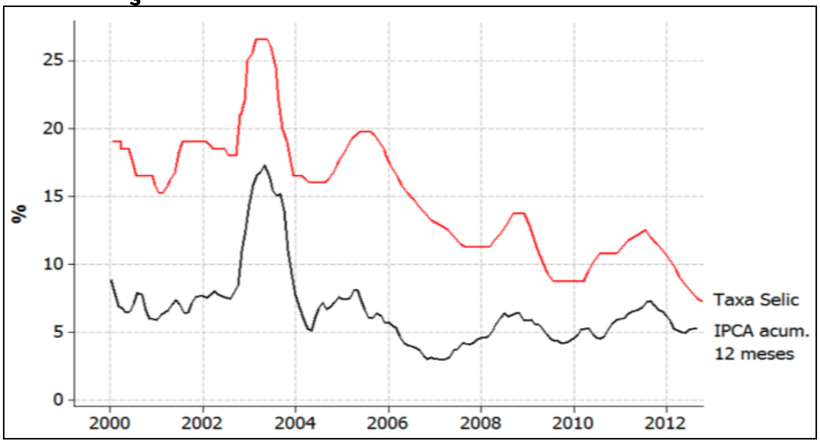


Fonte: Instituto Brasileiro de Geografia e Estatística

d) IPCA – Índice de preços ao consumidor amplo:

O IPCA (Índice de Preços ao Consumidor Amplo), medido mensalmente pelo IBGE (Instituto Brasileiro de Geografia e Estatística), foi criado em 1980 com o objetivo de oferecer a variação dos preços para o público final. Na prática, acaba sendo considerado o índice de inflação brasileira. Na Figura 11 tem-se a evolução da Taxa SELIC8 e do IPCA acumulado entre os anos 2000 e 2012.

Figura 11- Evolução do IPCA acumulado entre os anos 2000 e 2012.



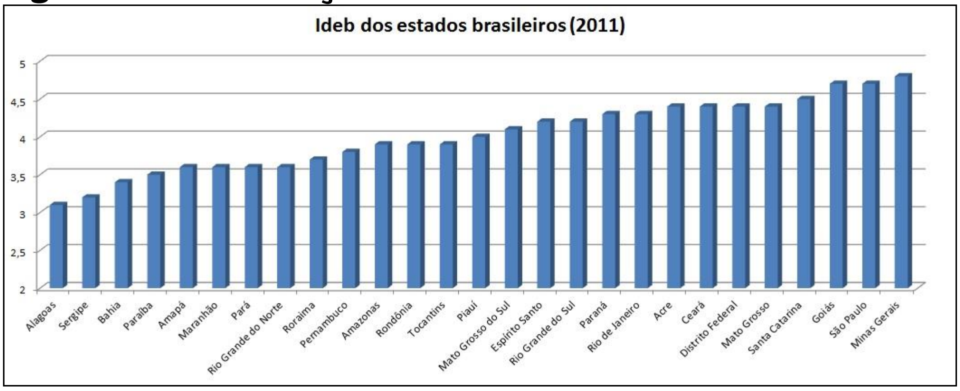
A pesquisa de preços é realizada em estabelecimentos comerciais, prestadores de serviços, domicílios (para verificar valores de aluguel) e concessionárias de serviços públicos. São considerados nove grupos de produtos e serviços: alimentação e bebidas; artigos de residência; comunicação; despesas pessoais; educação; habitação; saúde e cuidados pessoais; transportes e vestuário. Eles são subdivididos em outros itens. Ao todo, são consideradas as variações de preços de 465 subitens. O indicador reflete o custo de vida de famílias nas regiões metropolitanas de São Paulo, Rio de Janeiro, Belo Horizonte, Porto Alegre, Curitiba, Salvador, Recife, Fortaleza e Belém, além do Distrito Federal e do município de Goiânia.

e) IDEB – Índice de desenvolvimento da educação básica

O IDEB é avaliado pelo MEC – Ministério da Educação a cada dois anos e apresentado numa escala que vai de zero a dez. No total, o IDEB estabelece notas para cerca de 50 mil escolas públicas do país. Na Figura 12, tem-se a distribuição dos IDEBs de todos os estados brasileiros para o ano de 2011.

8 Os bancos tomam dinheiro emprestado pela Taxa SELIC.

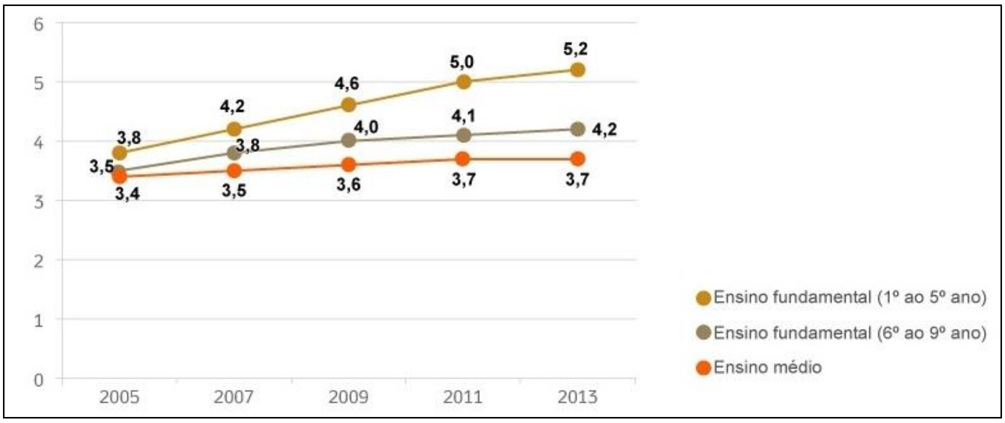
Figura 12 - Distribuição do IDEB entre os estados Brasileiros



Fonte: MEC, 2011

Mesmo que nos últimos 20 anos a dimensão educação tenha avançado mais que as outras duas dimensões do IDH-M, há ainda um longo caminho a ser percorrido. Na Figura 13, tem-se a evolução do IDEB do Ensino Fundamental e do Ensino Médio entre os anos 2005 e 2013.

Figura 13 – Evolução do IDEB médio do Brasil entre os anos 2005 e 2013.

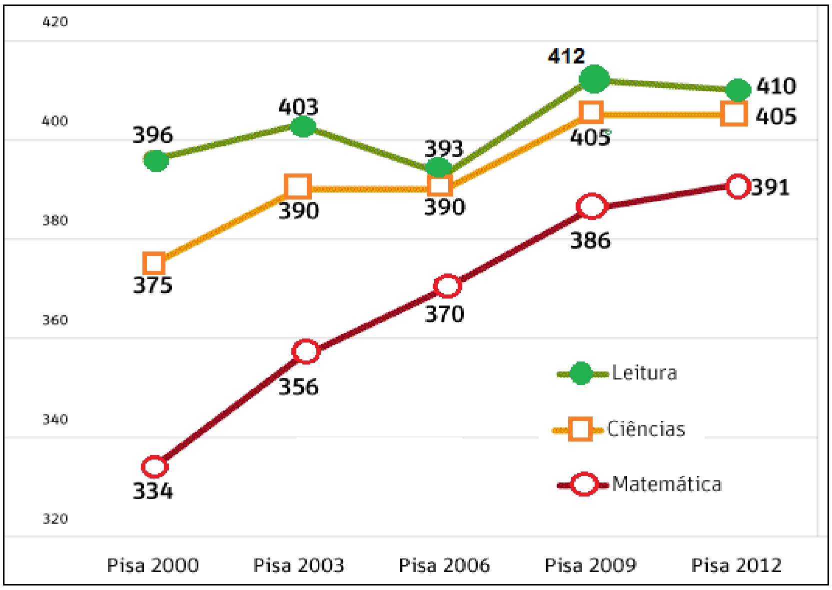


Fonte: Elaboração do autor

f) Indicador PISA

O *Programme for International Student Assessment* (Pisa) - Programa Internacional de Avaliação de Estudantes - é uma iniciativa de avaliação comparada, aplicada a estudantes na faixa dos 15 anos, idade em que se pressupõe o término da escolaridade básica obrigatória na maioria dos países. O programa é desenvolvido e coordenado pela Organização para Cooperação e Desenvolvimento Econômico (OCDE). Em cada país participante há uma coordenação nacional. No Brasil, o PISA é coordenado pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). Apesar de todos os avanços na área educacional, o país ficou na 58º posição entre os 65 países avaliados no PISA 2012 (Figura 14).

Figura 14– Evolução das notas dos estudantes no PISA entre os anos 2000 e 2012



Fonte: Elaborado pelo autor

Além desses indicadores apresentados existem outros que podem ser utilizados. Entre eles temos: índice de densidade populacional, GINI, índice de analfabetismo, índice de inovação, índice de mortalidade infantil, índice de desemprego e índice de urbanização.

Veja os vídeos abaixo para aprender mais sobre indicadores sociais brasileiros:
a) Gini: https://www.youtube.com/watch?v=bMoVlRamzek
b) IDEB: https://www.youtube.com/watch?v=0czbRa6ll5Y
c) Índice de Pobreza e desigualdade:
https://www.youtube.com/watch?v=EX23UmPrwOo
d) Índice IBOVESPA: https://www.youtube.com/watch?v=-ZFuIS59Yy4
e) O que são indicadores: https://www.youtube.com/watch?v=2Ns1Bnmhrn4

LISTA DE EXERCÍCIOS 1:

1- Faça uma comparação entre indicadores (renda, PISA, IDH, População, Área) entre Brasil, China, África do Sul, Rússia e Índia.

2- Organize a listagem dos 10 países que são maiores produtores do mundo de: carne bovina, carne suína, laranja, algodão, frangos, carros, bicicletas, soja, maçãs, açúcar, café, motocicletas, jatos comerciais, aço, petróleo.

3- Compare, construa gráficos e atualize os indicadores indicados a seguir para a América do Sul:

País	População 2008 (milhões)	PIB 2007 (milhões de U$)	PIB per capita U$ - 2007	Áreas (km^2)	IDH 2013 (0-100)
Argentina	41	260.122	13.300	2.766.890	81
Bolívia	10	13.292	4.000	1.098.580	66
Brasil	191	1.313.590	10.300	8.511.965	74
Chile	17	163.914	13.900	756.950	82
Colômbia	46	202.630	6.700	1.138.910	71
Equador	14	45.789	7.200	283.560	71
Guiana Francesa	0,21	-	6.000	91.000	86
Guiana	1,2	2.920	3.800	214.999	62
Paraguai	7	27.082	7.800	406.750	67
Peru	30	219.015	4.500	1.285.220	73
Suriname	0,48	4.073	7.800	163.270	64
Uruguai	3,4	37.188	11.600	176.220	79
Venezuela	27	334.575	12.200	912.050	76

4- Analise por meio do Atlas – PNUD a evolução do IDH das regiões brasileiras. http://www.atlasbrasil.org.br/2013/

2- Gráficos e Indicadores

Neste capítulo vamos mostrar como são construídos histogramas, gráficos e indicadores com apoio de ferramentas estatísticas. Inicialmente vamos apresentar como são construídos os histogramas.
Como exemplo, suponha que um pesquisador esteja interessado em conhecer qual a distribuição da estatura dos estudantes de uma escola. Para tanto, ele mediu 40 alunos obtendo a seguinte Tabela 2.

Tabela 1- Representação de 40 estaturas de estudantes.

162	163	148	166	169	154	170	166
164	165	159	175	155	163	171	172
170	157	176	157	157	165	158	158
160	158	163	165	164	178	150	168
166	169	152	170	172	165	162	164

Fonte: Elaborado pelo autor

Os dados apresentados dessa forma não possibilitam que sejam percebidos os padrões e frequências. A elaboração de um histograma possibilita uma compreensão melhor das informações. Inicialmente deve-se calcular qual a diferença entre o maior e o menor valor de estatura. Chamamos essa grandeza de Amplitude. Para organizar os dados e verificar quais as estaturas que mais se repetem é importante escolher o número de classes de análise (k). Considerando que N = 40 dados e utilizando-se da equação proposta por Herbert STURGES é possível determinar o número ideal de classes:

$$k = 1 + 3{,}322 \, . \log(N)$$

Substituindo-se os valores na equação tem-se após o arredondamento 6 classes.

$$k = 1 + 3{,}322 \, . \log(40) = 6{,}3$$

A maior estatura é de 178 e a menor estatura é de 148. Isso significa uma amplitude de 30. Considerando-se as 6 classes tem-se o intervalo de 5 cm em cada classe pois (30 / 6 = 5). A Tabela 2 a seguir representa a frequência com que ocorre a distribuição das estaturas.

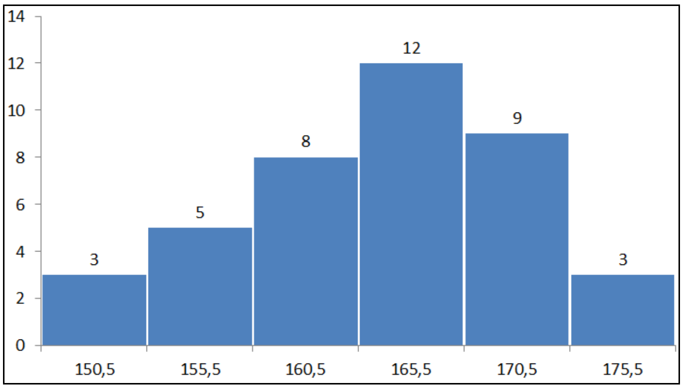
Tabela 1- Representação de 40 estaturas de estudantes.

Intervalo	Frequência	Freq. Relativa	Valor médio
148 |... 153	3	0,075	150,5
153 |... 158	5	0,125	155,5
158 |... 163	8	0,2	160,5
163|... 168	12	0,3	165,5
168|... 173	9	0,225	170,5
173 |... 178	3	0,075	175,5

Fonte: Elaborado pelo autor

O histograma representado pela Figura 15 permite a visualização rápida de como os dados estão distribuídos e quais são as estaturas mais comuns.

Figura 15- Histograma para distribuição das estaturas de uma turma de estudantes.



Fonte: Elaborado pelo autor

Para fins de simplificação, na Tabela 3 são apresentados alguns exemplos de números de classes obtidos a partir da equação de Sturges.

Tabela 3 – Número de classes obtidas por meio da equação de Sturges.

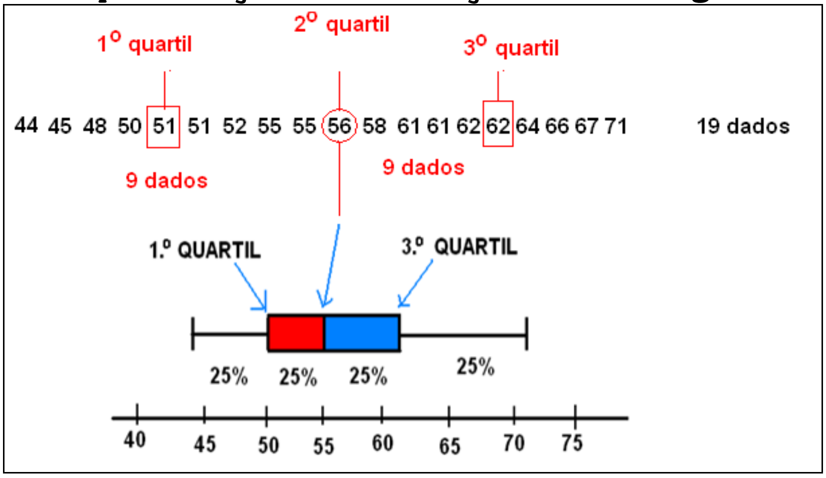
Número de dados	Número Aproximado de classes
20	5
40	6
60	7
80	7
100	8
1000	11

Fonte: Elaborada pelo autor

Outra maneira de se representar um conjunto de dados é por meio de DIAGRAMAS DE CAIXA, também conhecidos por BOX-PLOT. Os dados são divididos em duas partes (50% para cada lado), tendo o valor central chamado de MEDIANA. Cada parte também é dividida em 2 (25% =

quartil). Na Figura 16 é possível visualizar como um diagrama de caixa é construído para um conjunto de 19 dados.

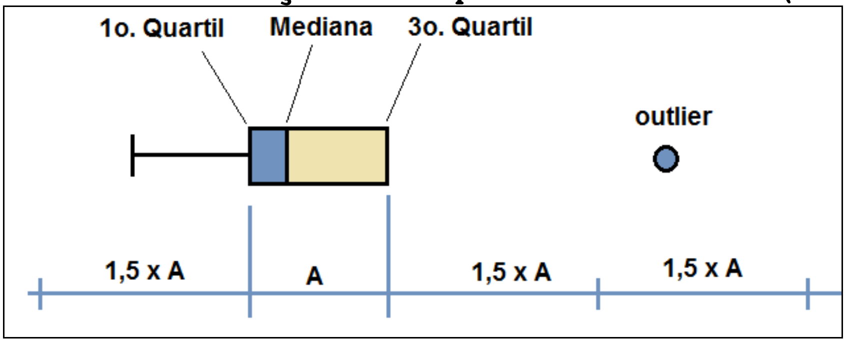
Figura 16- Representação da construção de um diagrama de caixa.



Fonte: Elaborado pelo autor

Para identificar possíveis pontos fora da curva, conhecidos como *outliers* adota-se o seguinte procedimento. Calcula-se qual é a diferença entre o terceiro e o primeiro quartil, representado por “A” na Figura 4. Qualquer medida que estiver acima de uma vez e meia dessa distância em relação ao primeiro ou terceiro quartil constitui-se em um *outlier,* conforme descrito na Figura 17.

Figura 17 – Identificação de um ponto fora da curva (*outlier*)



Fonte: Elaborado pelo autor

Além do histograma e do diagrama de caixa, também é comum a representação dos dados em diagramas de ramos e folhas, conforme ilustrado na Figura 18.

Figura 18– Exemplo de interpretação de um diagrama de ramos e folhas.

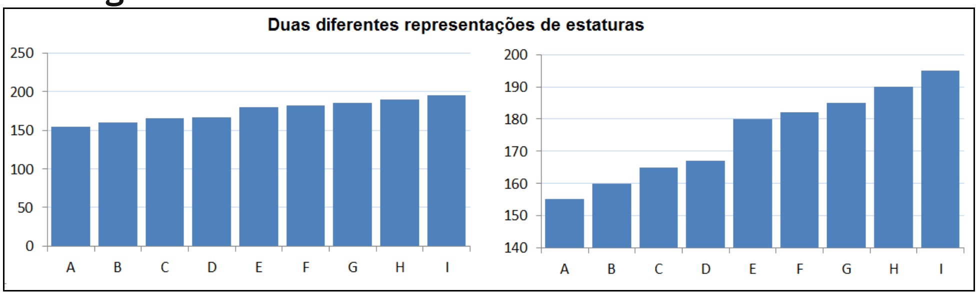
0		
1	7 8	17 18
2	0 5 8	20 25 28
3	4 4 7 9	34 34 37 39
4	1	41

Fonte: Elaborado pelo autor

Como é possível perceber, a forma como representamos os fenômenos é importante porque permite facilitar a compreensão dos dados. A construção de gráficos e indicadores pode ser realizada por diversos aplicativos tais como: Excel, Planilhas ODS – BR-office, R, Matlab, Wolfram[9] entre outros.

As representações gráficas também podem ser utilizadas com o objetivo de influenciar a compreensão dos menos atentos. Um dos erros mais comuns é a alteração das escalas nos gráficos, conforme ilustrado na Figura 19.

Figura 19- Gráficos resultantes de diferentes escalas.



Fonte: Elaborado pelo autor

No gráfico da esquerda parece que os estudantes têm estaturas muito mais próximas do que na realidade. Observamos que o gráfico da esquerda tem a sua escala (eixo y) iniciando no ponto zero, enquanto que o gráfico da esquerda tem a sua escala iniciando em 140 cm. O aluno mais baixo tem 155 cm de estatura enquanto que o mais alto 195 cm. Uma diferença de 40 cm – chamada de amplitude.

Na Figura 20, tem-se um erro de apresentação no gráfico veiculado em um telejornal, que representa a inflação entre os anos 2009 e 2013.

[9] https://www.wolframalpha.com/examples/Statistics.html

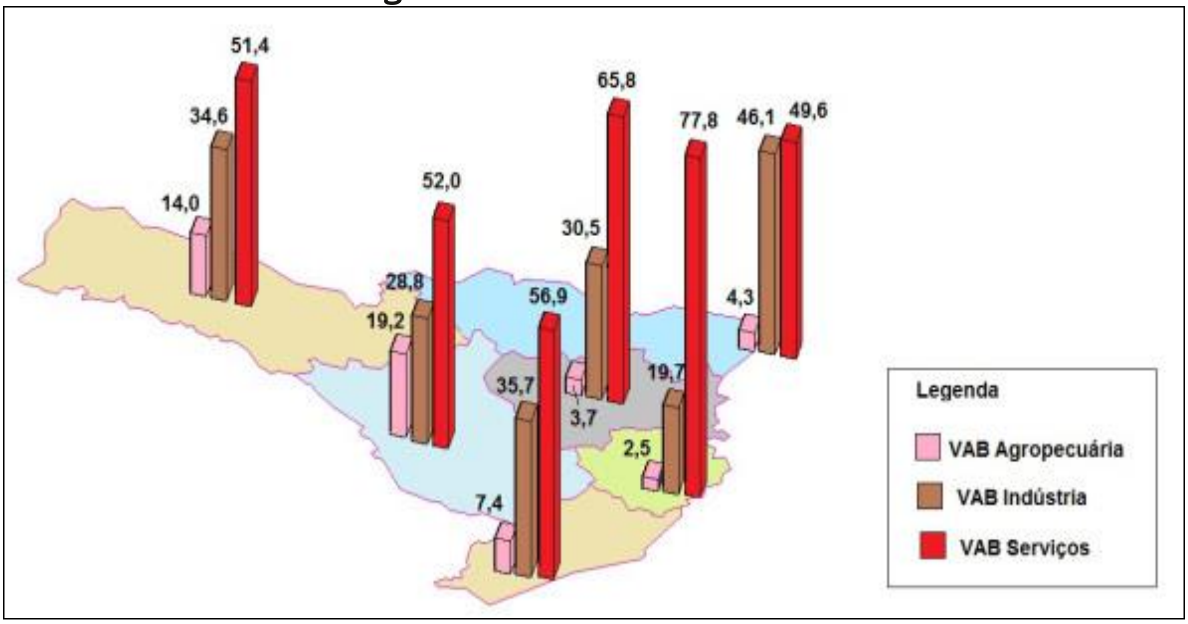
Figura 20- Gráfico com erro nas escalas[10].

Fonte: Globo News

Por esse motivo, a análise das escalas é fundamental para que as primeiras impressões não prejudiquem nossa interpretação.

Além dos histogramas, diagramas de caixa também são comuns a utilização de gráficos estilizados, no formato de pizza (setores), de radar e de linhas[11]. Na Figura 21 tem-se a ilustração de um gráfico estilizado representando a distribuição do Valor Adicionado Bruto entre as seis mesorregiões catarinenses. Esse gráfico foi construído com auxílio do software AutoCAD.

Figura 21- Mapa estilizado representando o Valor Adicionado Bruto das mesorregiões do Estado de Santa Catarina

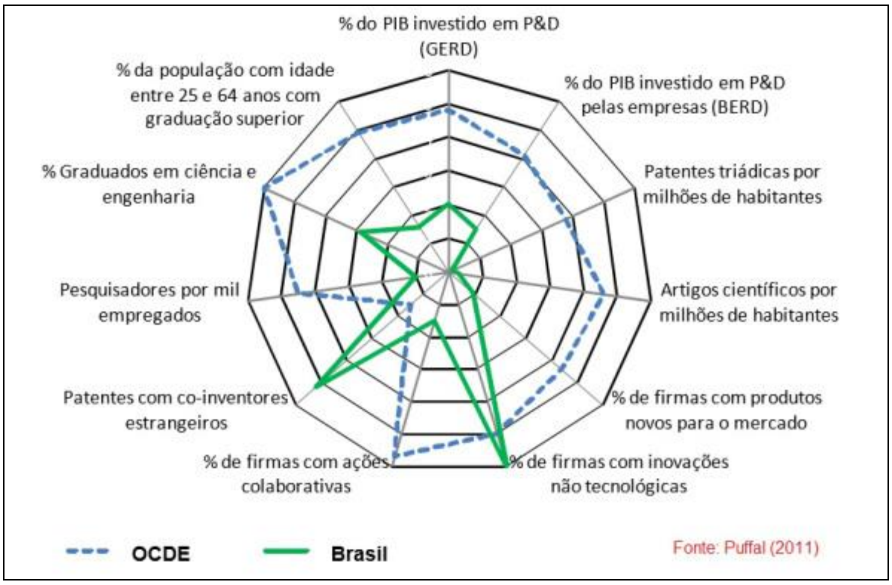


Fonte: IBGE, 2010

[10] Fonte: http://gizmodo.uol.com.br/mentir-visualizacao-dados/
[11] Veja mais em: http://univesptv.cmais.com.br/estatistica-aula-04-apresentacao-de-dados-tabelas-e-graficos

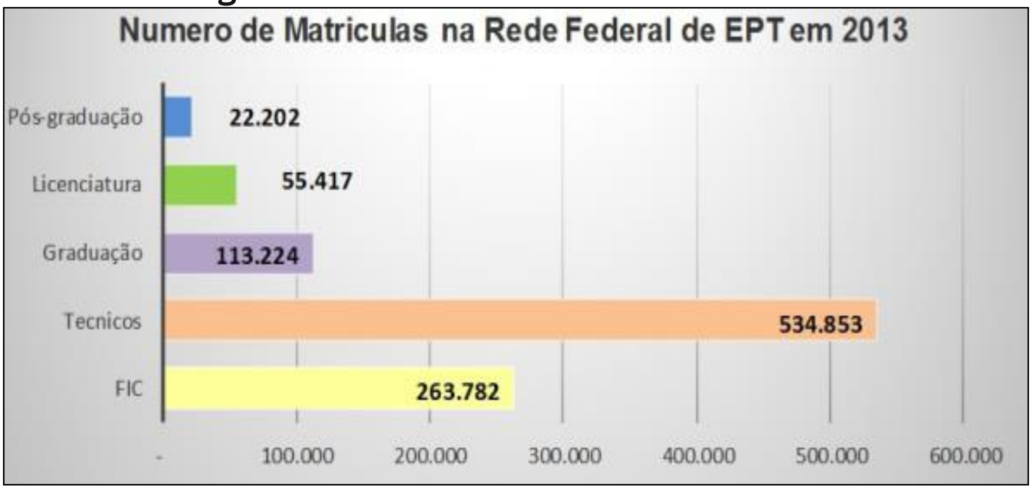
Na Figura 22 tem-se um gráfico tipo radar mostrando alguns comparativos entre o Brasil e o conjunto de países da OCDE.

Figura 22 - Gráfico do tipo radar representando dados do Brasil e dos países da OCDE (2011).



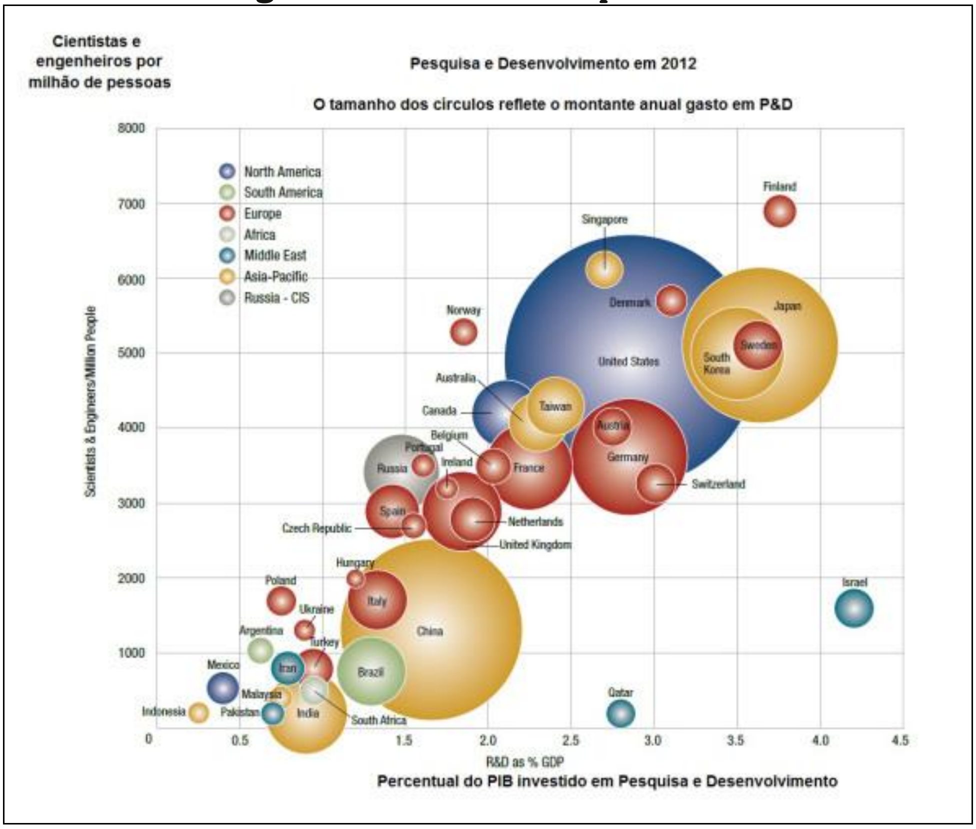
Na Figura 23, tem-se a ilustração de um gráfico de barras horizontais mostrando a distribuição de matrículas na Rede Federal EPT no ano de 2013.

Figura 23- Gráficos de barras horizontais.



Na Figura 24, tem-se um gráfico de bolhas mostrando a relação entre o número de cientistas e engenheiros por milhão de pessoas e o percentual de PIB investido em Pesquisa e Desenvolvimento em alguns países selecionados.

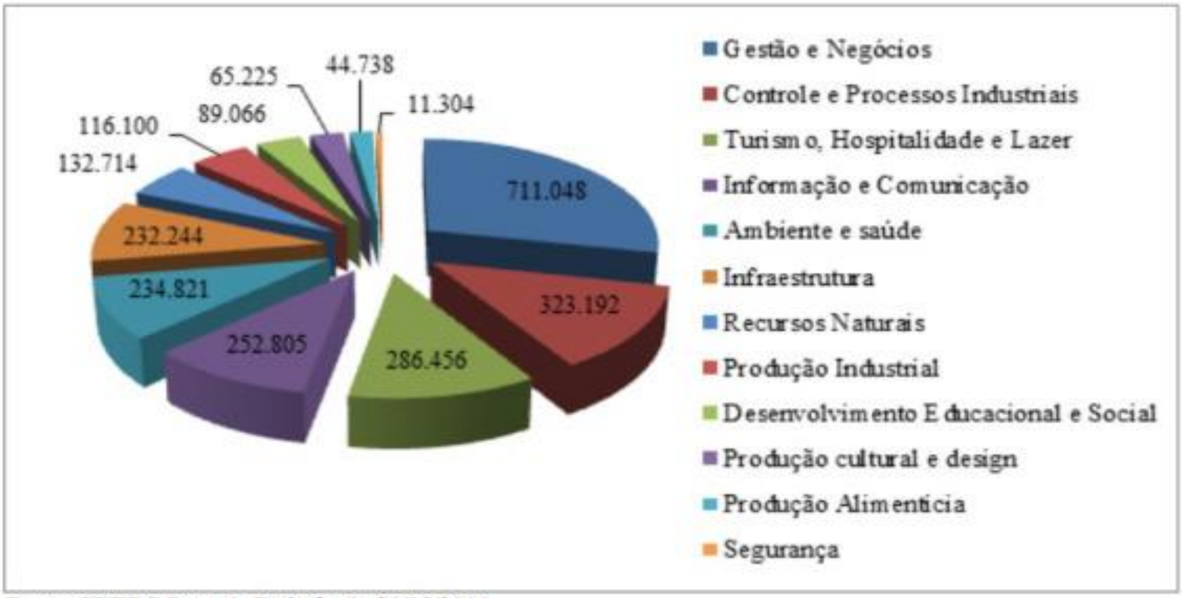
Figura 24- Gráfico do tipo bolha.



Fonte: http://battelle.org/docs/default-document-library/2012_global_forecast.pdf

Na Figura 25, tem-se um gráfico de pizza ou de setores, representando as matrículas em cursos de Formação Inicial e Continuada no Brasil do PRONATEC no ano de 2014.

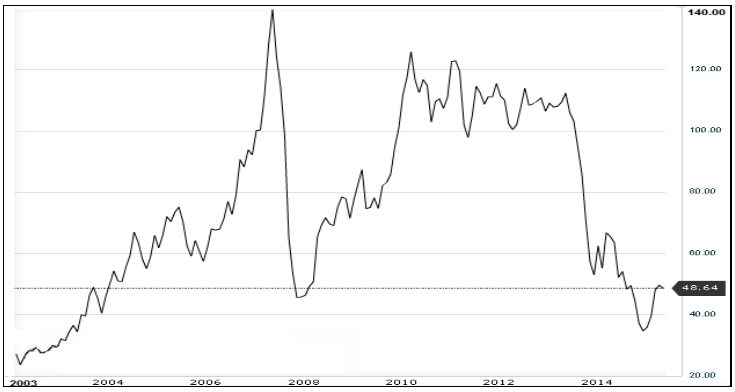
Figura 25 - Gráfico do tipo pizza representando matrículas nos cursos FIC –



Fonte: SETEC Data de Referência 31/05/2014

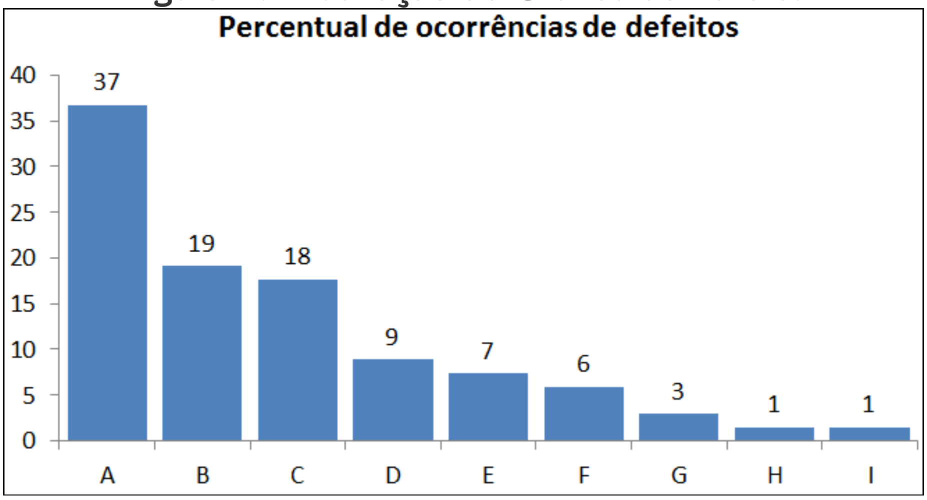
Na Figura 26, tem-se um gráfico de linha representado a evolução do preço de barril de petróleo (em U$) entre os anos 2003 e 2016.

Figura 26- Gráfico de linha representando a evolução do preço do barril de petróleo entre os anos de 2002 a 2016. [12]



Um tipo de gráfico também muito utilizado é o Diagrama de Pareto, que é conhecido como princípio 80-20. De acordo com Pareto, 80% das consequências decorrem de 20% das causas. Esta lei foi proposta por Joseph M. Juran, que deu esse nome como homenagem ao economista italiano Vilfredo Pareto. Algumas aplicações desse princípio: se uma empresa tem 100 clientes, em geral 20 deles são responsáveis por 80% dos lucros; mais de 80% das descobertas científicas são decorrentes do trabalho de 20% dos cientistas; 80% da riqueza do mundo está concentrada em 20% das pessoas; quando um avião cai é provável que 20% das causas sejam responsáveis por 80% dos defeitos e assim por diante. Na Figura 27 tem-se uma curva ABC representativa do Diagrama de Pareto. O conhecimento dos defeitos mais frequentes é importante para investimento de tempo e recursos na solução daquilo que é mais prioritário.

Figura 27- Ilustração do Gráfico de Pareto.



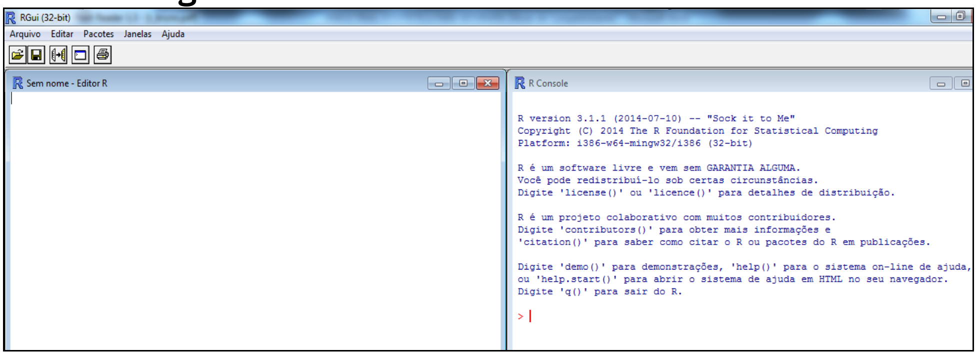
Fonte: Elaborado pelo autor

12 http://br.investing.com/commodities/brent-oil-advanced-chart

Os Histogramas e gráficos apresentados também podem ser construídos por meio do Software Estatístico R, que pode ser adquirido gratuitamente no link: http://www.vps.fmvz.usp.br/CRAN/ . Ele foi criado pelos professores Ross Ihaka e Robert Gentleman na Universidade de Auckland – na Nova Zelândia com a colaboração de pesquisadores de vários outros países. Trata-se de uma linguagem de programação especializada em computação de dados e que faz parte da filosofia de GNU – *General Public License*. Por ser gratuito e de fácil utilização vem se tornando um dos programas mais populares no mundo da estatística.

Após realizar download do programa, você verá uma tela de abertura conforme ilustrado na Figura 28– parte da direita. A tela de script (esquerda) auxilia a entrada das expressões necessárias. Basta acionar CONTROL R para que a expressão escrita na parte esquerda seja processada na tela da parte direita.

Figura 28– Tela de abertura do software R.



Fonte: Elaborado pelo autor

Durante a utilização do software é possível consultar a sintaxe de algum comando ou obter mais informações sobre determinada função. Para isso o R conta com o comando help. A sintaxe do comando é a seguinte:> help(comando) #sintaxe

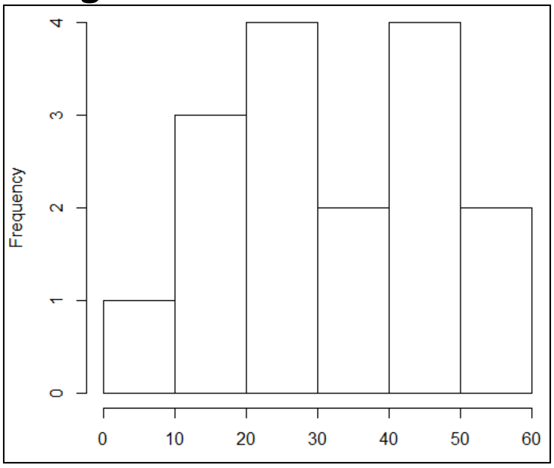
A seguir, serão apresentadas algumas aplicações do R na construção de gráficos.

Exemplo 1- Ao se digitar os comandos abaixo tem-se o histograma (Figura 29).

tempo<-c(50,40,41,17,11,7,22,44,28,21,19,23,37,51,54,42)

hist(tempo).

Figura 29- Histograma construído com uso do software R

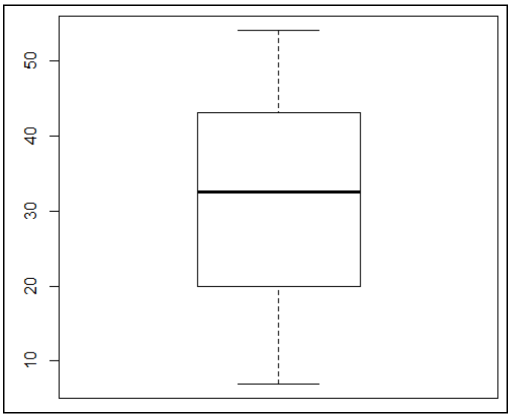


Fonte: Elaborado pelo autor

Exemplo 2- Ao se digitar os comandos abaixo tem-se o diagrama de caixa (Figura 30)

```
tempo<-c(50,40,41,17,11,7,22,44,28,21,19,23,37,51,54,42)
boxplot(tempo)
```

Figura 30- Diagrama de Caixa construído com uso do software R

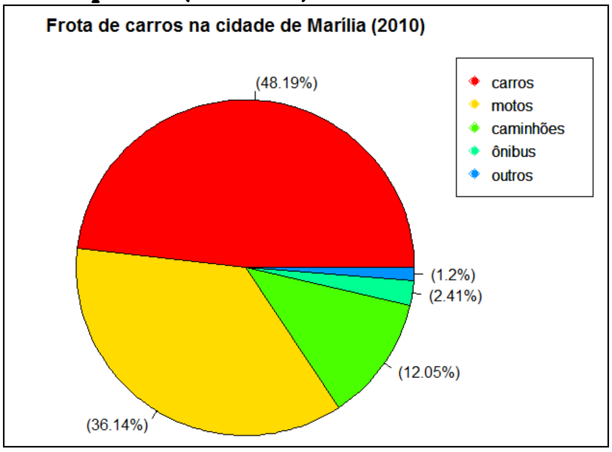


Fonte: Elaborado pelo autor

Exemplo 3- Ao se digitar os comandos abaixo tem-se gráfico de pizza (Figura 31):

```
frota<-c(80000, 60000, 20000,4000,2000)
names(frota)<-c("carros","motos","caminhões","ônibus","outros")
pie(frota)
porc<-round(frota*100/sum(frota),2) #arredonda a porcentagem)
rotulos<-paste("(",porc,"%)",sep="")
pie(frota, main="Frota de carros na cidade de Marília (2010)",labels=rotulos,
col=rainbow(7))
legend(1,1,names(frota),col = rainbow(7),pch=rep(20,6))
```

Figura 31- Gráfico de pizza (setores) construído com uso do software R.

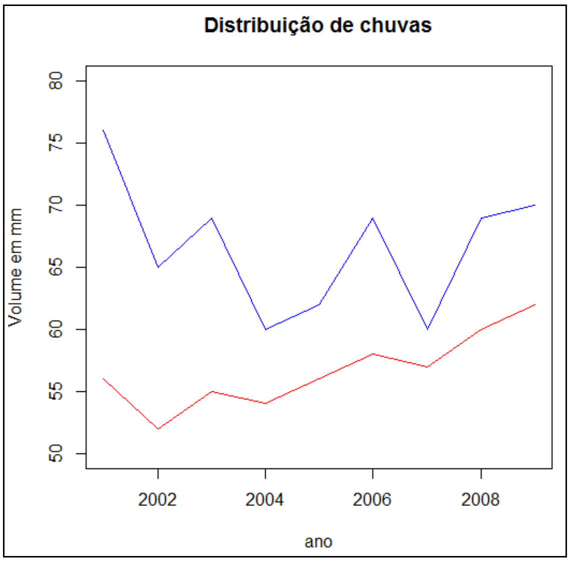


Fonte: Elaborado pelo autor

Exemplo 4- Ao se digitar os comandos abaixo tem-se gráfico de linha (Figura 32):

```
ano<-2001:2009
cidadea<-c(76,65,69,60,62,69,60,69,70)
cidadeb<-c(56,52,55,54,56,58,57,60,62)
plot(ano, cidadea,type="l",main="Distribuição de chuvas",xlab="ano",ylab="Volume em mm",col="blue",ylim=c(50,80))
lines(ano, cidadeb,col="red")
```

Figura 32- Gráfico de linhas construído com uso do software R

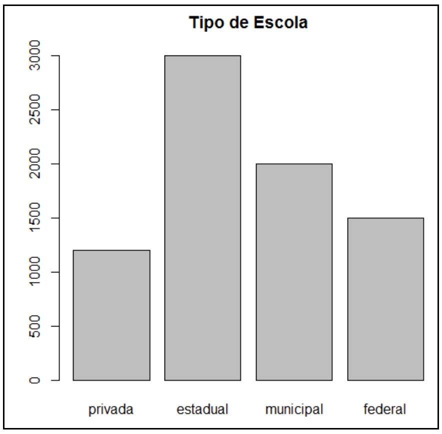


Fonte: Elaborado pelo autor

Exemplo 4- Ao se digitar os comandos abaixo tem-se gráfico de colunas (Figura 33):

```
alunos<-c(1200,3000,2000,1500)
escola<-c("privada","estadual","municipal","federal")
barplot(alunos, names.arg=escola, type="l",main="Tipo de Escola")
```

Figura 33- Gráfico de barra construído com uso do software R



Fonte: Elaborado pelo autor

LISTA DE EXERCÍCIOS 2

1- Ordene os dados. Indique o 1º, 2º e 3º quartil. Desenhe o diagrama de caixa.

11, 12, 4, 2, 3, 4, 11, 8, 5, 15, 20, 21

2- O quadro seguinte representa as estaturas (em cm) de 25 alunos de uma classe. Construa o histograma representativo.

155	163	148	166	169
164	165	159	175	155
170	165	176	157	157
150	150	160	165	164
166	169	152	170	190

3- Represente a distribuição do tamanho dos municípios catarinenses por meio de gráficos de barras e de setores.

Número de habitantes	Quantidade de municípios em SC	% de municípios
Até 5 mil	108	37
De 5 mil a 10 mil	64	22
De 10 mil a 20 mil	60	20
De 20 mil a 50 mil	34	12
Maior que 50 mil	27	9
Total	293	100

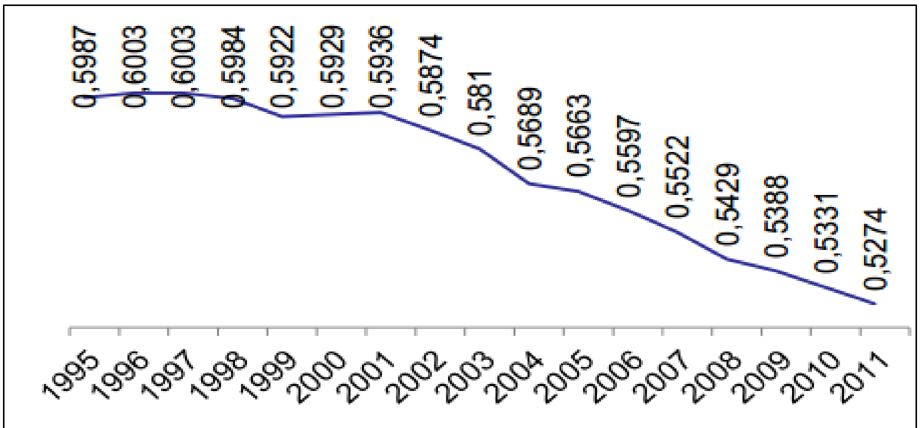
4- Analise a distribuição de municípios e a população do Estado de Santa Catarina por meio de um gráfico de setores e de barras.

Mesorregião	Número de cidades	População
Norte Catarinense	26	1.212.843
Vale do Itajaí	54	1.508.980
Grande Florianópolis	21	994.095
Serrana	30	406.741
Oeste Catarinense	118	1.200.712
Sul Catarinense	44	925.065

5- Represente o número de empresas instaladas nas cidades catarinenses por meio de um gráfico de setores.

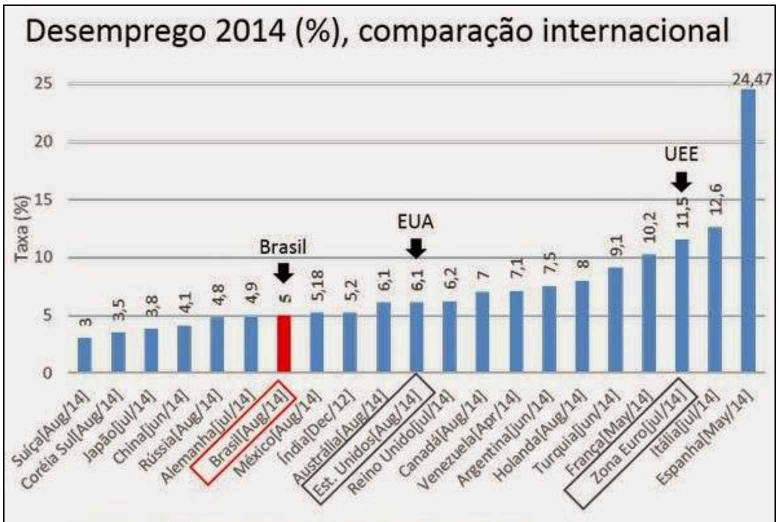
Cidade	Número de empresas
Palhoça	4852
Jaraguá do Sul	7105
Lages	5634
Itajaí	9380
Chapecó	8544
Criciúma	8660
São José	9632
Blumenau	18305
Florianópolis	24746
Joinville	19571

6- Interprete e reconstrua com outra escala o gráfico da evolução do Índice Gini médio do Brasil.



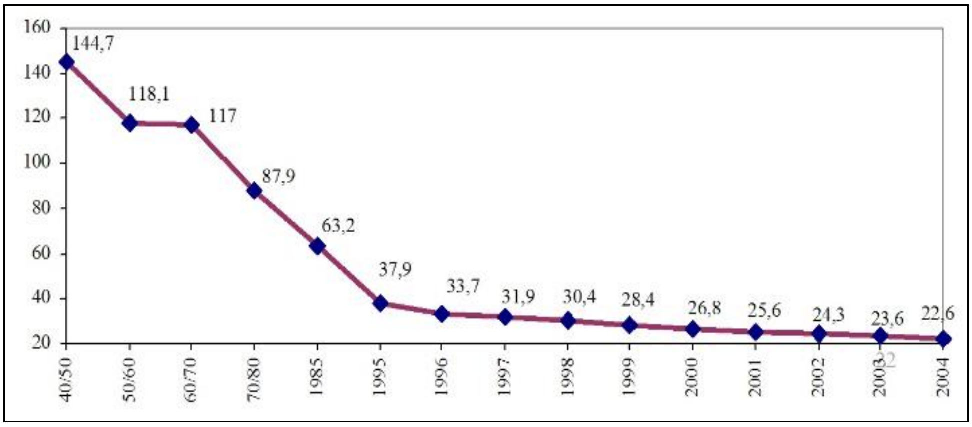
Fonte: IPEA a partir dos microdados da PNAD

7- Analise os dados comparativos para o desemprego para o ano de 2014. Construa um diagrama de caixa a partir das informações do gráfico.



Fonte: http://pt.tradingeconomics.com/brazil/unemployment-rate
http://www.ibge.gov.br/home/estatistica/indicadores/trabalhoerendimento/pme_nova/

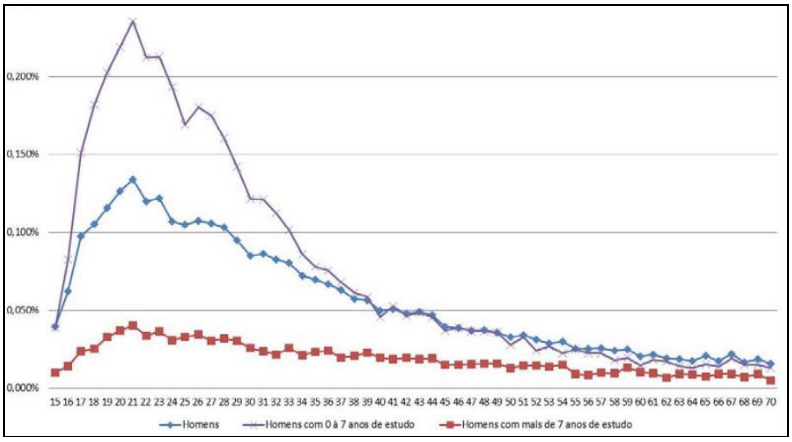
8- Analise a conveniência da escala utilizada no gráfico que mostra a redução da mortalidade infantil no Brasil (mortos por mil nascidos vivos).



9- Avalie o gráfico que representa o número de mortes em acidades de trânsito no Brasil. Represente os dados em um gráfico de barras.

Fonte:
http://www.vias-seguras.com/os_acidentes/estatisticas/estatisticas_nacionais

10- Avalie criticamente o gráfico que representa a probabilidade de mortes violentas de acordo com a faixa etária e nível de escolaridade.



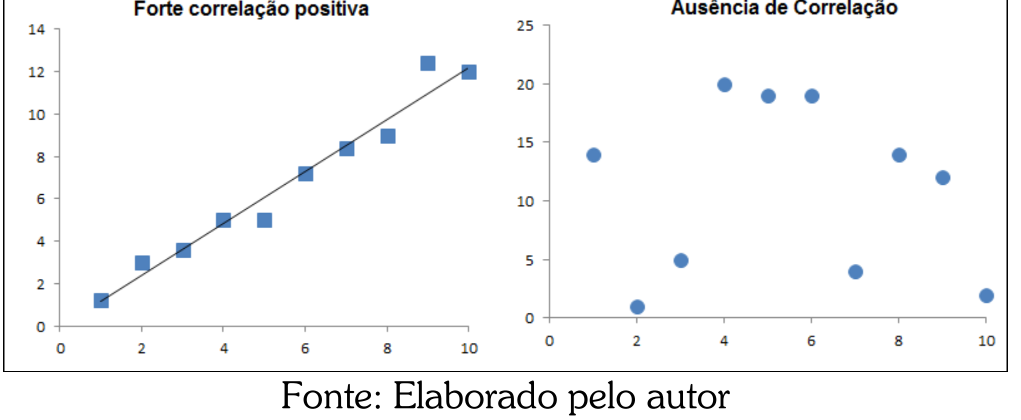
Fonte:
http://infogbucket.s3.amazonaws.com/arquivos/2016/03/22/atlas_da_violencia_2016.pdf

3- Correlações

Você já parou para pensar se existe uma correlação entre o peso (massa corporal) e a estatura dos estudantes de uma determinada turma? E entre horas de estudo e resultados nas provas? Ou entre a temperatura no verão e a venda de cervejas? Ou entre tempo de exposição na televisão de uma marca e resultado nas vendas?
Existe uma correlação entre duas variáveis quando uma delas está de alguma forma relacionada com a outra. Quando a alteração no valor de uma variável (chamada independente) provoca alterações no valor da outra variável (chamada dependente).
Nos exemplos acima é possível identificar com facilidade a relação de causa e efeito entre as variáveis. A variável venda de cerveja é uma variável dependente da variável independente temperatura. Essa relação de causa e efeito nem sempre existe. Por isso é importante sempre identificar se determinado fato realmente tem relação direta com outro. Quando isso não ocorre temos uma correlação chamada de "espúria".
Quando analisamos uma correlação simples entre duas grandezas temos quatro possibilidades. Pode existir correlação positiva, forte correlação negativa, forte correlação positiva ou ausência de correlação. Na Figura 34 são ilustradas duas dessas situações.
Para avaliar a força de uma correlação o cientista K. Pearson definiu o valor chamado "R", que pode ser calculado matematicamente. Quanto mais próximo de 1 (unidade) mais forte é a correlação. As correlações fracas têm valores de "R" menores que 0,5.

$$R = \frac{\sum (x_i - \bar{x}).(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 . \sum (y_i - \bar{y})^2}}$$

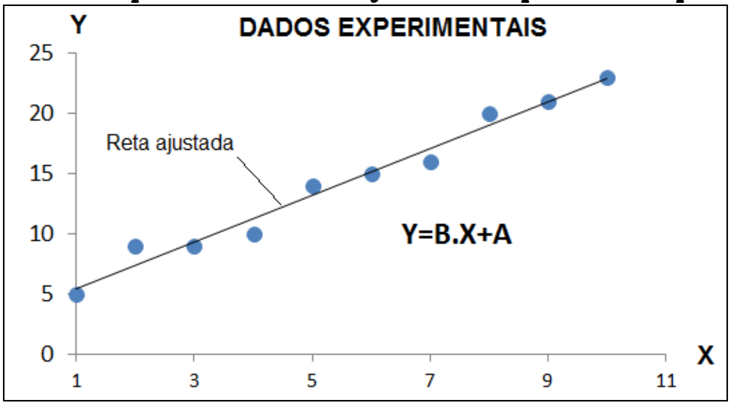
Figura 34- Diferentes correlações possíveis.



Fonte: Elaborado pelo autor

Muitas vezes os dados experimentais apontam para a existência de uma relação entre as variáveis dependente e independente. Mas para se estabelecer uma curva de ajuste que mais representa o fenômeno em estudo é importante utilizar técnicas matemáticas chamadas de regressão. Quando a relação é linear é possível, com pouco esforço, descobrir a correlação existente entre as variáveis dependente (Y) e independente (X). Na Figura 35, tem-se representados um conjunto de pontos experimentais e uma reta de ajuste dada pela equação Y = B.X + A. Essa é uma função do primeiro grau com coeficiente angular B e coeficiente linear A.

Figura 35- Reta que melhor se ajusta aos pontos experimentais.



Fonte: Elaborado pelo autor

$$B = \frac{N.\sum X.Y - \left[\left(\sum X\right)\sum Y\right]}{N.\left(\sum X^2\right) - \left(\sum X\right)^2} \qquad A = \frac{\sum Y}{N} - B.\frac{\sum X}{N}$$

Imagine como exemplo, que um médico tenha anotado ao longo dos anos as idades e as estaturas de uma criança, obtendo as seguintes informações:

X - Idade (anos)	Y - Estatura (cm)
6	70
8	110
10	130
12	150

Para esse caso é possível perceber que conforme a criança vai ficando mais velha, sua estatura aumenta, ou seja, existe uma relação direta de causalidade.

X	Y	X.Y	X^2
6	70	420	36
8	110	880	64
10	130	1300	100
12	150	1800	144

Soma =36	Soma =460	Soma = 4400	Soma =344

$$B = \frac{N.\sum X.Y - \left[\left(\sum X\right)\sum Y\right]}{N.\left(\sum X^2\right) - \left(\sum X\right)^2} = \frac{(4.4400) - (36.460)}{(4.344) - (36^2)} = 13$$

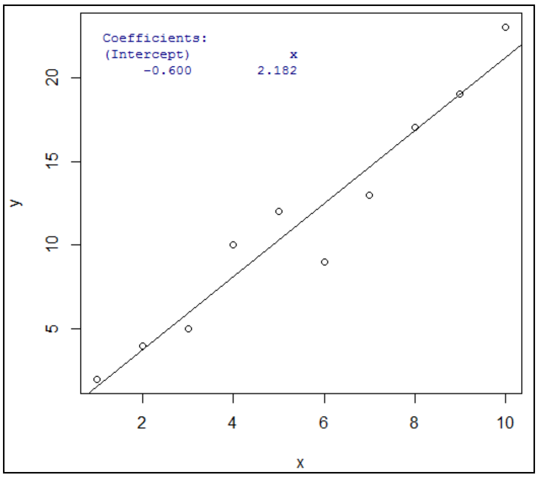
$$A = \frac{\sum Y}{N} - B.\frac{\sum X}{N} \Rightarrow A = \left(\frac{460}{4}\right) - 13.\left(\frac{36}{4}\right) = -2$$

A equação que correlaciona a estatura e a idade da criança é: Estatura = 13 x Idade – 2. O cálculo de R^2 fornece 0,96, o que possibilita afirmar que existe uma forte correlação[13]. Com o auxílio do software R é possível encontrar as correlações mais diversas.

Como exemplo digite os comandos e observe o valor dos coeficientes da reta ajustada (Figura 36). A equação que se ajusta exatamente ao conjunto de pontos experimentais é Y = 2,18.X - 0,6.

```
x<-c(1,2,3,4,5,6,7,8,9,10)
y<-c(2,4,5,10,12,9,13,17,19,23)
lm(y~x)
plot(x,y)
equacao<-lm(y~x)
abline(equacao)
```

Figura 36– Correlação linear elaborada no software R.



Fonte: Elaborado pelo autor

Quando determinamos uma correlação matemática entre uma variável dependente e outra independente é possível fazer a extrapolação de dados. Ou seja, podemos prever determinado fenômeno a partir de uma série histórica. Na realidade, com a existência de grande volume de dados disponíveis na internet é possível aplicar a técnica chamada de “BIG DATA”

[13] Recomendamos o vídeo com exemplo resolvido: https://www.educreations.com/lesson/view/estatistica-aula-31-correlacao-entre-idade-e-altur/19584560/

ou “mineração de dados” para avaliar padrões de comportamento das pessoas.

Se uma determinada loja sabe exatamente do que gostamos então ela pode customizar o atendimento. Diariamente deixamos dezenas de pistas sobre o que gostamos quando fazemos pesquisas na internet, quando curtimos publicações no *Facebook*. Esse banco de dados tem sido disputado por grandes empresas. A criação de modelos matemáticos cada vez mais sofisticados permite que se façam inferências futuras a partir de dados do passado.

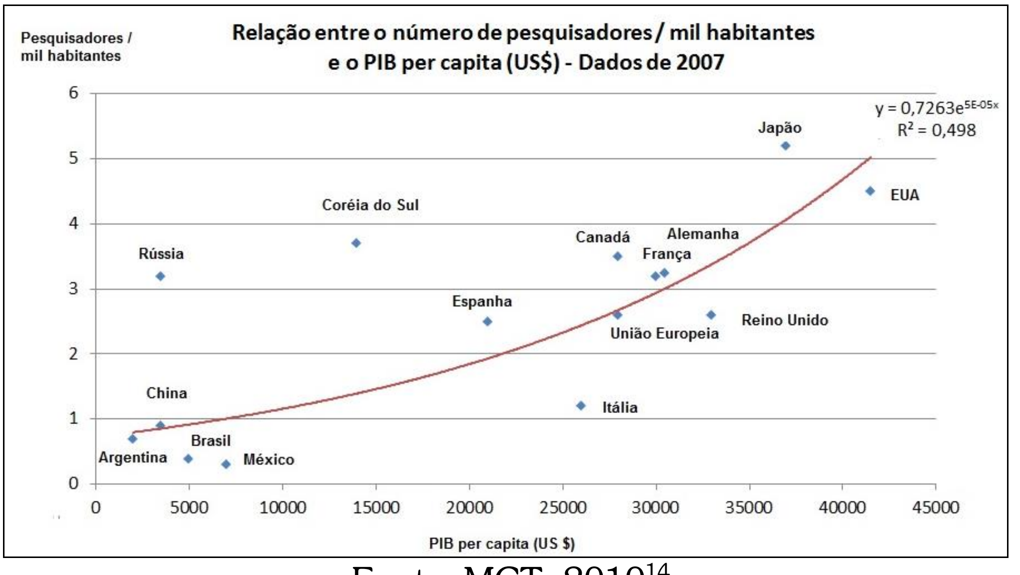
Mas nem toda correlação é simples como as apresentadas anteriormente. Há casos em que as correlações somente podem ser calculadas por meio de ferramentas computacionais. Um exemplo é o valor de venda de uma casa, onde o preço final depende de uma série de fatores como área construída, tempo de uso, localização, número de banheiros. Apenas a área construída não permite explicar o preço final. Nesse caso há programas como R – um software livre fácil de usar que possibilita que se encontrem as equações mais adequadas para cada caso.

Há também um tipo de correlação que não representa uma relação direta de causa e efeito. No entanto, a primeira vista é possível que pareça que sim. Nesse caso tem-se uma “correlação espúria”. Dois eventos distintos podem não ter relação alguma entre si. No entanto, por uma questão do acaso, mostram intima relação estatística. O fato de dois fenômenos ocorrerem ao mesmo tempo não permite a inferência de que um seja causado pelo outro. Um exemplo real é sobre os estudos sobre a paralisia infantil. Inicialmente os cientistas verificaram a existência de uma forte associação positiva entre o número de casos da doença por semana e o número de vendas de refrigerante na mesma semana. Nesse caso algumas pessoas começaram a estabelecer uma relação direta de causalidade. Mas isso é um absurdo que pode ser percebido por meio de perguntas simples: “o refrigerante causa pólio?” ou “a pólio aumenta a vontade de beber refrigerante?” À luz do nosso conhecimento atual, estas perguntas são sem sentido. No entanto, para estudos recentes, com doenças ainda pouco estudadas, por exemplo, perguntas similares podem não parecer tão absurdas. Um exemplo atual é a relação entre a microcefalia em recém-nascidos e os casos de Zica vírus. No final de 2015, quando a relação foi estabelecida, não havia ainda estudos científicos e número de casos suficientes para sustentar a afirmação. Ainda hoje há contestações das conclusões apresentadas pela Organização Mundial da Saúde.

O estatístico e geneticista inglês Ronald Fisher (1890-1962) provou na década de 30 que existia uma correlação positiva entre a população da cidade de Oldenburg e o número de cegonhas. Ele mostrou que a população e o número de cegonhas aumentaram ao longo do período de estudo. O resultado não significa que o crescente número de cegonhas causou o aumento observado na população. Na verdade, uma coisa não provoca a outra, mas as duas são causadas por uma terceira: o aumento da população.

Na Figura 37 tem-se uma possível correlação não linear entre o número de pesquisadores por mil habitantes em relação ao PIB per capita de países com mais de 30 milhões de habitantes. É preciso ficar atento para a relação de causa e efeito. Na promoção do desenvolvimento há um conjunto de fatores e causas econômicas, políticas e sociais que não podem ser relegadas a um segundo plano. Os países mais ricos investem mais em P&D por quê são mais ricos ou se tornaram mais ricos por quê investiram mais em P&D?

Figura 37 Relação entre o número de pesquisadores / mil habitantes e o PIB per capita de países com mais de 30 milhões de habitantes



Fonte: MCT, 2010[14]

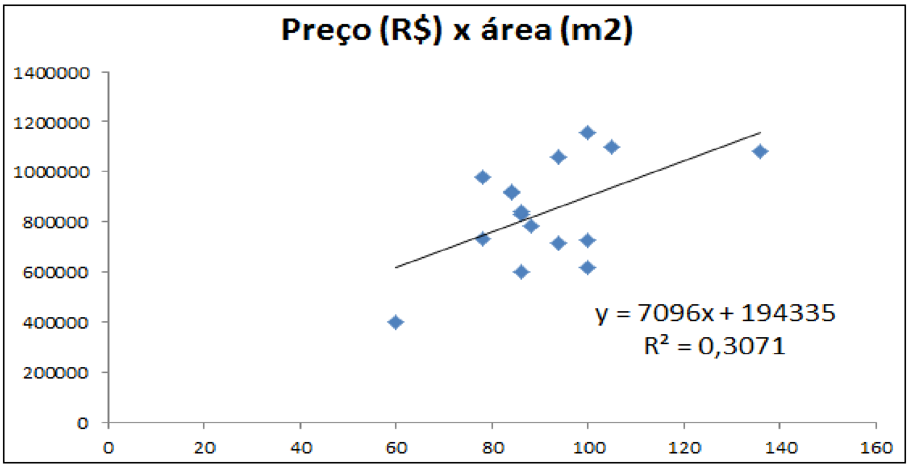
Para determinação da existência ou não de correlações, normalmente são utilizados softwares especializados. Existem regressões simples e regressões múltiplas, quando há uma variável independente e diversas outras dependentes. Essas são as mais comuns na realidade. Como exemplo, tem-se o custo de um imóvel como decorrente de sua área construída e de seu tempo de vida. O custo é a variável dependente da área e do tempo de vida. Trata-se de um evento onde Y = variável dependente e X1 e X2 = variáveis independentes.

14 http://www.mct.gov.br/upd_blob/0203/203406.pdf

Preço (R$) Y	Área (m^2) X1	Idade (anos) X2
400000	60	9
832000	86	10
1100000	105	8
727000	100	11
784000	88	8
1158400	100	9
1080000	136	9
840000	86	10
920000	84	11
713000	94	6
620000	100	14
600000	86	13
733000	78	10
915000	84	8
980000	78	6
1060000	94	4

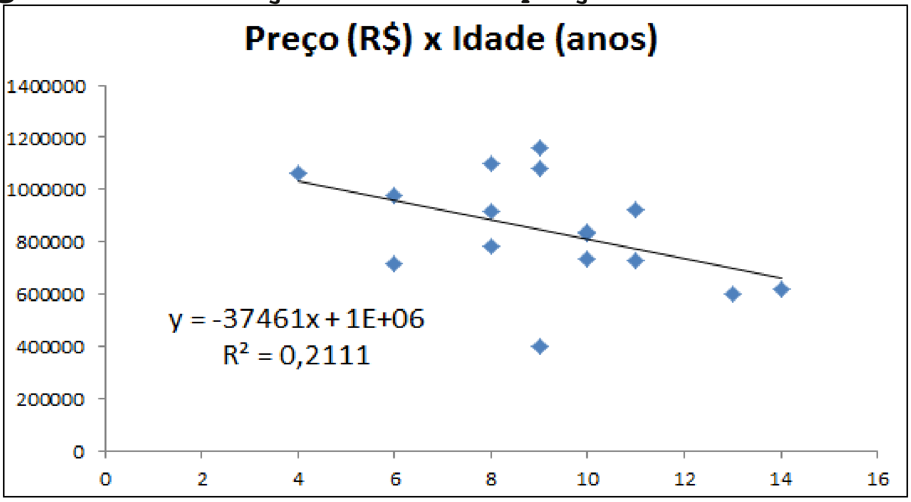
Nas Figuras 38 e 39, tem-se a representação das correlações entre preço e área e entre o preço e o tempo de uso da amostra de imóveis. Observe a partir do valor de R^2 que as correlações isoladas são fracas.

Figura 38– Correlação linear entre preço e área dos imóveis.



Fonte: Elaborado pelo autor

Figura 39– Correlação linear entre preço e idade dos imóveis.



Fonte: Elaborado pelo autor

A partir do Software Estatístico R é possível avaliar a correlação entre essas variáveis. Basta escrever as expressões a seguir:

y<-c(400000, 832000,1100000,727000,784000,1158400,1080000,840000,920000,713000,620000,600000,733000,915000,980000,1060000)
x1<-c(60, 86,105,100,88,100,136,86,84,94,100,86,78,84,78,94)
x2<-c(9,10,8,11,8,9,9,10,11,6,14,13,10,8,6,4)
model<-lm(y~x1+x2)
anova (model)
lm(formula=y~x1+x2)

```
Analysis of Variance Table

Response: y
          Df     Sum Sq    Mean Sq F value  Pr(>F)
x1         1 1.9952e+11 1.9952e+11  8.5338 0.01191 *
x2         1 1.4626e+11 1.4626e+11  6.2558 0.02653 *
Residuals 13 3.0394e+11 2.3380e+10
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> lm(formula=y~x1+x2)

Call:
lm(formula = y ~ x1 + x2)

Coefficients:
(Intercept)           x1           x2
     532625         7258       -38695
```

Preço do imóvel = R$ 53.2625 + (7.258 x Área) – (38.695 x Idade)

Se quisermos saber aproximadamente o custo de um apartamento de 100 metros quadrados e com 5 anos de idade basta substituir esses valores na equação obtida da regressão múltipla. Nesse caso o valor do imóvel custaria aproximadamente R$ 1.064.950,00.

LISTA DE EXERCÍCIOS 3

1- Calcule a correlação que relaciona a idade e a altura de uma criança.

Idade (anos)	Altura (cm)
6	70
8	110
10	130
12	150
14	155
15	180

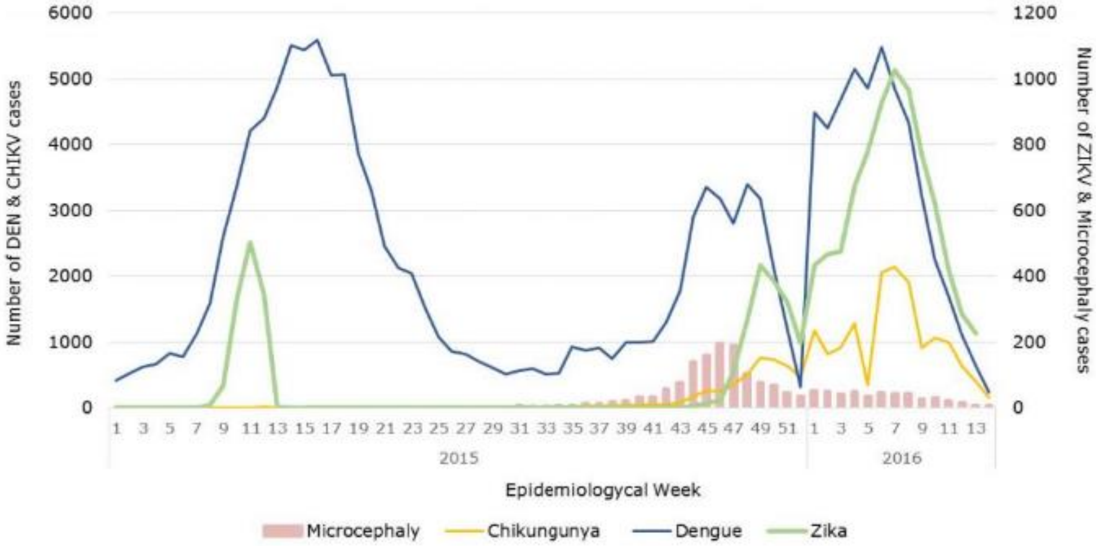
2- O dono de uma lanchonete anotou quanto de refrigerantes (em litros) ele vende ao longo dos dias de acordo com a temperatura. Qual a relação entre estas duas informações?

Temperatura (°C)	Refrigerantes vendidos (litros)
15	22
20	25
25	28
27	30
30	32
31	31
32	33
35	50

3- Um pesquisador está estudando a relação entre os preços de uma casa, o tamanho dos terrenos e o número de quartos. Analisando uma amostra de propostas de vendas em sites específicos ele anotou os valores médios das casas e as respectivas áreas dos terrenos e número de quartos. Qual a correlação entre essas 3 variáveis?

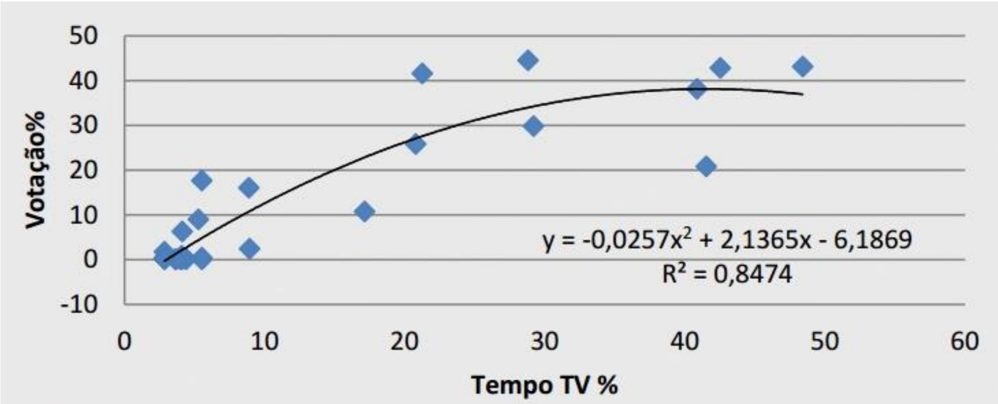
Preço da casa	Área do Lote (m^2)	Número de quartos
130.000	5000	3
134.000	5500	2
159.000	6000	4
164.000	6500	3
132.000	5200	2
125.000	5400	1
146.000	5700	3
168.000	6100	4
171.000	6300	4
187.000	6400	5

4- Analise criticamente as curvas que representam o número de casos de microcefalia, Chikungunya, dengue e Zica vírus.



http://www.pbs.org/newshour/updates/how-many-zika-infected-infants-will-develop-microcephaly-and-other-faqs/

5- Avalie criticamente a correlação polinomial que associa percentual de tempo de televisão e percentual de votação nas eleições.



Fonte: Elaborado pelo autor

4- Medidas de Tendência Central

Um conjunto de dados pode ser descrito por meio de alguns números representativos chamados de **"Medidas de Tendência Central ou Medidas de Centralidade"**. Entre elas temos a Media Aritmética, a Moda e a Mediana.

a) Média Aritmética é a mais usada dentre todas as médias, face à sua aplicabilidade a situações práticas. Podemos calcular a média aritmética de várias maneiras, dependendo apenas da forma em que os dados se encontram. Podemos utilizar a média simples ou a média ponderada.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Onde n = total de dados e x_i representam todos os elementos do conjunto de dados.

Quando os dados estão agrupados em intervalos de classe, convenciona-se que todos os valores incluídos em um determinado intervalo de classe coincidem com o seu ponto médio, e determina-se a média aritmética ponderada pela fórmula:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i \cdot f_i}{\sum_{i=1}^{n} f_i}$$

onde: x_i é o ponto médio de cada classe i.

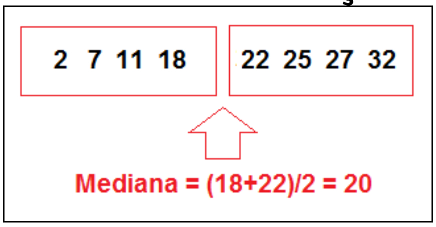
Além do cálculo da média simples também é comum o uso da média ponderada. Uma aplicação simples é quando se tem pesos diferentes nas notas das provas. Se um aluno tirou 10,0 na prova de peso 1 e 4,0 na prova de peso 2 então sua nota final será:

$$\bar{x} = \frac{10.1 + 4.2}{3} = 6$$

b) moda – é o valor da amostra que mais aparece (de maior frequência). Uma amostra pode ser: amodal, unimodal, bimodal, trimodal ou multimodal.

c) mediana – A mediana de uma amostra é aquele valor que ocupa a posição central do rol, isto é, a mediana é o valor que divide a amostra em duas partes iguais. A mediana pode não pertencer a amostra. Isso acontece no caso representado na Figura 40. A mediana divide os dados em 2 partes iguais. Mesmo não existindo o número 20 na sequência, esse é o valor da mediana, obtida por meio da média entre os números 18 e 22.

Figura 40 – Ilustração da forma de obtenção do valor da mediana.



A mediana tem uma vasta aplicação estatística porque é menos sensível aos valores extremos do conjunto de dados. Como exemplo: Uma turma tem as seguintes massas: 70, 80, 60, 90, 50, 55, 85. A média é calculada em 70kg. Mas se, ao invés de 90 a última massa fosse de 200kg a média da turma passaria a ser 86kg. Mas a mediana nos dois casos não se altera.

50 55 60 70 80 85 90	mediana = 70 kg e média = 70kg
50 55 60 70 80 85 200	mediana = 70kg e média = 86kg

Essa característica torna a mediana uma medida de tendência central importante para análises estatísticas. Muitas vezes a renda média dos moradores de uma cidade é de R$ 3.000,00, mas a mediana dos rendimentos é de R$ 600,00. Ou seja, metade dos moradores da cidade recebe menos que R$ 600,00.

Para avaliar o quanto os dados se dispersam em relação às medidas de tendência central tem-se a **variância**, calculada a partir da somatória de todos os desvios em torno da média aritmética ao quadrado. Por definição, o desvio padrão é calculado pela raiz quadrada da variância. Para uma população de tamanho “N” a variância é calculada pela equação:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

Quando o interesse for o cálculo da variância de uma amostra de dados é convencional o uso da expressão:

$$s^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{(N-1)}$$

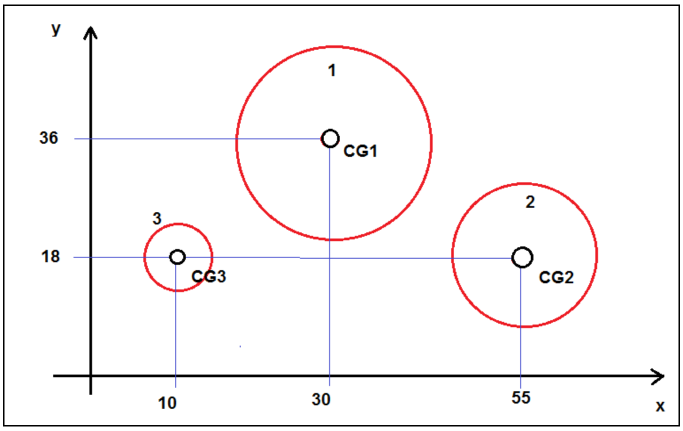
Um conceito muito utilizado na Geografia é o de centroide ou centro de massa de diversas populações distribuídas no espaço. As coordenadas x e y do centroide são calculadas pela equação:

$$x_{CG} = \frac{\sum_{i=1}^{n}(x_i.P_i)}{\sum_{i=1}^{n}P_i} \quad \text{e ainda} \quad y_{CG} = \frac{\sum_{i=1}^{n}(y_i.P_i)}{\sum_{i=1}^{n}P_i}$$

Exemplo 1:

Suponha que um novo centro de eventos está sendo planejado para uma determinada região. Nela há 3 comunidades residenciais e o centróide é um dos critérios para localização porque garante a equidistância (Figura 41). Suponha que a comunidade 1 tenha coordenada central (x=30, y=36)km e população de 20 mil pessoas. A comunidade 2 tem coordenada central de (x=55,y=18)km e população de 12 mil pessoas. Já a comunidade 3 tem coordenada central de (x=10,y=18)km e população de 5 mil pessoas. Qual é o centroide da população?

Figura 41– Cálculo do centróide da população de 3 comunidades



$$x_{CG} = \frac{\sum_{i=1}^{n}(x_i.P_i)}{\sum_{i=1}^{n}P_i} = \frac{(10.5000)+(30.20000)+(55.12000)}{37000} = 35{,}4km$$

$$y_{CG} = \frac{\sum_{i=1}^{n}(y_i.P_i)}{\sum_{i=1}^{n}P_i} = \frac{(18.5000)+(36.20000)+(18.12000)}{37000} = 27{,}73km$$

Como é possível observar os valores de 35,4km e 27,73km representam o ponto médio entre as comunidades. Esse valor também é conhecido como centro de massa.

Classes	Frequência Fi
39,5 a 44,5	3
44,5 a 49,5	8
49,5 a 54,5	16
54,5 a 59,5	12
59,5 a 64,5	7
64,5 a 69,5	3
69,5 a 74,5	1

Exemplo 2:

Um pesquisador anotou a frequência do volume de chuvas de uma cidade (em mm). Calcule a média e o desvio padrão.

Para resolver o problema é necessário preencher a tabela:

Classes	Frequência (F_i)	x_i (valor médio)	$x_i \cdot F_i$	$x_i^2.F_i$
39,5 a 44,5	3	42	126	5292
44,5 a 49,5	8	47	376	17672
49,5 a 54,5	16	52	832	43264
54,5 a 59,5	12	57	684	38988
59,5 a 64,5	7	62	434	26908
64,5 a 69,5	3	67	201	13467
69,5 a 74,5	1	72	72	5184
Soma	50		2725	150775

Nesse caso, a variância pode ser calculada por meio da expressão:

$$S^2 = \frac{\sum_{i=1}^{k} x_i^2 \times f_i - \left(\frac{\left(\sum_{i=1}^{k} x_i \times f_i \right)^2}{N} \right)}{N-1} = \frac{150755 - \left(\frac{2725^2}{50} \right)}{49} = 46{,}17$$

O desvio padrão é calculado pela raiz quadrada de 46,17 resultando em 6,79. A média dos dados é calculada por 2725 / 50 = 54,5mm. Nesse caso, o coeficiente de variação das medidas (CV) é calculado por 6,79/54,5 = 0,125 ou 12,5%. O coeficiente de variação é usado para analisar a dispersão em termos relativos a seu valor médio. Dessa forma, podemos dizer que o coeficiente de variação é uma forma de expressar a variabilidade dos dados excluindo a influência da ordem de grandeza da variável.

Exemplo 3

Dado um conjunto de massas de uma turma de estudantes, calcule qual é a média, o desvio padrão e o Coeficiente de Variação (CV): 63, 55, 78, 82, 95, 60, 82, 75,74, 76, 80, 90

Peso (x_i)	Média $\bar{x}$	($\bar{x}$ - x_i)	($\bar{x}$ - x_i)^2
63	75,8	-12,8	163,84
55	75,8	-20,8	432,64
78	75,8	2,2	4,84
82	75,8	6,2	38,44
95	75,8	19,2	368,64
60	75,8	-15,8	249,64
82	75,8	6,2	38,44
75	75,8	-0,8	0,64
74	75,8	-1,8	3,24
76	75,8	0,2	0,04
80	75,8	4,2	17,64
90	75,8	14,2	201,64
Soma			1519,68

$$\bar{x} = \frac{\sum_{i=1}^{n} (63 + 55 + 78 + 82 + 95 + 60 + 82 + 75 + 74 + 76 + 80 + 90)}{12} = 75{,}8$$

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N} = \frac{1519{,}68}{12} = 126{,}64$$

O desvio padrão é calculado em 11,25. Já o coeficiente de variação (CV) = 11,25/75,8 = 0,148 ou em termos percentuais: 14,8%.

Exemplo 4

Considere o conjunto de dados do exemplo 3. Identifique qual é a mediana, o primeiro quartil e o terceiro quartil: 63, 55, 78, 82, 95, 60, 82, 75,74, 76, 80, 90

Os dados devem inicialmente ser ordenados:

55, 60, 63, 74, 75, 76, 78, 80, 82, 82, 90, 95

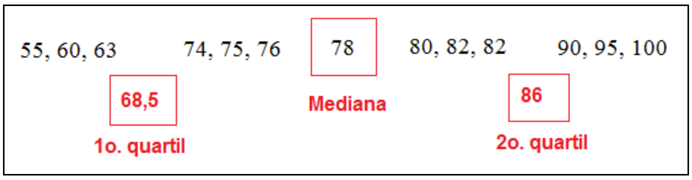
A mediana divide o conjunto de dados em 2 partes iguais. Considerando que na sequência há 12 números os dados podem ser separados da seguinte forma:

55, 60, 63, 74, 75, 76	78, 80, 82, 82, 90, 95

Nesse caso a mediana é a média entre número 76 e 78, ou seja: 77. Esse também é o segundo quartil. O primeiro quartil divide a primeira metade dos números em 2 partes iguais. Como não há esse número, utiliza-se o valor médio entre 63 e 74 que é igual a 68,5. O terceiro quartil também é calculado da mesma forma, sendo o valor médio entre 82 e 82 que é o próprio número 82.

Exemplo 5[15]

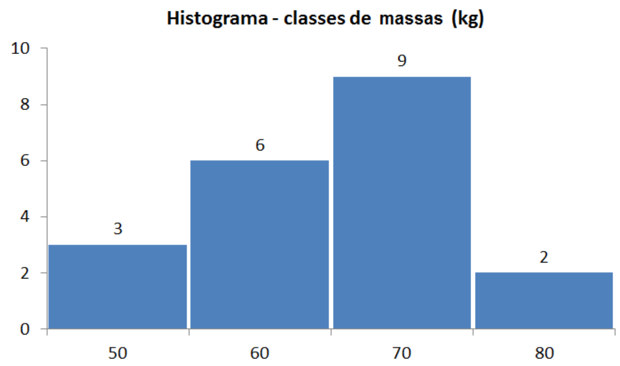
No exemplo anterior, suponha que o número 100 tenha sido acrescentado à série. Nesse caso o primeiro quartil, a mediana e o terceiro quartil seriam determinados da seguinte forma: 55, 60, 63, 74, 75, 76, 78, 80, 82, 82, 90, 95, 100



https://www.educreations.com/lesson/view/estatistica- -26-construcaode-dagramade-caix/19492821/

LISTA DE EXERCÍCIOS 4

1- Dado um histograma das massas de uma turma de estudantes, qual a moda e o terceiro quartil?



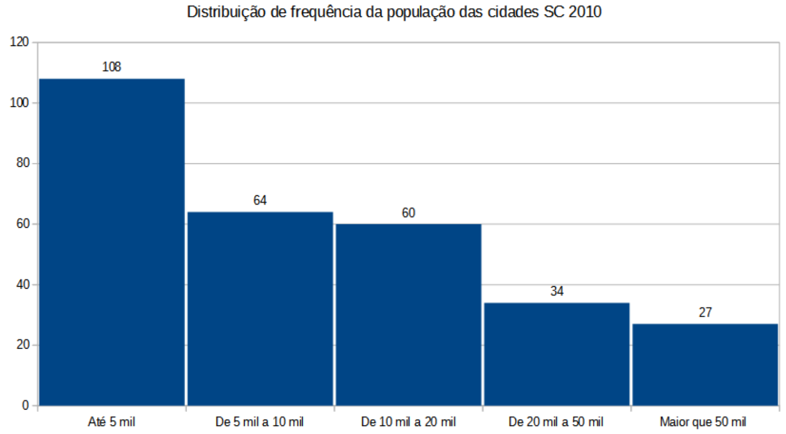
2- As notas de uma turma de alunos são mostradas na tabela. Qual a média e a mediana?

Nota	Quantidade
2	2
4	4
6	12
8	6
10	2

3- Na tabela são descritos os rendimentos médio e mediano do conjunto de trabalhadores formais das 10 maiores cidades catarinenses. Desenhe o diagrama de caixa do rendimento mediano. Calcule qual o rendimento médio dos trabalhadores dessas 10 cidades.

Cidade	Rendimento médio	Rendimento mediano
Palhoça	2637,92	670,00
Jaraguá do Sul	2963,43	737,50
Lages	2304,91	530,00
Itajaí	3025,88	732,13
Chapecó	2701,80	630,00
Criciúma	3170,88	740,00
São José	2706,76	705,00
Blumenau	3225,36	766,67
Florianópolis	3968,30	958,34
Joinville	3072,39	700,00

4- Considerando o Histograma representativo do tamanho dos municípios catarinenses, calcule qual o valor médio da população.



5- Calcule a média e o desvio padrão do PIB per capita das mesorregiões catarinenses.

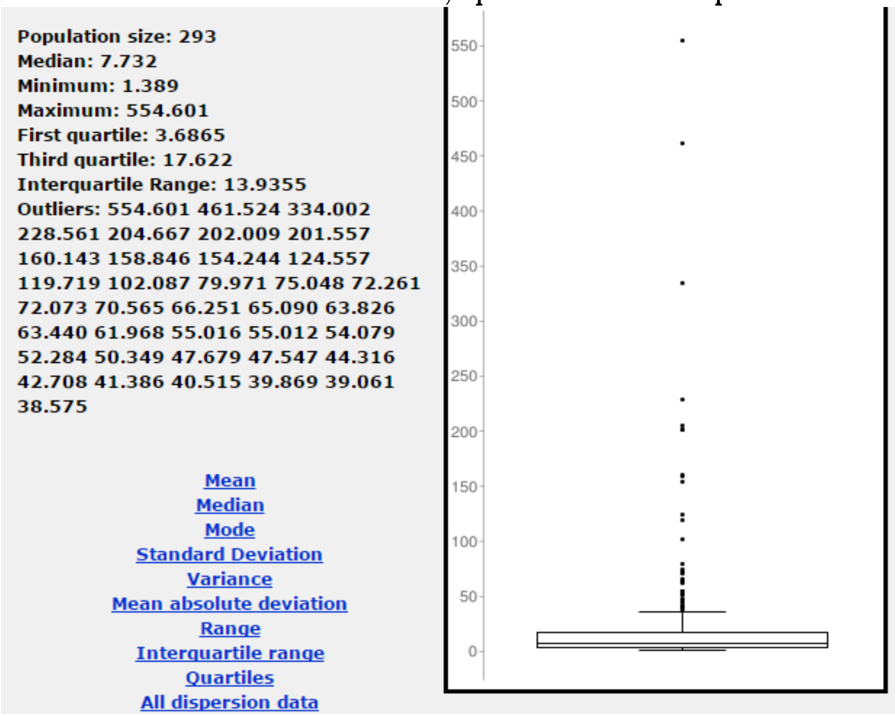
Mesorregião	PIB per capita R$
Norte Catarinense	22.613,00
Vale do Itajaí	20.131,00
Grande Florianópolis	16.470,00
Serrana	16.087,00
Oeste Catarinense	18.142,00
Sul Catarinense	17.821,00

6- Um novo centro de eventos está sendo planejado para uma determinada região. Nela há 3 comunidades residenciais e o centróide é um dos critérios para localização porque garante a equidistância. Suponha que a comunidade 1 tenha coordenada central (x=30, y=36)km e população de 20 mil pessoas. A comunidade 2 tem coordenada central de (x=60,y=20)km e população de 12 mil pessoas. Já a comunidade 3 tem coordenada central de (x=12,y=18)km e população de 5 mil pessoas. Calcule o centroide da população e da renda. Suponha que a comunidade 1 tenha renda total de 2 milhões de reais, a comunidade 2 tenha renda total de 12 milhões de reais e a comunidade 3 de 20 milhões de reais.

7- Os gastos mensais de uma amostra de famílias são descritos por meio de uma tabela, onde também estão descritas suas rendas. Qual a correlação existente entre a renda e o gasto mensal dessas famílias?

Quantia Gasta por semana (R$) Y	Renda da família (R$) X	X.Y	
120	6500		
68	3500		
35	3000		
60	4400		
100	8000		
91	7700		
44	3200		
71	3900		
89	4400		
113	7700		

8- Ao receber a informação de uma distribuição de população na forma de um diagrama de caixa ilustrado abaixo, que conclusões podem ser obtidas?



Fonte: http://www.alcula.com/calculators/statistics/box-plot/

9- Calcule a média, a amplitude, a mediana e o desvio padrão do conjunto de dados:

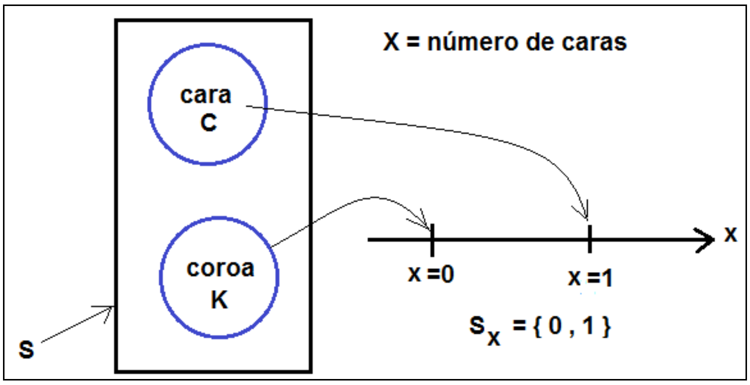
29, 35, 17, 30, 231, 6, 27, 35, 23, 29, 13

RECOMENDAÇÃO:
https://www.youtube.com/watch?v=2oGiQ3VySzU UNIVESP TV – CURSO DE ESTATÍSTICA

5- Distribuição de probabilidades

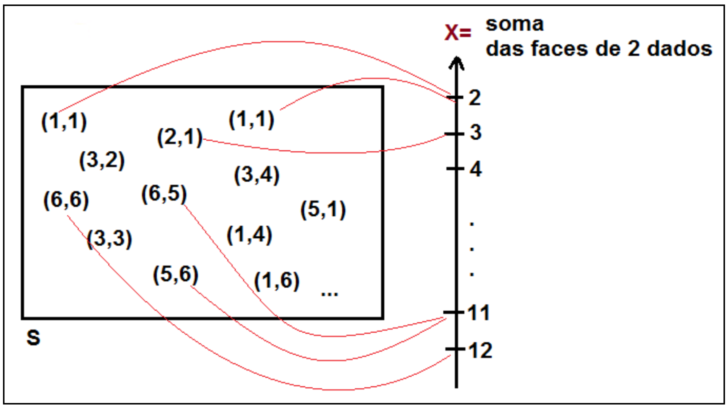
Uma variável aleatória tem um valor único (determinado aleatoriamente) para cada resultado de um experimento. A palavra aleatória indica que em geral só conhecemos aquele valor depois do experimento ser realizado. Como exemplo, quando lançamos uma moeda honesta sabemos a priori que a probabilidade de sair cara é 0,5 e a probabilidade de sair coroa é também 0,5. Mas não sabemos de antemão o resultado que sairá. Podemos chamar X = Variável Aleatória número de CARAS no lançamento de uma moeda. Nesse caso se sair coroa (K) o valor de X = 0 e se sair cara (C) o valor de X =1, conforme ilustrado na Figura 42.

Figura 42- Ilustração da variável aleatória X = número de caras.



Na Figura 43 tem-se também o exemplo do espaço amostral decorrente da soma das faces de 2 dados jogados simultaneamente. Seja a variável aleatória X = soma das faces dos 2 dados. O valor de X varia de 2 até 12.

Figura 43- Ilustração da variável aleatória X = soma das faces de 2 dados.



As variáveis aleatórias podem ser **discretas**, que assumem valores inteiros ou podem ser **contínuas**, que podem assumir infinitos valores dentro de um intervalo de números reais. Como exemplo de variável aleatória discreta tem-se o número de caras que pode ser obtido em 20 lançamentos de uma

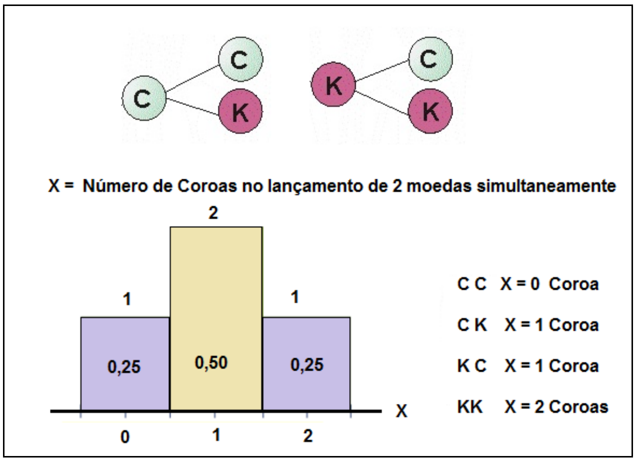
moeda, ou o número de faces pares no lançamento de 10 dados honestos. São exemplos de variáveis aleatórias contínuas as estaturas dos estudantes de uma determinada escola ou a massa corporal dos moradores de uma cidade.
Uma vez definida uma variável aleatória é importante definir Função de Probabilidade da variável aleatória discreta X, que a cada valor de X associa sua probabilidade de ocorrência. A soma de todos os valores de uma distribuição de probabilidades deve ser igual a 1, ou seja, $\sum P(x) = 1$, onde "X" toma todos os valores possíveis. Outra propriedade importante é que a probabilidade de ocorrência de um evento deve ser $0 \leq P(x) \leq 1$ para todo "X". No exemplo do lançamento de um dado honesto, todas as faces têm a mesma probabilidade de ocorrência (1/6). Logo:

$$\sum_{i=1}^{6}\left(\frac{1}{6}+\frac{1}{6}+\frac{1}{6}+\frac{1}{6}+\frac{1}{6}+\frac{1}{6}\right)=1$$

Quando lançamos duas vezes uma moeda honesta podemos ter nenhuma coroa, uma coroa ou duas coroas. Nesse caso trata-se de uma variável aleatória discreta (que assumem valores 0,1,2,3...n).

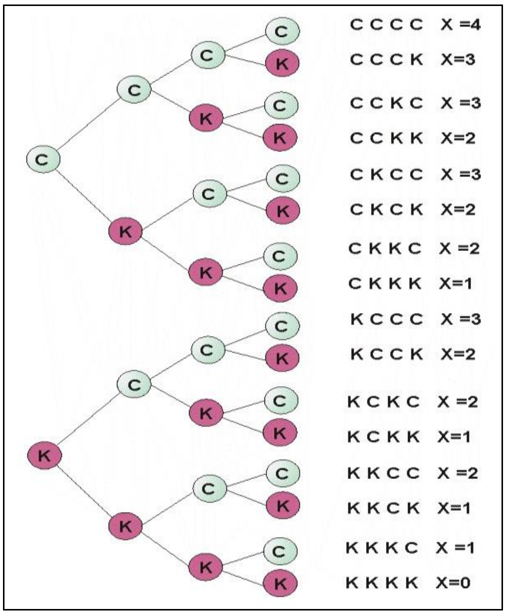
Se chamarmos de X = número de coroas temos então a seguinte distribuição de probabilidades: X = 0 quando não sair nenhuma coroa, X = 1 quando sair apenas 1 coroa e X=2 quando sair duas coroas. Na Figura 44 tem-se a representação da distribuição de probabilidades decorrentes dos dois lançamentos.

Figura 44- Distribuição de probabilidade decorrente de 2 lançamentos de uma moeda.



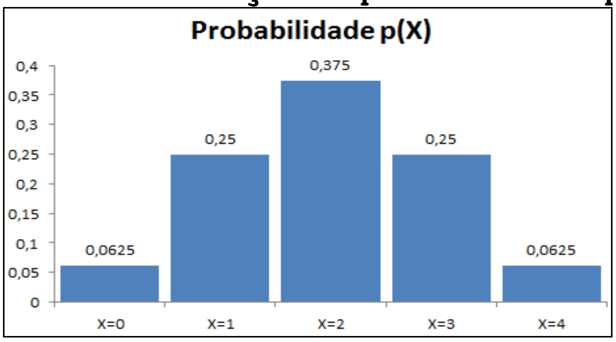
Se a moeda fosse lançada 4 vezes em sequência, a árvore de possibilidades poderia ser representada na Figura 45. Note que X = número de caras. Para o evento CCCC tem-se X=4, ou seja, o evento sair 4 caras em quatro lançamentos. Sua probabilidade de ocorrência é de 1/16 ou 0,0625 (6,25%). O evento X =1 aparece 4 vezes entre as 16 possibilidades. Logo sua probabilidade de ocorrência é 4/16 ou 0,25 (25%).

Figura 45- Distribuição de probabilidade decorrente dos 4 lançamentos de uma moeda.



A distribuição de probabilidades desse exemplo pode ser visualizada na Figura 46.

Figura 46- Distribuição de probabilidades p(x).



Quando um evento é dado em termos de suas probabilidades de ocorrência é possível se calcular uma média, também conhecida como Valor Esperado E(X) e a Variância VAR(X).

Como exemplo, seja uma variável aleatória X que representa em média o número total de dias de sol por semana na cidade de Florianópolis ao longo do ano.

A distribuição de probabilidades de ocorrência de X é dada por p(X):

Xi	p(Xi)
0	0,30
1	0,20
2	0,15
3	0,10
4	0,05
5	0,05
6	0,10
7	0,05

Nesse caso, o Valor Esperado e a Variância são calculados da seguinte forma:

$$E(X) = \sum_{i=1}^{N} (X_i . p(X_i)) \quad \text{e} \quad VAR(X) = E(X^2) - (E(X))^2$$

Xi	p (Xi)	Xi . p(Xi)	X^2	Xi^2.p(Xi)
0	0,30	0	0	0
1	0,20	0,20	1	0,20
2	0,15	0,30	4	0,60
3	0,10	0,30	9	0,90
4	0,05	0,20	16	0,80
5	0,05	0,25	25	1,25
6	0,10	0,60	36	3,60
7	0,05	0,35	49	2,45
Somatório		2,20	Somatório	9,80

Logo, o valor esperado E(X) = 2,20 e a Variância VAR(X) = 9,80 – $(2,20)^2$ = 4,96.

Em Estatística há diversos tipos de funções de distribuição de probabilidades. São exemplos de funções de distribuições de probabilidades discretas a Binomial e Poisson. São exemplos de funções de distribuição de probabilidades contínuas a Exponencial, T de Student, Normal e Qui-Quadrado.

a) Distribuição Binomial

No caso do lançamento da moeda um número elevado de vezes fica difícil calcular as probabilidades por meio do diagrama de árvore. Nesse caso usamos a função distribuição de probabilidades Binomial.

Considerando a variável aleatória X que representa o número de sucessos em N testes independentes, a distribuição denominada **Binomial** será dada por:

$$P(X = x) = \binom{n}{x} p^x.(1 - p)^{n-x}$$

Onde "p" é probabilidade de sucesso do evento em estudo e "q" = (1-p) é a probabilidade de fracasso do evento. Para as distribuições binomiais é possível calcular a média e o desvio padrão como sendo: μ (média) = n.p e desvio padrão:

$$\sigma = \sqrt{n \cdot p \cdot q}$$

Exemplo 1:

Suponha que um pesquisador esteja interessado em avaliar as chances de ocorrência de nenhuma inundação na cidade nos próximos 5 anos. Sabe-se que a probabilidade anual de ocorrência de inundações é de 20% ou 0,2. Seja X = número de inundações nos 5 anos. Esse valor pode ser de 0, 1, 2, 3, 4 e 5. Ou seja, durante os cinco anos observados pode não acontecer nenhuma inundação, mas também podem acontecer 1, 2, 3, 4 ou 5. Para fins de estatística diz-se que a probabilidade de sucesso, ou de ocorrência do evento observado é:

$$P(X = 0) = \binom{5}{0} 0,2^0.(1 - 0,2)^5$$

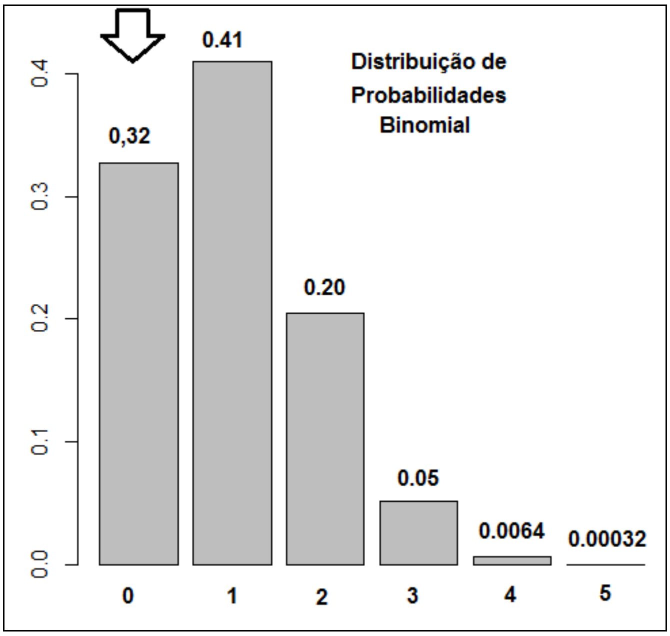
$$P(X = 2) = \frac{5.4.3.2.1!}{(0!)(5.4.3.2.1!)} \times 0,2^0(0,8)^5 = 1 \times 1 \times 0,32 = 0,32$$

Ou seja, há 32% de chances de não ocorrer enchente nos 5 anos observados.

No software R, o cálculo da probabilidade em questão seria obtido apenas com a expressão: dbinom(0,5,0.2)=0,32768. Para a construção do gráfico de distribuição de frequência de probabilidades basta digitar os seguintes comandos no R (Figura 47):

```
x<-0:5
fx<-dbinom(x,5,0.2)
plot(x,fx,type="h")
barplot(fx)
```

Figura 47– Distribuição de frequências de probabilidades binomiais.



Ao analisar a distribuição de frequências de probabilidades ilustrada na Figura 47, percebemos que há 41% de probabilidade de ocorrer 1 enchente nos 5 anos de análise. A probabilidade de ocorrência de 3 enchentes nesse período é de apenas 5%.

Exemplo 2:
Suponha que um determinado gene ocorra em 20% de uma população. Se uma amostra aleatória de 7 pessoas é selecionada ao acaso, qual é a probabilidade de encontrarmos nesse conjunto exatamente 3 pessoas com o gene? Sabemos que a probabilidade de sucesso (presença do gene) = 0,2. Logo p=0,2 e q=0,8. Na equação binomial tem-se:

$$P(X=3)=\binom{7}{3}.0{,}2^3.0{,}8^4=\frac{7.6.5.4!}{(3!).(4!)}\times 0{,}2^3.(0{,}8)^4=0{,}11 \quad \Rightarrow 11\%$$

Exemplo 3:
Considere que o Departamento de Estatística do Trabalho de um município estimou que 20 % da força de trabalho está desempregada. Uma amostra de 14 trabalhadores é obtida deste município. Calcule a probabilidade de 3 pessoas da amostra estarem desempregadas.
Considere a probabilidade de encontrar uma pessoa desempregada como sendo p = 0,2. Considere N=14 e q = 0,8. Substituindo esses valores na equação Binomial temos:

$$P(X=3)=\binom{14}{3}.0{,}2^3.0{,}8^{11}=\frac{14!}{3!\times 11!}\times 0{,}2^3.(0{,}8)^{11}=0{,}25 \quad \Rightarrow 25\%$$

Observamos que na equação para avaliar a probabilidade binomial é utilizada uma expressão comum na análise combinatória. Como exemplo, podemos combinar 4 objetos (C,B,S,T) em grupos de 2 objetos cada de 6 formas distintas: CB, CS,CT,BS,BT e ST.

$$\binom{4}{2} = \frac{4!}{2!(4-2)!} = \frac{4.3.2.1}{2.1.2.1} = \frac{24}{4} = 6$$

A distribuição binomial tem sua média deslocada para a direita quando a probabilidade de sucesso é mais próxima de 1, conforme demonstrado no Exemplo 4, resolvido com apoio do software R.

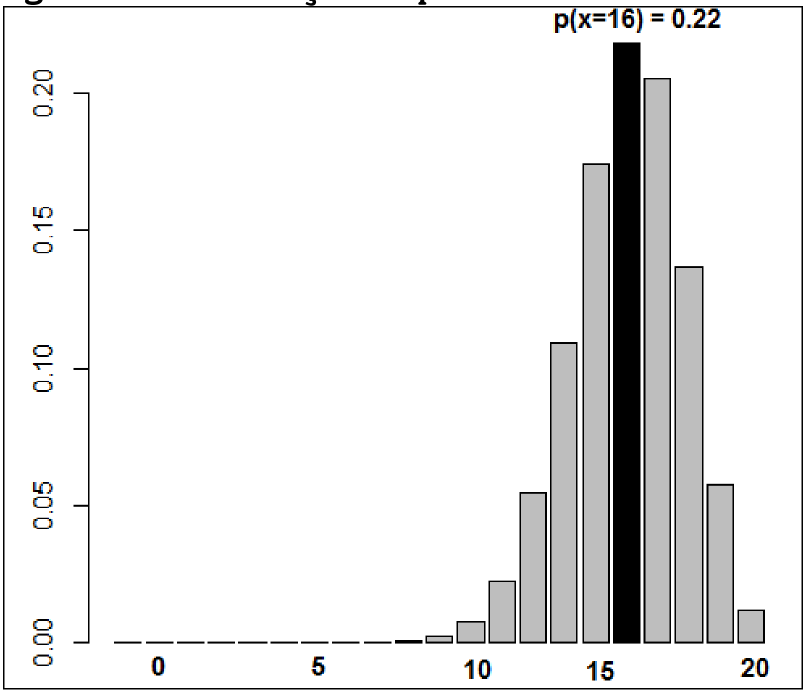
Exemplo 4:

Suponha que a chance de se encontrar uma peça sem defeito em uma linha de produção de uma indústria é de 80% (p=0,8 é a probabilidade de sucesso). Um estagiário selecionou aleatoriamente 20 peças para análise. Qual a probabilidade de se encontrar exatamente 16 peças boas nas 20 peças da amostra?

Ao digitar os comandos a seguir no software R tem-se a Figura 48:

```
x<-0:20
fx<-dbinom(x,20,0.8)
plot(x,fx,type="h")
dbinom(16,20,0.8)
barplot(fx)
```

Figura 48- Distribuição de probabilidades binomial.



Caso o interesse fosse saber qual a probabilidade de encontrarmos mais que 16 peças boas, bastaria somar a probabilidade p(X=17) + p(X=18) + p(X=19) + p(X=20) = 0,205+0,137+ 0,0576 + 0,0115 = 0,41 ou 41%.
Se o interesse fosse conhecer qual a probabilidade de encontrarmos menos que 17 peças boas nas 20 amostras: p(X < 17)=1-[p(X=17)+p(X=18)+p(X=19)+P(X=20)] =0,59 ou 59%.

b) Distribuição de Poisson

Em diversas situações nas quais estamos interessados no número de ocorrências de uma determinada variável em um dado intervalo contínuo (tempo ou espaço) utilizamos a distribuição de probabilidades de Poisson. Como exemplos de aplicação de Poisson temos as seguintes estimativas: número de chamadas telefônicas recebidas por minuto, número de mensagens que chegam a um servidor por segundo, número de acidentes por dia, número de defeitos por m^2 entre tantos outros exemplos.

$$P(x) = \frac{\lambda^x * e^{-\lambda}}{x!}$$

Onde λ é o número médio de ocorrências no intervalo e X é o número de ocorrências que desejamos calcular.

Exemplo 1:
Suponha que em um cruzamento acontecem em média 3 acidentes por mês. Qual é a probabilidade de ocorrência de 5 acidentes em um mês qualquer?

Nesse caso tem-se que a probabilidade é calculada como sendo 10%:

$$P(x) = \frac{\lambda^x * e^{-\lambda}}{x!} = \frac{3^5 * e^{-3}}{5!} = \frac{3^5 * 2{,}718^{-3}}{5.4.3.2.1} = 0{,}10$$

Exemplo 2:

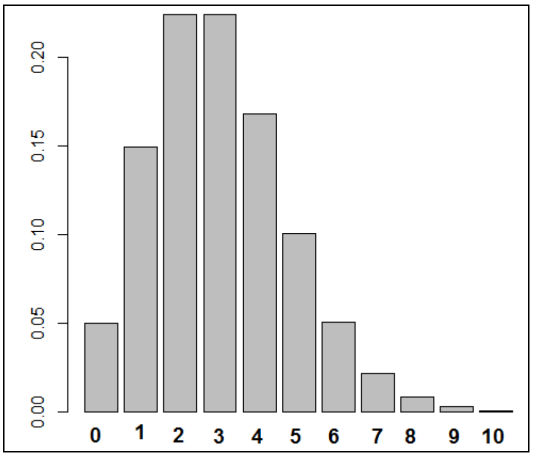
Uma delegacia de polícia recebe uma média de 5 solicitações por hora. Qual a probabilidade de que ela receba 2 solicitações em uma determinada hora selecionada aleatoriamente?

A solução é obtida a partir da consideração de que a média de chamadas λ =5 e o número de sucessos desejados X =2. A equação de Poisson fica:

$$P(X = 2) = \frac{5^2 * 2{,}718^{-5}}{2!} = 0{,}084 \Rightarrow 8{,}4\%$$

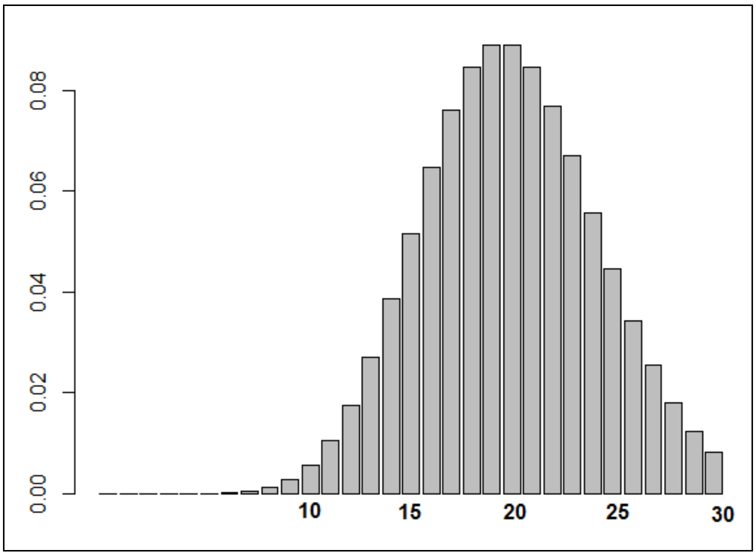
A distribuição de Poisson também pode ser modelada no software R. Se nosso interesse for calcular a probabilidade de ocorrer de 0 (zero) até 10 acidentes no mês em um cruzamento que tem média de 3 acidades digitamos no R:
dpois(0:10,3)
barplot(dpois(0:10,3)).
Como resultado, temos a distribuição de frequências de Poisson indicada na Figura 49:

Figura 49- Distribuição das frequências de probabilidades discretas de Poisson.



Observamos que o formato da distribuição de Poisson varia muito de acordo com o valor de λ. Na Figura 50 tem-se uma distribuição com λ=20. Digite no software R: dpois(0:30,20) e barplot(dpois(0:30,20)).

Figura 50- Distribuição das frequências de probabilidades discretas de Poisson.

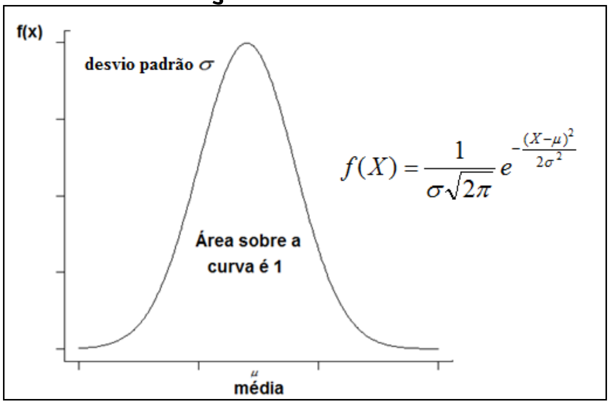


Ao contrário de uma variável aleatória discreta, uma *variável aleatória contínua* pode assumir qualquer valor fracionário dentro de um intervalo definido de valores. Por isso não podemos enumerar todos os possíveis valores da variável com os valores de probabilidade correspondentes. O tempo de vida de um rolamento, as massas das pessoas, a vida útil dos pneus e a estatura das pessoas são exemplos de variáveis aleatórias contínuas.

d) Distribuição Normal

A mais importante distribuição de probabilidade contínua é a NORMAL (também conhecida como curva de Gauss-Laplace). A curva que representa a distribuição normal de probabilidade tem uma forma de sino e é considerado um modelo matemático representativo de inúmeros fenômenos encontrados na natureza (Figura 51).

Figura 51- Ilustração de uma curva NORMAL.

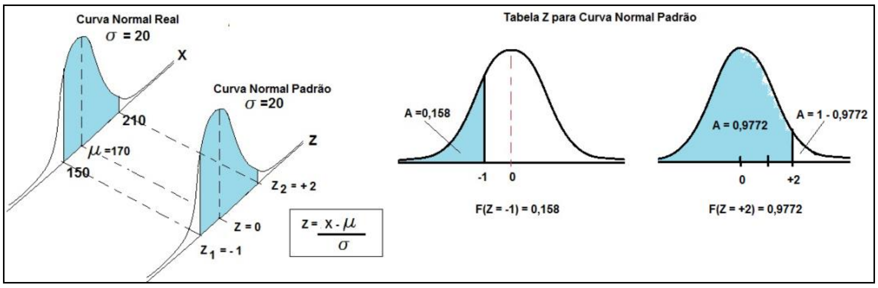


Observamos que os valores da variável aleatória X mais próximos da média ocorrem com maior frequência. Os valores simétricos da variável X em

relação à média ocorrem com mesma frequência e a área sobre a curva tem valor unitário 1. Existe simetria entre os dois lados da curva.
Para facilitar os cálculos há tabelas para distribuição normal padrão, que tem média "ZERO". Para se transformar uma curva normal real em uma curva normal padrão faz-se o procedimento indicado no exemplo 1 (Figura 52).

Exemplo 1:
Suponha que em um dado município a população tenha estatura com média 170cm e desvio padrão de 20cm. A Curva Normal real que representa essa distribuição de estaturas deve ser transformada em uma Curva Normal Padrão Z, com média igual a 0 (zero). As áreas sobre a curva de distribuição normal padrão Z são tabeladas e por isso são utilizadas para a realização dos cálculos da distribuição normal real X.

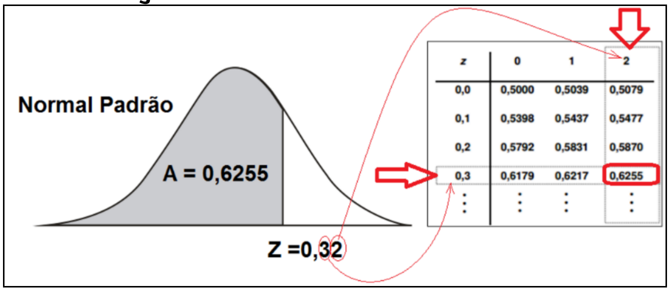
Figura 52- Transformação da Curva Normal Real na Curva Normal padronizada (tabelada).



Para saber a probabilidade de encontrar uma pessoa com estatura menor que 150 nessa população tem-se F(Z=-1)=0,158 ou 15,8% (Tabela Z disponível no Anexo). A probabilidade de encontrar uma pessoa com estatura menor que 210 é calculada a partir de F(Z=2)= 0,9772 ou 97,72%. Se desejarmos saber a probabilidade de se encontrar uma pessoa com estatura entre 150 e 210 basta calcular a diferença entre essas 2 áreas: 0,9772 – 0,158 = 0,819 ou aproximadamente 82%.

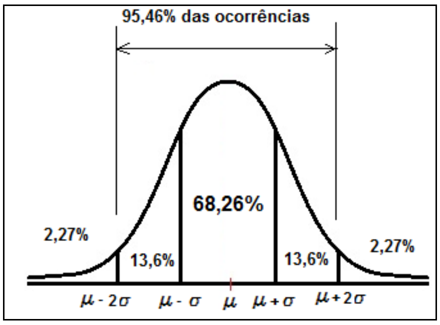
Conforme ilustrado na Figura 53, na Tabela Z padrão, a probabilidade é equivalente à área sombreada sobre a curva que fica à esquerda do valor de Z.
Se, por exemplo, Z = 0,32 então na tabela da curva normal padrão é possível encontrar o valor da probabilidade como sendo 0,6255, que significa que 62,55% dos fenômenos em estudo ocorrem até esse valor de Z = 0,32 (Figura 54).

Figura 53 – Ilustração do uso da Tabela Normal Padronizada.



Existe uma probabilidade de 95,46% de que uma determinada característica esteja presente entre -2 e +2 desvios-padrão ao redor da média. Ou seja, a maioria das frequências se situa ao redor da média entre de -2 desvios-padrões e +2 desvios-padrão. Na Figura 54 tem-se a representação de como as frequências se distribuem em uma curva normal.

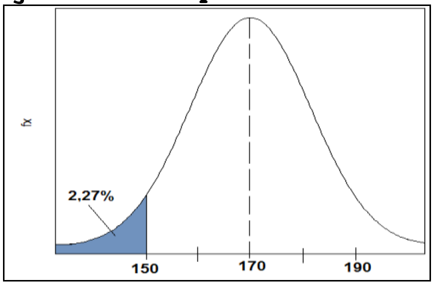
Figura 54- Características da curva normal.



Exemplo 2:
Suponha que a vida útil dos pneus de uma determinada marca se distribua normalmente com média μ= 100 meses e desvio padrão σ=20 meses. Nesse caso, 68,26% dos pneus terão vida útil estimada entre 80 e 120 meses. Apenas 15,87 % deles terão vida útil maior que 120 meses. Por simetria, apenas 15,87% deles terão vida útil inferior a 80 meses.

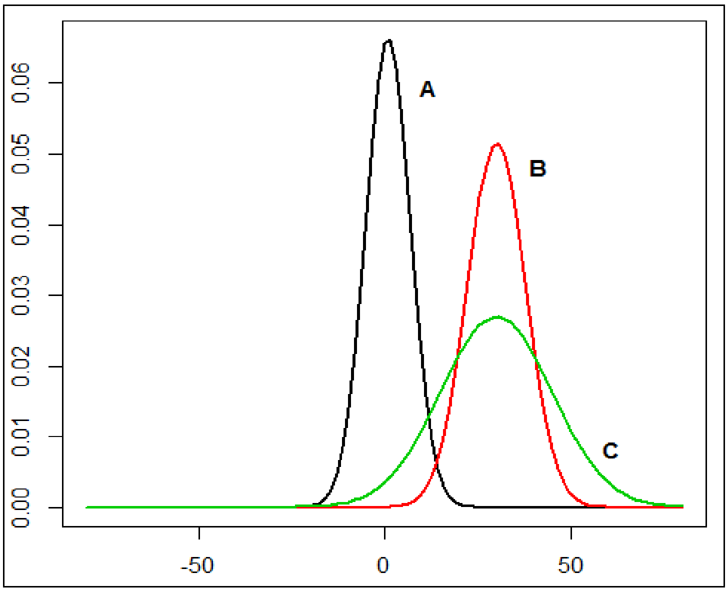
Exemplo 3:
Suponha que a estatura média de uma população é de 1,70m com desvio padrão de 0,10m, pode-se afirmar que aproximadamente 95,44% das pessoas terão estatura entre 1,50m e 1,90m (1,50 +/- 2 desvios-padrão).
A distribuição da estatura da população do exemplo acima poderia ser plotada no software R utilizando-se os comandos: x<-seq(80,250,len=170); fx<-dnorm(x, 170,10) e plot(x,fx,type="l"). Se quisermos conhecer a probabilidade de encontrarmos na população uma pessoa com estatura menor que 1,50m digitamos: pnorm(150, mean = 170, sd = 10). A resposta é 0.02275013 ou 2,27% (Figura 55).

Figura 55– Distribuição normal para estaturas de uma população.



Uma característica importante das curvas normais é que elas são mais ou menos achatadas em relação à média dependendo do desvio padrão. Quanto maior o desvio padrão, mais dispersos os resultados e isso tem influência no formato da curva normal conforme ilustrado na Figura 56. A curva B tem desvio padrão menor que a curva C, mas ambas têm a mesma média.

Figura 56– Características de diversas curvas normais.



Essas 3 curvas normais foram construídas no R a partir dos comandos:

```
curve(dnorm(x,mean=1,sd=sqrt(36)),lwd=2,from=-80,to=80)
curve(dnorm(x,mean=30,sd=sqrt(60)),col=2,lwd=2,add=T)
curve(dnorm(x,mean=30,sd=sqrt(220)),col=3,lwd=2,add=T)
```

Exemplo 4:
Um determinado índice analisado no exame de sangue de uma população é distribuído normalmente com média 200 e desvio padrão 50. Qual é a probabilidade de encontrar na população uma pessoa com índice entre 120 e 230? Calcula-se:

$$Z_1 = \frac{X_1 - \mu}{\sigma} = \frac{120 - 200}{50} = \frac{-80}{50} = -1{,}60$$

logo, A área correspondente a Z_1=-1,60 é 0,0548

a	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
-3,0	0,0013	0,0013	0,0013	0,0012	0,0012	0,0011	0,0011	0,0011	0,0010	0,0010
-2,9	0,0019	0,0018	0,0018	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014
-2,8	0,0026	0,0025	0,0024	0,0023	0,0023	0,0022	0,0021	0,0021	0,0020	0,0019
-2,7	0,0035	0,0034	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027	0,0026
-2,6	0,0047	0,0045	0,0044	0,0043	0,0041	0,0040	0,0039	0,0038	0,0037	0,0036
-1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233
-1,8	0,0359	0,0351	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
-1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
-1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
-1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
-1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0721	0,0708	0,0694	0,0681

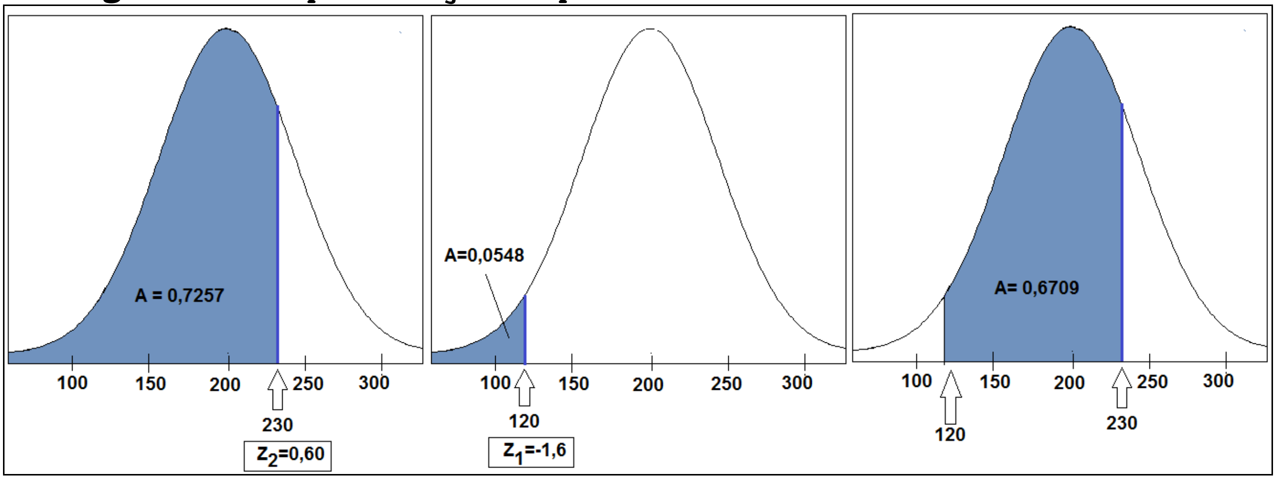
$$Z_2 = \frac{X_2 - \mu}{\sigma} = \frac{230 - 200}{50} = \frac{30}{50} = 0{,}60$$

logo, A área correspondente a Z_2 =1 é 0,7257

a	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133

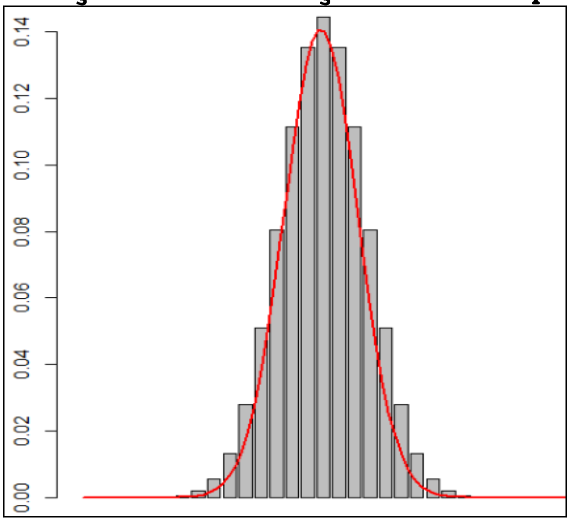
Graficamente podemos visualizar a área de interesse, que representa a probabilidade do evento de interesse ocorrer. O valor de 0,6709 é resultado da área 0,7257 menos a área 0,0548, obtidas da Tabela Normal Padrão. Observamos que quando a segunda área (centro) é subtraída da primeira (esquerda) a resultante é o intervalo mostrado no gráfico da direita (Figura 57).

Figura 57 – Representação da probabilidade de ocorrência de evento.



Quando o número de observações ou tentativas for relativamente grande, a distribuição de probabilidade normal pode ser utilizada para aproximações das probabilidades binomiais, conforme ilustrado na Figura 58.

Figura 58- Aproximação da distribuição binomial pela curva normal.

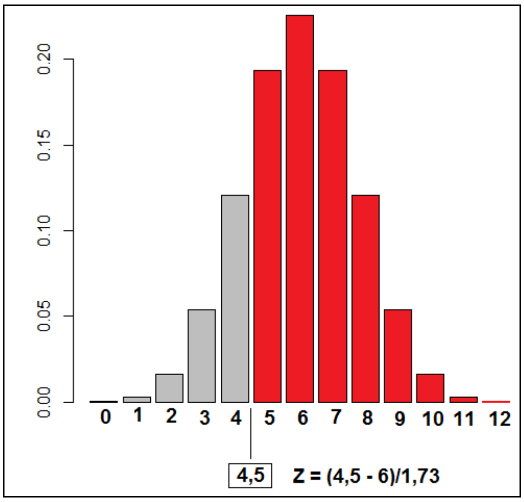


Como é possível observar, quando o número de lançamentos cresce, a curva de distribuição de frequências se aproxima da curva normal, possibilitando que ela seja utilizada nos cálculos como forma de simplificação. Nesse caso utiliza-se a média e o desvio padrão da distribuição binomial para cálculo dos parâmetros já conhecidos da distribuição normal padronizada. A média da distribuição normal é n.p e a variância é n.p.q.

Como exemplo, vamos supor que sejam lançadas 12 moedas simultaneamente. Qual seria a probabilidade de sair mais que 4 caras. Nesse caso, poderia se calcular p(X=5) + p(X=6)+..p(X=12) ou ainda calcular 1 – [p(X=0)+p(x=1)+p(x=2)+p(x=3)+p(x=4)]. Um modo mais fácil é fazer a aproximação com a curva normal.

Na Figura 59, adota-se o valor de X como sendo 4,5 (correção de 0,5). A distribuição binomial tem média igual a n.p = 12.0,5=6 e variância = n.p.q = 12.0,5.0,5 =3. Logo o desvio padrão é aproximadamente 1,73.

Figura 59- Aproximação da distribuição binomial pela curva normal.



Com esses valores é possível calcular um valor de Z correspondente e utilizar a curva normal para encontrar a probabilidade desejada. Com Z = -0,86 tem-se na Tabela Normal Padrão uma probabilidade de 0,194.

$$Z_2 = \frac{4,5 - 6}{1,73} = -0,86$$

Esse valor é a área da curva normal padrão acumulada de - ∞ até 1,73. Mas nosso interesse é exatamente a área do lado direito desse valor. Logo a distribuição para P(X>4 CARAS) é calculada como sendo 1 – 0,194 ~ 0,80 ou aproximadamente 80%.

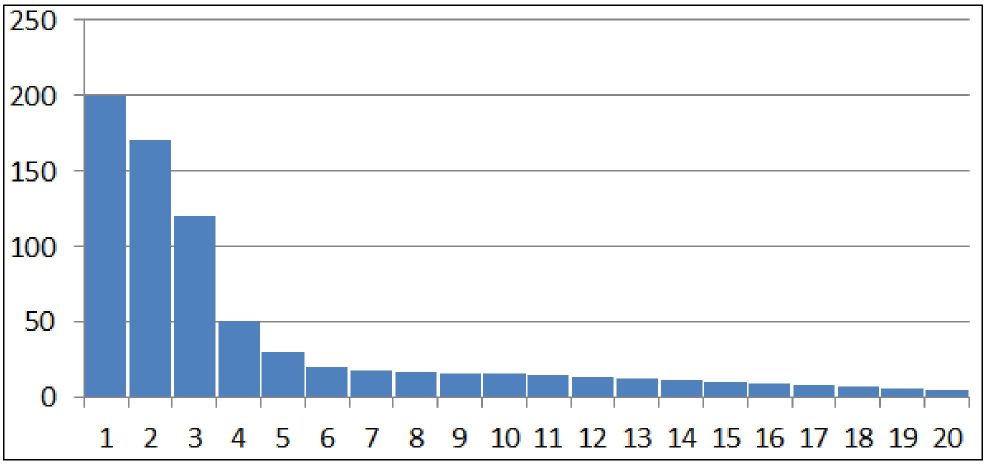
a	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
-3,0	0,0013	0,0013	0,0013	0,0012	0,0012	0,0011	0,0011	0,0011	0,0010	0,0010
-2,9	0,0019	0,0018	0,0018	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014
-2,8	0,0026	0,0025	0,0024	0,0023	0,0023	0,0022	0,0021	0,0021	0,0020	0,0019
-2,7	0,0035	0,0034	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027	0,0026
-1,0	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
-0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
-0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
-0,7	0,2420	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,2177	0,2148
-0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
-0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
-0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121

e) Distribuição de probabilidades exponencial

A distribuição exponencial é muito utilizada para descrever fenômenos como tempo de queima de componentes eletrônicos. Também é um bom modelo matemático para se explicar o motivo da probabilidade de uma pessoa frequentar um parque reduz conforme aumenta a distância dele até sua residência.

Como exemplo prático, vamos supor que um pesquisador tenha coletado as distâncias percorridas todos os dias pelos estudantes para chegarem a uma escola e obtido uma média de 7km. Ao construir o histograma da frequência – Figura 60 - de distribuição das distâncias ele percebeu que uma função de distribuição exponencial seria um modelo matemático adequado para esse caso.

Figura 60 – Histograma de frequência das distâncias percorridas pelos estudantes.



Sabendo-se que a probabilidade de um aluno frequentar a escola cai com a distância e obedece a equação abaixo, calcule qual é a probabilidade de um estudante que resida a mais de 15km frequentar a escola em estudo.

$$P(X > Xo) = e^{-\lambda . x}$$

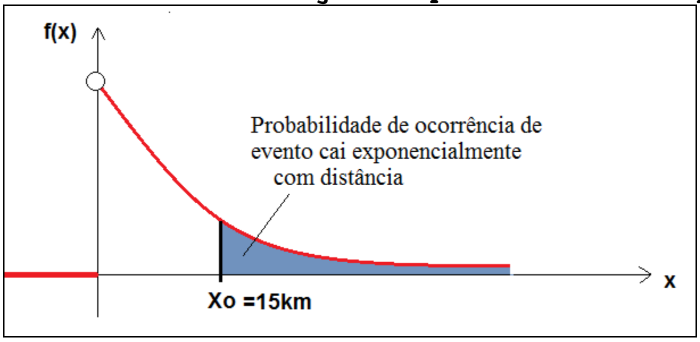
Onde λ = 1/distância média. No exemplo $\lambda = 0{,}1428$

A partir da equação é possível afirmar que a probabilidade de um estudante que reside a mais de 15km frequentar a escola do exemplo é de apenas 11%.

$$P(X > 15) = 2{,}71828^{-0{,}1428*15} = 0{,}11$$

Se no exemplo o objetivo fosse calcular a probabilidade de um estudante que reside a mais de 5 km frequentar a escola, teríamos então como resultado 48%. Isso acontece porque a probabilidade é equivalente à área sobre a curva da função exponencial, conforme mostrado na Figura 61.

Figura 61 – Curva de distribuição de probabilidade exponencial.



Esse tipo de informação é importante para se planejar a localização mais adequada para escolas, hospitais, postos de saúde, supermercados etc.

Como outro exemplo, suponha que um componente eletrônico tenha vida útil média estimada em 1200 horas. Nesse caso, qual seria a probabilidade dele queimar antes de 1000 horas?

Esse é um caso típico de distribuição de probabilidades exponencial, onde λ=1/1200=0,00083.

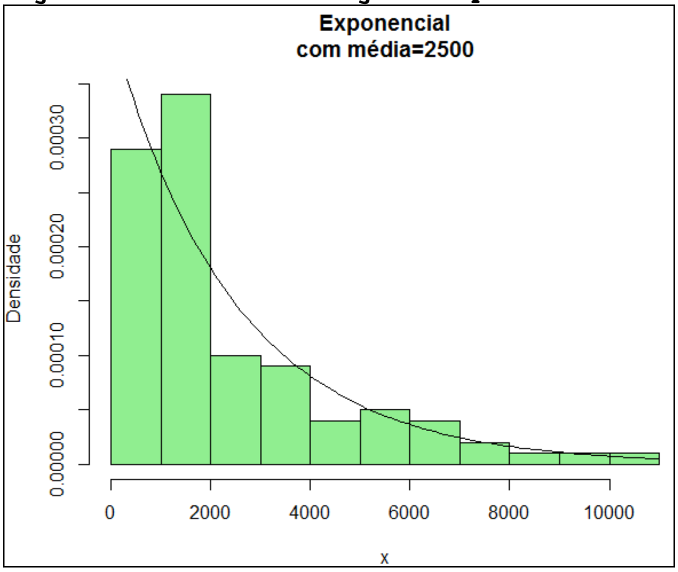
$$P(X > 1000) = 2{,}71828^{-0{,}00083\ *\ 1000} = 2{,}71828^{-0{,}8333} = 0{,}43$$

Logo, a probabilidade do componente queimar antes de 1000 horas é calculada por 1 – 0,43 = 0,57 ou aproximadamente 57%. Esse cálculo é necessário porque desejamos calcular a probabilidade (área) de ocorrer o evento da esquerda e não a da direita.

Essa distribuição de probabilidades pode ser modelada pelo software R. Como exemplo vamos supor um que um equipamento tenha vida média de 2500 horas. Digite os comandos a seguir no R e obtenha a distribuição correspondente na Figura 62.

```
x=rexp(100,1/2500)
hist(x,probability=TRUE,
col="lightgreen",main="Exponencial com média=2500",ylab="Densidade")
curve(dexp(x,1/2500),add=T)
```

Figura 62- Ilustração de uma distribuição de probabilidades exponencial.

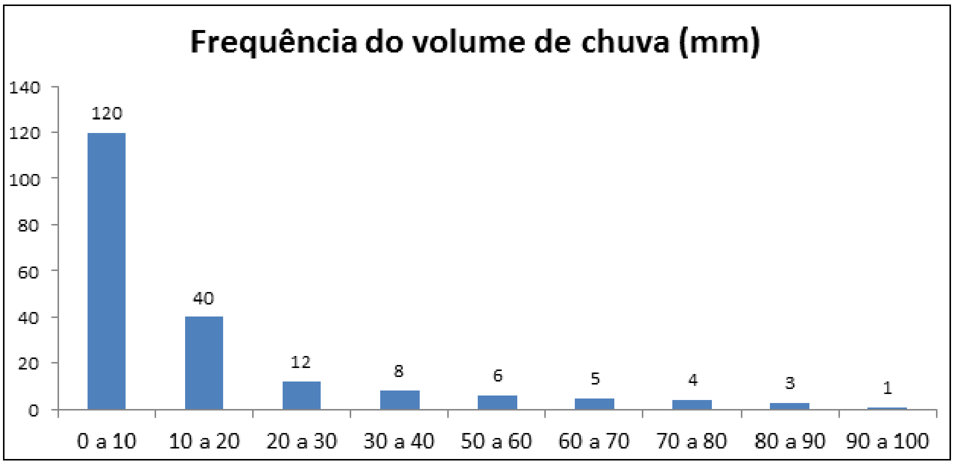


Essa distribuição pode ser relacionada com o modelo proposto por Von Thünen (1783 - 1850) na obra “O Estado Isolado”, onde a dimensão espacial foi aplicada para a solução de problemas de natureza econômica. Von Thünen[16] propôs um modelo no qual as atividades agrícolas dispersas ao redor de um centro urbano, são agrupadas formando cinturões ou anéis, que têm sua localização determinada, principalmente, pela distância da cidade central. As atividades agrícolas que ocupam áreas próximas ao centro urbano possuem altos custos de transporte ou um alto valor de retorno por unidade de área. Já as atividades localizadas em áreas distantes possuem um custo baixo de transporte ou necessitam de uma maior extensão de terra para produzir.

16 http://www.feweb.vu.nl/gis/ModellingLand-UseChange/ExerciseVonThunen.pdf

LISTA DE EXERCÍCIOS 5

1- Um pesquisador anotou a frequência e a quantidade de chuva diária em milímetros em determinada localidade. Para essa situação, que tipo de modelo de distribuição de probabilidades poderia ser utilizado? Como seria possível estimar a quantidade de chuva média?



2- Suponha que a temperatura para o mês de janeiro de uma determinada cidade possa ser modelada por uma distribuição Gaussiana caracterizada por μ=22,2ºC e desvio padrão σ= 4,4ºC. Nesse caso, qual seria a probabilidade de que em um determinado mês de janeiro a temperatura seja menor que 21,4ºC?

3- A probabilidade anual de inundações em uma comunidade é de 0,20. Qual a probabilidade de acontecerem 3 inundações nos próximos 10 anos?

4- Considere que em um cruzamento ocorre um assalto a cada dez dias. Qual é a probabilidade de ocorrência de três assaltos durante o período de 30 dias?

5- Construa uma curva normal com a ajuda do Software Estatístico R para o tempo demandado pelos ônibus para percorrer um determinado trecho. O tempo foi modelado por uma gaussiana de média de 12 minutos e desvio padrão de 3 minutos. Qual a probabilidade de um ônibus demorar mais de 15 minutos? Qual a probabilidade de um ônibus demorar entre 5 e 10 minutos?

RECOMENDAÇÃO:
https://www.youtube.com/watch?v=j3Zbup0KMxY
Distribuição de Probabilidades UNIVESP TV

6- Técnicas de Amostragem

É comum se dizer que não precisa provar um bolo inteiro para se conhecer seu sabor. Basta provar uma amostra. Essa é a ideia por trás das amostras aleatórias utilizadas em análises estatísticas. Uma amostra é uma parte representativa da população, isto é, a amostra deve possuir as mesmas características básicas da população. Se um pesquisador desejar saber qual é a estatura média dos alunos de uma determinada escola de Ensino Médio, basta escolher uma amostra aleatória e representativa desses alunos. Segundo Barbeta (2011), para se calcular o tamanho mínimo de componentes de uma amostra pode ser utilizada a equação:

$$n = \frac{N * n_o}{N + n_o}$$

Onde "N" é tamanho da população; "n" o tamanho da amostra e "no" é uma primeira aproximação para o tamanho da amostra calculado por 1/Erro amostral ao quadrado.

Exemplo: Em uma empresa com 10.000 funcionários, desejamos estimar o percentual de pessoas que são favoráveis a um determinado treinamento. Qual deve ser o tamanho da amostra para que o erro da pesquisa seja menor que 4%?

$$no = \frac{1}{0{,}04^2} = 625$$

logo, calculamos

$$n = \frac{10000 * 625}{10000 + 625} = 599$$

Quando o número total da população é desconhecido pode-se calcular o tamanho mínimo da amostra para ser representativa a partir da seguinte equação simplificada:

$$n = \frac{1{,}96^2 . 0{,}5^2}{Erro^2}$$

Como exemplo, suponha que seja necessário calcular a quantidade de eleitores que devem ser consultados em uma pesquisa. Considerando uma margem de erro de 5% temos: 384 pessoas. Para uma margem de erro de 2% tem-se necessidade de se consultar 2401 eleitores. Por isso, nas pesquisas eleitorais para presidente são entrevistadas, em geral, 2500 eleitores para se obter resultados com margem de erro de 2% e Nível de Confiança de 95%.

As amostras podem ser dos seguintes tipos: aleatória simples (sistemática, estratificada, estratificada proporcional, agrupamento) e não aleatórias.

a) Amostragem Casual ou Aleatória Simples – é equivalente a um sorteio aleatório. Nesse tipo de amostragem é necessário que os elementos da população sejam numerados e sorteados a partir de um programa ou de uma tabela de números aleatórios.

b) Amostragem Sistemática – em uma linha de produção, a cada dez itens produzidos podemos retirar um item para fazer parte de uma amostra da produção diária. Neste caso, estaríamos fixando o tamanho da amostra em 10% da população. Como exemplo, suponha que uma fábrica possui em estoque 450 computadores ordenados. O setor de controle de qualidade da fábrica deseja obter uma amostra formada por 25 unidades. Pode-se, neste caso, usar o seguinte procedimento: como 450/25 = 18, escolhe-se por sorteio casual um número de 1 a 18 (inclusive), o qual indica o primeiro elemento sorteado para a amostra; os demais serão periodicamente considerados de 18 em 18. Assim, se o número sorteado for o 4, toma-se, o 40 computador, o 220, o 400 etc., até completar a amostra. A amostragem sistemática necessita que os elementos da população a ser estudada já se encontrem ordenados. São exemplos prédios de uma rua, produtos dentro de uma linha de produção, prontuários médicos, os alunos inscritos em uma faculdade, etc. Para a seleção dos elementos que farão parte da amostra, será elaborado um sistema pelo pesquisador. Exemplo: Em uma rua há 900 casas. Desejamos escolher uma amostra de 50 delas para entrevistar os moradores. Divide-se 900 por 50 e obtém-se 18. Sorteamos a primeira casa e depois contamos 18 casas para obtermos a próxima até que todas as 50 sejam selecionadas. A escolha da primeira casa pode ser realizada a partir de uma tabela de números aleatórios.

c) Amostragem por Agrupamento – Quando a população apresenta ocorrência natural de subgrupos, cada um deles com características similares. Dividida a população em grupos, chamados de agrupamentos e selecione todos os membros de um ou mais agrupamentos (mas não todos). Exemplo – População de domicílios de uma cidade, os quarteirões formam os agrupamentos de domicílios.

d) Amostragem Estratificada Proporcional – na maioria das vezes a população se divide em estratos. Exemplo: uma turma de engenharia tem 66 alunos, onde 57 são meninos e 9 são meninas. Tem-se dois estratos nesta população (sexo masculino e feminino), logo para uma amostra de 10% da população tem-se 1 menina e 6 meninos. Para determinação da intenção de votos dos eleitores brasileiros é comum os institutos de pesquisas utilizarem a amostragem estratificada com sorteio aleatório dos entrevistados. Para chegar a eles, o conjunto da população adulta do país é dividida em cinco sub-universos, que representam as regiões Sul, Sudeste, Nordeste, Norte e Centro-Oeste. Em cada sub-universo os municípios são agrupados segundo a localização geográfica e nível socioeconômico. Em cada grupo são sorteados os municípios. Por sorteios sucessivos, chega-se ao bairro, à rua e ao indivíduo.

A pesquisa por amostragem para avaliar as intenções de voto para presidente foi utilizada pela primeira vez em 1932. A revista *Literary Digest* fez uma pesquisa sobre as intenções de voto dos seus leitores. Das 20 milhões de cédulas enviadas junto com a revista, 3 milhões foram devolvidas para a redação, apontando como virtual vencedor o candidato Franklin Roosevelt. Na eleição seguinte os resultados dessa pesquisa falharam enquanto o prof. George Gallup previu o resultado correto da eleição utilizando uma amostra de apenas 3 mil eleitores. Daí a preocupação com os estratos que compõem a população. Apesar de a amostra ter sido pequena, foi mais adequada que a amostra de 10 milhões de leitores da revista.

Para conhecer os estratos que existem na população brasileira os Institutos de pesquisa utilizam dados disponíveis no Tribunal Regional Eleitoral e no IBGE.

Na Tabela 5 tem-se a distribuição dos eleitores de acordo com o nível de instrução. Do total de 143,7 milhões de eleitores em maio de 2016, apenas 8,2 milhões possuem Ensino Superior completo. Um contingente de 67 milhões de eleitores não tem o Ensino Fundamental completo.

Tabela 5- Distribuição de eleitores brasileiros de acordo com a escolaridade.

Grau de Instrução	Masculino(M)	%M/T	Feminino(F)	%F/T	Não Informado(N)	%N/T	Total(T)	%T/TT
ANALFABETO	3.460.724	47,168	3.867.797	52,716	8.468	0,115	7.336.989	5,104
ENSINO FUNDAMENTAL COMPLETO	5.095.771	49,338	5.226.204	50,601	6.347	0,061	10.328.322	7,185
ENSINO FUNDAMENTAL INCOMPLETO	21.978.092	50,953	21.138.672	49,007	17.564	0,041	43.134.328	30,008
ENSINO MÉDIO COMPLETO	10.357.869	42,411	14.058.660	57,565	5.915	0,024	24.422.444	16,990
ENSINO MÉDIO INCOMPLETO	13.401.454	48,247	14.372.159	51,742	2.930	0,011	27.776.543	19,324
LÊ E ESCREVE	8.619.481	50,472	8.393.460	49,148	64.859	0,380	17.077.800	11,881
NÃO INFORMADO	51.727	45,910	57.271	50,831	3.672	3,259	112.670	0,078
SUPERIOR COMPLETO	3.321.153	40,469	4.883.908	59,511	1.688	0,021	8.206.749	5,709
SUPERIOR INCOMPLETO	2.403.401	44,937	2.944.088	55,046	949	0,018	5.348.438	3,721
TOTAL(TT)	68.689.672	47,786	74.942.219	52,136	112.392	0,078	143.744.283	100,000

Quanto à faixa etária temos que 24 milhões de eleitores têm mais de 60 anos de idade. Esses dados podem ser representados por meio de um histograma, conforme já vimos anteriormente.

Tabela 6- Distribuição dos eleitores brasileiros de acordo com a faixa etária.

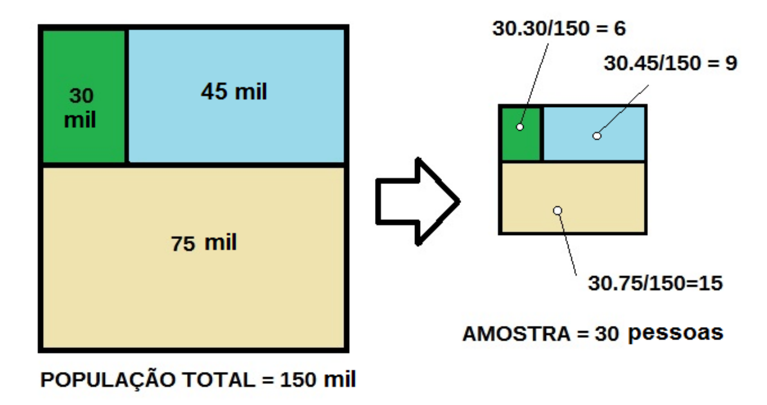
Faixa Etária	Masculino(M)	%M/T	Feminino(F)	%F/T	Não Informado(N)	%N/T	Total(T)	%/TT
Inválida	65.776	50,660	64.069	49,340	1	0,000	129.846	0,090
16 anos	449.223	49,900	450.975	50,100	0	0,000	900.198	0,620
17 anos	762.066	50,050	760.692	49,950	0	0,000	1.522.758	1,050
18 a 20 anos	4.214.246	49,170	4.355.966	50,830	0	0,000	8.570.212	5,910
21 a 24 anos	6.201.930	49,230	6.396.680	50,770	0	0,000	12.598.610	8,680
25 a 34 anos	15.676.600	48,440	16.684.165	51,560	0	0,000	32.360.765	22,300
35 a 44 anos	13.886.062	47,870	15.119.668	52,120	1.076	0,000	29.006.806	19,990
45 a 59 anos	16.316.862	47,240	18.182.095	52,640	39.636	0,110	34.538.593	23,810
60 a 69 anos	6.608.198	46,180	7.676.655	53,650	24.703	0,170	14.309.556	9,860
70 a 79 anos	3.219.048	44,620	3.979.114	55,150	16.554	0,230	7.214.716	4,970
Superior a 79 anos	1.727.349	43,870	2.194.940	55,750	14.756	0,370	3.937.045	2,710
TOTAL(TT)	69.127.360	47,640	75.865.019	52,290	96.726	0,070	145.089.105	100,000

A maior parte dos eleitores brasileiros vive na região Sudeste, que reúne 85 milhões dos habitantes do país. A região Sul tem população de 29 milhões. A região Nordeste 56 milhões, a Norte 17 milhões e a Centro-Oeste 15 milhões. Por esse motivo, a proporção de brasileiros entrevistados em cada região deve ser proporcional ao todo.

Os dados estatísticos mostram que a maior parte do eleitorado brasileiro é formada por mulheres (52 % do total). Um total aproximado de 300 mil eleitores votam no exterior.

Exemplo 1:
Em uma localidade com 150 mil habitantes (Figura 63), 45 mil têm menos de 20 anos de idade, 75 mil têm idades entre 30 e 50 anos e 30 mil têm mais de 50 anos de idade. Uma amostra de 30 habitantes desta população deve ser estabelecida com que proporções de idades?

Figura 63- Ilustração dos estratos de faixa etária (em milhares).



Logo as amostras são calculadas como: Amostra A = 30. 30/150 = 6 com mais de 50 anos de idade; Amostra B = 30. 45/150 = 9 com menos de 20 anos de idade e Amostra C = 30. 75/150 = 15 entre 30 e 50 anos de idade.

Exemplo2:
Uma das classificações úteis para questões de Marketing é em classes sociais. Analisando os diferentes critérios propostos para classificação empregados atualmente no Brasil, podemos generalizar as seguintes categorias[17]: Classe A: inclui as famílias com renda mensal igual ou maior que R$ 14.400,00. Classe B: inclui as famílias com renda mensal entre R$ 7.100,00 e R$ 14.399,00. Classe C: inclui as famílias com renda mensal entre R$ 2.600,00 e R$ 7.099,00. Classe D: inclui as famílias com renda mensal igual ou menor que R$ 2.599,00. Suponha que uma determinada população em estudo distribui-se nesses estratos, de acordo com as quantidades a seguir: Classe A: 60, Classe B: 90, Classe C: 120, Classe D: 480. Se nossa amostra é de 100 unidades adotamos o seguinte procedimento: a) soma dos estratos da população: 60 + 90 + 120 + 480 = 750 indivíduos. Como nossa amostra terá 100 indivíduos, 100/750 = 0,13. O fator 0,13 será multiplicado pelas quantidades de elementos de cada classe. Classe A: 60 x 0,13 = 8 unidades amostrais; Classe B: 90 x 0,13 = 12 unidades amostrais; Classe C: 120 x 0,13 = 16 unidades amostrais; Classe D: 480 x 0,13 = 64 unidades amostrais.

[17] Valores sujeitos à alteração anual.

LISTA DE EXERCÍCIOS 6

1- Considerando-se que a população brasileira pode ser estratificada por região, nível de escolaridade e por idade, quais seriam os estratos que você adotaria para uma pesquisa para presidente se a amostra para a pesquisa fosse de 2.400 pessoas?

2- Em uma empresa com 10.000 funcionários, desejamos estimar o percentual de pessoas que são favoráveis a um determinado treinamento. Qual deve ser o tamanho da amostra para que o erro da pesquisa seja menor que 4%?

3- Quantas pessoas devem ser entrevistadas para conhecermos a opinião dos 2.000 alunos de uma escola sobre a qualidade dos serviços da lanchonete?

7- Inferência Estatística

Uma das definições mais importantes na área de estatística é o Teorema Central do Limite[18]. Ele permite que se faça inferência a uma população a partir de amostras selecionadas aleatoriamente. Pelo Teorema, não importa qual é o formato da distribuição original de X, a distribuição de sua média se aproxima da distribuição normal a medida que o número de elementos da amostras cresce. Se $\overline{X}$ é a média de uma amostra aleatória de tamanho n, obtida de uma população com média μ e desvio padrão σ então:

$$Z = \frac{\overline{X} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)}$$

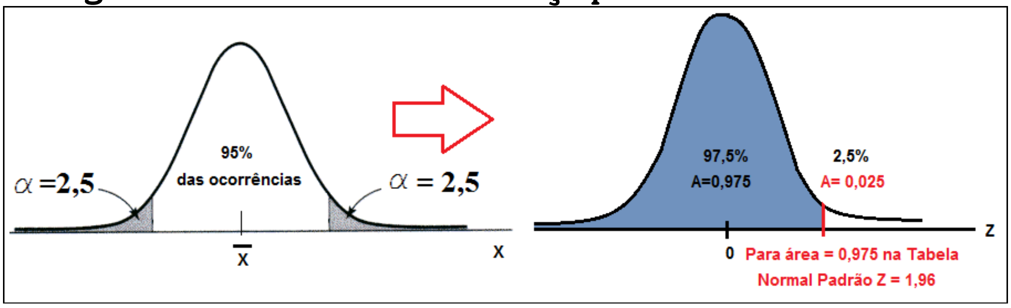
É uma Variável Aleatória cuja distribuição mais se aproxima da distribuição normal padronizada à medida que "n" tende ao infinito.

Dada uma população com desvio padrão "σ", a forma geral do INTERVALO DE CONFIANÇA para o valor médio da população "μ" (com nível de confiança estipulado) será:

$$\left[\overline{X} - \left(Z\frac{\sigma}{\sqrt{n}}\right)\right] < \mu < \left[\overline{X} + \left(Z\frac{\sigma}{\sqrt{n}}\right)\right]$$

O valor de Z depende do nível de confiança (NC) desejado. Para NC = 95% tem-se Z = 1,96 e para NC = 90% tem-se Z = 1,64.
Na Figura 64 é possível visualizar que 95% das ocorrências estão localizadas dentro do intervalo de confiança. Observe que o nível de confiança NC = 1-α (alfa). ALFA é o nível de significância. O valor de 1,96 é obtido na Tabela Normal Padronizada para área acumulada do lado esquerdo da curva normal igual a 0,975. Do lado direito tem-se uma área residual de 0,025. A soma total é igual a 1. Para encontrar o valor de -1,96 basta procurar na Tabela Normal Padronizada o valor de Z para a área de 0,025 acumulada do lado esquerdo.

Figura 64- Intervalo de confiança para médias amostrais.



[18] http://www.portalaction.com.br/probabilidades/732-teorema-central-do-limite

Como exemplo, suponha que uma população tenha estatura média desconhecida, mas desvio padrão conhecido e igual a 20 cm. Uma amostra de 25 pessoas tem suas estaturas medidas. A estatura média da amostra de 25 pessoas é calculada como sendo 170cm. Considerando-se que a estatura pode ser modelada pela distribuição normal e aplicando a expressão apresentada anteriormente tem-se que o intervalo de confiança da estatura média da população é

$$\left[170-\left(1{,}96\frac{20}{\sqrt{25}}\right)\right]<\mu<\left[170+\left(1{,}96\frac{20}{\sqrt{25}}\right)\right]$$

$$170-(7{,}84)<\mu<170+(7{,}84),\ \text{ou seja:}\quad 162{,}16<\mu<177{,}84$$

Esse intervalo de confiança tem um nível de confiança de 95% Ou seja, a partir de uma amostra de tamanho 25 é possível estimar o valor da média da estatura de toda a população com uma margem de erro de 7,84cm. Para reduzir essa margem de erro é preciso ampliar a amostra. Com uma amostra de tamanho 100, tem-se a margem de erro reduzida para 3,92cm. Isso acontece porque o tamanho da amostra aparece no denominador da expressão para cálculo do Intervalo de Confiança. O nível de confiança de 95% quer dizer que o resultado tem confiabilidade de 95%, ou seja, se 100 amostras fossem selecionadas, em 95 delas o resultado estaria dentro do intervalo de confiança calculado.

Vejam o exemplo: O tempo de deslocamento de todos os estudantes até uma determinada universidade pode ser modelado por uma gaussiana (distribuição normal) com desvio padrão de 8 minutos. Uma amostra de 20 estudantes foi entrevistada. O tempo médio para deslocamento desse grupo foi estimado em 80 minutos. Calcule o intervalo de confiança para a média de tempo de toda população de estudantes da universidade. Use o nível de confiança de 95% (Z = 1,96). Nesse caso basta substituir os dados na equação:

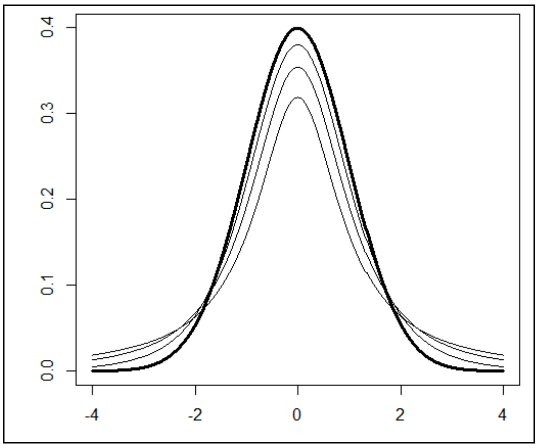
$$\left[\bar{X}-\left(Z\frac{\sigma}{\sqrt{n}}\right)\right]<\mu<\left[\bar{X}+\left(Z\frac{\sigma}{\sqrt{n}}\right)\right]\quad\Rightarrow\quad\left[80-\left(1{,}96\frac{8}{\sqrt{20}}\right)\right]<\mu<\left[80+\left(1{,}96\frac{8}{\sqrt{20}}\right)\right]$$

Obtemos que o intervalo de confiança para o tempo médio μ de deslocamento dos estudantes é de 76,5 minutos a 83,5 minutos com NC = 95%.

$$76{,}5<\mu<83{,}5$$

Quando não conhecemos o desvio padrão da população devemos calcular o desvio padrão da amostra e utilizar a Tabela T de Student[19] para obter o valor de "T" no lugar da variável "Z". Para obtenção de "T" tabelado usamos o nível de confiança desejado e o grau de liberdade GL = (n – 1). A distribuição T de Student tende para a curva normal quando o tamanho da amostra cresce conforme ilustrado na Figura 65.

Figura 65- Ilustração da relação entre a distribuição Normal e T de Student.



Vejam o exemplo: Um professor escolheu uma amostra de 12 alunos e perguntou qual era a distância percorrida para chegar até a escola (em quilômetros). Considere que as distâncias percorridas se apresentam distribuídas normalmente. Os valores foram listados abaixo. Calcule o intervalo de confiança para a média da distância percorrida pelos estudantes da turma com nível de confiança de 95%. As distâncias percorridas em km foram: 8,2 8,3 8,4 8,2 8,2 8,4 8,3 8,2 8,4 8,4 8,2 8,4.
Nesse caso, a distância média é calculada como sendo $\overline{X}$ =8,3km. Já o desvio padrão foi calculado como sendo s = 0,095. Para NC = 95% e GL = (n-1) = 11 tem-se T tabelado = 2,201 (T Student).
Logo o intervalo de confiança da média de distâncias percorridas pela população de estudantes da escola é calculado como segue:

$$\left[\overline{X} - \left(T\frac{s}{\sqrt{n}}\right)\right] < \mu < \left[\overline{X} + \left(T\frac{s}{\sqrt{n}}\right)\right]$$

$$\left[8{,}3 - \left(2{,}201\frac{0{,}095}{\sqrt{12}}\right)\right] < \mu < \left[8{,}3 + \left(2{,}201\frac{0{,}095}{\sqrt{12}}\right)\right]$$

[19] Student foi um pseudônimo utilizado por Willian Gosset para publicação de seus trabalhos

O intervalo de confiança para a média das distâncias percorridas é de $8{,}24\,km < \mu < 8{,}36\,km$ com nível de confiança de 95%. Na Tabela T de Student é preciso identificar o G.L = grau de liberdade e o nível de confiança. À medida que o grau de liberdade aumenta o valor de T tende ao mesmo valor de Z (distribuição normal).

Tabela 5 – Distribuição de Probabilidades T de Student – VER TABELA ANEXA.

g.l.	Nível de confiança, c	0,50	0,80	0,90	0,95	0,98	0,99
	Unicaudal, α	0,25	0,10	0,05	0,025	0,01	0,005
	Bicaudal, α	0,50	0,20	0,10	0,05	0,02	0,01
1		1,000	3,078	6,314	12,706	31,821	63,657
2		0,816	1,886	2,920	4,303	6,965	9,925
3		0,765	1,638	2,353	3,182	4,541	5,841
4		0,741	1,533	2,132	2,776	3,747	4,604
5		0,727	1,476	2,015	2,571	3,365	4,032
6		0,718	1,440	1,943	2,447	3,143	3,707
7		0,711	1,415	1,895	2,365	2,998	3,499
8		0,706	1,397	1,860	2,306	2,896	3,355
9		0,703	1,383	1,833	2,262	2,821	3,250
10		0,700	1,372	1,812	2,228	2,764	3,169
11		0.697	1.363	1.796	2.201	2.718	3.106
12		0,695	1,356	1,782	2,179	2,681	3,055
13		0,694	1,350	1,771	2,160	2,650	3,012

LISTA DE EXERCÍCIOS 7

1- Um pesquisador observou que o tempo médio de deslocamento dos trabalhadores de uma determinada empresa pode ser modelado por uma distribuição normal. Para realização de uma estimativa do tempo médio de deslocamento da população ele selecionou aleatoriamente 10 profissionais para entrevista. Os tempos gastos foram anotados em minutos. Nesse caso, qual seria o intervalo de confiança para o tempo médio de deslocamento da população de trabalhadores? Utilize nível de confiança de 95%.

Tempos anotados em minutos: 16 23 17 19 14 17 18 16 17 18

2- A estatura de uma amostra de estudantes foi anotada. Com nível de confiança de 95%, qual é o intervalo de confiança para a média da estatura de todos os estudantes da escola?

137 154 159 155 167 159 158 159 152 169
154 158 140 149 145 157 160 155 155 143
157 139 159 139 129 162 151 150 134 151

3- Um pesquisador observou que o tempo médio de admissão dos trabalhadores de uma determinada empresa pode ser modelado por uma distribuição normal. Para realização de uma estimativa do tempo médio de admissão de todos os trabalhadores da empresa ele selecionou aleatoriamente 12 profissionais para entrevista. Os tempos foram anotados em anos. Nesse caso, qual seria o intervalo de confiança para o tempo de admissão de todos os trabalhadores da empresa? Utilize nível de confiança de 90%.

Tempos anotados em anos: 16 23 17 19 14 17 18 16 17 18 12 19

4- Uma empresa empacotadora de café precisa garantir que seus pacotes de café estejam dentro dos limites fixados pela inspeção federal. Uma amostra de 9 pacotes foram avaliados. Sabe-se que desvio padrão da máquina é de 12g. As massas são indicadas abaixo:

983 992 1011 976 997 1000 1004 983 998

a) Nesse caso, qual será o intervalo de confiança das massas da máquina para níveis de confiança de 90, 95 e 99%?

b) Qual o tamanho da amostra para que a amplitude do intervalo de confiança seja de 2g com nível de confiança de 95%?

c) Se o desvio padrão da máquina fosse desconhecido, qual seria o Intervalo de confiança considerado?

8- Testes de Hipóteses

Os testes de hipóteses foram criados no início do século XX pelo geneticista e estatístico Sir Ronald Aylmer Fisher (1890 – 1962) e se tornaram a referência quando o objetivo é avaliar, por exemplo, se um determinado procedimento médico alternativo produz realmente resultados melhores.

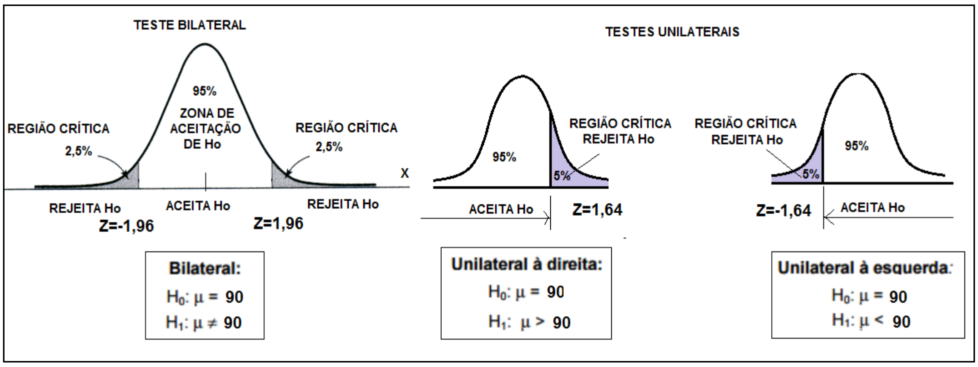
Como exemplo prático, vamos supor que uma determinada região do país é conhecida por ter uma população obesa. A distribuição de probabilidade do peso dos homens dessa região entre 20 e 30 anos é normal com média de 90 kg e desvio padrão de 10 kg. Um endocrinologista propõe um tratamento para combater a obesidade que consiste de exercícios físicos, dietas e ingestão de um medicamento. Ele afirma que com seu tratamento o peso médio da população da faixa em estudo diminuirá em um período de três meses. Para avaliar se o tratamento deu certo é possível formular duas hipóteses: Ho, chamada de **Hipótese Nula** e que diz que a média dos pesos dos homens em estudo após o tratamento não mudou nada e ficou em 90kg e H1, chamada de **Hipótese Alternativa**, que diz que a média dos pesos é diferente que 90kg. Também é possível a análise da Hipótese Alternativa como menor que 90kg. O objetivo do Teste de Hipóteses é mostrar se a Hipótese Alternativa H1 é aceitável ou não. Mas esse tipo de análise também é suscetível a dois tipos de erros: Erro tipo 1, quando rejeitamos Ho quando de fato Ho é verdadeira e Erro tipo II quando não rejeitamos Ho, quando de fato Ho é falsa (Figura 66).

Figura 66- Tipos de erros no Teste de Hipóteses.

Decisão tomada	Situação real	
	Ho é verdadeira	Ho é falsa
Não rejeitar Ho	Decisão correta	Erro tipo II
Rejeitar Ho	Erro tipo I	Decisão correta

Isso ocorre porque toda análise envolve um nível de confiança e uma região crítica onde os resultados não podem ser avaliados com precisão. A Hipótese nula não pode ser rejeitada se o valor do Z ou T calculado estiver fora da região crítica. Na Figura 67 tem-se a zona de aceitação de Ho para testes bilaterais e testes unilaterais.

Figura 67 – Tipos de testes de Hipóteses.



Devemos calcular o Z de teste (ou Z calculado) ou T de teste (quando o desvio padrão não é conhecido) e comparar esse valor com os resultados obtidos a partir da Tabela Z ou T de Student para determinado nível de confiança (Figura 68).

Figura 68– Procedimento para realização de Testes de Hipóteses

Quando H1 for:	Rejeitar Ho se:
$\mu < \mu_0$	$Z_{calc} < -Z_{\alpha}$
$\mu > \mu_0$	$Z_{calc} > Z_{\alpha}$
$\mu \neq \mu_0$	$Z_{calc} < -Z_{\alpha/2}$ $Z_{calc} > Z_{\alpha/2}$

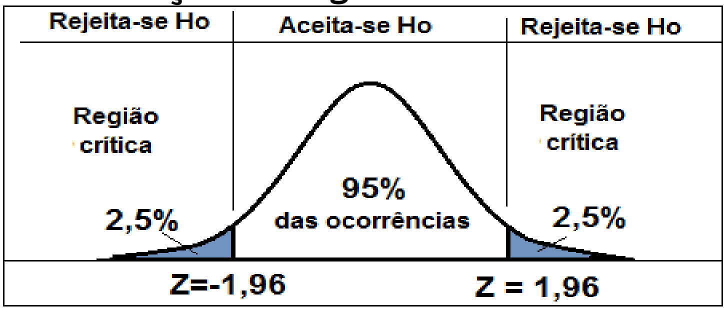
Desvio padrão conhecido	Desvio padrão desconhecido
$Z_{calc} = \dfrac{\overline{X} - \mu_0}{\left(\dfrac{\sigma}{n}\right)}$	$T_{calc} = \dfrac{\overline{X} - \mu_0}{\left(\dfrac{s}{n}\right)}$

Se nível de confiança for de 95% e o teste for bilateral, então Z tabelado é de 1,96. Se o nível de confiança for de 90% e o teste for bilateral, então Z tabelado para comparar com Z calculado será de 1,64. O cálculo da estatística de teste T de Student é utilizado quando não se conhece o desvio padrão de uma determinada população.

Exemplo 1:

Uma pesquisa divulgou que o volume de chuvas em uma região para o mês de junho é de 330mm com um desvio padrão típico de 10mm. Uma amostra com 35 dias da série histórica foi analisada. O valor médio do volume de chuvas foi de 333mm. Com estes dados é possível afirmar que a média do volume de chuvas para o período é mesmo 330mm? Use o nível de significância de α=0,05 (que é correspondente ao Nível de Confiança a 95%).
Solução: Considere Ho (Hipótese Nula) como sendo μ = 330mm e H1 (Hipótese Alternativa) como sendo $\mu \neq$ 330mm. Como temos o desvio padrão s= 10mm usamos a estatística de teste Z. Nesse caso é um teste bilateral e deve-se rejeitar a Hipótese nula se Z calculado for maior ou menor que Z tabelado para nível de confiança de 95% para as duas extremidades da curva normal (Figura 69). Nesse caso tem-se Z = -1,96 e Z = 1,96.

Figura 69 – Ilustração das regiões críticas em testes bilaterais.



Logo, a Hipótese nula será rejeitada se o valor de Z, calculado a partir da média das 35 medições, obedecer uma das seguintes condições: $Z < -1{,}96$ ou $Z > 1{,}96$ que são consideradas regiões críticas para o teste bilateral.

$$Z = \frac{\overline{X} - \mu}{\dfrac{\sigma}{\sqrt{n}}} = \frac{333 - 330}{\dfrac{10}{\sqrt{35}}} = 1{,}77$$

Como o valor de Z calculado não está na região crítica de rejeição de Ho, não é possível rejeitar a Hipótese Nula com nível de confiança de 95%. A média do volume de chuvas pode ser sim de 333mm. Há 5% de chance de que essa decisão seja errada.

O Teste de Hipóteses também pode ser realizado quando se tem 2 amostras de duas populações diferentes. Nesse caso é necessário avaliar os parâmetros X1 (média da amostra 1), s1 (desvio padrão da amostra 1) e X2 (média da amostra 2) e s2 (desvio padrão da média 2)
A Hipótese Nula é a diferença das duas médias populacionais. A estatística de teste para avaliação da rejeição ou não da Hipótese Nula é calculada conforme as equações demonstradas na Figura 70. Mais uma vez usa-se a estatística de teste T quando não se conhece o desvio padrão da população, mas apenas da amostra.

Figura 70- Equações para Testes de Hipóteses de duas médias

H_0: $\mu_1 - \mu_2 = \delta$

Hipóteses alternativas	Se σ <u>for</u> conhecido, rejeite H_0 se	Se σ <u>não for</u> conhecido, rejeite H_0 se
$\mu_1 - \mu_2 < \delta$	$Z < -z_\alpha$	$T < -t_\alpha$
$\mu_1 - \mu_2 > \delta$	$Z > z_\alpha$	$T > t_\alpha$
$\mu_1 - \mu_2 \neq \delta$	$Z < -z_{\alpha/2}$ ou $Z > z_{\alpha/2}$	$T < -t_{\alpha/2}$ ou $T > t_{\alpha/2}$

$$Z = \frac{\overline{X}_1 - \overline{X}_2 - \delta}{\sqrt{\dfrac{\sigma_{X_1}^2}{n_1} + \dfrac{\sigma_{X_2}^2}{n_2}}} \qquad T = \frac{\overline{X}_1 - \overline{X}_2 - \delta}{\sqrt{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}} \cdot \sqrt{\frac{n_1 . n_2 (n_1 + n_2 - 2)}{n_1 + n_2}}$$

$$\nu = n_1 + n_2 - 2$$

Fonte: Albertazzi, 2012

Quando não se tem o desvio padrão da população então deve-se calcular os desvios-padrão das amostras e usar a estatística de teste T de Student.

Exemplo 2:
Os moradores de duas cidades vizinhas conhecidas pelas suas baixas temperaturas disputam o título de cidade mais fria no inverno. A cidade A diz que sua temperatura média é de 2 graus Celsius inferior que a da cidade B. Uma amostra de 15 temperaturas de cada cidade são tomadas em uma determinada semana. As temperaturas obtidas foram de 13,34 graus com desvio padrão de 0,297 e 15,22 graus e 0,208. Com nível de confiança de 95% o que é possível afirmar?
Considerando-se a Hipótese Nula formulada como sendo Ho: μA - μA = 2ºC e a Hipótese Alternativa como sendo μA- μB < 2ºC. Nesse caso a Hipótese Nula será rejeitada se o valor da Estatística de Teste T calculada for menor que -1,701 (obtida da Tabela T de Student para nível de confiança de 95% e 28 graus de liberdade). Observe que 28 = (15+15-2).

$$T = \frac{15,22 - 13,34 - 2,00}{\sqrt{(15-1).0,208^2 + (15-1).0,297^2}} . \sqrt{\frac{15.15.(15+15-2)}{15+15}} = -1,28$$

Como o valor de T calculado é maior que T tabelado não é possível rejeitar a hipótese nula Ho. Pode se afirmar com 95% de nível de confiança que a temperatura média de inverno da cidade A é menor que a temperatura média de inverno da cidade B em 2 graus Celsius.

Exemplo 3:
Um estudante fez um ensaio para determinar a influência da corrente de alimentação na qualidade da imagem. Para tal, realizou seis ensaios com a corrente de 1A (ampere) e seis outros ensaios com a corrente de 2A. Para cada ensaio, calculou um coeficiente de qualidade, encontrando os resultados da tabela abaixo. Quanto maior o valor do coeficiente, melhor é qualidade da imagem. Com 95% de probabilidade é possível afirmar que a corrente de alimentação do laser diodo tem influência na qualidade da imagem?

Corrente	Ensaio 1	Ensaio 2	Ensaio 3	Ensaio 4	Ensaio 5	Ensaio 6
1A	208,6	209,0	208,1	208,3	209,2	208,3
2A	202,1	197,9	200,4	200,7	203,0	203,1

Solução: É necessário se calcular a média de coeficientes obtidos com a corrente de 1A e a média dos coeficientes obtidos com corrente de 2A. Com esses valores é necessário se formular a hipótese nula. Nesse caso adota-se que μ1 - μ2=0. A Hipótese alternativa é que a diferença μ1 - μ2 > 0. Nesse

caso, a Hipótese Nula só poderá ser rejeitada se a Estatística de Teste T calculada for superior ao valor de T tabelado para nível de confiança de 95% e grau de liberdade = 10 (6 ensaios + 6 ensaios – 2).
Nesse caso T calculado é de 9,39 que é superior ao T tabelado (1,812), o que permite afirmar com nível de confiança de 95% que a corrente elétrica interfere na qualidade da imagem.

Outro tipo de Teste de Hipóteses muito importante é o do "Qui-quadrado" ou "c^2". O procedimento utilizado anteriormente é muito parecido. Primeiro calcula-se um determinado c^2 de Teste Estatístico e compara-se com um valor de c^2 crítico obtido da Tabela da Distribuição Qui-Quadrado (Anexo). Faz-se a comparação para rejeitar ou aceitar a Hipótese nula. Esse teste também é conhecido como Teste de Contingência.

Exemplo 4:

Vamos supor que uma indústria produza refrigerantes do tipo A, tipo B e do tipo C. O objetivo do departamento de marketing é avaliar se a venda destes produtos está relacionada ao gênero do consumidor. Foram selecionados aleatoriamente 150 consumidores para responder um questionário sobre a preferência pelos refrigerantes do tipo A, B ou C. Os resultados das frequências observadas são tabelados a seguir:

Gênero	Tipo A	Tipo B	Tipo C	Total
Mulheres	20	40	20	80
Homens	30	30	10	70
Total	50	70	30	150

Considere Ho = hipótese nula o caso em que a preferência não tenha relação com o gênero do consumidor e H1 = hipótese alternativa o caso em que a preferência dependa do gênero. Calcule as frequências esperadas para cada uma das células da tabela acima.

$$Freq\ esperada = \frac{soma\ da\ linha \times soma\ da\ coluna}{soma\ geral} = \frac{80 \times 50}{150} = 26,67$$

Gênero	Tipo A	Tipo B	Tipo C	Total
Mulheres	26,67	37,33	16	80
Homens	23,33	32,67	14	70
Total	50	70	30	150

O cálculo de X^2 é realizado pela equação:

$$X^2 = \sum_{i=1}^{n}\left(\frac{(freq\ observada - freq\ esperada)^2}{freq\ esperada}\right) = \left[\left(\frac{(20-26,67)^2}{26,67}\right) + \ldots \left(\frac{(10-14)^2}{14}\right)\right] = 6,13$$

O grau de liberdade é calculado pela equação: (nº de linhas -1).(nº de colunas -1) = 2. Na tabela para QUI quadrado (Figura 71), com GL = 2 e nível de confiança de 95% tem-se: X^2crítico = 5,99.

Figura 71 – Obtenção da estatística de teste Qui-quadrado tabelado.

G.L.	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
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
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
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
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
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
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

Como o valor de X^2 crítico tabelado é menor que X^2 calculado (5,99 < 6,13) a hipótese nula deve ser rejeitada. Logo, com 95% de nível de confiança a hipótese alternativa é aceita e a preferência pelos refrigerantes do tipo A, B e C depende sim do gênero do consumidor.

A seguir, apresentamos mais um exemplo.

Renda R$	Número de filhos				Total
	0	1	2	Mais de 2	
Menos de 2000	15	27	50	43	135
De 2000 a 5000	25	30	12	8	75
Mais de 5000	8	13	9	10	40
Total	48	70	71	61	250

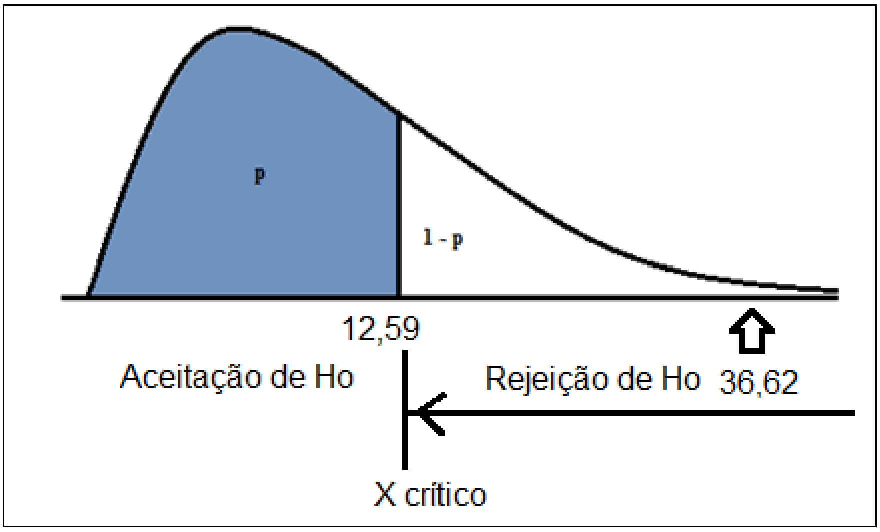
A Hipótese Nula é que o número de filhos e a renda são independentes. Já a Hipótese Alternativa é que existe dependência entre essas duas grandezas. Para cada célula da tabela deve ser calculado o valor esperado conforme o modelo:

$$E_{11} = \frac{48 \times 135}{250} = 25{,}92$$

Vamos verificar se há dependência entre a renda e o número de filhos em famílias de uma cidade. Suponha que, a partir de 250 famílias escolhidas ao acaso, tenhamos a tabela:

Renda R$	Número de filhos				Total
	0	1	2	Mais de 2	
Menos de 2000	25,92	37,80	38,34	32,94	135
De 2000 a 5000	14,40	21,00	21,30	18,30	75
Mais de 5000	7,68	11,20	11,36	9,76	40
Total	48	70	71	61	250

Figura 73 – Ilustração da região de rejeição de Ho na curva Qui-quadrado.



Vídeo recomendado da Khan Academy para Teste de Contingência usando Qui-Quadrado https://www.youtube.com/watch?v=snUTmyRrbG4

A estatística Qui-Quadrado é calculada pela expressão:

$$\chi^2 = \frac{(15-25{,}92)^2}{25{,}92} + \frac{(25-14{,}4)^2}{14{,}4} + \frac{(8-7{,}68)^2}{7{,}68} + \ldots\ldots + \frac{(10-9{,}76)^2}{9{,}76} = 36{,}62$$

A partir da determinação do grau de liberdade = 2 x 3 = 6 . Na tabela X^2 , com nível de confiança de 95% temos X^2 tabelado = 12,6 (Figura 72).

Figura 72- Obtenção do valor de X^2 tabelado. VER TABELA ANEXA.

G.L.	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
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
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
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
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
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
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

Como X^2 calculado é maior que X^2 tabelado rejeitamos a Hipótese nula. Com 95% de nível de confiança podemos afirmar que não existe independência entre a renda e o número de filhos.

LISTA DE EXERCÍCIOS 8:

1- Um pesquisador tem interesse de saber se a preferência de uso do transporte público está relacionada com o gênero. Ele entrevista 400 pessoas e obteve as informações tabeladas. Existe influência do gênero na escolha do transporte público?

Usa transporte público	Homens	Mulheres
Usuários	92	88
Não usuários	108	112

2- Uma pesquisa divulgou que o volume de chuvas em uma região para o mês de junho é de 320mm com um desvio padrão típico de 20mm. Uma amostra com 25 dias da série histórica foi analisada. O valor médio do volume de chuvas foi de 340mm. Com estes dados é possível afirmar que a média do volume de chuvas para o período é mesmo 320mm? Use o nível de significância de 0,05.

3- Os moradores de duas cidades vizinhas conhecidas pelas suas baixas temperaturas disputam o título de cidade mais fria no inverno. A cidade A diz que sua temperatura média é de 5°C inferior que a da cidade B. Uma amostra de 16 temperaturas de cada cidade são tomadas em uma determinada semana. As temperaturas obtidas foram de 16°C com desvio padrão de 2°C e 14°C e desvio padrão de 4°C. Com nível de confiança de 95% o que é possível afirmar?

4- Avalie se os níveis de renda de duas cidades estão associados com NC = 99%. Foram pesquisados 400 moradores ao todo.

	A	B	C	D	Total
X	28	42	30	24	124
Y	44	78	78	76	276
Total	72	120	108	100	400

Seja Ho = as variáveis são independentes e H1 = as variáveis são dependentes.

VÍDEO RECOMENDADO: TESTE DE HIPÓTESE – UNIVESP TV

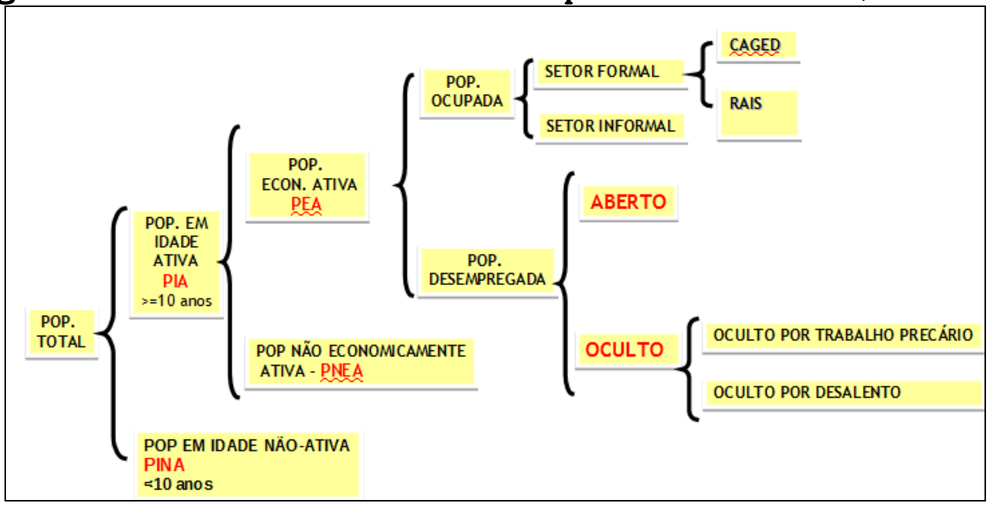
https://www.youtube.com/watch?v=9zMREPL93WA

ANEXO A- Banco de Dados

A obtenção de estatísticas confiáveis para análise socioespaciais pode ser realizada por meio de diversos sistemas / banco de dados tais como RAIS, CAGED, IBGE (SIDRA, IBGE-Cidades, PNAD, Censos) entre outros.

Na Figura 74 é possível visualizar que somente parte da população economicamente ativa está ocupada. E do volume de trabalhadores ocupados, apenas parte está no setor formal. O banco de dados do CAGED e do RAIS possibilita o acesso a esses dados em detalhes. No ano de 2014, havia no país 49 milhões de vínculos ativos em 31 de dezembro.

Figura 74 – Detalhamento do recorte para uso da RAIS / CAGED.



a)- RAIS - Relação Anual de Informações Sociais

Foi criada no ano de 1975 pelo Decreto nº 76.900/75 como um registro anual obrigatório dos empregos formais para todos os estabelecimentos brasileiros. Quando o estabelecimento não tem nenhum vínculo empregatício no período ela deve declarar a RAIS NEGATIVA.

Os dados da RAIS podem ser organizados de forma desagregada até em nível municipal e em nível de subatividades econômicas e ocupações. As ocupações são organizadas a partir da CBO – Catálogo Brasileiro de Ocupações. As atividades econômicas por meio da CNAE. Podem também ser consultados para determinado ano, região, ocupação e atividade econômicas. É possível conhecer número de empregos, as admissões e os desligamentos. Esses dados podem ser separados por gênero, faixa etária, grau de instrução, rendimento e faixas de rendimentos. Também podemos conhecer a média salarial e a faixa etária dos trabalhadores.

Com essas informações é possível entre diversas aplicações a análise da evolução do mercado de trabalho e subsidiar os investimentos públicos na formação profissional.

Conhecer o significado de algumas expressões é importante para o uso da RAIS:

a) Vínculos empregatícios: São consideradas como vínculos as relações de trabalho dos celetistas, dos estatutários, dos trabalhadores regidos por contratos temporários, por prazo determinado, e dos empregados avulsos, quando contratados por sindicatos. Na RAIS é possível selecionar apenas os vínculos ativos no dia 31 de dezembro de um determinado ano.

b) Trabalhador celetista e por prazo determinado: O trabalhador celetista é aquele cuja relação de emprego é regida pela CLT, independentemente de o empregador ser do setor público ou privado. Os trabalhadores por prazo determinado, regidos pela Lei nº 9.601, são aqueles que podem ser contratados por um período máximo de dois anos, desde que esse tipo de contrato tenha sido previsto em convenção ou em acordo coletivo.

c) Tamanho do estabelecimento: o tamanho do estabelecimento é determinado pelo número de empregos nele existente em 31 de dezembro do ano-base.

d) Atividade econômica: Desde o ano de 1994, o Ministério do Trabalho e Emprego assumiu a atividade econômica declarada pelo estabelecimento, captada de acordo com o novo Código de Atividade Econômica – CNAE/95. Para manter a comparabilidade dos dados anuais, foi elaborada uma compatibilização com o código do IBGE. Em 2002, foi realizada a primeira revisão da CNAE95, que foi denominada de CNAE 1.0. No caso dos estabelecimentos com mais de uma atividade econômica, é considerada a atividade principal. As atividades relativas a depósito e a escritórios administrativos ou de representação de empresas são colocadas na atividade principal da matriz. A partir do ano base 2006, a RAIS vem captando informações segundo o novo código de Atividade Econômica – CNAE 2.0.

e) Classificação Brasileira de Ocupações[20] – CBO: A CBO é o documento que codifica os títulos e as características das ocupações do mercado de trabalho brasileiro.

f) Faixa etária: O enquadramento dos vínculos na faixa etária considera os anos completos em 31 de dezembro.

g) Admissão e desligamento: Os conceitos de admissão e desligamento utilizados na RAIS referem-se às alterações de emprego ocorridas no estabelecimento, incluindo as transferências de empregados, de um estabelecimento para outro, da mesma empresa.

[20] http://www.mtecbo.gov.br/cbosite/pages/home.jsf

h) Por admissão entende-se toda entrada de trabalhador no estabelecimento no ano, qualquer que seja sua origem e, por desligamento, toda saída de pessoa cuja relação de emprego com o estabelecimento cessou durante o ano por qualquer motivo (demissão, aposentadoria, morte), seja por iniciativa do empregador ou do empregado. As entradas e saídas por transferências aparecem incluídas, respectivamente, nas admissões e nos desligamentos.

i) Remuneração média mensal em salário mínimo: A remuneração média mensal em salários mínimos é definida como a média aritmética das remunerações individuais no mês de referência, convertidas em salários mínimos, no período vigente do ano-base. Integram essa remuneração os salários, ordenados, vencimentos, honorários, vantagens, adicionais, gratificações, etc. Está excluída a remuneração do 13° salário.

b) CAGED - Cadastro Geral de Empregados e Desempregados

O CAGED foi criado em 1965 pela Lei n° 4.923 e se constitui de uma fonte de informação de âmbito nacional e de periodicidade mensal. Foi criado como instrumento de acompanhamento e de fiscalização do processo de admissão e de dispensa de trabalhadores regidos pela CLT, com o objetivo de assistir os desempregados e de apoiar medidas contra o desemprego. A partir de 1986, passou a ser utilizado como suporte ao pagamento do seguro-desemprego e, mais recentemente, tem contribuído para o planejamento da abertura de cursos de requalificação profissional do volume de trabalhadores demitidos.

O CAGED apresenta desagregações idênticas às da RAIS, em termos geográficos, setoriais e ocupacionais, possibilitando a realização de estudos que indicam as tendências mais atuais. Para uso do banco de dados é importante que sejam detalhados os significados das expressões:

a) Flutuação do emprego: relaciona-se a movimentação das admissões e desligamentos em um determinado período.

b) Variação absoluta (saldo): Indica a diferença entre admissões e desligamentos (a – d).

c) Variação relativa: É obtida mediante a divisão da variação absoluta sobre o estoque do primeiro dia do mês. Mostra o desempenho do emprego em termos percentuais.

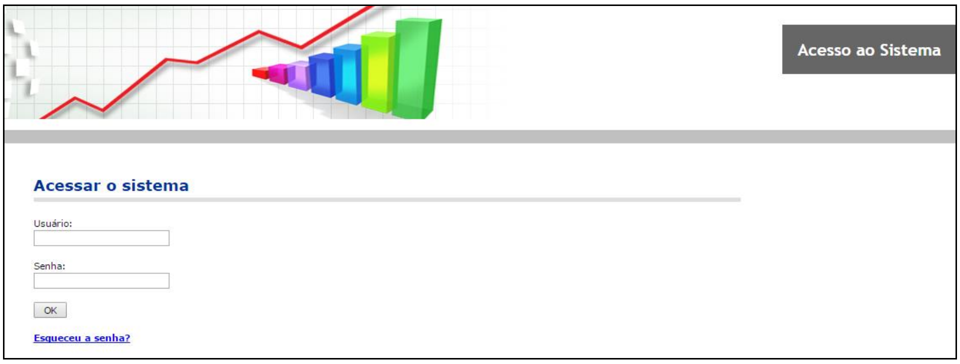
d) Índice de emprego: Indica a evolução do emprego formal. É calculado tendo-se por base o encadeamento das variações relativas mensais.

e) Taxa de rotatividade: Mede o percentual dos trabalhadores substituídos mensalmente em relação ao estoque vigente no primeiro dia do mês, em nível geográfico e setorial, mas não em nível ocupacional.
Assim como a RAIS, o CAGED oferece informações referentes aos estabelecimentos e aos empregados.

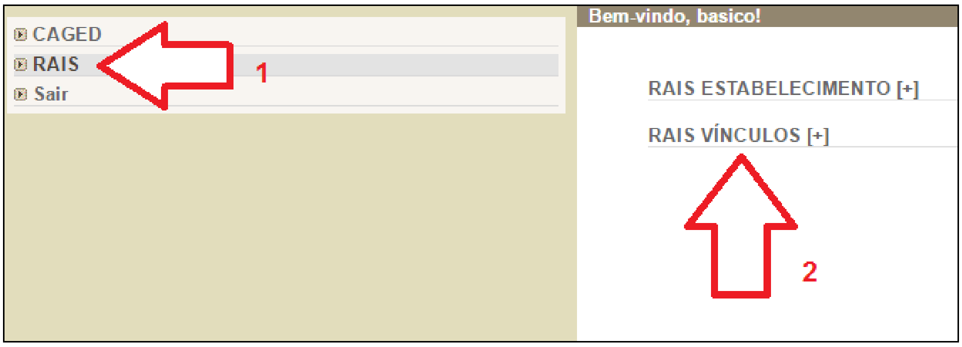
A seguir serão apresentados alguns exemplos de extração de dados da Plataforma RAIS / CAGED

Exemplo 1:
Quais são os empregos com vínculos ativos em 31/12 por estabelecimentos no ano de 2014 no Estado de Santa Catarina nos Grandes Setores do IBGE?

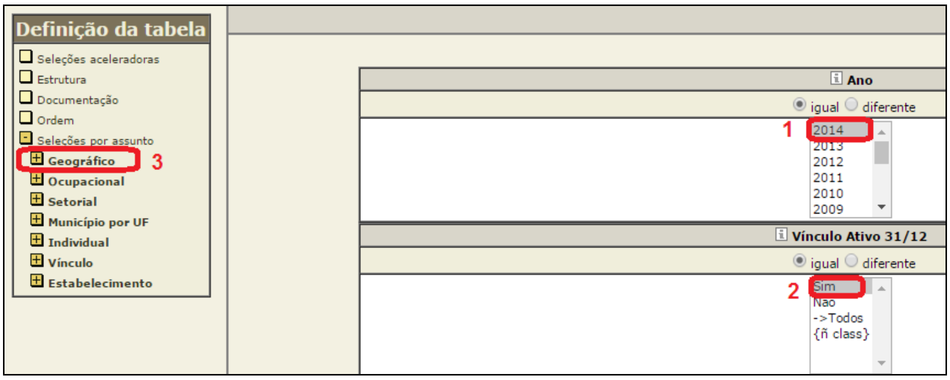
Basta digitar o endereço eletrônico e inserir o nome de usuário e senha: http://bi.mte.gov.br/bgcaged/login.php. Utilize como usuário: basico e como senha 12345678



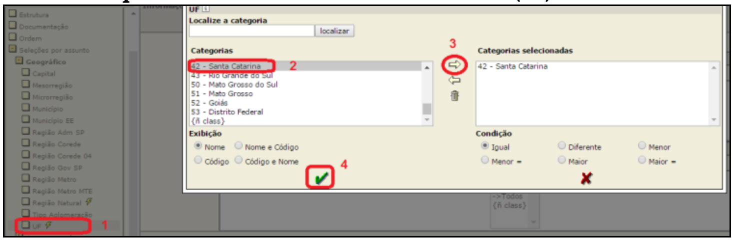
Clique em 1 e depois em 2 para acessar a base de dados dos empregos formais da RAIS.



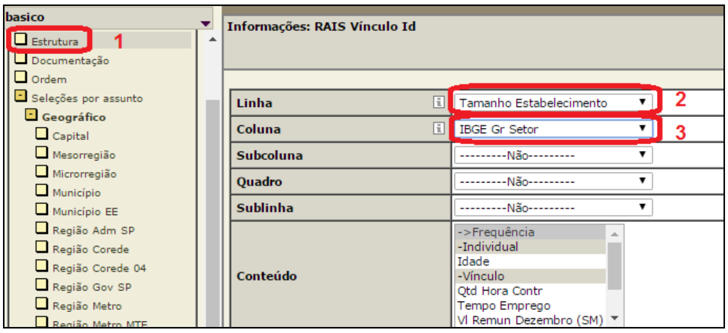
Clique sobre o link: ANO CORRENTE A 2002. Escolha a opção VÍNCULOS ATIVOS EM 31/12 ou TODOS OS VÍNCULOS. Acesse a aba SELEÇÕES POR ASSUNTO e clique sobre o link GEOGRÁFICO.



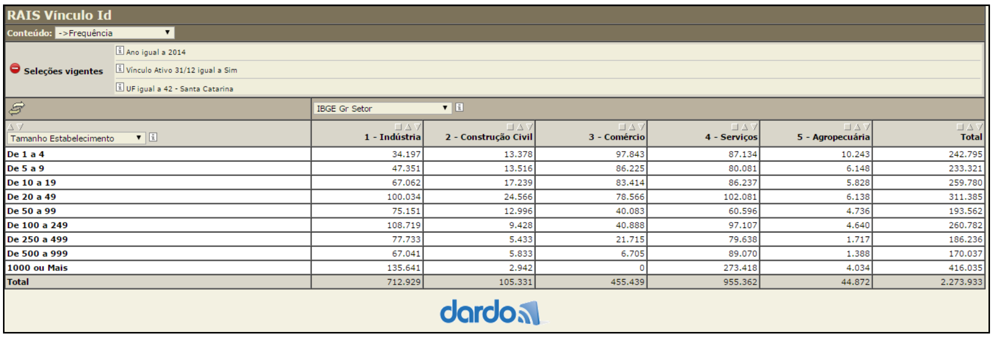
Selecione a opção UF (1). Escolha o Estado de Santa Catarina (2). Clique na seta (3) e finalmente clique no desenho CERTO (4).



Pode-se definir no MENU ESTRUTURA como o relatório será apresentado. Escolha TAMANHO DO ESTABELECIMENTO na Linha e GRANDE SETOR IBGE na coluna.



Clique no ícone em formato de um RAIO para a consulta ser concluída. Observe o resultado, que pode ser exportado no formato CSV ao clicar no LINK correspondente no menu superior.



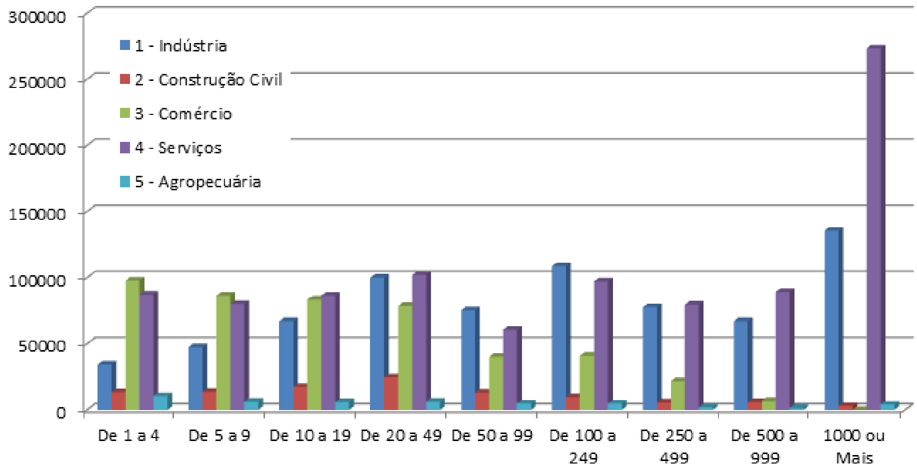
RAIS Vínculo Id

Conteúdo: ->Frequência

Seleções vigentes: Ano igual a 2014; Vínculo Ativo 31/12 igual a Sim; UF igual a 42 - Santa Catarina

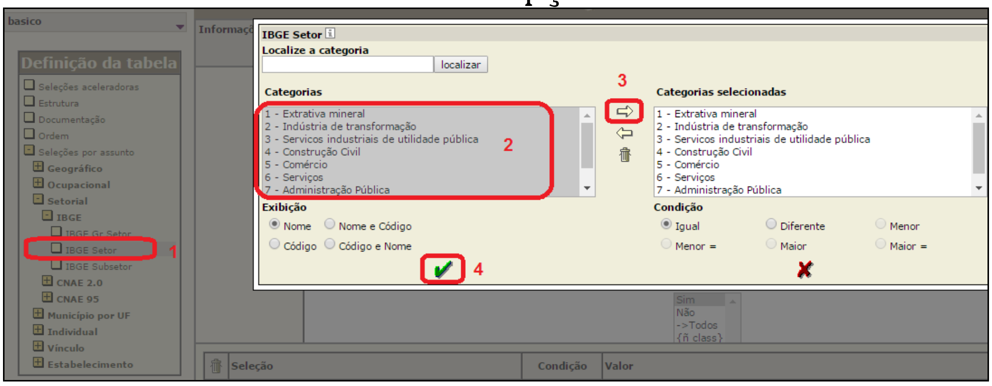
IBGE Gr Setor

Tamanho Estabelecimento	1 - Indústria	2 - Construção Civil	3 - Comércio	4 - Serviços	5 - Agropecuária	Total
De 1 a 4	34.197	13.378	97.843	87.134	10.243	242.795
De 5 a 9	47.351	13.516	86.225	80.081	6.148	233.321
De 10 a 19	67.062	17.239	83.414	86.237	5.828	259.780
De 20 a 49	100.034	24.566	78.566	102.081	6.138	311.385
De 50 a 99	75.151	12.996	40.083	60.596	4.736	193.562
De 100 a 249	108.719	9.428	40.888	97.107	4.640	260.782
De 250 a 499	77.733	5.433	21.715	79.638	1.717	186.236
De 500 a 999	67.041	5.833	6.705	89.070	1.388	170.037
1000 ou Mais	135.641	2.942	0	273.418	4.034	416.035
Total	712.929	105.331	455.439	955.362	44.872	2.273.933

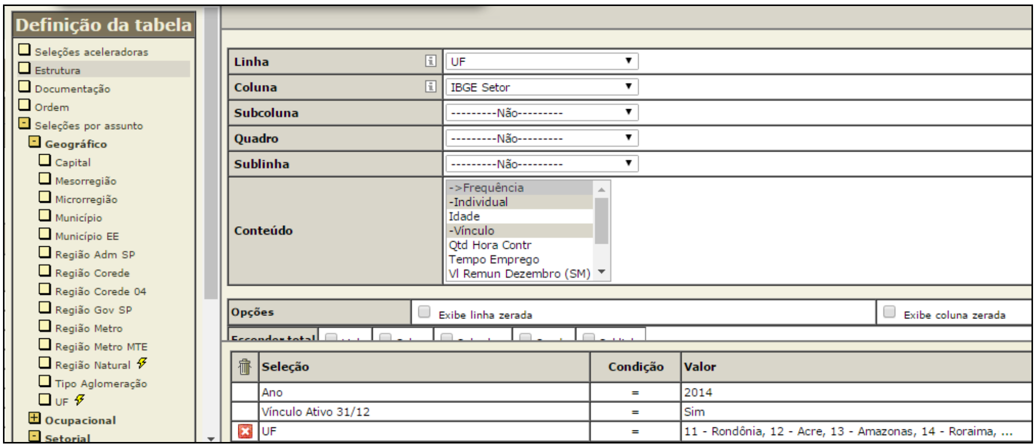


Exemplo 2:

Qual foi o estoque de trabalhadores no Brasil por setor de atividade em 2014? Devemos selecionar ANO CORRENTE 2014. No menu GEOGRÁFICO acesse a variável UF e selecione todos os estados. No Menu Setorial selecione a variável SET IBGE e selecione todas as opções.

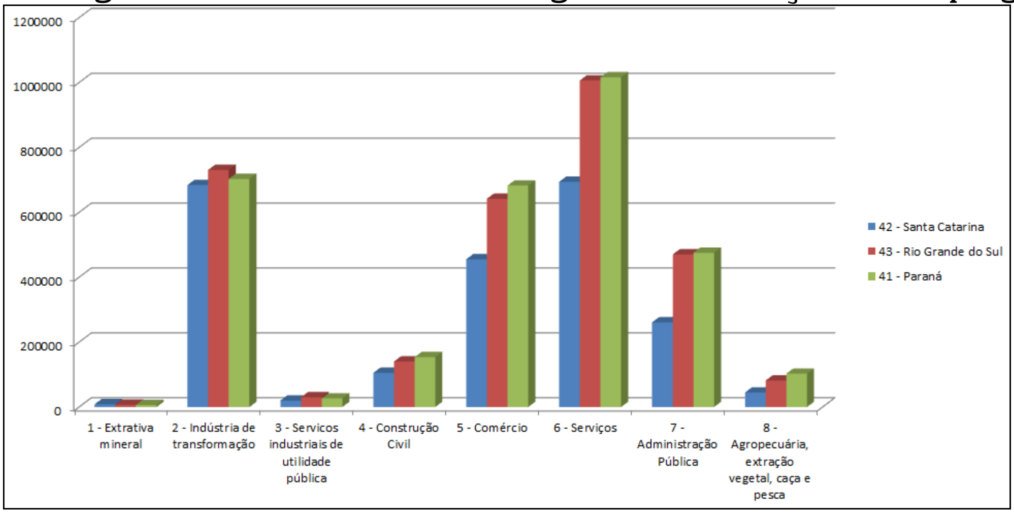


No menu ESTRUTURA em linha selecione UF e na coluna IBGE SETOR.



O resultado pode ser visualizado a seguir:

Para a região Sul do Brasil temos a seguinte distribuição de empregos:



Acesse os exemplos resolvidos passo a passo no link:

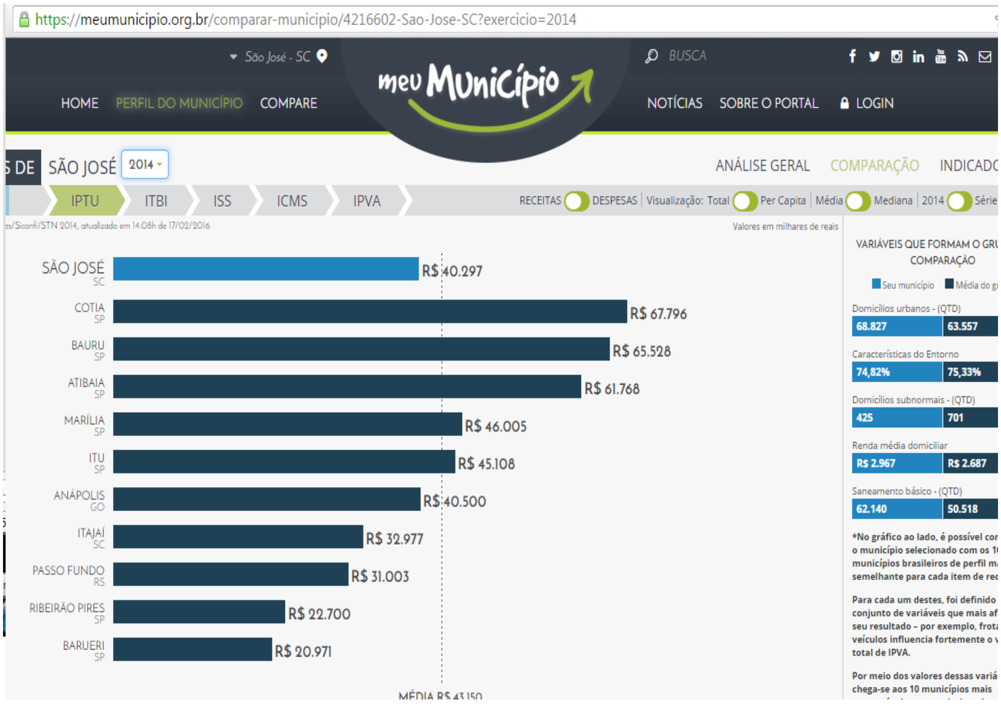
https://estatisticaparageografia.wordpress.com/banco-de-dados-rais/

d- Outras bases de dados

Além da RAIS / CAGED há diversos bancos de dados disponíveis para consulta pela internet. Entre as principais temos:

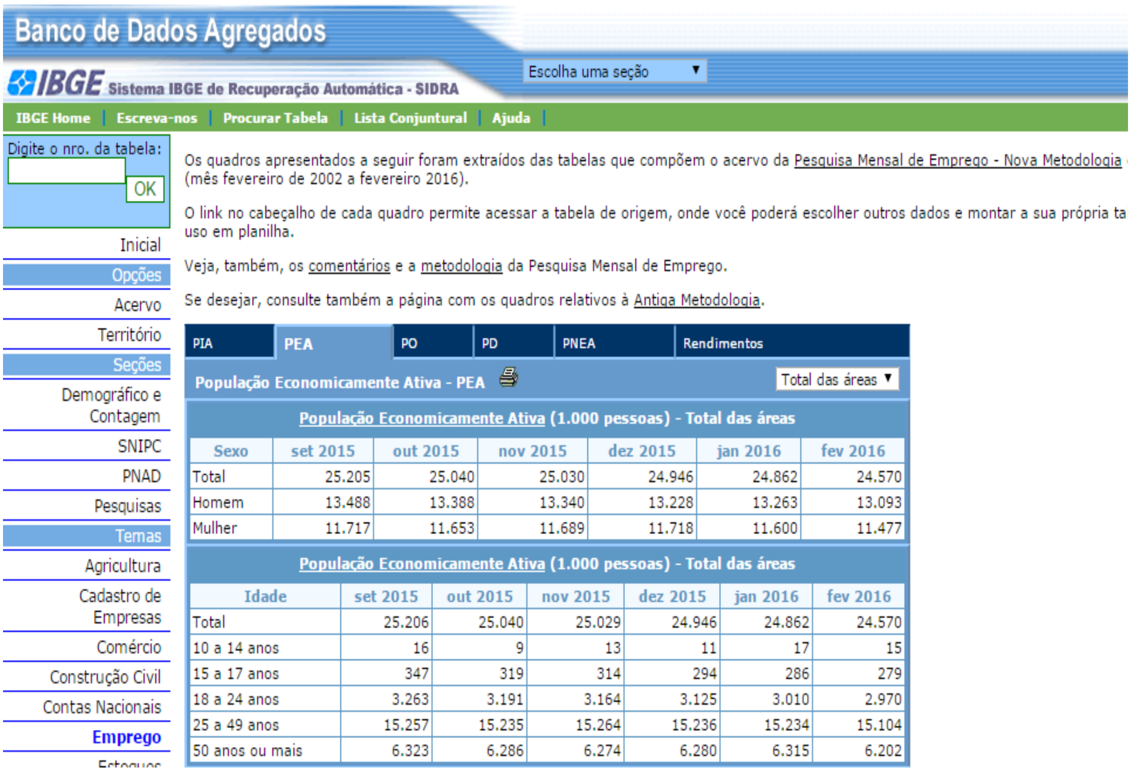
Banco de dados sobre os municípios brasileiros

https://meumunicipio.org.br/perfil-municipio/4216602-Sao-Jose-SC



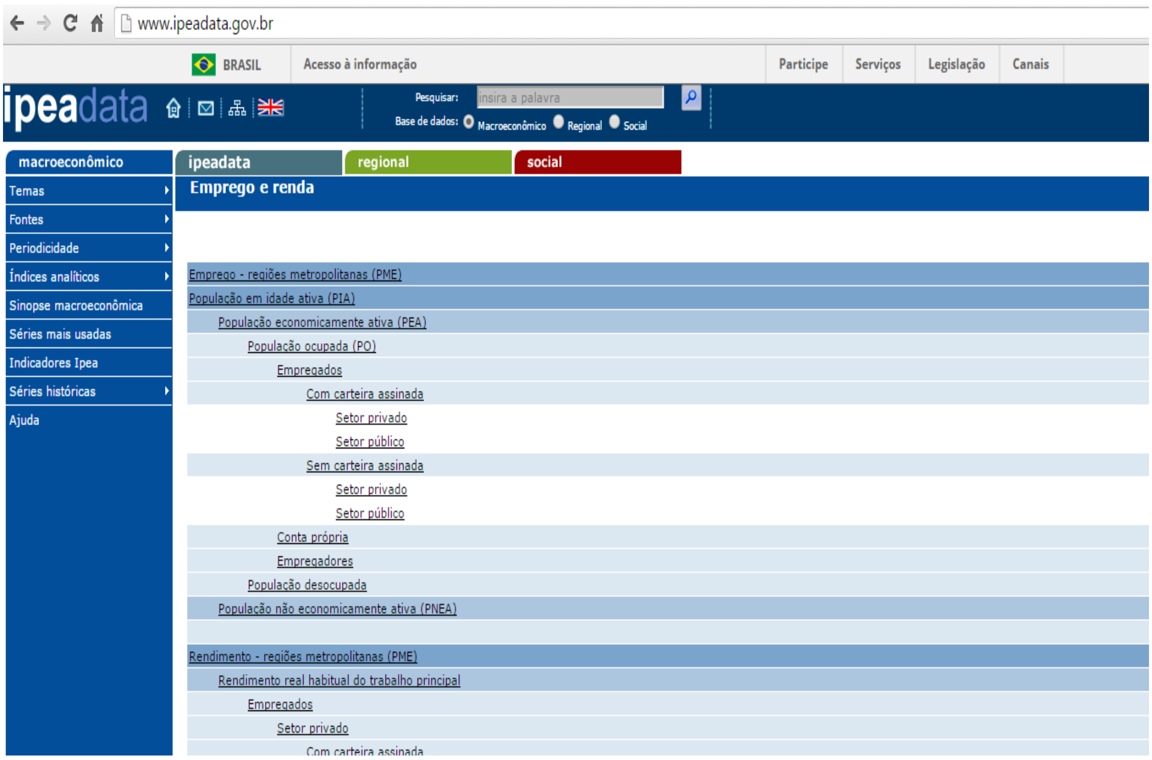
SIDRA – Sistema IBGE de Recuperação Automática

http://www.sidra.ibge.gov.br/



IPEADATA

http://www.ipeadata.gov.br/

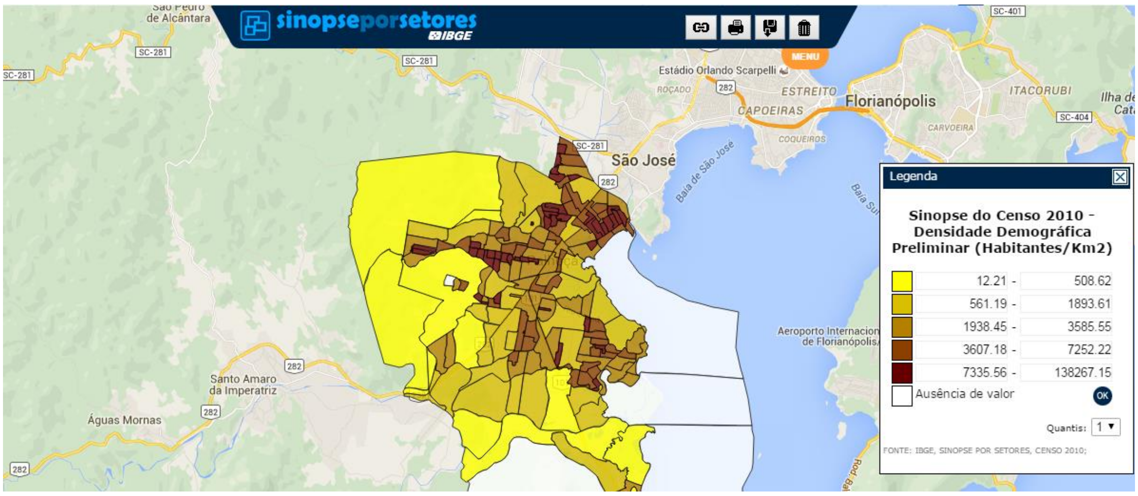


IBGE CIDADES

http://www.cidades.ibge.gov.br/xtras/home.php

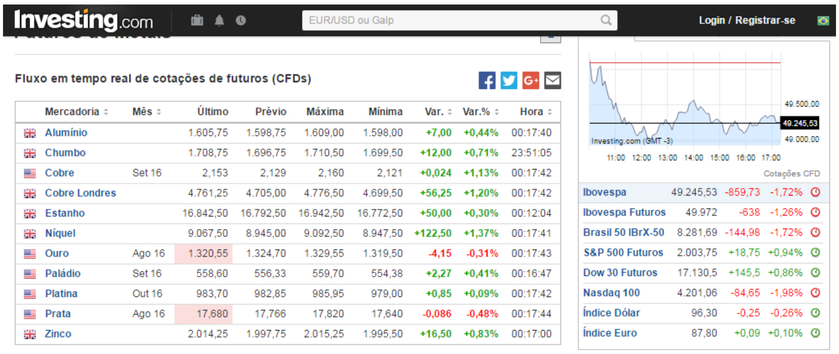
SINOPSE POR SETORES CENSITÁRIOS

http://www.censo2010.ibge.gov.br/sinopseporsetores/



INVESTIMENTOS

http://br.investing.com/commodities/metais



Investing.com

EUR/USD ou Galp

Login / Registrar-se

Fluxo em tempo real de cotações de futuros (CFDs)

Mercadoria	Mês	Último	Prévio	Máxima	Mínima	Var.	Var.%	Hora
Alumínio		1.605,75	1.598,75	1.609,00	1.598,00	+7,00	+0,44%	00:17:40
Chumbo		1.708,75	1.696,75	1.710,50	1.699,50	+12,00	+0,71%	23:51:05
Cobre	Set 16	2,153	2,129	2,160	2,121	+0,024	+1,13%	00:17:42
Cobre Londres		4.761,25	4.705,00	4.776,50	4.699,50	+56,25	+1,20%	00:17:42
Estanho		16.842,50	16.792,50	16.942,50	16.772,50	+50,00	+0,30%	00:12:04
Níquel		9.067,50	8.945,00	9.092,50	8.947,50	+122,50	+1,37%	00:17:41
Ouro	Ago 16	1.320,55	1.324,70	1.329,55	1.319,50	-4,15	-0,31%	00:17:43
Paládio	Set 16	558,60	556,33	559,70	554,38	+2,27	+0,41%	00:16:47
Platina	Out 16	983,70	982,85	985,95	979,00	+0,85	+0,09%	00:17:42
Prata	Ago 16	17,680	17,766	17,820	17,640	-0,086	-0,48%	00:17:44
Zinco		2.014,25	1.997,75	2.015,25	1.995,50	+16,50	+0,83%	00:17:00

49.500,00
49.245,53
49.000,00
Investing.com (GMT -3)
11:00 12:00 13:00 14:00 15:00 16:00 17:00
Cotações CFD

Ibovespa	49.245,53	-859,73	-1,72%
Ibovespa Futuros	49.972	-638	-1,26%
Brasil 50 IBrX-50	8.281,69	-144,98	-1,72%
S&P 500 Futuros	2.003,75	+18,75	+0,94%
Dow 30 Futuros	17.130,5	+145,5	+0,86%
Nasdaq 100	4.201,06	-84,65	-1,98%
Índice Dólar	96,30	-0,25	-0,26%
Índice Euro	87,80	+0,09	+0,10%

ALICEWEB – MDIC

http://aliceweb.mdic.gov.br//index/home

Arranjos Produtivos Locais:

http://portalapl.ibict.br/menu/itens_menu/apls/apl_o_que_sao.html

Portal da Transparência:

http://www.portaltransparencia.gov.br/

ANEXO B - EXERCÍCIOS[21]

1- Em 16 de junho de 2016 ocorreu o Feirão de Empregos de Florianópolis. Como seria possível quantificar o total de pessoas na fila? Se você tivesse que descrever o perfil dessas pessoas qual estratégia você utilizaria? Qual o contexto socioeconômico desse evento ?

2- Um novo centro de eventos está sendo planejado para uma determinada região. Nela há 3 comunidades residenciais e o centro médio é um dos critérios para localização. Suponha que a comunidade 1 tenha coordenada central (30, 36)km e população de 20 mil pessoas. A comunidade 2 tem coordenada central de (55,18)km e população de 12 mil pessoas. Já a comunidade 3 tem coordenada central de (10,18)km e população de 5 mil pessoas. Qual é o centro médio ponderado? Se o critério fosse a renda e não o tamanho da população, qual seria o novo centro médio. Suponha que a comunidade 1 tenha renda total de 2 milhões de reais, a comunidade 2 tenha renda total de 12 milhões de reais e a comunidade 3 de 20 milhões de reais. Considere a equação abaixo, onde Pi pode ser tanto população quanto renda.

$$x_{CG} = \frac{\sum_{i=1}^{n} (x_i . P_i)}{\sum_{i=1}^{n} P_i} \quad \text{e ainda} \quad y_{CG} = \frac{\sum_{i=1}^{n} (y_i . P_i)}{\sum_{i=1}^{n} P_i}$$

[21] Parte das soluções desses exercícios encontra-se no blog: http:estatisticaparageografia.wordpress.com

3- Uma nova escola está sendo construída pela prefeitura para atender as crianças de uma região. Considerando que a probabilidade de uma criança da região se matricular nessa nova escola segue uma distribuição de probabilidade exponencial, calcule quais as chances de uma criança que mora a 10km de distância estudar na nova escola. É conhecido que o valor esperado de distância dessa distribuição é de 4km.

4- Os gastos mensais de uma amostra de famílias são descritos por meio de uma tabela, onde também estão descritas suas rendas. Qual a correlação existente entre a renda e o gasto mensal dessas famílias?

Quantia Gasta por semana (R$) Y	Renda da família (R$) X	X.Y	
120	6500		
68	3500		
35	3000		
60	4400		
100	8000		
91	7700		
44	3200		
71	3900		
89	4400		
113	7700		

5- Um pesquisador está estudando a relação entre os preços de uma casa, o tamanho dos terrenos e o número de quartos. Analisando uma amostra de propostas de vendas em sites específicos ele anotou os valores médios das casas e as respectivas áreas dos terrenos e número de quartos.

Preço da casa	Área do Lote (m2)	Número de quartos
130.000	5000	3
134.000	5500	2
159.000	6000	4
164.000	6500	3
132.000	5200	2
125.000	5400	1
146.000	5700	3
168.000	6100	4
171.000	6300	4
187.000	6400	5

6- A partir da população das cidades catarinenses (PNAD, 2014) elabore um histograma da quantidade de habitantes. Os dados estão disponíveis no IBGE.

7- Se uma multinacional quisesse instalar uma fábrica em cada uma das 5 cidades que mais cresceram em Santa Catarina em termos populacionais nos últimos 14 anos, quais seriam essas cidades? Dados PNAD 2014 e IBGE 2000.

8- Se uma grande multinacional quisesse instalar uma fábrica em cada uma das 5 cidades que mais cresceram em Santa Catarina em termos econômicos nos últimos anos, quais seriam essas cidades?

9- O que é COEFICIENTE DE VARIAÇÃO DOS DADOS e qual sua importância?

10- Calcule a média, a amplitude, a mediana e o desvio padrão do conjunto de dados:

29, 35, 17, 30, 231, 6, 27, 35, 23, 29, 13

11- A probabilidade anual de inundações em uma comunidade é de 0,10. Qual a probabilidade de acontecerem 3 inundações nos próximos 10 anos?
12- Considere que em um cruzamento ocorrem um assalto a cada dez dias. Qual a probabilidade de ocorrência de três assaltos durante o período de 25 dias?

13- Uma doença acontece aleatoriamente no espaço com um caso incidente a cada 10 quilômetros quadrados. Qual a probabilidade de se encontrarem quatro casos em uma área de 30 quilômetros quadrados?

14- O tempo de deslocamento ao trabalho é normalmente distribuída com média de 30 minutos e desvio padrão de 10 minutos. Qual a probabilidade de uma pessoa escolhida ao acaso na população tenha tempo de deslocamento maior que 40 minutos?

15- Qual a probabilidade do tempo de deslocamento se situar entre 20 e 30 minutos?

16- Ordene os dados. Indique o 1º, 2º e 3º quartil. Desenhe o diagrama de caixa. Calcule a média e a mediana dos dados. Determine qual o desvio padrão.

11, 12, 4, 2, 3, 4, 11, 8, 5, 15, 20, 21

17- Calcule a correlação que relaciona a idade e a altura de uma criança.

Idade (anos)	Altura (cm)
6	70
8	110
10	130
12	150
14	155
15	160

18- O dono de uma lanchonete anotou quanto de refrigerantes (em litros) ele vende ao longo dos dias de acordo com a temperatura. Qual a relação entre estas duas informações?

Temperatura	Litros
15	22
20	25
25	28
27	30
30	32
31	31
32	33
35	35

19- Os dados a seguir representam as alturas (em cm) de 25 alunos de uma classe. Construa o histograma e calcule a média e o desvio padrão.

155	163	148	166	169
164	165	159	175	155
170	165	176	157	157
150	150	160	165	164
166	169	152	170	190

20- Qual a reta ajustada que melhor representa a correlação entre as grandezas X e Y representadas abaixo?

X	5	7	7	10	6	7	9
Y	7	9	8	10	5	7	8

21- Calcule a média, a mediana e a moda dos dados apresentados a seguir:

80, 94, 86, 88, 84, 85, 85, 91, 93

22- Calcular a média e o desvio padrão dos dados apresentados por meio da tabela de classes / frequência:

Classe x	Frequência
150 a 155	2
156 a 160	4
161 a 165	6
166 a 170	15
171 a 175	6
176 a 180	4
181 a 185	3

23- Construir o diagrama de caixa (Box-plot) dos dados:

12, 16, 13, 9, 18, 15, 14, 21, 7, 10, 11, 20, 5, 18, 37, 16, 17

24- As notas de turma de alunos são mostradas na tabela. Qual a média e a mediana?

Nota	Quantidade
2	2
4	4
6	12
8	6
10	2

25- Uma caixa possui 10 peças, mas 4 delas são defeituosas. Selecionando-se aleatoriamente 2 bolas sem reposição, qual a probabilidade de obtermos 2 peças boas ?

26- Um dado equilibrado é lançado. Qual a probabilidade de sair a face o número 4, se já temos a informação de que a face que saiu é par ?

27- Considere 3 lançamentos seguidos de uma moeda honesta. Qual a probabilidade de sair exatamente 2 cara nesses 3 lançamentos?

28- Uma caixa tem 5 bolas brancas e 2 bolas pretas. Selecionando-se aleatoriamente (por sorteio) 2 bolas sem reposição, qual a probabilidade de sair 2 bolas pretas?

29- Considere que dois dados honestos sejam lançados juntos. Em cada jogada, calcula-se a soma dos resultados. Qual a probabilidade de que a soma seja 5 ou 7 ?

30- Um piloto tem probabilidade de vencer uma corrida calculada em 1/10. Qual a probabilidade do piloto vencer duas corridas em 5 ?

31- Uma urna tem bolas numeradas de 1 a 20. Sorteamos uma bola aleatoriamente. Qual a probabilidade de sair um número múltiplo de 2 ou de 3 ?

32- Ao lançar um dado muitas vezes, uma pessoa percebeu que a face 6 saia com o triplo de frequência da face 1 e que as outras faces saiam com a frequência esperada de um dado não viciado. Qual a frequência da face 1?

33- Uma pesquisa é realizada com 10.000 consumidores sobre a preferência por tipo de sabão em pó. Verificou-se que 7.500 usam a marca X. 4.500 usam a marca Y. 2.000 utilizam as duas marcas. Foi sorteada uma pessoa entre as 10.000 e verificou-se que ela usa a marca X. Qual a probabilidade dessa pessoa também ser usuária da marca Y?

34- Em um colégio 10% dos homens e 8% das mulheres têm mais que 1,80m de altura. O total de homens é de 60% dos estudantes. Se um estudante é escolhido aleatoriamente e tem mais que 1,80m de altura, qual é a probabilidade de que seja mulher?

35- Uma cidade tem 50.000 pessoas e 3 jornais em circulação: A, B e C. Sabe-se que 15000 pessoas lêem o jornal A, 10000 pessoas lêem o jornal B, 8000 lêem o jornal C, 6000 lêem os jornais A e B, 4000 lêem os jornais A e C, 3000 lêem os jornais B e C, 2.000 lêem os jornais A, B e C. Uma pessoa é escolhida aleatoriamente. Qual é probabilidade de que ela leia pelo menos um jornal? Qual a probabilidade de que ela leia apenas 1 jornal?

36- Um casal pretende ter 4 filhos. Qual a probabilidade de nascerem EXATAMENTE dois meninos?

37- Uma empresa de aluguel de carros anota o número de carros alugados. Em um determinado período, a probabilidade de alugar 10 carros é de 30%, a de alugar 11 carros é e 30%, de alugar 12 carros é de 35% e de alugar 13 carros é de 15%. Calcule o número médio de carros alugados por semana.

38- Uma pesquisa realizada com 1.000 estudantes, sendo 500 mulheres e 500 homens, mediu o tempo de reação para frear um carro em milisegundos. O valor médio obtido tanto para homens quanto para mulheres foi de 150ms com um desvio padrão de 25ms. Considerando que o tempo de reação obedece a uma distribuição normal, qual é a probabilidade de encontrar uma pessoa com tempo de maior que 200ms?

39- Em uma rede de computadores, em 20% dos dias ocorre alguma falha. Considere a variável aleatória X = número de dias com falha na rede. Considere o período de observação de 10 dias e suponha que os eventos são independentes. Qual a probabilidade de ocorrer mais que 6 dias e falhas na rede, considerando os 10 dias de observação?

40- Uma fábrica de cimentos necessita encher sacos com peso médio de 50kg. No entanto, a massa é normalmente distribuída com desvio padrão de 1kg. Selecionando-se um saco de cimento aleatoriamente, qual a probabilidade de que ele tenha massa menor que 49kg?

41- Uma máquina produz discos de diâmetro médio de 3cm com desvio padrão de 0,08cm. As peças que se afastam por mais de 0,16cm do diâmetro médio são consideradas com defeito. Qual o percentual de peças consideradas defeituosas?

42- A vida média de uma marca de televisão é de 10 anos com desvio padrão de 1,5 anos. A campanha de lançamento diz que todos os produtos que tiverem defeito dentro do prazo de garantia serão trocados por novos. Se você fosse o gerente de produção, qual seria o tempo de garantia que você especificaria para ter no máximo 5% de trocas?

43- Uma empresa produz resistores com resistência média de 60 ohms e desvio padrão de 4 ohms. A resistência é normalmente distribuída. Qual a probabilidade de encontrarmos resistores com resistência inferior a 50 ohms?

44- A vida útil de um tipo de lâmpada é normalmente distribuída com valor médio de 1.000h e desvio padrão de 50h. Ao selecionarmos uma lâmpada aleatoriamente, qual a probabilidade de que ela queime entre 500 e 600 horas?

45- Um cruzamento tem uma média de 5 acidentes por mês. Qual a probabilidade de ocorrer 4 acidentes em um mês qualquer?

46- Um taxista recebe em média 5 chamadas a cada hora. Qual a probabilidade de não receber nenhuma chamada em uma determinada hora?

47- Um time de futebol joga 8 partidas. Assumindo que a probabilidade de vitória em cada jogo é de 40%, qual é a probabilidade de que o time vença exatamente 4 jogos?

48- Um posto de gasolina atende em média 8 clientes por hora. Qual a probabilidade de que apenas 4 clientes sejam atendidos em uma hora?

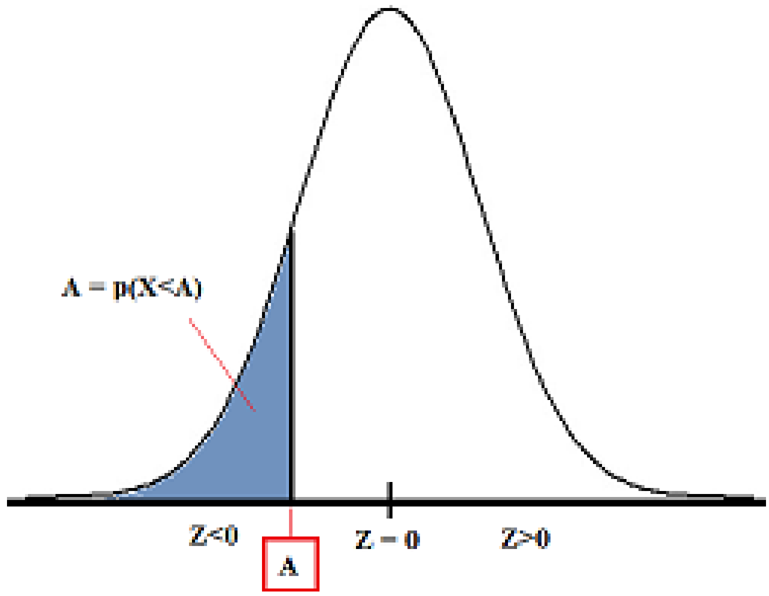
49- Suponha que em uma linha de produção a probabilidade de se obter uma peça defeituosa seja de 5%. Toma-se uma amostra de 30 peças para serem inspecionadas. Qual a probabilidade de se obter na amostra mais que 2 peças defeituosas?

50- Suponha que numa linha de produção a probabilidade de se obter uma peça defeituosa é de 10%. Toma-se uma amostra de 10 peças para serem inspecionadas. Qual a probabilidade de se obter duas peças defeituosas?

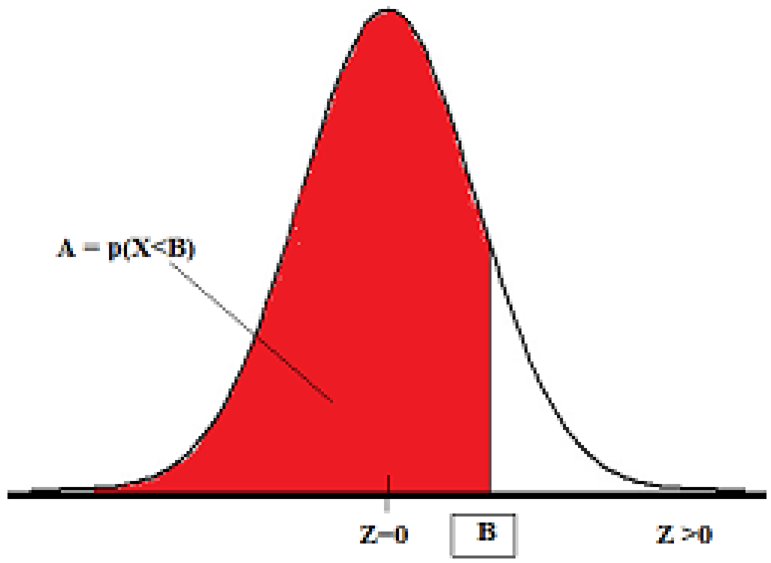
ANEXO C - TABELAS

CURVA NORMAL PADRONIZADA – VALORES SIMÉTRICOS

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3,0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990



a	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
-3,0	0,0013	0,0013	0,0013	0,0012	0,0012	0,0011	0,0011	0,0011	0,0010	0,0010
-2,9	0,0019	0,0018	0,0018	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014
-2,8	0,0026	0,0025	0,0024	0,0023	0,0023	0,0022	0,0021	0,0021	0,0020	0,0019
-2,7	0,0035	0,0034	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027	0,0026
-2,6	0,0047	0,0045	0,0044	0,0043	0,0041	0,0040	0,0039	0,0038	0,0037	0,0036
-2,5	0,0062	0,0060	0,0059	0,0057	0,0055	0,0054	0,0052	0,0051	0,0049	0,0048
-2,4	0,0082	0,0080	0,0078	0,0075	0,0073	0,0071	0,0069	0,0068	0,0066	0,0064
-2,3	0,0107	0,0104	0,0102	0,0099	0,0096	0,0094	0,0091	0,0089	0,0087	0,0084
-2,2	0,0139	0,0136	0,0132	0,0129	0,0125	0,0122	0,0119	0,0116	0,0113	0,0110
-2,1	0,0179	0,0174	0,0170	0,0166	0,0162	0,0158	0,0154	0,0150	0,0146	0,0143
-2,0	0,0228	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188	0,0183
-1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233
-1,8	0,0359	0,0351	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
-1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
-1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
-1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
-1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0721	0,0708	0,0694	0,0681
-1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,0838	0,0823
-1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
-1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
-1,0	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
-0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
-0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
-0,7	0,2420	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,2177	0,2148
-0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
-0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
-0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
-0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
-0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
-0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
0,0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641

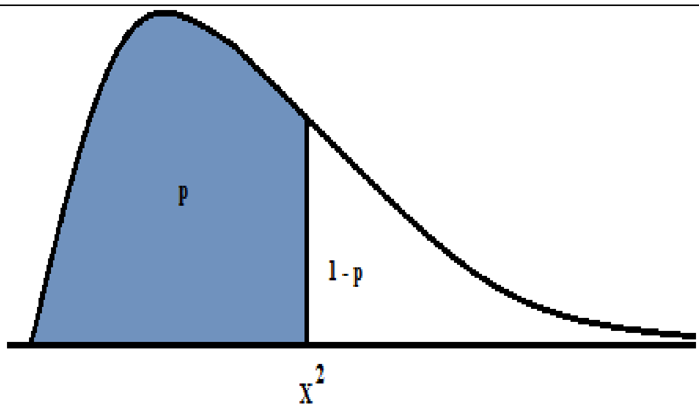


a	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990

TABELA PARA DISTRIBUIÇÃO T STUDENT

gl	Teste Unilateral								
	15%	10%	5%	2,5%	2%	1%	0,5%	0,1%	0,05%
	Teste Bilateral								
	30%	20%	10%	5%	4%	2%	1%	0,2%	0,1%
1	1,9626	3,0777	6,3137	12,7062	15,8945	31,8210	63,6559	318,2888	636,5776
2	1,3862	1,8856	2,9200	4,3027	4,8487	6,9645	9,9250	22,3285	31,5998
3	1,2498	1,6377	2,3534	3,1824	3,4819	4,5407	5,8408	10,2143	12,9244
4	1,1896	1,5332	2,1318	2,7765	2,9985	3,7469	4,6041	7,1729	8,6101
5	1,1558	1,4759	2,0150	2,5706	2,7565	3,3649	4,0321	5,8935	6,8685
6	1,1342	1,4398	1,9432	2,4469	2,6122	3,1427	3,7074	5,2075	5,9587
7	1,1192	1,4149	1,8946	2,3646	2,5168	2,9979	3,4995	4,7853	5,4081
8	1,1081	1,3968	1,8595	2,3060	2,4490	2,8965	3,3554	4,5008	5,0414
9	1,0997	1,3830	1,8331	2,2622	2,3984	2,8214	3,2498	4,2969	4,7809
10	1,0931	1,3722	1,8125	2,2281	2,3593	2,7638	3,1693	4,1437	4,5868
11	1,0877	1,3634	1,7959	2,2010	2,3281	2,7181	3,1058	4,0248	4,4369
12	1,0832	1,3562	1,7823	2,1788	2,3027	2,6810	3,0545	3,9296	4,3178
13	1,0795	1,3502	1,7709	2,1604	2,2816	2,6503	3,0123	3,8520	4,2209
14	1,0763	1,3450	1,7613	2,1448	2,2638	2,6245	2,9768	3,7874	4,1403
15	1,0735	1,3406	1,7531	2,1315	2,2485	2,6025	2,9467	3,7329	4,0728
16	1,0711	1,3368	1,7459	2,1199	2,2354	2,5835	2,9208	3,6861	4,0149
17	1.0690	1.3334	1.7396	2.1098	2.2238	2.5669	2.8982	3.6458	3.9651
18	1,0672	1,3304	1,7341	2,1009	2,2137	2,5524	2,8784	3,6105	3,9217
19	1,0655	1,3277	1,7291	2,0930	2,2047	2,5395	2,8609	3,5793	3,8833
20	1,0640	1,3253	1,7247	2,0860	2,1967	2,5280	2,8453	3,5518	3,8496
21	1,0627	1,3232	1,7207	2,0796	2,1894	2,5176	2,8314	3,5271	3,8193
22	1,0614	1,3212	1,7171	2,0739	2,1829	2,5083	2,8188	3,5050	3,7922
23	1,0603	1,3195	1,7139	2,0687	2,1770	2,4999	2,8073	3,4850	3,7676
24	1,0593	1,3178	1,7109	2,0639	2,1715	2,4922	2,7970	3,4668	3,7454
25	1,0584	1,3163	1,7081	2,0595	2,1666	2,4851	2,7874	3,4502	3,7251
26	1,0575	1,3150	1,7056	2,0555	2,1620	2,4786	2,7787	3,4350	3,7067
27	1,0567	1,3137	1,7033	2,0518	2,1578	2,4727	2,7707	3,4210	3,6895
28	1,0560	1,3125	1,7011	2,0484	2,1539	2,4671	2,7633	3,4082	3,6739
29	1,0553	1,3114	1,6991	2,0452	2,1503	2,4620	2,7564	3,3963	3,6595
30	1,0547	1,3104	1,6973	2,0423	2,1470	2,4573	2,7500	3,3852	3,6460
35	1,0520	1,3062	1,6896	2,0301	2,1332	2,4377	2,7238	3,3400	3,5911

TABELA DE DISTRIBUIÇÃO Qui-QUADRADO



	Nível de Confiança													
GL	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
60	35,534	37,485	40,482	43,188	46,459	52,294	59,335	66,981	74,397	79,082	83,298	88,379	91,952	60
70	43,275	45,442	48,758	51,739	55,329	61,698	69,334	77,577	85,527	90,531	95,023	100,425	104,215	70
80	51,172	53,540	57,153	60,391	64,278	71,145	79,334	88,130	96,578	101,879	106,629	112,329	116,321	80
90	59,196	61,754	65,647	69,126	73,291	80,625	89,334	98,650	107,565	113,145	118,136	124,116	128,299	90
100	67,328	70,065	74,222	77,929	82,358	90,133	99,334	109,141	118,498	124,342	129,561	135,807	140,170	100

TABELA DE NÚMEROS ALEATÓRIOS – GERADAS NO EXCEL

	A	B	C	D	E	F	G	H	I	J	K	L
1	2	8	8	1	5	4	0	1	1	3	3	9
2	8	5	9	1	8	8	1	0	7	3	9	0
3	9	3	3	8	6	8	9	7	8	5	4	8
4	4	0	2	3	4	9	4	5	3	9	2	1
5	0	9	6	4	4	9	3	3	8	8	8	0
6	1	7	4	4	8	6	4	8	5	7	6	0
7	8	1	0	0	6	4	6	3	5	8	2	1
8	9	0	1	0	5	1	1	6	0	1	6	2
9	9	6	2	4	3	3	2	7	2	1	7	4
10	8	6	9	4	7	2	6	6	1	9	4	0
11	9	4	1	2	1	8	6	6	9	3	2	5
12	6	2	5	7	5	9	4	9	6	8	1	1
13	3	7	3	4	2	8	4	4	0	5	3	6
14	0	6	8	2	9	2	8	1	3	8	0	5
15	7	6	2	3	5	1	1	6	5	1	1	5
16	7	5	7	3	9	0	1	9	8	1	0	5
17	5	4	5	0	3	8	7	9	4	4	0	2
18	3	8	0	2	4	2	3	1	9	1	6	8
19	2	8	5	3	8	9	9	3	0	3	7	1
20	9	2	7	2	4	5	8	9	6	2	6	9
21	5	2	0	6	7	5	9	1	1	7	0	0
22	5	8	6	5	7	1	5	9	6	2	8	7
23	7	6	2	4	7	6	3	0	3	2	3	7
24	9	1	3	1	0	9	6	3	7	9	6	2
25	0	6	8	6	0	0	4	7	3	0	5	4
26	0	5	8	1	1	2	2	7	0	9	1	1
27	9	8	8	1	8	9	6	0	7	7	1	4
28	3	7	9	0	0	9	5	1	1	1	3	2
29	4	7	3	6	5	1	4	5	9	8	4	9

REFERÊNCIAS:

BRASIL. Ministério da Educação. Instituto Nacional de Estudos e Pesquisas Educacionais. Censo da educação superior. Disponível em: < http://portal.inep.gov.br/>. Acesso em: 07 jun. 2016.

______. Ministério do Trabalho e Emprego. Classificação Brasileira de Ocupações. CBO 2002. Disponível em: <http://www.mtecbo.gov.br/cbosite/pages/home.jsf> Acesso em: 26 jun. 2016.

BARBETTA, P. A. Estatística Aplicada às Ciências Sociais. Florianópolis: Ed. UFSC, 2011.

BUSSAB, W. Morettin, P. Estatística básica; 5ª ed. São Paulo. Saraiva, 206.

COSTA, S.F. (1992). Introdução Ilustrada à Estatística. 2 ed. São Paulo. Harbra.

CRESPO, Antonio A. Estatística Fácil; 19ª ed. São Paulo. Saraiva, 2009.

DOWNING, Douglas; CLARK, Jeffrey Estatística Aplicada (Série Essencial). 3ª ed. São Paulo. Saraiva, 2010.

FONSECA, J.S. e MARTINS, G.A. Curso de Estatística. 3ª ed. São Paulo. Ed. Atlas,1982.

FREUND, J.E. e SIMON, G.A. Estatística Aplicada. Ed. Bookman, 1999.

HAZAN, Samuel. Fundamentos de Matemática Elementar: Combinatória e Probabilidade. 7 ed. São Paulo. Ed. Atual, 2004.

KAZMIER, Leonard J. Estatística Aplicada à Economia e Administração. Makron, 1982.

GONÇALVES Junior, A.A. Estatística e Metrologia. Notas de aula. Florianópolis. 2012.

LARSON, Ron; FARBER, Betsy Estatística aplicada; 2ª ed. São Paulo: Pearson Prentice Hall, 2004.

MEYER, P.L. Probabilidade: Aplicações à Estatística: 2 ed. Rio de Janeiro: Livros Técnicos e Científicos, 1983.

MONTGOMERY, Douglas C.; RUNGER, George C. Estatística Aplicada e Probabilidade para Engenheiros; 4ª ed. Rio de Janeiro: LTC, 2009.

MORETTIN, P. A.; BUSSAB, W. O. Estatística Básica. São Paulo. Saraiva, 2010.

ROGERSON, Peter A. Métodos estatísticos para a geografia: um guia para o estudante. 7 ed. Porto Alegre. Bookman, 2012.

SALSBURG, David. Uma Senhora Toma Chá...como a estatística revolucionou a ciência no século XX. Rio de Janeiro. Ed. Sahar. 2009.

SPIEGEL, Murray R Estatística. 3ª ed. São Paulo. Pearson, 1994.

STEVENSON, W.J. Estatística Aplicada à Administração. Editora HARBRA, 1986.

STEWART, Ian. Dezessete equações que mudaram o mundo. Rio de Janeiro. Zahar, 2013.

TRIOLA, Mario F. Introdução á Estatística. 7a ed., Rio de Janeiro. LTC, 1999.

SITES DE INTERNET CONSULTADOS

1- VEDUCA – CURSO DE ESTATÍSTICA
https://www.youtube.com/watch?v=VPrM1O--uKk
2- METÓDOS QUANTITATIVOS EM MEDICINA – USP
https://www.youtube.com/watch?list=PLKN-Hz0lVZ-JSq2_ZtaUl2CRdsfqJg7ln&v=U_ivNXumrhw
3- CURSO DE ESTATÍSTICA UNIVESP - TV
https://www.youtube.com/watch?v=K1MXYc_89D8
4- CURSO DE ESTATÍSTICA – IFPR
https://www.youtube.com/watch?v=nK-cHaBNVeQ
5- APRENDA USAR O SOFTWARE R
https://www.youtube.com/watch?time_continue=2&v=wYXpbu-Y370
6- HANS ROSLING
https://www.youtube.com/watch?time_continue=4&v=jbkSRLYSojo
7- COMO PREVER O FUTURO
https://www.youtube.com/watch?v=gAifa_CVGCY
8- CURSO DO M.I.T (EUA)
https://www.youtube.com/watch?list=PLQ3khvAsNhargDx0dG1cQXOrA2u3JsFKc&v=j9WZyLZCBzs

9- REPORTAGEM DA GLOBO NEWS SOBRE BIG DATA:
http://www.youtube.com/watch?v=LsMt5jp1a9k
10 – O PRAZER DA ESTATÍSTICA:
http://www.youtube.com/watch?v=AfYVOsuT-EI
11- O QUE É ESTATÍSTICA:
http://www.youtube.com/watch?v=-Wm9cxiXUe0
12- VOCAÇÃO – ESTATÍSTICA:
http://www.youtube.com/watch?v=vwo3GzKuNXo
13- AULAS DE ESTATÍSTICA E PROBABILIDADE DO VEDUCA:
http://www.veduca.com.br/play/7026
14- KHAN ACADEMY:
https://www.khanacademy.org/math/probability/independent-dependent-probability/old_prob_videos/v/introduction-to-random-variables?playlist=Statistics
15- DISTRIBUIÇÃO DE PROBABILIDADE BINOMIAL
http://www.youtube.com/watch?v=ConmIDAzRqI&feature=youtu.be
16- O CRESCIMENTO DA POPULAÇÃO MUNDIAL – ANÁLISE ESTATÍSTICA
http://www.youtube.com/watch?v=RuGTZEXh6yw
17- AULA DE ESTATÍSTICA DA RNP:
Curso Estatística RNP
18- ESTATÍSTICA DESCRITIVA:
http://www.youtube.com/watch?v=l2MyLvp82Rg

19 – TEOREMA DO LIMITE CENTRAL 2:
http://www.youtube.com/watch?v=zEwT_fIpSBE
20- AS MELHORES ESTATÍSTICAS QUE VOCÊ JÁ VIU.
http://www.youtube.com/watch?v=HQPSRHncJLo
21- ESTATÍSTICAS E O PODER DA MÁQUINA DE LAVAR ROUPA
http://www.youtube.com/watch?v=khsq7nHAveA
22- COEFICIENTE DE CORRELAÇÃO E REGRESSÃO
http://www.youtube.com/watch?v=ODGzDA4zAq8
23- COMO SÃO REALIZADAS AS PESQUISAS ELEITORAIS:
http://www.yotube.com/watch?v=mWI8QM-HoeU&feature=youtu.be

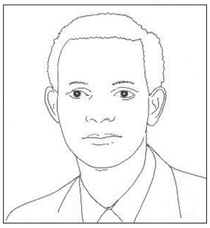
EXERCÍCIOS RESOLVIDOS - EDUCREATIONS

- Aula 43 - Uso Da Curva Normal
- Aula 44 - Distribuição Normal
- Aula 45 - Distribuição Binomial
- Aula 46 - Probabilidade Binomial
- Aula 47 - Distribuição Binomial
- Aula 48 - Distribuição Binomial
- Aula 49 - Exercícios Resolvidos - Binomial E Probabilidade Condicional
- Aula 50 - Exercícios
- Aula 51 - Probabilidade
- Aula 52 - Inferência Estatística

EXERCÍCIOS RESOLVIDOS - EDUCREATIONS

– Aula 1 - Exercícios de Probabilidade
– Aula 2 - Exercícios de Probabilidade
– Aula 3 - Probabilidade de eventos não exclusivos
– Aula 4 - Probabilidade Condicional 1
– Aula 5 - Probabilidade Condicional 2
– Aula 6 - Probabilidade Condicional 3
– Aula 7 - Probabilidade Condicional 4
– Aula 8 - Probabilidade Condicional 5
– Aula 9 - Exercícios Gerais
– Aula 10 - Aplicando distribuição de probabilidades binomial 1
– Aula 11 - Cálculo de probabilidades usando diagrama de Veen
– Aula 12 - Distribuição probabilidades binomial
– Aula 13 - Distribuição de probabilidades binomial
– Aula 14 - Organização de dados e construção de diagrama de caixa (Quartil e Box Plot)
– Aula 15 - Cálculo de probabilidades usando curva normal
– Aula 16 - Calculando probabilidades com curva normal
– Aula 17 - Organização de dados em quartis e construção de diagrama de caixa
– Aula 18 - Probabilidade de obter bolas da mesma cor de uma urna
– Aula 19 - Média e desvio padrão a partir de um histograma
– Aula 20 - Poisson
– Aula 21 - Distribuição Normal
– Aula 22 - Distribuição normal
– Aula 23 - Média, moda e diagrama de caixa
– Aula 24 - Distribuição de Poisson
– Aula 25 - Distribuição binomial
– Aula 26 - Construção de diagrama de caixa
– Aula 27 - Aproximação da distribuição binomial como uma normal
– Aula 28 - Teorema Do Limite Central
– Aula 29 - Exercício de probabilidade
– Aula 30 - Probabilidade binomial aplicada ao controle estatístico de processos
– Aula 31 - Correlação entre idade e altura de crianças
– Aula 32 - Distribuição de Poisson
– Aula 33 - Probabilidade de erros em um módulo
– Aula 34 - Diagrama De Veen
– Aula 35 - Eventos
– Aula 36 - Usando Curva Normal
– Aula 37 - Aproximação Normal

- Aula 38 - Aproximação Normal
- Aula 39 - Usando Curva Normal
- Aula 40 - Construção De Histograma
- Aula 41 - Construção De Histograma
- Aula 42 - Usando Curva Normal



Jesué Graciliano da Silva, natural de Marília (SP), é Engenheiro Mecânico graduado pela Universidade Federal de Santa Catarina, no ano de 1993. Sua carreira profissional iniciou 10 anos antes como desenhista em um escritório de engenharia, profissão que lhe permitiu custear seus estudos superiores. Possui especialização em Engenharia de Segurança do Trabalho pela UFSC (1994-1995) e Curso “Escola de Governo” pela UDESC (1995). Concluiu em 1999 o mestrado na UFSC, na área de Ciências Térmicas (POSMEC).

Desde 1993, é professor efetivo do atual Instituto Federal de Educação, Ciência e Tecnologia de Santa Catarina – Câmpus São José, onde atua na Área Técnica de Refrigeração e Condicionamento de Ar e no Curso de Engenharia de Telecomunicações, nas disciplinas de Projetos, Termodinâmica, Transferência de Calor, Mecânica dos Fluidos, Estatística, Mecânica dos Sólidos, Desenho Técnico e Instalações de Refrigeração e Ar-Condicionado.

De 2003 a 2006, foi Diretor do Câmpus São José. Atuou como Diretor de Gestão do Conhecimento do IFSC de fevereiro de 2008 a fevereiro de 2009. A partir de fevereiro de 2009, assumiu a função de Pró-Reitor de Desenvolvimento do IFSC. De junho a dezembro de 2011, atuou como Reitor pro tempore do Instituto Federal de Santa Catarina. De fevereiro a maio de 2012, atuou como Ouvidor-Geral do IFSC. De junho a outubro de 2012 atuou como Reitor *pro tempore* do IF-Farroupilha. De agosto de 2013 a janeiro de 2014 atuou como Reitor *pro tempore* do IF Paraná.

É autor dos livros “Introdução à Tecnologia da Refrigeração e Climatização” pela Editora Artliber e “Liderança Ética e Servidora” pela Editora do IFSC. É também coautor dos livros: “Do Discurso à Ação – uma experiência de gestão participativa na educação pública” - Editora Nova Letra, “Desenho Técnico para Refrigeração e Climatização” - Amazon, “Instalação de climatizadores tipo Splits na Prática” - Amazon, “Refrigeração e Climatização na Prática” - Amazon, e do livro-blog “Transformação do CEFET-SC em IFSC, concepções, conquistas e desafios”. Atualmente, está realizando doutorado no Programa de Pós-Graduação em Geografia – Área de Desenvolvimento Regional e Urbano na UFSC. Sua pesquisa é sobre a expansão da Rede Federal de Educação Profissional, Científica e Tecnológica no Brasil e suas implicações socioespaciais no Estado de Santa Catarina.