MAF 105 - Estatística Básica

Prof. Fernando de Souza Bastos

Instituto de Ciências Exatas e Tecnológicas
Universidade Federal de Viçosa
Campus UFV - Florestal

07/08/2018

Sumário

Introdução

Em alguma fase de seu trabalho, o pesquisador depara-se com o problema de analisar e entender um conjunto de dados relevante ao seu particular objeto de estudos. Ele necessitará trabalhar os dados para transformá-los em informações, para compará-los com outros resultados, ou ainda para julgar sua adequação a alguma teoria.

Não Tem Como Escapar dos Dados!

Introdução

BIG
DATA
STORAGE
SHARING
DATA
TOOLS
TRANSFER
ANALYSIS
TERM
COLLECTION
ANY
BECOMES
BLANKET
INCLUDE
CAPTURE
DEPENDENCIES
HIGH
TRENDS
INSIGHT
LOGS
TYPES
HANDLE
THINGS
STREAM
ENABLE
TRIGGER
ENHANCED
CAPABILITIES
AREAS
WIRELESS
OUT
DENSITY
TIME
ENTERPRISE
BUSINESS
DISCOVERY
SOFTWARE
PHYSICS
HUNDREDS
RESEARCH
SIMULATIONS
MAKING
NETWORKS
MICROPHONES
PROCESSING
TAKE
DECISION
INTERNET
SCIENTISTS
PACKAGES
SPEED
PACKAGES
GENOMICS
FINANCE
PROCESS
REPORT
BIG
CURATE
OPTIMIZATION
TERABYTES
DATA
CAPTURE
INFORMATION
LARGE
CURATION
MANAGEMENT
PROCESSING
AND
VELOCITY
VARIETY
SENSOR
CHALLENGES
USING
TRADITIONAL
ON-HAND
3V
FOR
REQUIRE
PETABYTES
AMOUNT
IN
STATISTICS
INFORMATICS
DOMAIN
DIFFICULT
PROCESS
DATABASE
RELATIONAL
METEOROLOGY
RANGE
VELOCITY
GROWS
SIMULATION
EXPLORE
THAT
GARTNER
COMPLEX
ASSETS
VOLUME
THAT
CAMERAS
AND
BYTES
PROCESS
APPLICATIONS
IS A
ENVIRONMENTAL
TOLERABLE
CONNECTOMICS
PROBLEMS
BUSINESS INTELLIGENCE
FORMS
NEW
SOURCES
DOUG
LANEY
MANAGE
GOVERNMENT
SEARCH
FLOW
SYSTEMS
SIZE
VISUALIZATION

Introdução

- Google: são 3,5 bilhões de buscas por dia.

Introdução

- Google: são 3,5 bilhões de buscas por dia.
- YouTube: mais de 1 bilhão de usuários e são assistidas 100 milhões de horas de vídeos por dia.

Introdução

- Google: são 3,5 bilhões de buscas por dia.
- YouTube: mais de 1 bilhão de usuários e são assistidas 100 milhões de horas de vídeos por dia.
- Facebook: 1,5 bilhões de usuários ativos em um mês.

Introdução

- Google: são 3,5 bilhões de buscas por dia.
- YouTube: mais de 1 bilhão de usuários e são assistidas 100 milhões de horas de vídeos por dia.
- Facebook: 1,5 bilhões de usuários ativos em um mês.
- Instagram: 3,5 bilhões de likes e 80 milhões de fotos carregadas em um dia e 40 bilhões de fotos compartilhadas no total.

Introdução

- Google: são 3,5 bilhões de buscas por dia.
- YouTube: mais de 1 bilhão de usuários e são assistidas 100 milhões de horas de vídeos por dia.
- Facebook: 1,5 bilhões de usuários ativos em um mês.
- Instagram: 3,5 bilhões de likes e 80 milhões de fotos carregadas em um dia e 40 bilhões de fotos compartilhadas no total.
- WhatsApp: mais de 300 bilhões de mensagens por dia e 700 milhões de usuários por mês.

Introdução

- Google: são 3,5 bilhões de buscas por dia.
- YouTube: mais de 1 bilhão de usuários e são assistidas 100 milhões de horas de vídeos por dia.
- Facebook: 1,5 bilhões de usuários ativos em um mês.
- Instagram: 3,5 bilhões de likes e 80 milhões de fotos carregadas em um dia e 40 bilhões de fotos compartilhadas no total.
- WhatsApp: mais de 300 bilhões de mensagens por dia e 700 milhões de usuários por mês.
- Twitter: 500 milhões de tweets por dia e 316 milhões de usuários ativos por mês.

Introdução

- Internet: quase 950 milhões de sites e 3,2 bilhões de pessoas conectadas.

Introdução

- Internet: quase 950 milhões de sites e 3,2 bilhões de pessoas conectadas.
- Telefones celulares: mais de 7,5 bilhões.

Introdução

- Internet: quase 950 milhões de sites e 3,2 bilhões de pessoas conectadas.
- Telefones celulares: mais de 7,5 bilhões.
- Dispositivos conectados: serão 50 bilhões em 2050.

Introdução

- Internet: quase 950 milhões de sites e 3,2 bilhões de pessoas conectadas.
- Telefones celulares: mais de 7,5 bilhões.
- Dispositivos conectados: serão 50 bilhões em 2050.
- São gerados 2,5 exabytes ($10^{18} bytes$) de dados por dia que dobram a cada 40 meses.

A IDC (empresa líder em inteligência de mercado e consultoria nas indústrias de tecnologia da informação, telecomunicações e mercados de consumo em massa de tecnologia) estima que, do total de dados no mundo, 22% contêm informação útil. E apenas 5% foram analisados e utilizados de alguma forma. [citar EXAME]

Introdução

Como processar tanta informação? Como gerar informação a partir dos dados? Essa não é uma tarefa fácil, é necessário, anos de estudo e variadas competências, Estatística, Matemática, Ciência da Computação e diversas outras. Mas o profissional que tem conhecimento estatístico e sabe avaliar processos ou prever possíveis resultados de ações é um dos profissionais mais requisitados na atualidade e recebe os maiores salários!

Introdução

Engenheiro ou Cientista de Dados

- **O que faz:** combina habilidades em negócios e estatística. É o profissional responsável por solucionar problemas do negócio com técnicas de orientação a dados, bem como detectar tendências que podem ajudar nos resultados de uma empresa
- **Perfil:** qualificações estatísticas, matemáticas e curiosidade para fazer descobertas em big data
- **Salário:** R$ 9 mil a R$ 15 mil

Fonte: https://epocanegocios.globo.com/Carreira/noticia/2017/12/profissoes-que-estarao-em-alta-no-brasil-em-2018.html

Introdução

Você precisa ser especialista em Estatística ou Matemática ou mesmo ter feito uma graduação nestas áreas?

Introdução

Você precisa ser especialista em Estatística ou Matemática ou mesmo ter feito uma graduação nestas áreas? A resposta é **não**.

Introdução

Você precisa ser especialista em Estatística ou Matemática ou mesmo ter feito uma graduação nestas áreas? A resposta é **não**. Apesar dessas áreas permitirem uma compreensão mais abrangente, é possível aprender estes conceitos e aplica-los, ao longo da sua jornada de aprendizagem. Você não precisa aprender todos os tópicos relacionados à Estatística ou Matemática.

Fonte: http://datascienceacademy.com.br/blog/cientista-de-dados-por-onde-comecar-em-8-passos/

Introdução

Estatística é a ciência que nos ajuda a tomar decisões e tirar conclusões na presença de variabilidade. A estatística é uma maneira de raciocinar que pode ajudar você a tomar decisões mais bem fundamentadas. A estatística ajuda você a solucionar problemas que envolvem decisões que estão baseadas em dados que tenham sido coletados.

Introdução

Ressalto que a Estatística é a ciência que ensina a "ESCUTAR"os dados, não é uma ciência para provar alguma coisa em relação ao que você deseja que os dados digam!

Introdução

Ressalto que a Estatística é a ciência que ensina a "ESCUTAR"os dados, não é uma ciência para provar alguma coisa em relação ao que você deseja que os dados digam!

A estatística é a arte de torturar os números até que eles confessem!
*Darrell Huff's - Como mentir com estatísticas (1954)

Introdução

Na primeira parte deste curso estaremos interessados na redução, análise e interpretação dos dados sob consideração, adotando um enfoque que chamaremos de análise exploratória de dados (AED). Nessa abordagem tentaremos obter dos dados a maior quantidade possível de informação, que indique modelos plausíveis a serem utilizados numa fase posterior, a análise confirmatória de dados (ou inferência estatística).

Introdução

Tradicionalmente, uma análise descritiva de dados limita-se a calcular algumas medidas de posição e variabilidade, como a média e variância, por exemplo. Contrária a essa tendência, uma corrente mais moderna, utiliza principalmente técnicas gráficas, em oposição a resumos numéricos. Isso não significa que sumários não devam ser obtidos, mas uma análise exploratória de dados não deve se limitar a calcular tais medidas.

Modelos

Fundamentalmente, quando se procede a uma análise de dados, busca-se alguma forma de regularidade ou padrão ou, ainda, um modelo, presente nas observações.

Modelos

Exemplo

Imagine que estejamos estudando a relação entre rendimentos e gastos de consumo de um conjunto de indivíduos. Podemos obter um gráfico como o da Figura 1. O que se espera, intuitivamente, é que os gastos de um indivíduo estejam diretamente relacionados com os seus rendimentos, de modo que é razoável supor uma "relação linear" entre essas duas quantidades. Os pontos da Figura 1 não estão todos, evidentemente, sobre uma reta; essa seria o nosso padrão ou modelo. A diferença entre os dados e o modelo constitui os resíduos.

Modelos

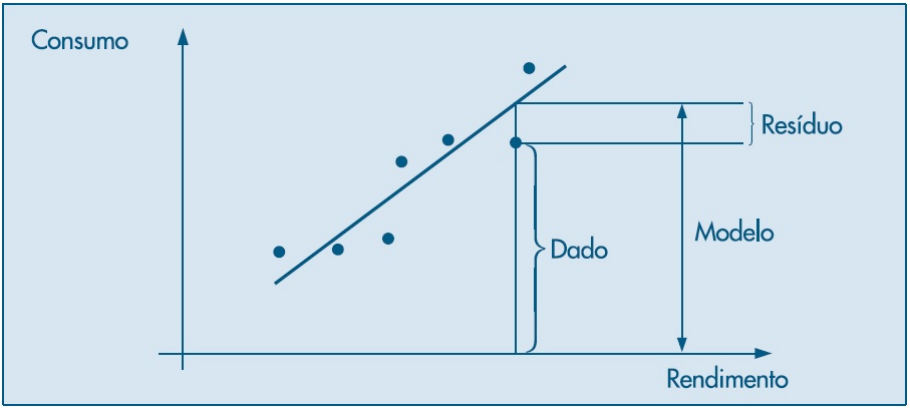


Figura: Relação entre consumo e rendimento.

Modelos

De modo esquemático:

$$\text{Dados} = \text{Modelo} \ + \ \text{Resíduos}, \quad \text{ou, ainda,} \quad D = M + R \qquad (1)$$

Modelos

De modo esquemático:

$$\text{Dados} = \text{Modelo} + \text{Resíduos}, \quad \text{ou, ainda,} \quad D = M + R \qquad (1)$$

A parte M é também chamada parte suave (ou regular ou, ainda, previsível) dos dados, enquanto R é a parte aleatória. A parte R é tão importante quanto M, e a análise dos resíduos constitui uma parte fundamental de todo trabalho estatístico. Basicamente, são os resíduos que nos dizem se o modelo é adequado ou não para representar os dados. De modo coloquial, o que se deseja é que a parte R não contenha nenhuma "suavidade", caso contrário mais "suavização" é necessária.

Modelos

Uma análise exploratória de dados busca, essencialmente, fornecer informações para estabelecer (1).

Técnicas Computacionais

O desenvolvimento rápido e constante na área de computação foi acompanhado pela introdução de novas técnicas de análise de dados, notadamente de métodos gráficos e de métodos chamados de computação intensiva.

Técnicas Computacionais

Para a implementação dessas técnicas, foram desenvolvidos programas estatísticos, atualmente usados em larga escala tanto no meio acadêmico como em indústrias, bancos, órgãos de governo etc. Esses programas podem ser genéricos ou específicos. Os programas genéricos (como o R, Minitab, Splus, SPSS, SAS etc.) são adequados para realizar uma gama variada de análises estatísticas. Os pacotes específicos são planejados para realizar análises particulares de uma determinada área. O Excel não é considerado um software estatístico mas pode utilizado em diversas ocasiões para análise estatística.

Técnicas Computacionais

Por outro lado, os programas podem exigir maior ou menor experiência computacional dos usuários. Alguns operam com menus, e seu uso é mais simples. Outros requerem maior familiaridade com o computador e são baseados em linguagens próprias.

Técnicas Computacionais

Além dos pacotes estatísticos, há outros pacotes de grande utilidade para realizar tarefas matemáticas. Dentre estes, mencionamos o Mathematica, o Maple, o Gauss e o MatLab. Existe também o Wolfram Alpha que é um mecanismo de conhecimento computacional que funciona on-line. Além disso, nesse curso, iremos utilizar o Latex como editor de texto.

Métodos Gráficos

Os métodos gráficos têm encontrado um uso cada vez maior devido ao seu forte apelo visual. Normalmente, é mais fácil para qualquer pessoa entender a mensagem de um gráfico do que aquela embutida em tabelas ou sumários numéricos.

Os gráficos são utilizados para:

- buscar padrões e relações;

Os gráficos são utilizados para:

- buscar padrões e relações;
- confirmar (ou não) certas expectativas que se tinha sobre os dados;

Os gráficos são utilizados para:

- buscar padrões e relações;
- confirmar (ou não) certas expectativas que se tinha sobre os dados;
- descobrir novos fenômenos;

Os gráficos são utilizados para:

- buscar padrões e relações;
- confirmar (ou não) certas expectativas que se tinha sobre os dados;
- descobrir novos fenômenos;
- confirmar (ou não) suposições feitas sobre os procedimentos estatísticos usados; e

Os gráficos são utilizados para:

- buscar padrões e relações;
- confirmar (ou não) certas expectativas que se tinha sobre os dados;
- descobrir novos fenômenos;
- confirmar (ou não) suposições feitas sobre os procedimentos estatísticos usados; e
- apresentar resultados de modo mais rápido e fácil.

Podemos usar métodos gráficos para plotar os dados originais ou outros dados derivados deles. Por exemplo, a investigação da relação entre as variáveis da Figura (1) pode ser feita por meio daquele diagrama de dispersão. Mas podemos também "ajustar" uma reta aos dados, calcular o desvio (resíduo) para cada observação e fazer um novo gráfico, de consumo contra resíduos, para avaliar a qualidade do ajuste.

Conjuntos de Dados

Na página da disciplina (https://maf105.github.io/) aparecem alguns conjuntos de dados que serão utilizados nos exemplos ou nos exercícios propostos. Aconselho a todos a reproduzir os exemplos, usando esses dados, bem como resolver os problemas, pois somente a efetiva manipulação de dados pode levar a um bom entendimento das técnicas apresentadas.

Os conjuntos de dados apresentados provêm de diferentes fontes, que são mencionadas em cada conjunto e depois explicitadas nas referências. usaremos para as análises estatísticas o software R, calculadora científica e planilhas do Excel.

Mãos à obra!!!

Tipo de Variáveis

Conhecer o tipo da variável é importante porque os métodos estatísticos que você pode utilizar em sua análise variam de acordo com o tipo. A natureza dos valores correspondentes aos dados associados à variável determina o seu respectivo tipo. Existem dois tipos principais para variáveis:

- **Variáveis categóricas** (ou **Variáveis qualitativas**) apresentam valores que podem somente ser posicionados em categorias tais como sim e não. "Você tem um perfil no Facebook?" (sim ou não) e que ano está cursando na faculdade (Primeiro Ano, Segundo Ano, Terceiro Ano ou Quarto Ano) são exemplos de variáveis categóricas.

Tipo de Variáveis

Conhecer o tipo da variável é importante porque os métodos estatísticos que você pode utilizar em sua análise variam de acordo com o tipo. A natureza dos valores correspondentes aos dados associados à variável determina o seu respectivo tipo. Existem dois tipos principais para variáveis:

- **Variáveis categóricas** (ou **Variáveis qualitativas**) apresentam valores que podem somente ser posicionados em categorias tais como sim e não. "Você tem um perfil no Facebook?" (sim ou não) e que ano está cursando na faculdade (Primeiro Ano, Segundo Ano, Terceiro Ano ou Quarto Ano) são exemplos de variáveis categóricas.
- **Variáveis numéricas** (ou **Variáveis quantitativas**) apresentam valores que representam quantidades.

Tipo de Variáveis

Dentre as variáveis qualitativas, ainda podemos fazer uma distinção entre dois tipos: variável qualitativa nominal, para a qual não existe nenhuma ordenação nas possíveis realizações, e variável qualitativa ordinal, para a qual existe uma ordem nos seus resultados.

Variáveis numéricas podem ser, ainda, identificadas como variáveis discretas ou variáveis contínuas.

Variáveis numéricas podem ser, ainda, identificadas como variáveis discretas ou variáveis contínuas.

Variáveis discretas apresentam valores numéricos que surgem a partir de um processo de contagem. "A quantidade de canais de TV a Cabo Premium que você assina" é um exemplo de uma variável numérica discreta, uma vez que a resposta corresponde a um entre uma quantidade finita de números inteiros. Você assina zero, um, dois ou mais canais. "A quantidade de itens que um consumidor adquire" também corresponde a uma variável numérica discreta, uma vez que você está contando o número de itens comprados.

Variáveis contínuas produzem respostas numéricas que surgem a partir de um processo de medição. O tempo que você espera pelo atendimento de um caixa no banco é um exemplo de variável numérica contínua, uma vez que a resposta pode assumir qualquer valor dentro dos limites de um continuum, ou de um intervalo, dependendo da precisão do instrumento de medição. Por exemplo, o seu tempo de espera poderia ser 1 minuto 1,1 minuto, 1,11 minuto ou 1,113 minuto, dependendo da precisão do dispositivo de medição utilizado. (Teoricamente, dois valores contínuos jamais poderiam ser absolutamente idênticos. No entanto, uma vez que nenhum dispositivo de medição é perfeitamente preciso, podem ocorrer valores contínuos idênticos para dois ou mais itens ou indivíduos.)

Em uma primeira análise, identificar o tipo da variável pode parecer fácil, embora algumas variáveis que você poderia desejar estudar possam ser categóricas ou numéricas, dependendo do modo como você as define. Por exemplo, "idade" aparentaria ser uma variável numérica evidente, mas o que acontece se você estiver interessado em comparar os hábitos de compra de crianças, adolescentes, pessoas de meia-idade e pessoas com idade para aposentadoria? Nesse caso, definir "idade" como uma variável categórica faria mais sentido. Mais uma vez, isso ilustra o ponto anterior de que, sem definições operacionais, as variáveis não têm nenhum significado.

Introdução

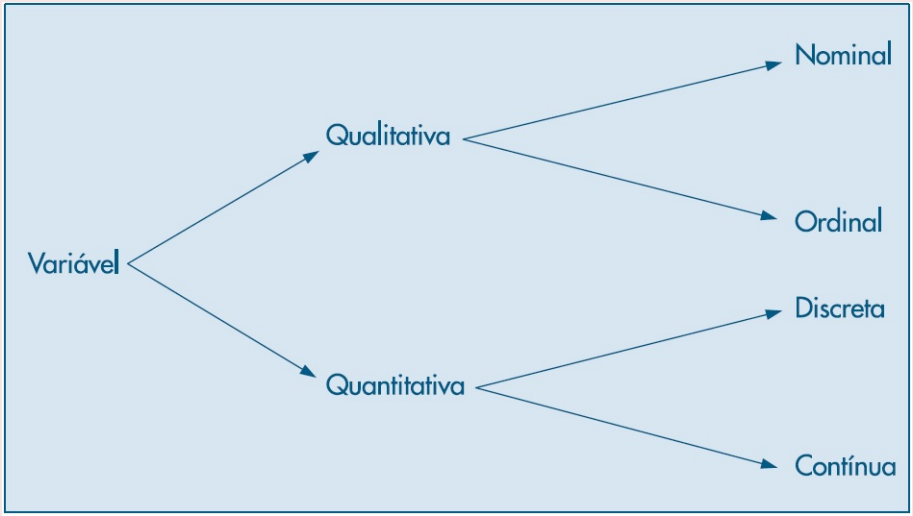


Figura: Classificação de uma variável.

Exemplo

Um pesquisador está interessado em fazer um levantamento sobre alguns aspectos socioeconômicos dos empregados da seção de orçamentos da Companhia MB. Usando informações obtidas do departamento pessoal, ele elaborou a Tabela abaixo.

Variável	Representação
Estado Civil	X
Grau de Instrução	Y
Número de filhos	Z
Salário	S
Idade	U
Região de Procedência	V

Distribuições de Freqüências

Quando se estuda uma variável, o maior interesse do pesquisador é conhecer o comportamento dessa variável, analisando a ocorrência de suas possíveis realizações. Vejamos uma maneira de se dispor um conjunto de realizações, para se ter uma idéia global sobre elas, ou seja, de sua distribuição.

Tabela: Freqüências e porcentagens dos 36 empregados da seção de orçamentos da Companhia MB segundo o grau de instrução.

Grau de instrução	Frequência n_i	Proporção f_i	Porcentagem $100f_i$
Fundamental	12	0,3333	33,33
Médio	18	0,5000	50,00
Superior	6	0,1667	16,67
Total	36	1,0000	100,00

Observando os resultados da segunda coluna, vê-se que dos 36 empregados da companhia, 12 têm o ensino fundamental, 18 o ensino médio e 6 possuem curso superior.

Uma medida bastante útil na interpretação de tabelas de freqüências é a proporção de cada realização em relação ao total, pois podem ser utilizadas quando se quer comparar resultados de duas pesquisas distintas.

Uma medida bastante útil na interpretação de tabelas de freqüências é a proporção de cada realização em relação ao total, pois podem ser utilizadas quando se quer comparar resultados de duas pesquisas distintas.

Por exemplo, suponhamos que se queira comparar a variável grau de instrução para empregados da seção de orçamentos com a mesma variável para todos os empregados da Companhia MB. Digamos que a empresa tenha 2.000 empregados e que a distribuição de freqüências seja a da próxima Tabela.

Tabela: Freqüências e porcentagens dos dos 2.000 empregados da seção de orçamentos da Companhia MB segundo o grau de instrução.

Grau de instrução	Frequência n_i	Proporção f_i	Porcentagem $100f_i$
Fundamental	650	0.325	32.50
Médio	1.020	0.51	51.00
Superior	330	0.165	16.50
Total	2.000	1,0000	100.00

Não podemos comparar diretamente as colunas das freqüências das Tabelas 1 e 2, pois os totais de empregados são diferentes nos dois casos. Mas as colunas das porcentagens são comparáveis, pois reduzimos as freqüências a um mesmo total (no caso 100).

A construção de tabelas de freqüências para variáveis contínuas necessita de certo cuidado. Por exemplo, a construção da tabela de freqüências para a variável salário, usando o mesmo procedimento acima, não resumirá as 36 observações num grupo menor, pois não existem observações iguais. A solução empregada é agrupar os dados por faixas de salário.

Tabela: Freqüências e porcentagens dos dos 2.000 empregados da seção de orçamentos da Companhia MB por faixa de salário.

Classe de Salários	Frequência n_i	Porcentagem $100f_i$
4.00|-8.00	10	27.78
8.00|-12.00	12	33.33
12.00|-16.00	8	22.22
16.00|-20.00	5	13.89
20.00|-24.00	1	2.78
Total	36	100.00

Procedendo-se desse modo, ao resumir os dados referentes a uma variável contínua, perde-se alguma informação. Por exemplo, não sabemos quais são os oito salários da classe de 12 a 16, a não ser que investiguemos a tabela original. Sem perda de muita precisão, poderíamos supor que todos os oito salários daquela classe fossem iguais ao ponto médio da referida classe, isto é, 14.

Note que estamos usando a notação $a| - b$ para o intervalo de números contendo o extremo a mas não contendo o extremo b. Podemos também usar a notação $[a, b)$ para designar o mesmo intervalo $a| - b$.

A escolha dos intervalos é arbitrária e a familiaridade do pesquisador com os dados é que lhe indicará quantas e quais classes (intervalos) devem ser usadas.

- Número pequeno de classes $\rightarrow$ perda de informação;
- Número grande de classes $\rightarrow$ perda da visão geral dos dados como um conjunto;
- A sugestão é usar de 5 a 15 classes com a mesma amplitude;

Para construir uma distribuição de frequências separando por classes uma determinada variável podemos utilizar:

- número de classes$(n_c) \approx \sqrt{n}$ ou usamos a regra de Sturges $n_c = \ln(n)$;
- Amplitude da classe $= \dfrac{AT}{n_c}$

em que $AT = \text{Maior valor} - \text{Menor valor}.$

Um procedimento alternativo para resumir um conjunto de valores, com o objetivo de se obter uma idéia da forma de sua distribuição, é o ramo-e-folhas. Uma vantagem deste diagrama é que não perdemos (ou perdemos pouca) informação sobre os dados em si.

Diagrama de ramos e folhas para variáveis contínuas

Quando o número de observações é relativamente grande, este diagrama pode ser útil.

Tabela: Diagrama de Ramos e Folhas da idade

Ramo	Folhas
2	0 3 5 6 6 7 8 9
3	0 1 1 2 2 3 3 4 4 5 5 6 6 7 7 8 9
4	0 0 1 1 2 3 3 4 6 8

Tabela: Diagrama de Ramos e Folhas dos Salários ($\times$ sal. Min)

Ramo	Folhas			
4	00	56		
5	25	73		
6	26	66	86	
7	39	44	59	
8	12	46	74	95
9	13	35	77	80
10	53	76		
11	06	59		
12	00	79		
13	23	60	85	
14	69	71		
15	99			
16	22		61	
17	26			
18	75			
19	40			
20				
21				
22				
23		30		

Algumas informações que se obtêm deste ramo-e-folhas são:

1. Há um destaque grande para o valor 23,30.

Algumas informações que se obtêm deste ramo-e-folhas são:

1. Há um destaque grande para o valor 23,30.
2. Os demais valores estão razoavelmente concentrados entre 4,00 e 19,40.

Algumas informações que se obtêm deste ramo-e-folhas são:

1. Há um destaque grande para o valor 23,30.
2. Os demais valores estão razoavelmente concentrados entre 4,00 e 19,40.
3. Um valor mais ou menos típico para este conjunto de dados poderia ser, por exemplo, 10,00.

Algumas informações que se obtêm deste ramo-e-folhas são:

1. Há um destaque grande para o valor 23,30.
2. Os demais valores estão razoavelmente concentrados entre 4,00 e 19,40.
3. Um valor mais ou menos típico para este conjunto de dados poderia ser, por exemplo, 10,00.
4. Há uma leve assimetria em direção aos valores grandes; a suposição de que estes dados possam ser considerados como amostra de uma população com distribuição simétrica, em forma de sino (a chamada distribuição normal), pode ser questionada.

A representação gráfica da distribuição de uma variável tem a vantagem de, rápida e concisamente, informar sobre sua variabilidade. Existem vários gráficos que podem ser utilizados. Vejamos alguns!

Gráficos para Variáveis Qualitativas

Existem vários tipos de gráficos para representar variáveis qualitativas. Vários são versões diferentes do mesmo princípio, logo nos limitaremos a apresentar dois deles: gráficos em barras e de composição em setores ("pizza" ou retângulos).

Tomemos como ilustração a variável Y: grau de instrução, exemplificada nas Tabelas 2.2 e 2.3. O gráfico em barras consiste em construir retângulos ou barras, em que uma das dimensões é proporcional à magnitude a ser representada (n_i ou f_i), sendo a outra arbitrária, porém igual para todas as barras. Essas barras são dispostas paralelamente umas às outras, horizontal ou verticalmente. Na próxima Figura temos o gráfico em barras (verticais) para a variável Y.

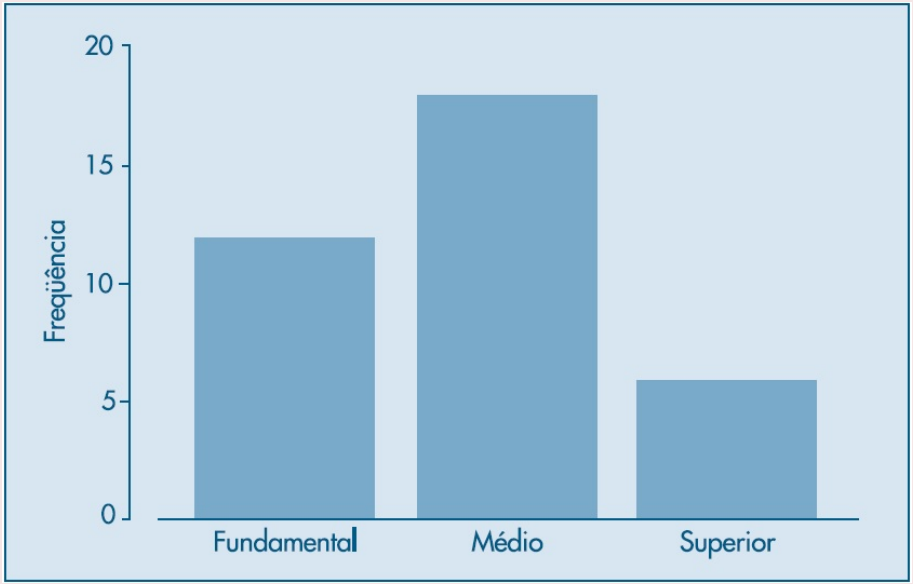


Figura: Gráfico em barras para a variável Y : grau de instrução.

Já o gráfico de composição em setores, sendo em forma de "pizza" o mais conhecido, destina-se a representar a composição, usualmente em porcentagem, de partes de um todo. Consiste num círculo de raio arbitrário, representando o todo, dividido em setores, que correspondem às partes de maneira proporcional. A próxima Figura mostra esse tipo de gráfico para a variável Y.

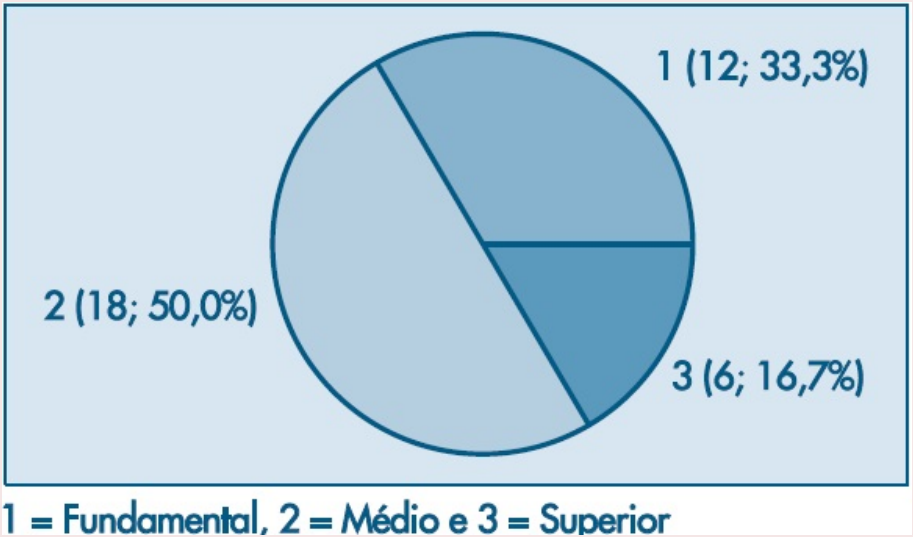


Figura: Gráfico em setores para a variável Y : grau de instrução.

Gráficos para Variáveis Quantitativas

Para variáveis quantitativas podemos considerar uma variedade maior de representações gráficas. Podemos considerar gráficos de barra, gráfico de pontos, gráficos com barras empilhadas, gráficos de dispersão e outros. Mas para resumir os dados, o mais utilizado é o histograma.

O histograma é um gráfico de barras contíguas, com as bases proporcionais aos intervalos das classes e a área de cada retângulo proporcional à respectiva frequência. Pode-se usar tanto a frequência absoluta, f_i, como a relativa, f_{ri}.

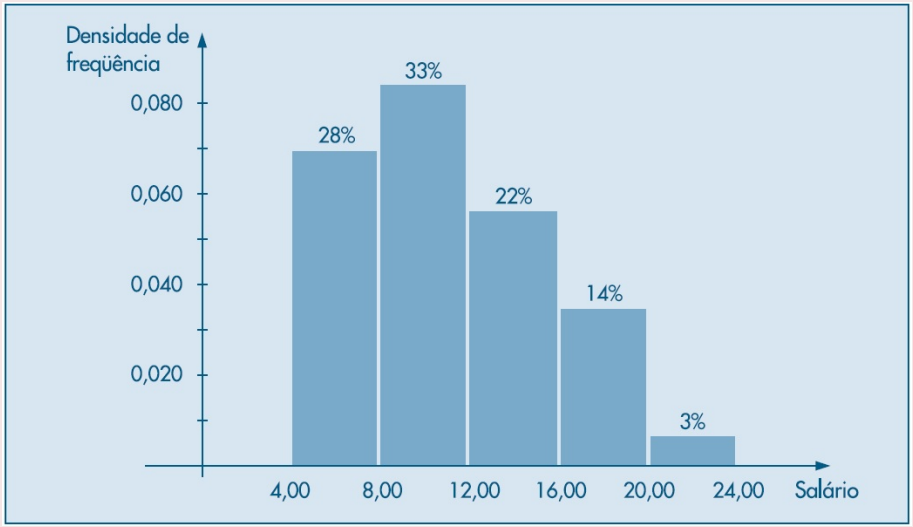


Figura: Histograma da variável S : Salários.

Para facilitar o entendimento, foi colocada acima de cada setor (retângulo) a respectiva porcentagem das observações (arredondada). Assim, por meio da figura, podemos dizer que 61% dos empregados têm salário inferior a 12 salários mínimos, ou 17% possuem salário superior a 16 salários mínimos.

Do mesmo modo que usamos um artifício para representar uma variável contínua como uma variável discreta, podemos usar um artifício para construir um histograma para variáveis discretas. A próxima Figura é um exemplo de como ficaria o histograma da variável Z, número de filhos dos empregados casados.

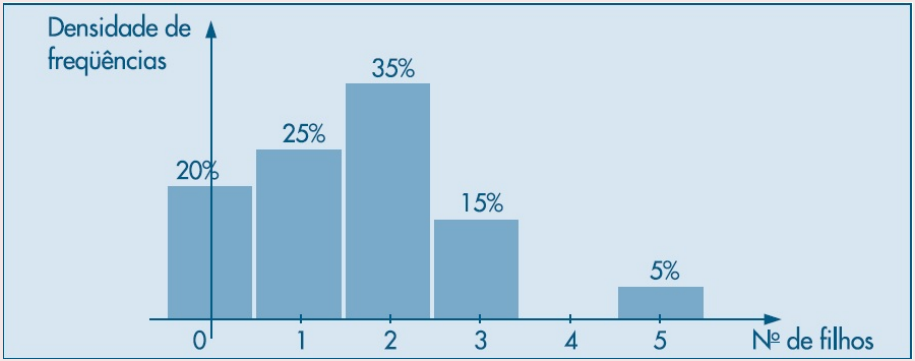


Figura: Histograma da variável Z : número de filhos.

Caso utilizemos as frequências acumuladas ou frequências acumuladas relativas para construir o histograma teríamos um histograma conhecido como **ogiva** ou **ogiva percentual**, respectivamente.

Vimos que o resumo de dados por meio de tabelas de freqüências e ramo-e-folhas fornece muito mais informações sobre o comportamento de uma variável do que a própria tabela original de dados. Muitas vezes, queremos resumir ainda mais estes dados, apresentando um ou alguns valores que sejam representativos da série toda. Quando usamos um só valor, obtemos uma redução drástica dos dados. Usualmente, emprega-se uma das seguintes medidas de posição (ou localização) central: média, mediana ou moda.

A moda é definida como a realização mais freqüente do conjunto de valores observados. Por exemplo, considere a variável Z, número de filhos de cada funcionário casado, resumida na Tabela de dados da companhia MB. Vemos que a moda é 2, correspondente à realização com maior freqüência, 7. Em alguns casos, pode haver mais de uma moda, ou seja, a distribuição dos valores pode ser bimodal, trimodal etc.

A mediana é a realização que ocupa a posição central da série de observações, quando estão ordenadas em ordem crescente. Assim, se as cinco observações de uma variável forem $3, 4, 7, 8$ e 8, a mediana é o valor 7, correspondendo à terceira observação. Quando o número de observações for par, usa-se como mediana a média aritmética das duas observações centrais. Acrescentando-se o valor 9 à série acima, a mediana será $(7 + 8)/2 = 7,5$.

Finalmente, a média aritmética, conceito familiar ao leitor, é a soma das observações dividida pelo número delas.

Neste exemplo, as três medidas têm valores próximos e qualquer uma delas pode ser usada como representativa da série toda. A média aritmética é, talvez, a medida mais usada. Contudo, ela pode conduzir a erros de interpretação. Em muitas situações, a mediana é uma medida mais adequada. Veremos algumas situações que ilustram tal afirmação.

Se $x_1, \cdots, x_n$ são os n valores (distintos ou não) da variável X, a média aritmética, ou simplesmente média, de X pode ser escrita como:

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n}}{x_i} \tag{2}$$

Agora, se tivermos n observações da variável X, das quais n_1 são iguais a x_1, n_2 são iguais a x_2, n_k iguais a x_k, então a média de X pode ser escrita como:

$$\bar{x} = \frac{n_1 x_1 + \cdots + n_k x_k}{n} = \frac{\sum_{i=1}^{k}}{n_i x_i} \tag{3}$$

Consideremos, agora, as observações ordenadas em ordem crescente. Vamos denotar a menor observação por $x_{(1)}$, a segunda por $x_{(2)}$, e assim por diante, obtendo-se

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n-1)} \leq x_{(n)}. \tag{4}$$

As observações ordenadas como em (4) são chamadas **estatísticas de ordem**. Com esta notação, a mediana da variável X pode ser definida como:

$$md(X) = \begin{cases} X_{(\frac{n+1}{2})}, & \text{se} \quad \mathsf{n} \quad \text{é impar}, \\ \frac{X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}}{2}, & \text{se} \quad \mathsf{n} \quad \text{é par}. \end{cases}$$

A mediana é uma medida mais robusta que a média, quando submetida a mudanças nos valores observados, ou a incorporação de mais observações no conjunto de dados original.

O resumo de um conjunto de dados por uma única medida representativa de posição central esconde toda a informação sobre a variabilidade do conjunto de observações. Por exemplo, suponhamos que cinco grupos de alunos submeteram-se a um teste, obtendo-se as seguintes notas:

- Grupo A (Variável X): 3,4,5,6,7
- Grupo B (Variável Y): 1,3,5,7,9
- Grupo C (Variável Z): 5,5,5,5,5
- Grupo D (Variável W): 3,5,5,7
- Grupo E (Variável V): 3,5,5,6,6

Vemos que $\bar{x} = \bar{y} = \bar{z} = \bar{w} = \bar{v} = 5$. A identificação de cada uma destas séries por sua média (5, em todos os casos) nada informa sobre suas diferentes variabilidades. Notamos, então, a conveniência de serem criadas medidas que sumarizem a variabilidade de um conjunto de observações e que nos permita, por exemplo, comparar conjuntos diferentes de valores, como os dados acima, segundo algum critério estabelecido.

Um critério freqüentemente usado para tal fim é aquele que mede a dispersão dos dados em torno de sua média, e duas medidas são as mais usadas: desvio médio e variância. O princípio básico é analisar os desvios das observações em relação à média dessas observações.

Um critério freqüentemente usado para tal fim é aquele que mede a dispersão dos dados em torno de sua média, e duas medidas são as mais usadas: desvio médio e variância.

$$Dm(X) = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n} \tag{5}$$

$$Var(X) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n} \tag{6}$$

O princípio básico é analisar os desvios das observações em relação à média dessas observações. Desvio é interpretado como o afastamento de uma observação em relação a uma determinada medida de posição.

Sendo a variância uma medida de dimensão igual ao quadrado da dimensão dos dados (por exemplo, se os dados são expressos em cm, a variância será expressa em cm^2), pode causar problemas de interpretação. Costuma-se usar, então, o desvio padrão, que é definido como a raiz quadrada positiva da variância.

$$dp(X) = \sqrt{Var(X)} \tag{7}$$

Ambas as medidas de dispersão (Dm e dp) indicam em média qual será o "erro" (desvio) cometido ao tentar substituir cada observação pela medida resumo do conjunto de dados (no caso, a média).

Coeficiente de Variação

$$CV(X) = \frac{dp(X)}{\bar{X}} \tag{8}$$

Mesmo o DP pode induzir à conclusões errôneas com relação à variabilidade. Suponha dois conjunto de dados $D_1 = \{10, 20, 30\}$ e $D_2 = \{10000, 10010, 10020\}$. Note que nestes casos $\bar{x}_1 = 20, dp(x) = 10, \bar{x}_2 = 10010$ e $dp(x_2) = 10$. Porém, em termos percentuais, o primeiro conjunto de dados é mais heterogênio.

Medidas Complementares para Análise de Dados

- Extremos: O menor e o maior valor do conjunto de dados;
- Quartis (Q)
  - 1° Quartil: deixa um quarto dos valores abaixo, e três quartos acima dele;
  - 2° Quartil = Mediana: deixa metade dos valores abaixo, e metade acima dele;
  - 3° Quartil: deixa três quartos dos valores abaixo, e um quarto acima dele;
- Intervalo Interquartil (pode ser considerada uma medida robusta de dispersão).

Os cinco valores, $x(1), q1, q2, q3ex(n)$ são importantes para se ter uma boa idéia da assimetria da distribuição dos dados. Para uma distribuição simétrica ou aproximadamente simétrica, deveríamos ter:

(a) $q_2 - x_{(1)} \approx x_{(n)} - q_2$;

(b) $q_2 - q_{(1)} \approx q_{(3)} - q_2$;

(c) $q_1 - x_{(1)} \approx x_{(n)} - q_3$;

(d) distâncias entre mediana e q_1, q_3 menores do que distâncias entre os extremos e q_1, q_3.

A diferença $q_2 - x_{(1)}$ é chamada dispersão inferior e $x_{(n)} - q_2$ é a dispersão superior. A condição (a) nos diz que estas duas dispersões devem ser aproximadamente iguais, para uma distribuição aproximadamente simétrica. A próxima Figura ilustra estes fatos para a chamada distribuição normal ou gaussiana.

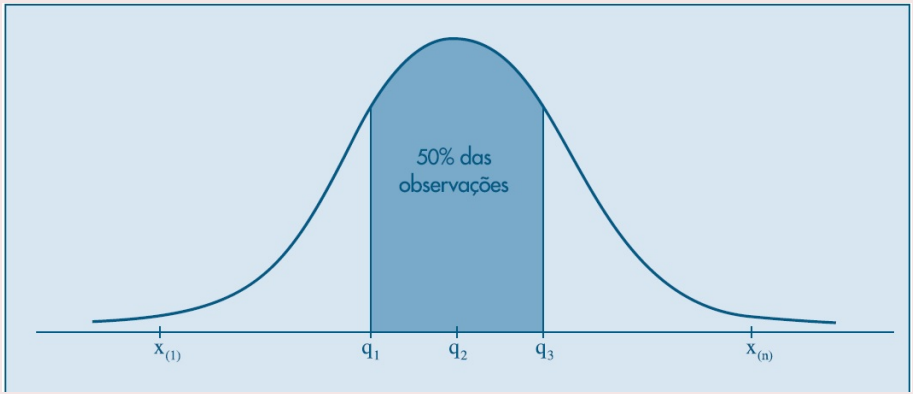
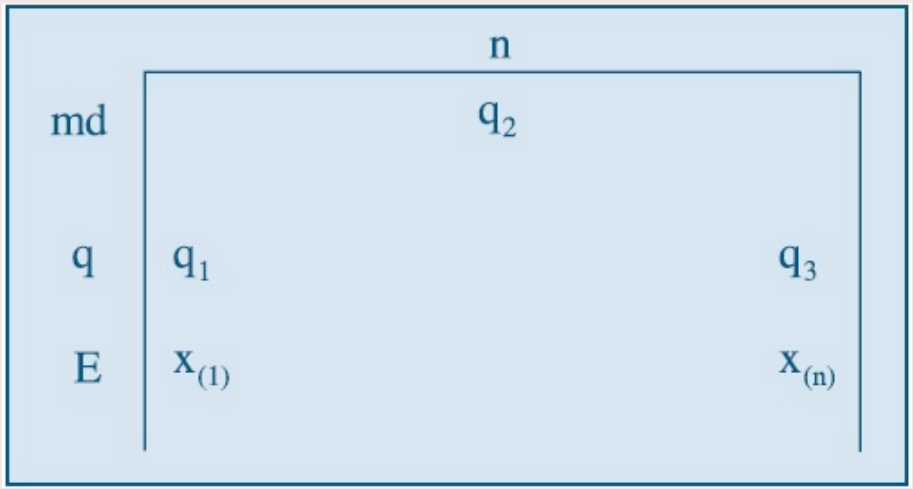
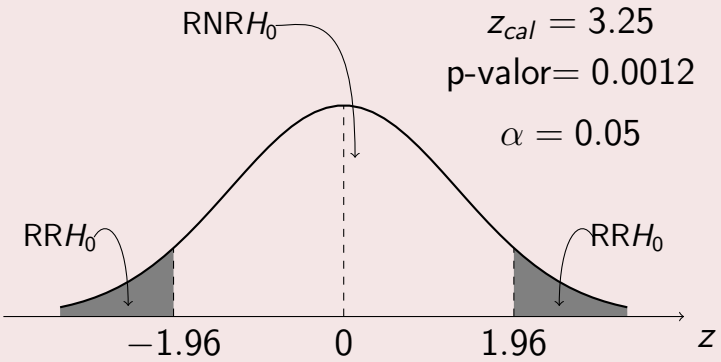


Figura: Uma distribuição simétrica: normal ou gaussiana.

As cinco estatísticas de ordem consideradas acima podem ser representadas esquematicamente como na próxima Figura, onde também incorporamos o número de observações, n. Representamos a mediana por md, os quartis por q e os extremos por E.




RNRH_0
$z_{cal} = 3.25$
p-valor$= 0.0012$
$\alpha = 0.05$
RRH_0
RRH_0
-1.96
0
1.96
z

Box-Plot

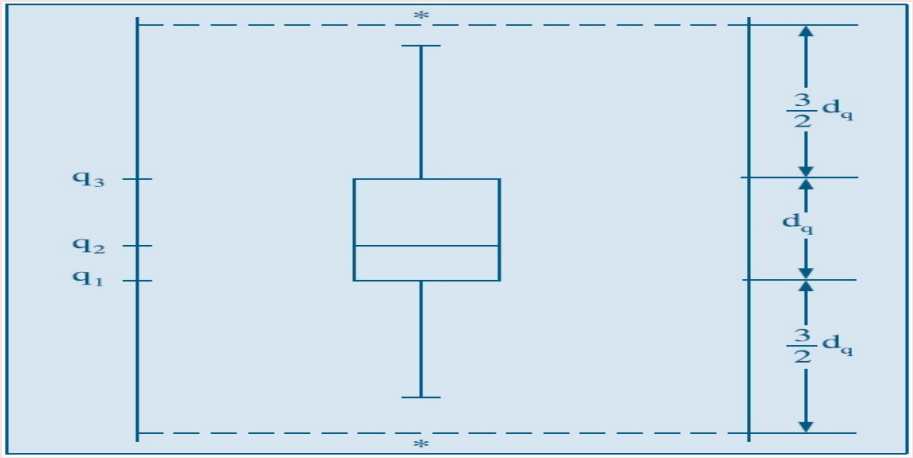


Figura: Esquema de um BoxPlot.

Box-Plot

O boxplot (gráfico de caixa) é um gráfico utilizado para avaliar a distribuição empírica do dados. O boxplot é formado pelo primeiro e terceiro quartil e pela mediana. Para construir este diagrama, consideremos um retângulo onde estão representados a mediana e os quartis. A partir do retângulo, para cima, segue uma linha até o ponto mais remoto que não exceda $LS = q_3 + (1,5)d_q$, chamado limite superior. De modo similar, da parte inferior do retângulo, para baixo, segue uma linha até o ponto mais remoto que não seja menor do que $LI = q_1 - (1,5)d_q$, chamado limite inferior.

Box-Plot

Os valores compreendidos entre esses dois limites são chamados valores adjacentes. As observações que estiverem acima do limite superior ou abaixo do limite inferior estabelecidos serão chamadas pontos exteriores e representadas por asteriscos. Essas são observações destoantes das demais e podem ou não ser o que chamamos de outliers ou valores atípicos.

O box plot dá uma idéia da posição, dispersão, assimetria, caudas e dados discrepantes. A posição central é dada pela mediana e a dispersão por d_q. As posições relativas de q_1, q_2, q_3 dão uma noção da assimetria da distribuição. Os comprimentos das caudas são dados pelas linhas que vão do retângulo aos valores remotos e pelos valores atípicos.

Assimetrias

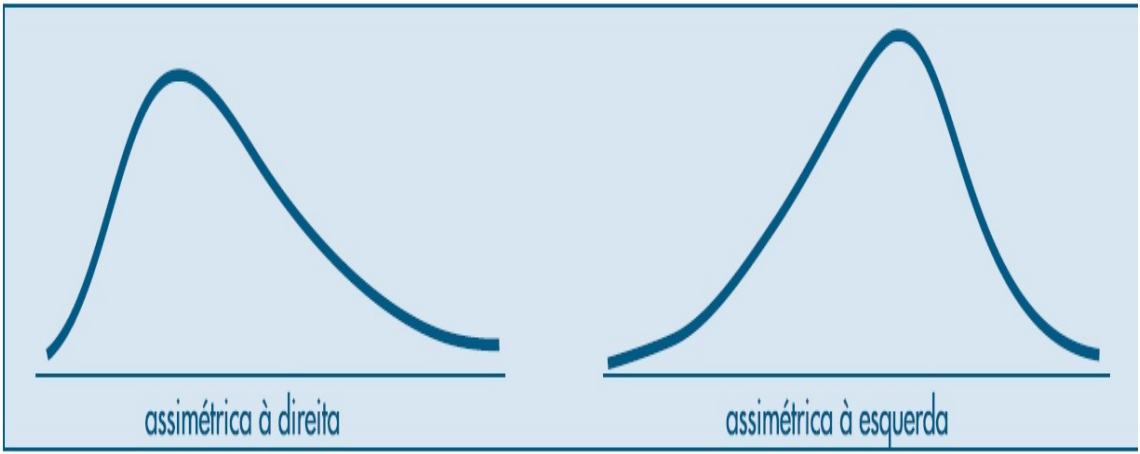


Figura: Distribuições assimétricas.

Transformações

Vários procedimentos estatísticos são baseados na suposição de que os dados provêm de uma distribuição normal (em forma de sino) ou então mais ou menos simétrica. Mas, em muitas situações de interesse prático, a distribuição dos dados da amostra é assimétrica e pode conter valores atípicos, como vimos em exemplos anteriores.

Se quisermos utilizar tais procedimentos, o que se propõe é efetuar uma transformação das observações, de modo a se obter uma distribuição mais simétrica e próxima da normal. Uma família de transformações frequentemente utilizada é

$$X^* = \begin{cases} X^p, & \text{se} \quad \text{p}>0 \\ \ln(X), & \text{se} \quad \text{p}=0 \\ -X^p, & \text{se} \quad \text{p}<0 \end{cases}$$

Normalmente, o que se faz é experimentar valores de p na sequência $\cdots, -3, -2, -1, -1/2, -1/3, -1/4, 0, 1/4, 1/3, 1/2, 1, 2, 3, \cdots$

Para cada valor de p obtemos gráficos apropriados (histogramas, desenhos esquemáticos etc.) para os dados originais e transformados, de modo a escolhermos o valor mais adequado de p. Para dados positivos, a distribuição dos dados é usualmente assimétrica à direita. Para essas distribuições, a transformação acima com $0 < p < 1$ é apropriada, pois valores grandes de x decrescem mais, relativamente a valores pequenos. Para distribuições assimétricas à esquerda, tome $p > 1$.

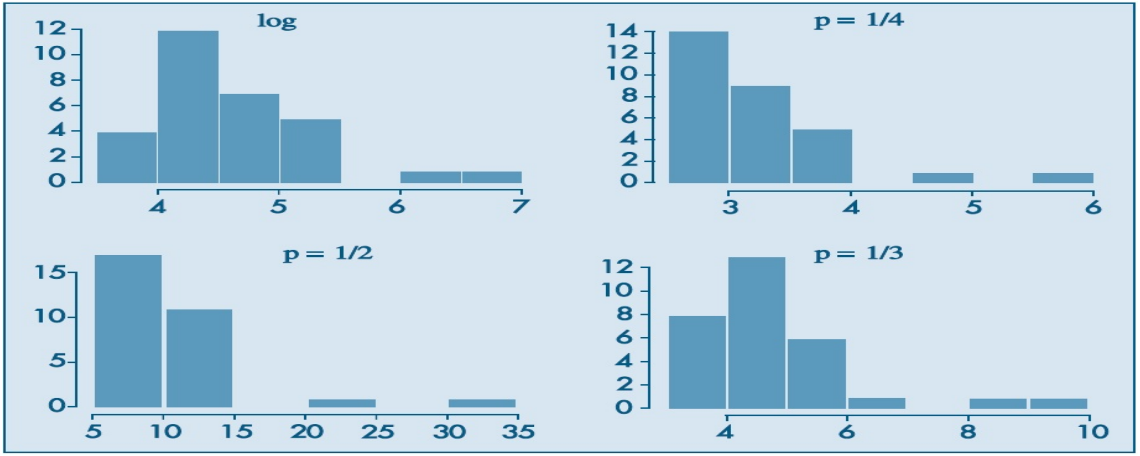


Figura: Exemplo de Histogramas para dados transformados.