Análise Estatística com Excel

Prof. Dr. Evandro Marcos Saidel Ribeiro
E-mail: esaidel@usp.br Home page: www.fearp.usp.br/~saidel

Módulo 1

1 Introdução

1.1 Apresentação geral dos tópicos do curso

Tópico 2 – Módulo 1: Estatística Descritiva
O estudo de estatística descritiva é importante para conhecer as principais medidas estatísticas, bem como os recursos de apresentação gráfica de dados. Obtemos informações a partir dos dados. Um recurso interessante é o **boxplot** (Figura 1.1), que informa medidas de centro, de variação e ainda proporciona uma visualização da distribuição dos valores. Valores de preços, receitas, tempos, etc. podem ser analisados através do Boxplot.

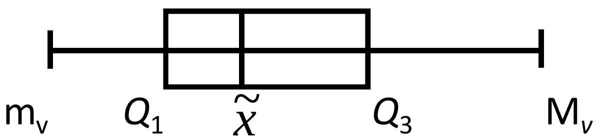


Figura 1.1 Boxplot

Tópico 3 – Módulo 1: Probabilidade
A chance ou possibilidade de que um evento venha a acontecer é importante para a tomada de decisão. Assim, o estudo de probabilidade está diretamente relacionado ao conhecimento ou histórico dos eventos que ocorrem na empresa. Por exemplo, se a taxa de faltas num determinado setor é igual a 20%, então podemos estimar a probabilidade de mais de 5 funcionários faltarem num determinado dia, ou em geral mais de x funcionários faltarem num determinado dia (ver figura 1.2).

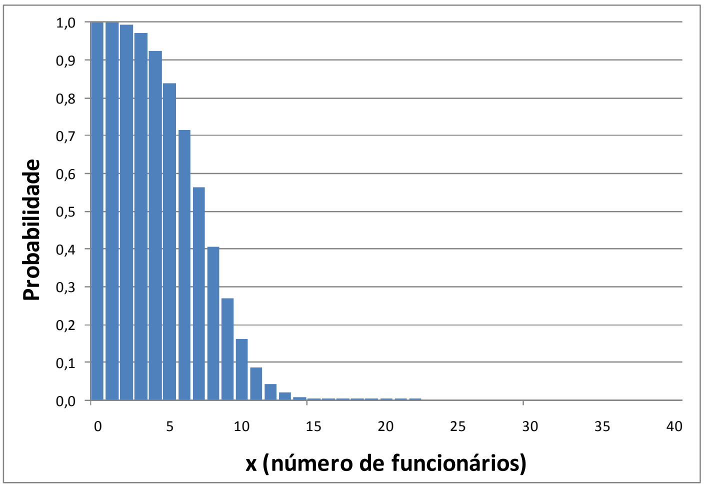


Figura 1.2 Probabilidade de faltar mais de x funcionários de um setor com 40 funcionários, num determinado dia. A taxa de faltas é de 20% ao dia.

O conhecimento de **distribuições de probabilidade** permite diversos tipos de análise, incluindo simulação.

Tópico 4 – Módulo 2: Estatística Inferencial
A Estatística Inferencial utiliza padrões observados em amostras para fazer inferências sobre a população da qual a amostra foi retirada. Estas inferências podem ser estimativas de intervalo de confiança, **testes de hipóteses**, descrição de associação entre dados (**correlação**), modelagem entre os dados (**análise de regressão**). Em geral são considerados valores com certa probabilidade de ocorrência, estudamos as distribuições de amostras e voltamos a atenção para os valores mais frequentes (ver Figura 1.3).

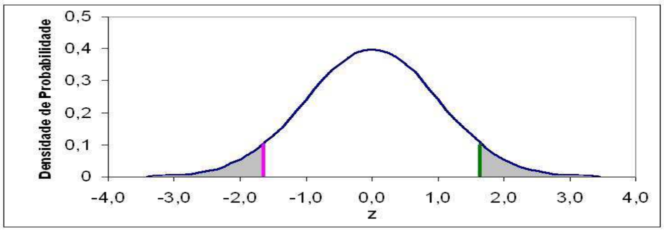


Figura 1.3 Intervalos de confiança e níveis de significância estatística são baseados em distribuições de amostras.

Tópico 5 – Módulo 2: Análise Multivariada de dados
A análise multivariada refere-se a métodos estatísticos que tornam possível a análise simultânea de medidas múltiplas para cada indivíduo (CORRAR, PAULO, DIAS FILHO, 2007). Modelos de análise de risco, análise de crédito empregam cada vez mais técnicas de análise multivariada de dados.
A Eq (1) ilustra o modelo de regressão múltipla: a variável Y pode ser explicada em termos das variáveis $x_1, x_2, \ldots, x_r$.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_r x_r + \varepsilon \qquad (1)$$

A base para muitas técnicas de análise multivariada é a matriz de covariâncias que pode ser utilizada na análise de risco.

Tópico 6 – Módulo 3: Análise de Séries Temporais
Através da análise de séries temporais estudamos tendências, variações sazonais, variações cíclicas, (Figura 1.4). Nesta revisão vamos estudar modelos de previsão; analisar como escolher um modelo de previsão mais apropriado; estudar índices de preços.

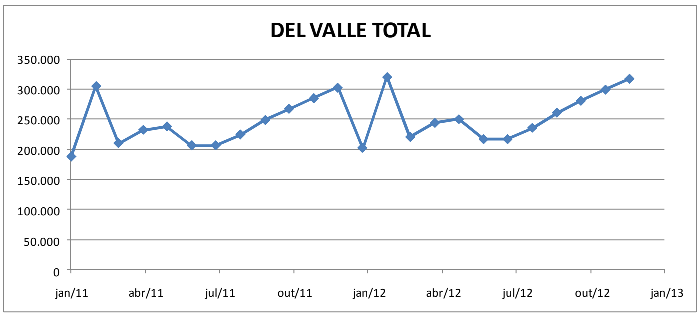


Figura 1.4 Série temporal de quantidades de um produto.

Tópico 7 – Módulo 3: Controle Estatístico de Processos
Alguns métodos para melhoria contínua de produtos têm sido desenvolvidos com ênfase cada vez maior na estatística, na melhoria de processos e na otimização do sistema como um todo. Neste treinamento iremos estudar gráficos de controle, (veja exemplo na Figura 1.5) um tipo especial de série temporal utilizada principalmente na gestão da qualidade. Os gráficos de controle, além de fornecer uma exposição visual dos dados que

representam os processos, têm o objetivo de separar as causas especiais de variação das causas comuns de variação.

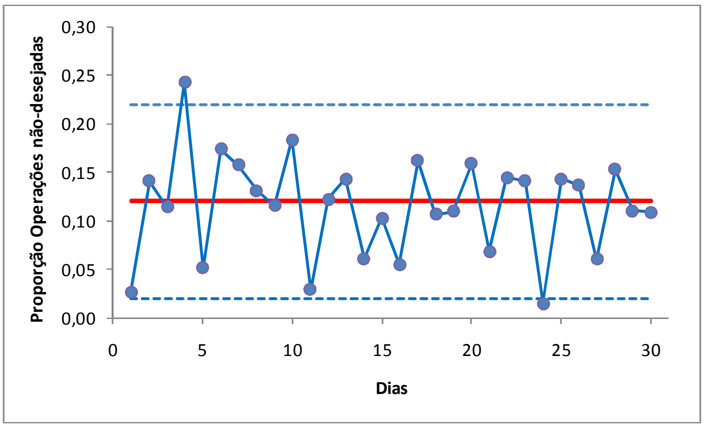


Figura 1.5 Gráfico de controle para análise da proporção de itens não-conformes.

1.2 Estatística e Excel na empresa

"O Microsoft Excel fornece um bom suporte para aplicar métodos estatísticos na tomada de decisão empresarial" (LEVINE, 2008).

O Excel é uma escolha atrativa pois

- não incorre custos adicionais para aquisição de softwares estatísticos;
- a maior parte dos usuários em empresas conhecem o Excel
- é fácil de utilizar e aprender;
- as funções gráficas e estatísticas utilizam dados de planilhas com de diversas aplicações na empresa;
- alguns gráficos produzem resultados visuais mais atraentes do que vários programas estatísticos.

Evidentemente, ao utilizar o Excel, você deve ser cauteloso em relação aos dados e ao método que será utilizado. O Excel possui ferramentas de análise de dados com as principais técnicas e métodos de inferência estatística. Neste curso vamos explorar a inferência estatística com funções do Excel e também utilizar as ferramentas de análise de dados. Um resumo das principais funções são apresentadas na Tabela Excel1 (a seguir) e na Tabela Excel2 (página 26). Além das funções de Estatística, as Funções de Procura e Referência e as Funções Matemáticas, são bastante utilizadas.

O software Excel possui várias versões, sendo que uma grande mudança foi observada no Excel 2007 com o conceito da faixa de opções. Este curso foi desenvolvido com os detalhes da versão do Office 2007, mas os recursos gerais podem ser utilizados nas outras versões. Alguns detalhes apresentados aqui já foram atualizados para o Excel 2010.

O arquivo ***EstatisticaExcel.xlsm*** contém os vários exemplos e exercícios desenvolvidos neste curso.

2 Estatística Descritiva

A Estatística Descritiva permite ao pesquisador uma melhor compreensão dos dados por meio de tabelas, gráficos e medidas-resumo, identificando tendências, variabilidade e valores extremos. Observem medidas de **centro**, **variação** e **distribuição**.

Tabela Excel1. Funções Excel, parte 1.

Função	Descrição
PROCV(*valor*; *matriz*; *coluna*; [tipo] **)**	Procura na *matriz* o *valor* e retorna o conteúdo da *coluna* especificada.
PROCH(*valor*; *matriz*; *linha*; [tipo] **)**	Procura na *matriz* o *valor* e retorna o conteúdo da *linha* especificada.
ÍNDICE(*matriz*; *lin*; *col* **)**	Examina a *matriz* e retorna o conteúdo da *linha* e *coluna*.
MÍNIMO(*matriz***)**	Retorna o valor mínimo.
MÁXIMO(*matriz***)**	Retorna o valor máximo.
CONT.VALORES(*matriz***)**	Nro de células não vazias.
CONCATENAR(*texto1*;*texto2*; ...**)**	Agrupa conteúdos.
FREQÜÊNCIA (*matriz1*;*matriz2***)**	Frequência de valores em *matriz1* que ocorrem de acordo com os limites definidos na *matriz2*.
MÉDIA(*valor1*; *valor2*;..**)**	Média aritmética
MED(*valor1*; *valor2*;..**)**	Mediana
MODO(*valor1*; *valor2*;..**)**	Moda
DESVPAD(...; ...; ...**)**	Desvio-padrão
VAR(...; ...; ...**)**	Variância
QUARTIL(...; k**)**	Quartil k
PERCENTIL(...; k/100**)**	Percentil k
DISTRBINOM(*x*; *n* ; *p* ;FALSO)	Distribuição binomial: *x* sucessos, *n* tentativas; *p* prob. de sucesso em 1 tent.
POISSON(*x*; *L* ;FALSO**)**	Distribuição de Poisson; *x* observações; *L* é a média por intervalo.
DIST.NORM(*x*; μ; σ;FALSO**)**	Distribuição Normal p(x), com média μ e desvio-padrão σ. Para P(x < x_0) utilizar Verdadeiro
DIST.NORMP(*z***)**	Distribuição normal padrão, P(z < z_0)
INV.NORM(P; μ; σ**)**	Para uma probabilidade P retorna o valor de x.
INV.NORMP(P**)**	Para uma probabilidade P retorna o valor de z.
ALEATÓRIO()	Sorteio de um número aleatório entre 0,0 e 1,0.
ALEATÓRIOENTRE(N_1;N_2**)**	Sorteio de um número aleatório entre N_1 e N_2. O resultado é um número inteiro.

No arquivo *EstatisticaExcel.xlsm* considere a planilha *Empresas*. Vamos estudar como transformar os dados das empresas em informações úteis para eventual tomada de decisão.

2.1 Histograma, Gráfico de Pareto, Polígono Frequência, Ogiva

Para construir um gráfico que resuma os dados, primeiramente devemos estudar as frequências de ocorrência das observações.

Histograma

Um gráfico de barras (colunas) no qual a escala horizontal contém **classes** e a vertical contém frequências.

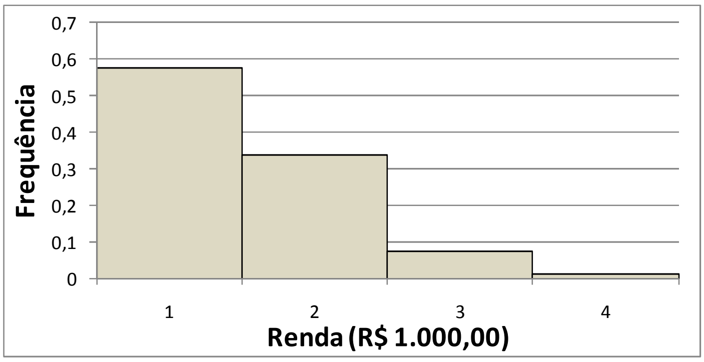


Figura 2.1 Histograma

Gráfico de Pareto

É um gráfico de barras para dados qualitativos, com barras (ou colunas) dispostas em ordem pela frequência.

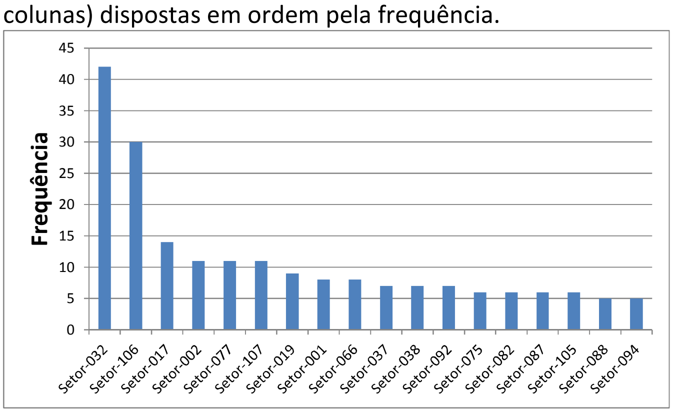


Figura 2.2 Gráfico de Pareto.

Polígono Frequência

O Polígono Frequência utiliza segmentos de retas ligados a pontos localizados nos pontos médios das classes. Veja Figura 2.3.

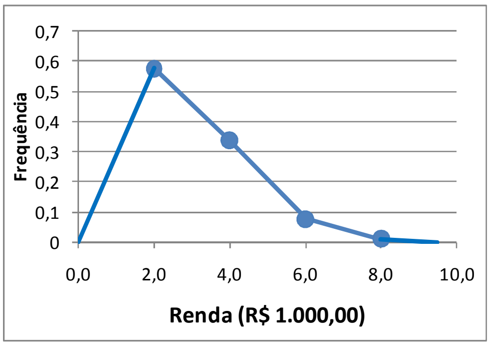


Figura 2.3 Polígono Frequência.

Gráfico Ogiva

Ogiva é um gráfico que representa frequências acumuladas. Ogiva utiliza as fronteiras da classe ao longo da escala horizontal.

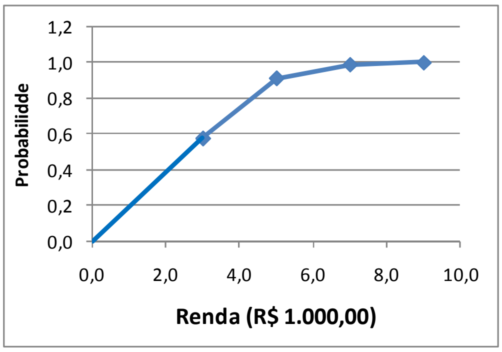


Figura 2.4 Gráfico Ogiva.

2.2 Distribuição de frequência

Uma **distribuição de frequência** é a listagem dos valores dos dados (individualmente ou por classes), juntamente com suas frequências correspondentes (ou contagens).

Construção de uma distribuição de Frequência

Nomear regiões no Excel que contenham os valores a serem analisados é uma boa prática. Utilize a **Caixa de Nome**. Ou selecione uma região com os nomes na primeira linha e digite: **Ctrl + Shift + F3**.

a) Decida sobre o número de classes n_c . Geralmente entre 5 e 20 (ou utilize a regra de *Sturges* $n_c = 1 + 3{,}3 \log n$).

b) Calcule a amplitude de cada classe,

$$A_c = \frac{M_v - m_v}{n_c},$$ sendo: A_c a amplitude da classe, M_v o maior valor (arredondado para cima), m_v o menor valor (arredondado para baixo) e n_c o número de classes. Arredonde o resultado.

c) Ponto inicial do primeiro intervalo de classes: m_v

d) O primeiro intervalo de classes terá o limite inferior m_v e o limite superior $m_v + A_c$.

e) Usando Limite Superior da 1ª classe e amplitude de classe, prossiga e liste os outros limites superiores.

f) Calcule os outros limites inferiores considerando os limites superiores das classes anteriores mais uma pequena variação (Ex. 0,0000001).

g) Encontre a frequência total para cada classe. No Excel utilize o comando FREQÜÊNCIA(matriz_dados;matriz_bin):

  I. a matriz de dados é a região com os valores da variável

  II. a matriz "bin" é a região com os **limites superiores** do intervalo de classes

  III. o comando é digitado para a primeira classe

  IV. para obter as frequências em cada classe selecione a coluna inteira que deseja obter a frequência e digite a tecla "**F2**" e em seguida: **[Ctrl] + [Shift] + [Enter]**

Pode ser que a variabilidade dos dados seja muito grande (isto ocorre muito em ciências sociais, em geral para dados financeiros). **Neste caso é conveniente considerar intervalos de classes logarítmicos**. Para tanto considere o menor e o maior valor e calcule o **LOG10** destes valores. Considere estes novos valores como sendo m_v e M_v. Construa os limites das classes (Lim) de acordo com os passos (b) até (f). Antes de calcular a frequência volte os valores dos limites das classes para o valor da variável fazendo, numa coluna, os novos limites **=10^(Lim)**. Retome o item (g) e calcule as frequências em cada classe.

Exercícios no Excel (*EstatisticaExcel.xlsm*; Planilha *Aula1*)

2.2.1 Intervalo de classes, distribuição de frequências. Com base nos dados de Ativo Total (AT) das empresas da planilha *Empresas*, defina intervalos de classes (lineares e depois logarítmicos) e faça uma tabela com Frequência Total, Frequência Relativa, Frequência Relativa Acumulada, em cada intervalo. Com base na tabela que melhor representa a distribuição, construa o gráfico Polígono Frequência e o Gráfico Ogiva.

2.2.2 Histograma e Gráfico de Pareto. A planilha *Frequencia*, apresenta dados de clientes de uma concessionária. Construa um histograma para a variável ND (número de dependentes). Construa um gráfico de Pareto para esta mesma variável.
2.2.3 Exercícios adicionais na planilha *Frequencia*.

2.3 Medidas de Centro, Variação, Distribuição

Nesta seção vamos estudar como obter medidas úteis de estatística descritiva.

Medidas de centro: Média, mediana, moda

Estas são as medidas de centro mais conhecidas e que geralmente atraem mais atenção, pois resumem todo o conjunto de dados em um único valor:

Media: Média aritmética (x-barra)

$\overline{x}$ =MÉDIA(...; ...; ...)

Mediana: Valor do meio para dados ordenados (x-til)

$\tilde{x}$ =MED(...; ...; ...)

Moda: Valor mais frequente

M =MODO(...; ...; ...) ou MODO.ÚNICO(...; ...; ...)

Exercícios no Excel (*EstatisticaExcel.xlsm*; Planilha *Aula1*)

2.3.1 Obtenha as medidas de centro, (Média, Mediana e Moda) para a variável Renda dos clientes da concessionária (do exercício 2.2.2). Obtenha as medidas para AT (Ex. 2.2.1)
Utilize também a função MÉDIASE:
=MÉDIASE(intervalo condição; critério; intervalo para média)

Medidas de Variação: Amplitude, desvio-padrão, variância, coeficiente de variação

Amplitude: Diferença entre o maior valor e o menor valor dentre o conjunto de dados.

$$A = M_v - m_v$$

Desvio-padrão: Uma dispersão média dos dados em torno da média.

$$s = \sqrt{\frac{\sum (x-\overline{x})^2}{n-1}} = \text{DESVPAD(...; ...; ...)}$$

Variância: Também é uma medida de dispersão média dos dados em torno da média, é o desvio padrão ao quadrado.

$$s^2 = \frac{\sum (x-\overline{x})^2}{n-1} = \text{VAR(...; ...; ...)}$$

Coeficiente de variação: Diferentemente das outras medidas de variação, o coeficiente de variação é uma medida relativa de variação, expressa na forma de porcentagem.

$$\text{CV} = \left(\frac{s}{\overline{x}}\right) \times 100\%$$

Exercícios no Excel (*EstatisticaExcel.xlsm*; Planilha *Aula1*)

2.3.2 Obtenha as medidas de variação, Amplitude, desvio-padrão, variância, coeficiente de variação, para a variável renda dos clientes da concessionária.

Alguns valores de estatísticas descritivas que vimos podem ser obtidos para intervalos com filtros, mas devemos utilizar outra função, pois mesmo que o filtro seja aplicado no intervalo os resultados das funções que vimos não se alteram. Para esta questão veja a planilha *Subtotal*.

❶ Arquivo de recuperação → ***EstatisticaExcel_01.xlsm***

Medidas de distribuição (dispersão):

Escore-z Diferença entre o valor x e a média aritmética, dividida pelo desvio-padrão. Informa quantos desvios a observação x está a partir da média dos valores.

$$z = \frac{(x-\overline{x})}{s}$$

Valores comuns observados em amostras costumam ter um escore-z entre -2 e 2, ou seja: $-2 < z < 2$.

Quartis Valores que divide os dados ordenados em 4 partes
=QUARTIL(Intervalo;quarto)
Exemplo: Q_1 = QUARTIL(...; 1) valor que separa 25% dos dados inferiores dos 75% superiores.

Percentis Valores que divide os dados ordenados em 100 partes
percentil k =PERCENTIL(Intervalo;k/100)
Exemplo: P_{68} = PERCENTIL(...; 68/100) valor que separa 68% dos dados inferiores dos 32% superiores.

Exercícios no Excel (*EstatisticaExcel_01.xlsm*; Planilha *Aula1*)

2.3.3 Obtenha as medidas de distribuição: Quartis – Q_1, Q_2 e Q_3, e os percentis – P_{18}, P_{50} , P_{85}, para a variável Renda dos clientes da concessionária. Transforme os valores obtidos para Escore-z.

2.4 Boxplot (ou diagrama de caixa)

Boxplot é um gráfico com um resumo de cinco números que fornecem uma boa descrição da distribuição dos valores.

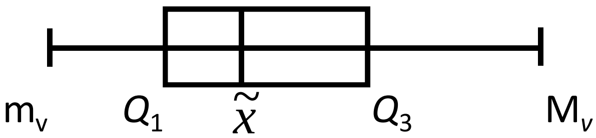


Figura 2.5 Boxplot

Exercícios no Excel (*EstatisticaExcel_01.xlsm*; Planilha *Aula1*)

2.4.1 Faça um Boxplot para a variável renda dos clientes da concessionária. (Boxplot no Excel não é automático)

BoxPlot procedimento no Excel – (A)

- Obtenha, para cada amostra, os cinco números necessários. Por exemplo, números para duas amostras de clientes inadimplentes (ST=1) e adimplentes (ST=0).
- Obtenha valores adicionais como indicado abaixo

	ST=0	ST=1
Max	8,2	4,5
Q3	4,6	2,4
Mediana	3,8	2,1
Q1	3,0	1,7
Min	1,3	1,2

Erro (+): Max - Q3	3,6	2,1
Q3-Mediana	0,8	0,3
Mediana-Q1	0,8	0,4
Q1	3,0	1,7
Erro (-): Q1-Min	1,7	0,5

- Selecione: Inserir > Gráficos – Colunas > 2D empilhadas
- Alterne entre Linha/Coluna e acerte a ordem das séries

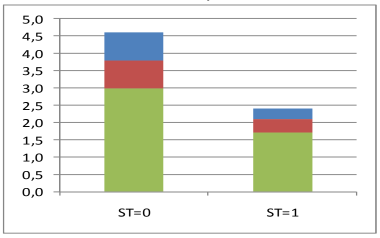


Figura 2.6 Boxplot no Excel? Gráfico Colunas/Barras

- Na coluna de baixo [de cima] adicione barras de erros negativas [positivas]: Ferram. de gráficos > Layout > Análise > barra de erros > mais opções >
  (x) Menos (x) Personalizado >
  valor do erro negativo [positivo]
  - selecionar células Erro(-) [Erro(+)]

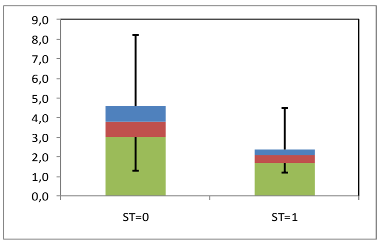


Figura 2.7 Boxplot no Excel: adição de barras de erros.

- Formate as cores de cada coluna (a de baixo fica sem cor).

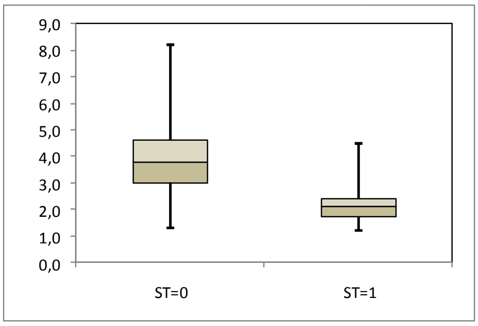


Figura 2.8 Boxplot no Excel, formatação final.

Este procedimento só é aplicado para variáveis com valores maiores ou iguais a zero. Para variáveis que assumem valores negativos podemos utilizar o procedimento mais geral a seguir.

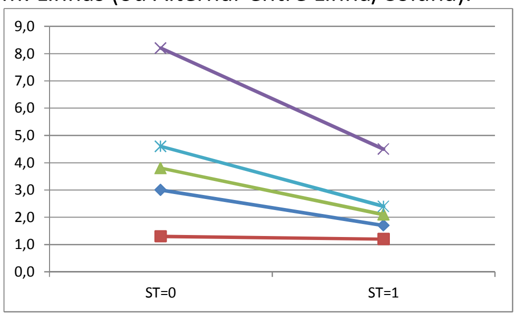
BoxPlot, procedimento no Excel – (B)

Vamos verificar um procedimento alternativo obter Boxplots

- O primeiro passo é obter o resumo dos cinco números na ordem indicada abaixo (diferente do procedimento anterior):

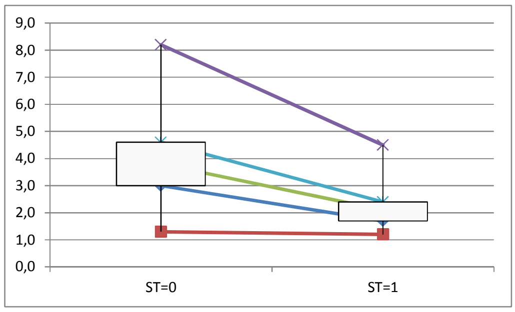
	ST=0	ST=1
Q1	3,0	1,7
Min	1,3	1,2
Mediana	3,8	2,1
Max	8,2	4,5
Q3	4,6	2,4

- Selecione as duas colunas com as estatísticas, incluindo os rótulos de dados.
- Selecione Inserir > Gráfico > Linhas > Linha com marcadores.
- Séries em: Linhas (ou Alternar entre Linha/Coluna).



- Selecione a área de plotagem e depois:
Ferramentas de Gráfico > Layout >
  Linhas > Linhas de máximo/mínimo
  Barras superiores/inferiores

Resultado:



- Remova as linhas:

- Clique com o botão direito na linha e depois em:

Formatar séries de dados > Cor da linha > Sem linha

- Clique em outra linha de clique em [F4], (F4 repete o comando)
- Altere os símbolos de acordo com a preferência.

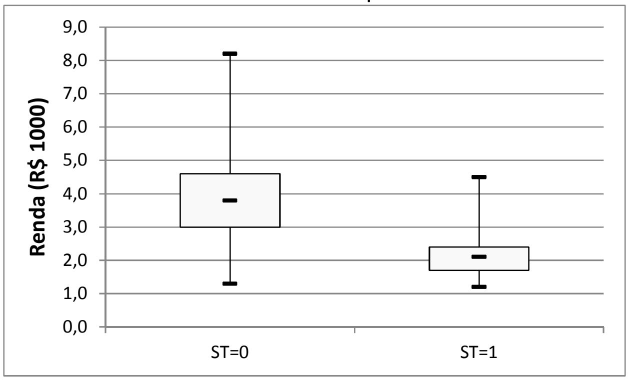


Figura 2.9 Boxplot a partir de gráfico de Linha, serve para qualquer conjunto de dados.

Exercícios no Excel (*EstatisticaExcel_01.xlsm*; Planilha *Aula1*)
2.4.2 Faça um Boxplot para a variável ROA (*Return On Assets*) e para o Ativo Total das Empresas na planilha *Empresas*.

❷ Arquivo de recuperação → ***EstatisticaExcel_02.xlsm***

3 Probabilidade

3.1 Introdução à probabilidade

Probabilidade: um valor numérico entre **zero** e **um** que representa a chance, eventualidade ou a possibilidade de que um determinado evento venha a ocorrer.

Probabilidades podem ser determinadas da seguinte maneira:

a. a priori: baseada na contagem de possíveis eventos
b. de forma empírica: baseada na observação
c. de forma subjetiva: estimativa que muda de pessoa para pessoa

Probabilidade de ocorrência de um evento A: $0 \le P(A) \le 1$
Evento complementar $\bar{A}$: $P(\bar{A}) + P(A) = 1$
Regra da adição $P(A \text{ ou } B) = P(A) + P(B) - P(A \text{ e } B)$
Regra da multiplicação: $P(A \text{ e } B) = P(A) \cdot P(B \mid A)$

Tabelas de contingência
Existem várias maneiras de apresentar resultados de contagens, um método é a tabela de contingência. Uma tabela de contingência apresenta os resultados de duas variáveis categóricas em tabulação cruzada.

Exercícios no Excel (*EstatisticaExcel_02.xlsm*; Planilha *Aula1*)
3.1.1 Considere a Tabela 3.1 como resultado da observação da produção de Tijolos e Telhas. Determine a probabilidade de que num sorteio aleatório

a. seja obtido um produto com defeito
b. seja obtido um tijolo
c. seja obtido um tijolo com defeito
d. seja obtido ou uma telha ou um produto sem defeito.

Tabela 3.1 Resultado da produção de Tijolos e Telhas.

Produto	Com defeito	Sem Defeito
Tijolo	6.000	84.000
Telha	3.000	27.000

3.2 Distribuições de probabilidades - variáveis discretas

Distribuição Binomial (Discreta)

Neste caso, $P(x)$ é probabilidade de se obter exatamente x sucessos em n tentativas.

$P(x)$ = DISTRBINOM(x; n ; p ;FALSO)

x é o número específico de sucessos em n tentativas
n é um número fixo de tentativas
p Probabilidade de um sucesso em uma única tentativa
q Probabilidade de um fracasso eu uma única tentativa
$q = 1 - p$

média: np
variância: npq

Exercícios no Excel
3.2.1 Se a taxa de faltas num setor é, em média, 20% ao dia e o setor tem 40 funcionários, então determine a probabilidade de faltar num mesmo dia mais do que x funcionários, sendo x um número de 0 a 40. O resultado é o gráfico apresentado na Figura 1.2 (página 2).

Distribuição de Poisson (Discreta)

Neste caso, $P(x)$ é probabilidade de se obter x ocorrências num determinado intervalo (pode ser intervalo de tempo, lugar, ...)

$P(x)$ = POISSON(x; L ;FALSO)

x é o número de ocorrências
L é a taxa média de ocorrências por intervalo

média: L
variância: L

Aplicações: Teoria de filas; Ordens de serviço

3.2.2 Se um processo de produção de *cookies* está sob controle espera-se que a o número médio de gotas de chocolates por *cookie* seja 6,0. Determine a probabilidade de que, numa inspeção, um determinado *cookie* apresente

a. cinco ou seis gotas de chocolate;
b. menos de cinco gotas de chocolate;
c. mais de oito gotas de chocolate;
d. Quais seriam as respostas dos itens a, b e c se o número médio de gotas de chocolates por cookie fosse igual a 4,5?

3.3 Distribuições de probabilidades - variáveis contínuas

Para variáveis contínuas, as distribuições são **densidades de probabilidades** $p(x)$. A probabilidade P é definida para um **intervalo de valores**.

Distribuição Normal (contínua)
Para a distribuição Normal: a densidade de probabilidade é dada no Excel: $p(x)$ = DIST.NORM(x; μ; σ;FALSO)
Sendo que: μ é a média e σ é o desvio padrão

A probabilidade é determinada a partir de p(x), para uma faixa ou regiões de valores de x. por exemplo, a probabilidade de x assumir valores menores do que um determinado valor x_0 é dada por:

$P(x < x_0)$ = DIST.NORM(x_0; μ; σ;VERDADEIRO)

Considerando desvios-padrão a partir da média obtemos intervalos com áreas sob a curva. A probabilidade de obter valores de x em determinada região é igual a área sob a curva nesta região. Na Figura 3.1 é apresentado um esboço da distribuição normal, $p(x)$, e as probabilidades, P, de se obter um valor de x entre determinados valores de desvios-padrão em torno da média.

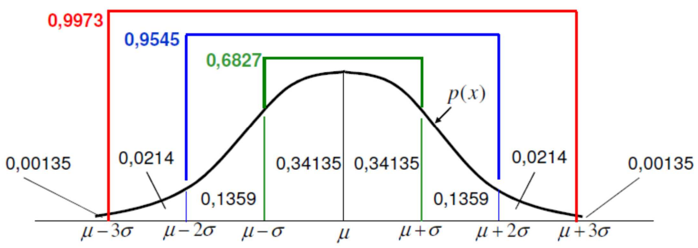


Figura 3.1 Distribuição Normal e probabilidades.

A distribuição Normal, ou Gaussiana é uma das mais utilizadas na estatística. Os dados podem ser "padronizados" em termos da média e do desvio-padrão para que possam ser comparados aos valores típicos da distribuição normal.

Esta padronização é feita transformando os valores de x em Escore-z, como o esquema abaixo.

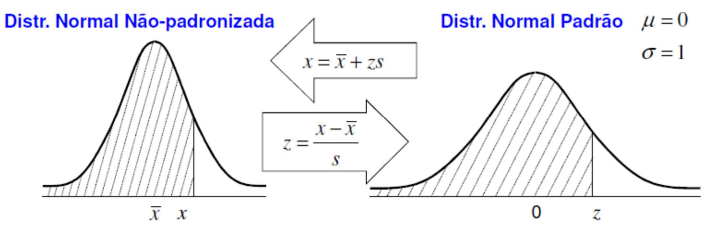


Figura 3.2 Qualquer distribuição normal pode ser transformada em distribuição normal padronizada, na qual a média é zero e o desvio-padrão é um.

Exercícios no Excel (*EstatisticaExcel_02.xlsm*; Planilha *Aula1*)
3.3.1 Uma empresa observou que a distância viajada por caminhão, a cada ano, é distribuída nos moldes da distribuição normal*, com média igual a 80.500 Km e um desvio-padrão igual a 19.300 Km.
a. Que proporção desses caminhões se pode esperar que viaje entre 55.000 Km e 80.500 Km no ano?
b. Que porcentagem de caminhões pode ser esperada que viaje abaixo de 48.000 Km ou acima de 96.500 Km no ano?
c. Quais seriam as respostas de a e b se o desvio-padrão fosse igual a 13.000 Km?

* Este exercício foi retirado de um livro texto, na prática distâncias viajadas seguem distribuições que se desviam muito da distribuição normal. Apresentam leis de potência.

❸ Arquivo de recuperação → ***EstatisticaExcel_03.xlsm***

3.4 Simulação (Números aleatórios e VBA no Excel)

A solução de problemas gerenciais pode ser viabilizada por várias técnicas quantitativas. A simulação é uma técnica que pode ser empregada quando várias outras apresentam limitações.

Por exemplo:
- Determinar a probabilidade de um produto ser lucrativo de acordo com vários cenários;
- Determinar uma quantia em estoque para que a demanda não atendida não ultrapasse determinado percentual;
- Conhecer o número mínimo de telefonistas para que apenas 3% das solicitações não sejam atendidas de imediato;
- Programar a produção, definir níveis de estoque, número de funcionários, planejar investimentos, ...

A base para a técnica de simulação está no **conhecimento das probabilidades** de ocorrência de diversos cenários.

Assim, um determinado evento pode ser simulado através de um **sorteio aleatório**. O sorteio é feito de acordo com as probabilidades de ocorrência de cada variável. Realizando **milhares** de sorteios os diversos cenários são revelados e uma estatística descritiva de todas as possibilidades revela a frequência dos diversos cenários.

Esta forma de realizar a simulação é conhecida como **Simulação de Monte Carlo**.

Exercícios no Excel (*EstatisticaExcel_03.xlsm*; Planilha *Simulacao*)
3.4.1 Utilize uma tabela de frequência de demanda diária e a geração de números aleatórios para simular a demanda diária. Utilize a função ALEATÓRIO() e PROCV.

3.4.2 Um analista está realizando uma simulação para investigar o comportamento do lucro nos próximos 12 meses. Para tanto ele precisa simular valores para vendas. Ele sabe que a variável vendas tem uma distribuição normal com média $ 10.000,00 e desvio padrão $ 800,00. Utilize a função ALEATÓRIO() para gerar uma Probabilidade P(vendas), ou seja um número entre 0 e um. Depois utilize a função INV.NORM para obter o valor de vendas.

3.4.3 A indústria de confecções VP quer determinar o lucro provável de seu produto principal (calça jeans) para o próximo ano. Após pesquisas de mercado e análises financeiras fez a seguinte estimativa:

Preço de venda (PV): $ **25,00** por unidade
Custos e despesas fixos (CDF): $ 100.000,00.

O Lucro é dado por:

Lucro = [(PV - CMP - CMO) x Demanda] – CDF

Sendo que

- o custo de mão de obra (CMO) pode variar de $8,00 à $12,00 por unidade de acordo com uma distribuição uniforme.
- o custo da matéria prima (CMP) depende do fornecedor de acordo com uma tabela de frequência a seguir.

Tabela 3.2: Frequência de CMP.

Fornecedor	Frequência	$/Unid
A	0,1	3,00
B	0,2	4,00
C	0,4	5,00
D	0,2	6,00
E	0,1	7,00

- a demanda tem uma distribuição normal com média igual a 13.000 unidades e um desvio padrão 3.800.

a. Faça uma simulação para o lucro anual

b. Obtenha vários valores para a simulação do lucro anual. Considere a “macro” para copiar e colar os valores simulados descrita no final deste exercício.

c. Faça uma análise estatística descritiva destes possíveis valores.

Copiar e colar resultados de simulações utilizando MACROS.

Para criar uma macro que copie e cole um determinado valor podemos utilizar o gravador de macros e depois inserir alguns comandos no código VBA.

Passo 1: Verifique se a guia “Desenvolvedor” aparece na faixa de opções do Excel. Se não aparece então, no Excel 2010, selecione:

> Arquivo > Opções > Personalizar Faixa de Opções

(x) Desenvolvedor

Crie a macro pelo gravador de macros: Na guia Desenvolvedor:

Passo 2: Clique em “Gravar macro”;

Passo 3: Copie o resultado da simulação (célula G55);

Passo 4: Cole Valores na célula P58;

Passo 5: Pare de gravar;

Passo 6: Aba o VBA e edite a macro para fazer o número de simulações indicadas na célula H57

> Desenvolvedor > Macros > NomeMacro > Editar

A macro final fica com o seguinte código:

```
Sub NomeMacro()
'
' valores Macro
'
' Atalho do teclado: Ctrl+s
'
Nmaximo = Cells(57,8)

 For i = 1 To Nmaximo
   Range("G55").Select
   Selection.Copy
   Cells(57 + i, 16).Select

    Selection.PasteSpecial Paste:=xlPasteValues,
     Operation:=xlNone,
     SkipBlanks _:=False,
     Transpose:=False
    Application.CutCopyMode = False
 Next i
End Sub
```

Sendo que

- Para obter 350 simulações o “For” deve ir até 350, isto é obtido através da célula (57,8) ou seja célula H57.
- “G55” neste caso é o resultado de uma simulação, que muda a cada atualização da planilha. Para outras simulações altere este “Range” para a célula correspondente (D142 na simulação 3.4.4)
- `Cells(57 + i, 16)` indica a célula na qual os resultados serão registrados. Linha (57 + i), Coluna 16 (ou P58, P59, ...).

3.4.4 As perdas de vendas de pneus por falta de estoque têm sido constantes. Abaixo segue uma tabela de frequências observadas para a demanda diária de pneus, bem como uma tabela de frequências observadas de prazos de entrega.

Demanda de pneus

Demanda diária (quantidade de pneus)	Frequencia Observ. em 300 dias
21	30
22	60
23	120
24	60
25	30

Prazo de entrega

Prazo de entrega (dias)	Frequencia Observ. em 210 dias
1	30
2	60
3	120

- Elabore um modelo de simulação da administração de estoques para os próximos 30 dias, levando em conta os dados históricos levantados.
- Calcule o percentual de satisfação da demanda do produto.
- Considere um estoque inicial igual a 40 unidades e o número de pedidos a receber igual a zero.
- Serão realizados pedidos de 50 unidades sempre que o estoque atingir um número de 30 unidades.

Sugestão: Faça a simulação de cada dia numa linha.

- Inicie o dia indicando qual é o dia que está sendo simulado, 1, 2, ... (col A, lin 110 da planilha *Simulacao*)
- Na coluna seguinte verifique o número de unidades recebidas neste dia e na próxima coluna coloque o estoque inicial;
- Na próxima coluna (col D) simule da demanda para este dia.
- Analise o estoque e a demanda. Na próxima coluna indique a demanda atendida.

- Na próxima coluna indique o estoque final.
- Considere uma coluna com “Estoque provisionado” que indicará se já existe pedido juntamente com o estoque final.
- Considere uma coluna com a informação se é necessária ou não a realização do pedido.
- Faça uma coluna (coluna I) com a simulação do prazo de entrega se neste dia houve pedido.
- Uma coluna (coluna J) para especificar o dia de entrega.

Resolução da simulação 3.4.4

Considere inicialmente:
D101 = 40; Estoque inicial
D102 = 0; Número de pedidos a receber
D104 = 50; Quantidade considerada em cada pedido
D105 = 30; Estoque exigido para realizar pedido
E83:G87 = tabela de frequência para simular a demanda
E92:G94 = tabela de frequência para simular o prazo

Primeira Linha ou “Dia 1” (linha 110 da planilha *Simulacao*):
B110 = D102
C110 = D101
D110 = PROCV(ALEATÓRIO();$E$83:$G$87;3;VERDADEIRO)
E110 = SE(D110>C110;C110;D110)
F110 = C110-E110
G110 = F110
H110 = SE(G110<=$D$105;"Sim";"Não")
I110 = SE(H110="Sim";PROCV(ALEATÓRIO();$E$92:$G$94;3;
VERDADEIRO);"")
J110 = SE(H110="Sim";A110+I110;"")

Segunda linha ou “Dia 2” (linha 111)
B111 = CONT.SE($J$110:J110;A111)*$D$104
C111 = F110+B111
G111 = G110-E111+SE(H110="Sim";$D$104;0)

Demais linhas
Cópia da linha 2 com o cuidado de fixar ou não algumas células

Uma vez feita a simulação para 30 dias verificar a demanda atendida nestes 30 dias, por exemplo nas células D140:D142. O valor obtido se refere à uma simulação, deve ser repetido muitas vezes.

A simulação terá sentido quando muitos resultados possíveis forem registrados e a partir destes resultados uma analise estatística descritiva for realizada.

Para este problema desenvolvemos uma macro da mesma forma da macro anterior e copiamos valores simulados.

Veja a análise estatística na planilha completa. Altere valores para pedidos, estoque limite para pedidos...

❹ Arquivo de recuperação → ***EstatisticaExcel_04.xlsm***

Módulo 2

4 Estatística Inferencial

A Estatística Inferencial utiliza padrões observados em amostras para fazer inferências sobre a população que a amostra representa. Estas inferências podem ser estimativas de intervalo de confiança, testes de hipóteses, descrição de associação entre dados (correlação), modelagem entre os dados (análise de regressão).

O Excel possui ferramentas de análise de dados com as principais técnicas e métodos de inferência estatística. Nesta aula vamos explorar a inferência estatística com funções do Excel e utilizar as ferramentas de análise de dados.

4.1 Amostragem

A finalidade da amostragem é permitir fazer inferências e generalizações acerca de características de uma população com base na análise de apenas alguns de seus elementos (CORRAR e THEÓPHILO, 2008).

Caso Oxford Cereals (**OC**) (LEVINE p208):
A Oxford Cereals abastece milhares de caixas de cereais durante um turno de oito horas. Aqui, você é responsável por monitorar a quantidade de cereal em cada caixa. Para serem coerentes com o conteúdo especificado na embalagem de cada caixa, as caixas devem conter uma média aritmética de 368 gramas de cereal. Em razão da velocidade do processo, a quantidade de cereal em cada caixa varia fazendo com que algumas caixas fiquem mal abastecidas enquanto outras hiperabastecidas. Se o processo não estiver funcionando de maneira apropriada, o peso médio das caixas pode se desviar demasiadamente do valor especificado no rótulo.

A pesagem de todas as caixas consumiria muito tempo, pode ser extraída uma amostra de caixas. Para a amostra selecionada você planeja pesar cada caixa e calcular a média aritmética para a amostra. **É preciso determinar a probabilidade de que essa média tenha sido extraída aleatoriamente de uma população cuja média é igual a 368 gramas**.

Com base em sua análise você deve decidir entre manter, alterar ou interromper o processo.

O processo de amostragem começa pela definição da **grade**. A grade é uma lista de itens que compõem a população.

Utiliza-se n para representar o tamanho da amostra e N para representar o tamanho da grade.

Amostras aleatórias simples
Cada item de uma grade tem a mesma chance de ser selecionado em comparação com cada um dos outros itens. A chance de selecionar qualquer membro da grade em particular na primeira seleção é $1/N$.

Amostragem com reposição: Depois de selecionar um determinado item através de um sorteio, você devolve este item para a grade onde ele passa a ter a mesma probabilidade de vir a ser selecionado novamente. O processo de seleção é repetido até se obter o tamanho desejado da amostra, n.

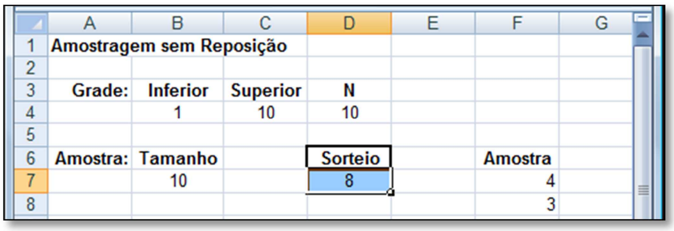
EXCEL: No Excel o sorteio pode ser feito com o comando ALEATÓRIOENTRE(i;j), sendo i = número do item inicial e j = número do item final na grade. Por exemplo, para itens numerados de 187 à 433, realizar uma amostra de tamanho 15 com reposição.

EXCEL solução. Em 15 células diferentes digite:
=ALEATÓRIOENTRE(187;433)

Exercícios no Excel (*EstatisticaExcel_04.xlsm*; Planilha *Aula2*)
4.1.1 Obtenha uma amostra aleatória simples **com reposição**, de tamanho n = 40 para uma grade de itens de 1 a 3000.

Amostragem sem reposição: Uma vez selecionado um determinado item você não pode selecioná-lo novamente. Na primeira seleção a chance de selecionar um determinado item é $1/N$, na segunda seleção é $1/(N-1)$ e assim por diante. O processo continua até selecionar uma amostra desejada, com tamanho n.

EXCEL: Para realizar amostragem sem reposição é necessário desenvolver uma macro. A planilha *AmostragemSR*, contém uma macro desenvolvida neste curso, para esta finalidade.



	A	B	C	D	E	F	G
1	Amostragem sem Reposição						
2							
3	Grade:	Inferior	Superior	N			
4		1	10	10			
5							
6	Amostra:	Tamanho		Sorteio		Amostra	
7		10		8		4	
8						3	

EXCEL solução. Para obter uma amostra aleatória sem reposição informe os limites da grade nas células B4 e C4. Digite o tamanho da amostra n na célula B7 de forma a não ser maior do que N (valor calculado em D4). Para executar obter a amostra digite **Ctrl+t**. A amostra (valores sorteados ao acaso sem repetição) é obtida na coluna F, a partir da célula F7.

Exercícios no Excel (*EstatisticaExcel_04.xlsm*; Planilha *Aula2*)
4.1.2 Obtenha uma amostra aleatória simples sem reposição, de tamanho n = 40, para uma grade de itens de 1 a 3000. Utilize a planilha *Aula2*, B36:E45 para copiar os 40 valores obtidos na planilha *AmostragemSR*.

Distribuições de amostragens
Em muitas aplicações queremos fazer estimativas a partir de amostras, por exemplo, **estimar a média da população a partir da média de uma amostra**.

Os itens selecionados mudam muito de uma amostra para outra, assim é natural selecionarmos amostras com diferentes médias. Ou seja, **temos uma distribuição de médias para as amostras**.

O erro padrão da média aritmética, $\sigma_{\overline{X}}$, é igual ao desvio padrão na população dividido pela raiz quadrada do tamanho da amostra:

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}.$$

Caso Oxford Cereals (**OC**) (LEVINE p230):
Considerando que o desvio padrão do processo de abastecimento de cereais seja 15 gramas, calcule o erro padrão da média se você selecionar aleatoriamente uma amostra de 25 caixas sem reposição, dos milhares de caixas abastecidas durante determinado turno.

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{25}} = \frac{15}{5} = 3\text{ g.}$$

A variação das médias (para amostras de tamanho 25) **é bem menor do que a variação nas caixas individuais**.

Se você selecionar aleatoriamente uma amostra, que tipo de resultado você espera para a média?

A média de todas as médias amostrais possíveis é igual à média da população, chamamos esta média de $\mu_{\overline{X}}$. Então $\mu_{\overline{X}} = \mu$.

TEOREMA DO LIMITE CENTRAL
Mesmo que a distribuição de valores individuais na população não seja normal (gaussiana) a distribuição das médias amostrais será descrita aproximadamente por uma normal, à medida que o tamanho da amostra fique grande o suficiente.

A partir da distribuição normal podemos encontrar a área abaixo de qualquer valor de *x* fazendo a conversão para valores de Z.

Assim, podemos analisar o quanto um valor específico de média obtido, $\overline{X}$, está próximo ou não da média da população, em termos do erro padrão amostral:

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

Caso Oxford Cereals (**OC**) (LEVINE p231):
Se você selecionar aleatoriamente uma amostra de 25 caixas, o que você acha da média aritmética do peso de 25 caixas ser 200 gramas? e 365 gramas? O desvio padrão e média da população são 15 g e 368 g respectivamente.

4.2 Intervalos de confiança

A **inferência estatística** é o processo que utiliza resultados de amostras para extrair conclusões sobre as características de uma população.

Suponha que não sabemos o valor da média da população, podemos a partir da média amostral, estimar um **intervalo de confiança** para a média da população. O intervalo está baseado na distribuição de médias amostrais. O nível de confiança está baseado no percentual de valores de médias amostrais incluídas no intervalo. É simbolizado por $NC = (1-\alpha)\times 100\%$, sendo α a proporção de valores que se encontra fora do intervalo de confiança.

Intervalo de confiança para médias (σ conhecido)

Quando se conhece o desvio padrão da população, o intervalo de confiança para a média da população é dado por:

$$\bar{X} - E \le \mu \le \bar{X} + E$$

sendo $E = Z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$ o erro de amostragem e $Z_{\alpha/2}$ valor correspondente a uma área acumulada de $(1-\alpha/2)$ a partir da distribuição normal padronizada. O tamanho da amostra pode ser estimado com base no nível de confiança e no erro máximo:

$$n = Z^2_{\alpha/2}\frac{\sigma^2}{E^2}$$

o valor obtido deve ser arredondado para o número inteiro acima. **Quando temos uma população finita** (ou seja de tamanho N), o tamanho da amostra para um determinado erro será definido por

$$n = \frac{z^2_{\alpha/2}\sigma^2 N}{(N-1)E^2 + z^2_{\alpha/2}\sigma^2}$$

e o erro é definido por

$$E = Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

Exercícios no Excel (*EstatisticaExcel_04.xlsm*; Planilha *Aula2*)

4.2.1 A partir da amostra obtida no exercício 4.1.1 utilize os valores de pesos de caixas de cereais (planilha *OxfordCereals*) e determine um intervalo de confiança para a média da população, sabendo que o desvio padrão para a população de caixas é igual a 15 g. Interprete o resultado obtido.

Intervalo de confiança para médias (σ desconhecido)

Quando **não se conhece** o desvio padrão da população, o intervalo de confiança para a média da população é dado por:

$$\bar{X} - t_{\alpha/2}\frac{s}{\sqrt{n}} \le \mu \le \bar{X} + t_{\alpha/2}\frac{s}{\sqrt{n}}$$

sendo s o desvio padrão da amostra e $t_{\alpha/2}$ o valor correspondente a uma área acumulada de $(1-\alpha/2)$ a partir da distribuição *t* de Student com $n-1$ graus de liberdade.

Exercícios no Excel (*EstatisticaExcel_04.xlsm*; Planilha *Aula2*)

4.2.2 A partir da amostra obtida no exercício 4.1.2 utilize os valores de pesos de caixas de cereais (planilha OxfordCereals) e determine um intervalo de confiança para a média da população. Considere desconhecido o desvio padrão da população.

4.2.3 Refaça os exercícios 4.2.1 e 4.2.2 com as funções do Excel:

INT.CONFIANÇA.NORM(alfa; desv. pad. ; tamanho)

INT.CONFIANÇA.T(alfa; desv. pad. ; tamanho)

Estas funções fornecem o **Erro** estatístico a partir do nível de confiança “alfa”, do desvio padrão “desv. pad.” e do “tamanho” da amostra, especificados no argumento da função.

Intervalo de confiança para proporção

Uma extensão de intervalo de confiança para dados categóricos. Vamos estimar a proporção de itens que têm certa característica de interesse. A proporção desconhecida da população é representada pela letra grega π, e a proporção da amostra por $p = X/n$, sendo X o número de itens que apresentam a característica de interesse e n o tamanho da amostra. A estimativa do intervalo de confiança para a proporção da população é dada por:

$$p - E \le \pi \le p + E$$

sendo $E = Z_{\alpha/2}\sqrt{\dfrac{p(1-p)}{n}}$.

O número de elementos para realizar uma amostragem pode ser definido em termos do nível de confiança e do erro máximo desejado:

$$n = Z^2_{\alpha/2}\frac{\pi(1-\pi)}{E^2}$$

e neste caso π é a expectativa da proporção da população. Para garantir uma amostra com erro máximo igual a E, utiliza-se $\pi = 0{,}5$ na expressão acima. **Para uma população finita** de tamanho N, o tamanho da amostra será definido por

$$n = \frac{z^2_{\alpha/2}\pi(1-\pi)N}{(N-1)E^2 + z^2_{\alpha/2}\pi(1-\pi)}$$

e o erro é definido por

$$E = z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\sqrt{\frac{N-n}{N-1}}$$

Intervalo unilateral para proporção:

Ex. Para estimar taxas de não-conformidade com o controle interno, auditores extraem uma amostra aleatória de faturas de vendas e determinam a frequência com que as mercadorias foram retiradas sem uma guia de autorização de saída de material do almoxarifado. Comparam os resultados com uma taxa tolerável previamente estabelecida, que corresponde a proporção máxima permitida de itens na população que estão fora dos padrões de conformidade. É comum utilizar um intervalo de confiança unilateral. Ou seja, o auditor obtém a proporção p de itens em não conformidade e calcula o limite superior dado por

$$\text{Limite superior} = p + Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\sqrt{\frac{N-n}{N-1}}$$

onde N é o número de elementos na população. Caso este limite superior seja maior do que a taxa de exceção tolerável, o auditor conclui que a taxa de controle para não-conformidade está demasiadamente elevada.

Exercícios no Excel (*EstatisticaExcel_04.xlsm*; Planilha *Aula2*)

4.2.4 Suponha que, em uma auditoria mensal, você selecione 400 faturas de vendas de uma população de 10.000 faturas.
Na amostra de 400 faturas de vendas, 20 delas estão violando o controle interno. Caso a taxa de exceção tolerável para esse controle interno seja 6%, o que você deveria concluir? Utilize um nível de confiança de 95%.
Qual é o limite de violação permitidas, nesta amostra de 400 faturas de forma a não violar o controle interno?

❺ Arquivo de recuperação → ***EstatisticaExcel_05.xlsm***

4.3 Testes de Hipóteses

Teste de hipóteses é um método de inferência estatística. A partir de uma amostra procura-se examinar a validade de uma declaração sobre um parâmetro da população que investigada.

Exercícios no Excel (*EstatisticaExcel_05.xlsm*; Planilha *Aula2*)

4.3.1-a A Oxford Cereais deseja determinar se o processo de abastecimento de cereais está operando adequadamente (ou seja, se a média permanece dentro dos 368 g especificados). Suponha que uma amostra de 25 caixas de cereais indique uma média aritmética da amostra, $\overline{X}$ igual a 372,5 g, e o desvio padrão da população seja pressuposto como igual a 15 g.

Procedimento para teste de hipóteses

1 Etapa da pesquisa: Escreva a afirmativa original em símbolos matemáticos utilize =, ≠, >, <, ≤, ≥.
2 Etapa da pesquisa: Escreva o oposto da afirmativa original em símbolos matemáticos.
3 Etapa estatística: A hipótese nula, H_0, é (entre 1 e 2) a afirmativa que contém a igualdade, **mas utilize somente o símbolo de igual para H_0**.
4 Etapa estatística: A hipótese alternativa H_1, é a outra.
5 Etapa estatística: Calcule a Estatística teste e determine o valor-P (P-value): Probabilidade associada à estatística teste.
6 Etapa estatística: Se valor-P < α rejeite H_0; caso contrário, não rejeite H_0.
7 Etapa da pesquisa: Retome a afirmativa original e estabeleça uma conclusão.

Testes para uma amostra

Teste z para médias (σ conhecido)

Estatística teste:

$$Z_{teste} = \frac{\overline{X}-\mu}{\dfrac{\sigma}{\sqrt{n}}}$$

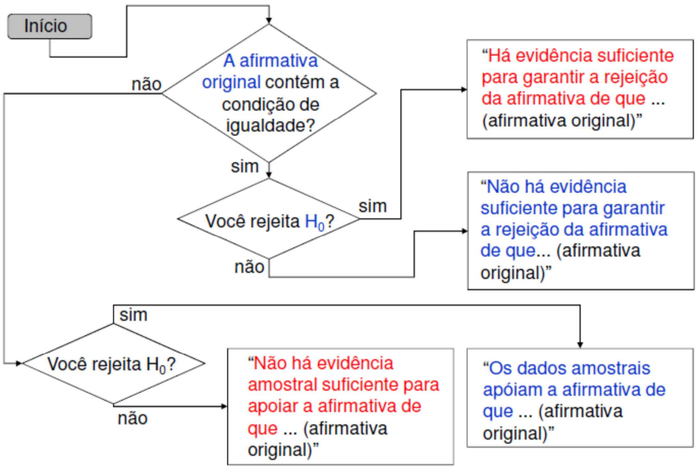


Fig. Esquema para estabelecer conclusão no teste de hipóteses.

Exercícios no Excel (*EstatisticaExcel_05.xlsm*; Planilha *Aula2*)

4.3.1-b Utilize a amostra obtida em 4.1.2 e a função do Excel:
=TESTE.Z(B48:E57;E74;15)

Teste t para médias (σ desconhecido)

Estatística teste:

$$t_{teste} = \frac{\overline{X}-\mu}{\dfrac{s}{\sqrt{n}}}$$

t_{teste} é o valor obtido a partir da distribuição t de Student com $n-1$ graus de liberdade.

Teste z para proporção

Estatística teste:

$$Z_{teste} = \frac{p-\pi}{\sqrt{\dfrac{\pi(1-\pi)}{n}}} \quad \text{ou} \quad Z_{teste} = \frac{X-n\pi}{\sqrt{n\pi(1-\pi)}}$$

sendo X o número de sucessos na amostra.

Testes para duas amostras

Nesta etapa estendemos o teste de hipóteses para procedimentos que comparam estatísticas oriundas de duas amostras, extraídas de duas populações.

Caso BLK Alimentos (**BLK**) (LEVINE p324):

O tipo de disposição de mercadorias utilizado em um supermercado afeta as vendas dos produtos? Você deseja comparar o volume de vendas de refrigerantes da marca BLK quando o produto é exposto em uma localização normal de prateleira com o volume de vendas quando o produto é exposto em posições especiais de ponta de corredor. Para testar a eficácia da exposição em ponta de corredor você seleciona 20 filiais da cadeia de supermercados que apresentam volumes de vendas similares. Depois designa aleatoriamente 10 dentre as 20 para um grupo (grupo 1) e as outras 10 para outro grupo (grupo 2). Os gerentes das 10 lojas do grupo 2 passam a utilizar a exposição em locais especiais em ponta de corredor. Ao final de uma semana são registradas as vendas do refrigerante da marca BLK. De que modo você determinaria se as vendas do refrigerante BLK nas lojas que fizeram uso da exposição promocional de ponta de corredor são iguais àquelas ocorridas quando o refrigerante é exposto em prateleiras com localização regular? Como utilizar esta informação para incrementar as vendas dos refrigerantes BLK?

Teste t de variância agrupada para a diferença entre duas médias (populações independentes)

Hipóteses: H_0: $\mu_1 = \mu_2$ ou $\mu_1 - \mu_2 = 0$

H_1: $\mu_1 \neq \mu_2$ ou $\mu_1 - \mu_2 \neq 0$

Estatística teste:

$$t_{teste} = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

sendo $s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1-1)+(n_2-1)}$ e gl = $n_1 + n_2 - 2$.

No Excel obtenha diretamente o valor-P para este teste:
=TESTE.T(Dados1;Dados2; 2 – opção bi-caudal; 2 – tipo de teste)

Exercícios no Excel (*EstatisticaExcel_05.xlsm*; Planilha *Aula2*)

4.3.2 Caso BLK: Uma amostra com volumes de vendas semanais de refrigerantes BLK a partir de 10 filiais que utilizam exposição em prateleiras regulares. Outra amostra a partir de 10 filiais que utilizam exposição em pontas de corredor. Teste a afirmativa de que as vendas semanais de refrigerantes BLK é a mesma quando é utilizada uma localização regular e quando é utilizada uma localização em ponta de prateleira. Os registros de volumes de vendas estão no arquivo Excel.

Teste t em pares (populações dependentes)

Hipóteses: H_0: $\mu_D = 0$, onde $\mu_D = \mu_1 - \mu_2$

H_1: $\mu_D \neq 0$

Estatística teste:

$$t_{teste} = \frac{\overline{D} - \mu_D}{\frac{s_D}{\sqrt{n}}}$$

sendo $\overline{D}$ a média das n diferenças observadas e s_D o desvio padrão das n diferenças observadas.

No Excel obtenha diretamente o valor-P para este teste:
=TESTE.T(Dados1;Dados2; 2 – opção bi-caudal; 1 – tipo de teste)

Exercícios no Excel (*EstatisticaExcel_05.xlsm*; Planilha *Aula2*)

4.3.3 A Associação Automobilística da América (AAA) conduziu um teste de milhagem para comparar a milhagem de combustível em testes de direção realizados por membros da AAA na vida real com testes de direção feitos na cidade e em auto-estradas conforme o padrão do governo. São obtidos resultados, para um mesmo carro, da milhagem de combustível nos dois tipos de teste de direção. Determine se existe alguma diferença entre a média da milhagem de combustível de acordo com os dois tipos de testes de direção. Utilize os dados de milhas por galão na planilha *Aula2*, célula B129.

Os testes de hipóteses também estão disponibilizados na análise de dados do Excel. Para habilitar a análise de dados:

- Arquivos > Opções > Suplementos
  Suplementos do Excel [ir]
  [x] Ferramentas de Análise
  [x] Solver

Com isto a faixa de opções exibe, na aba "Dados" a categoria "Análise" com as opções "Análise de Dados" e "Solver".

❻ Arquivo de recuperação → ***EstatisticaExcel_06.xlsm***

Exercícios adicionais (na planilha *Aula2*, célula A147)

4.2.5 (LEVINE 8.78) Um auditor precisa estimar a percentagem de vezes em que a empresa deixa de seguir um procedimento de controle interno. É selecionada uma amostra de tamanho 50, de uma população de 1000 itens, e em sete casos o procedimento de controle interno não foi seguido.

a. Construa uma estimativa para o intervalo de confiança unilateral de 90% correspondente à proporção da população de itens na qual o procedimento de controle interno não foi seguido.
b. Se a taxa de exceção tolerável for 0,15, o que o auditor deve concluir?

4.2.6 (LEVINE 8.80) Uma loja que vende móveis de quarto está realizando um inventário de artigos para cama (colchões, estratos de molas e estruturas) em estoque. Uma auditora da loja deseja estimar a média aritmética do valor das camas em estoque naquele momento. Ela deseja ter 99% de confiança de que sua estimativa para a média aritmética do valor está correta, dentro dos limites de $\pm$\$100. Com base em experiências passadas, ela estima que o desvio-padrão para o valor de uma cama seja igual a \$200,00.

a. Que tamanho de amostra ela deve selecionar?
b. Utilizando o tamanho de amostra selecionado em (a), uma auditoria foi conduzida, com os seguintes resultados:

$\bar{x} = \$1654{,}27$ $\qquad$ $S = \$184{,}62$

Construa uma estimativa para o intervalo de confiança de 99% do valor total das camas em estoque ao final do mês, no caso de existirem 258 camas em estoque.

4.2.7 LEVINE (8.81) Uma característica importante da qualidade em um processo de enchimento de saquinhos de chá é o peso do chá em cada saquinho. Neste exemplo, o peso no rótulo da embalagem indica que a média aritmética da quantidade correspondente a 5,5 gramas de chá em um saquinho. Se os saquinhos tiverem seu peso inferior ao especificado, surgem dois problemas. Em primeiro lugar, os consumidores podem não conseguir que o chá fique tão forte quanto desejam. Em segundo lugar, a empresa pode estar violando normas de peso e medidas. Por outro lado, se a média aritmética da quantidade de chá em um saquinho exceder o peso especifico no rótulo, a empresa estará desperdiçando produto. Conseguir uma quantidade exata de chá em cada saquinho é uma questão problemática, em razão da variação da temperatura e da umidade dentro da fábrica, de diferenças na densidade do chá e da operação extremamente rápida de abastecimento da máquina (aproximadamente 170 saquinhos por minuto). Os dados a seguir correspondem aos pesos, em gramas, de uma amostra de 50 saquinhos de chá produzidos em uma hora por uma única máquina:

5,65	5,44	5,42	5,40	5,53	5,34	5,54
5,45	5,52	5,41	5,57	5,40	5,53	5,54
5,55	5,62	5,56	5,46	5,44	5,51	5,47
5,40	5,47	5,61	5,53	5,32	5,67	5,29
5,49	5,55	5,77	5,57	5,42	5,58	5,58
5,50	5,32	5,5	5,53	5,58	5,61	5,45
5,44	5,25	5,56	5,63	5,50	5,57	5,67
5,36						

a. Construa uma estimativa para o intervalo de confiança de 99% para a média aritmética da população relativa ao peso dos saquinhos de chá.
b. A empresa está atendendo aos requisitos estabelecidos no rótulo, de que a média aritmética da quantidade de chá em um saquinho é igual a 5,5 gramas?

4.3.4 (LEVINE 9.55) Um artigo declarou que uma ida típica a um supermercado leva uma média aritmética de 22 minutos. Suponha que, para testar essa afirmativa, você selecione uma amostra de 50 consumidores em um supermercado local. A média aritmética do tempo de compra para uma amostra de 50 consumidores é de 25,36 minutos, com um desvio-padrão de 7,24 minutos. Utilizando o nível de significância de 0,1, existem evidências de que a média aritmética do tempo de compra no supermercado seja diferente do valor alegado de 22 minutos?

4.3.5 (LEVINE 9.57) Um fabricante de chocolate utiliza máquinas para embalar as barras à medida que elas se movimentam ao longo de uma linha de produção. Embora as embalagens apresentem no rótulo 227 gramas, a empresa deseja que as embalagens contenham 232 gramas, de modo tal que efetivamente nenhuma das embalagens contenha menos de 232 gramas. Uma amostra de 50 embalagens é selecionada periodicamente, e o processo de embalagem é interrompido caso existam evidências de que a média aritmética da quantidade embalada seja diferente de 232 gramas. Suponha que em uma determinada amostra com 50 embalagens a média aritmética da quantidade contida seja 231,3 gramas, com um desvio-padrão amostral igual a 1,45 gramas.
a. Existem evidências de que a média aritmética da população seja diferente de 232 gramas? (Utilize um nível de significância de 0,05).
b. Determine o valor-*P* e interprete o seu significado.

4.3.6 (LEVINE 9.58) Considerando uma amostra (feita em seis cadeias de cinemas) de preços (em dólares) para dois ingressos, incluída a taxa de serviços de emissão pela Internet, uma embalagem grande de pipocas e dois refrigerantes médios, os resultados obtidos foram

36,15	31,00	35,05	40,25	33,75	43,00

a. Utilizando o nível de significância de 0,05, existem evidências de que a média aritmética do preço seja diferente de $35,00?
b. Determine o valor-*P* e interprete o seu significado.
c. Qual pressuposto sobre a distribuição da população é necessário em (a) e (b)
d. Você acredita que o pressuposto declarado em (c) está sendo seriamente violado?

4.3.7 (LEVINE 9.60) Os dados a seguir representam a quantidade de refrigerante abastecido em uma amostra de 50 garrafas de dois litros consecutivas. Os resultados, apresentados horizontalmente na ordem em que as garrafas estão sendo abastecidas, foram (em litros):

2,109	2,086	2,066	2,075	2,065	2,057	2,052
2,044	2,036	2,038	2,031	2,029	2,025	2,029
2,023	2,020	2,015	2,014	2,013	2,014	2,012
2,012	2,012	2,012	2,005	2,003	1,999	1,996
1,997	1,992	1,994	1,986	1,984	1,984	1,973
1,975	1,971	1,969	1,966	1,967	1,963	1,957
1,951	1,951	1,947	1,941	1,941	1,938	1,908
1,894						

a. no nível de significância de 0,05, existem evidências de que a média aritmética da quantidade de refrigerante abastecida seja diferente de 2 litros?
b. Determine o valor-*P* em (a) e interprete o seu significado.

5 Análise Multivariada de Dados

5.1 Introdução à análise multivariada de dados

A análise multivariada refere-se a métodos estatísticos que tornam possível a análise simultânea de medidas múltiplas para cada indivíduo (CORRAR, PAULO, DIAS FILHO, 2007).

Análise multivariada de dados é uma extensão natural das análises univariadas e bivariadas. A análise multivariada é utilizada para estudar modelos em que todas as variáveis sejam aleatórias e inter-relacionadas, de modo que seus diferentes efeitos não possam ser interpretados de forma separada (FAVERO et. al. 2009).

Qualquer procedimento para análise de dados depende de pelo menos quatro definições iniciais:

- Número de variáveis a serem analisadas ao mesmo tempo
- Nível de mensuração das variáveis de interesse. As variáveis podem ser métricas (quantitativas) ou não-métricas (qualitativas)
- Interesse descritivo ou inferencial da análise. Médias, frequências, dispersão, testes de hipóteses.
- Interesse na inter-relação ou relação de dependência das variáveis em estudo. Exemplos técnicas de inter-relação: análise de agrupamentos, fatorial, escalonamento

multidimensional. Exemplos de técnicas de dependência: regressão linear e não linear, discriminante, regressão logística, análise de sobrevivência.

A base para a análise multivariada de dados está na correlação entre variáveis e/ou observações. Desta forma, no próximo tópico estudamos correlações entre variáveis e algumas aplicações.

5.2 Correlação

A análise de correlação investiga a relação entre duas populações de duas ou mais variáveis. Assim, existe uma correlação entre duas variáveis quando uma delas está relacionada com a outra de alguma maneira.

Coeficiente de correlação linear de Pearson

A correlação linear entre duas variáveis pode ser expressa numericamente através do coeficiente de correlação linear de Pearson, r. Considerando duas variáveis, x e y, o coeficiente de correlação de Pearson é dado por:

$$r=\frac{n\left(\sum xy\right)-\left(\sum x\right)\left(\sum y\right)}{\sqrt{[n\left(\sum x^2\right)-\left(\sum x\right)^2][n\left(\sum y^2\right)-\left(\sum y\right)^2]}}$$

sendo n o número de pares de observações na amostra analisada.

No Excel: r = PEARSON(valores_x;valores_y)
r = CORREL(valores_x;valores_y)

Para a população o coeficiente é denominado ρ (rho).

O coeficiente de correlação de Pearson pode assumir valores de -1 à +1. Quando r se aproxima de -1 ou +1 a **relação linear entre as variáveis numéricas passa a ser mais forte**. Próximo de zero a relação linear é fraca. O sinal indica se as variáveis estão positivamente correlacionadas (maiores valores de x geralmente fazem par com maiores valores de y), ou negativamente correlacionadas (maiores valores de x geralmente fazem par com menores valores de y). A existência de forte correlação não implica um efeito de causalidade. Indica somente tendências nos dados.

Exercícios Excel (*EstatisticaExcel_06.xlsm; AnaliseMultivariada*)

5.2.1 Considere os dados do exercício 4.3.3. Calcule o coeficiente de correlação r para analisar a correlação entre a milhagem calculada pelos proprietários e a milhagem de acordo com os padrões do governo americano.

Matriz de Correlação

Caso Omni Foods (LEVINE p499):

A Omni Foods, uma empresa de produtos alimentícios, planeja lançar uma nova barra energética a OmniPower. O mercado já contém diversas barras de sucesso, o gerente de marketing deseja maximizar o efeito dos seus planos de marketing. Em particular, precisa saber o efeito que o preço e as promoções internas da loja terão sobre as vendas de OmniPower. Trata-se de uma análise multivariada.

Exercício no Excel

5.2.2 Considere uma amostra de 34 lojas para um estudo de teste de mercado para a barra OmniPower. Veja na planilha *AnaliseMultivariada* as variáveis volume de vendas, preço de uma barra medido em centavos e o orçamento mensal para despesas com promoções internas da loja. Determine uma matriz de correlação e analise os valores dos coeficientes de correlação linear de Pearson para cada par de variáveis.

Utilize também a ferramenta de análise de dados.

5.3 Covariância

A covariância também é uma medida da força de uma relação linear entre duas variáveis numéricas (X e Y). A covariância da amostra é dada por

$$\operatorname{cov}(X,Y)=s_{X,Y}=\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)\left(Y_i-\bar{Y}\right)}{n-1}.$$

No Excel: =COVARIAÇÃO.S(valores_x; valores_y)

A covariância para $Y=X$ é a variância de X, ou seja s_X^2.

No Excel: = VAR.A(valores_x)

O valor da covariância tem uma deficiência como medida da relação linear entre duas variáveis. Como ela pode assumir qualquer valor, fica difícil de fazer comparações. Para comparações é melhor utilizar o coeficiente de correlação de Pearson, que é uma covariância com valores entre -1 e 1.

Porém, a covariância tem aplicações em análise de risco e retorno de ações. Considere X e Y como retorno de ações, se quando X estiver abaixo (acima) da média Y também estiver então a covariância será positiva. Se quando X estiver abaixo (acima) da média Y estiver acima (abaixo) da média então a covariância será negativa.

Matriz de Covariância

Exercício no Excel

5.3.1 Considere uma amostra Taxas de retorno trimestral de quatro títulos, planilha **AnaliseMultivariada**. Calcule uma matriz de Covariância

$$\tilde{COV}=\begin{vmatrix} S_{RA}^2 & S_{RA,RB} & S_{RA,RC} & S_{RA,RD} \\ S_{RB,RA} & S_{RB}^2 & S_{RB,RC} & S_{RB,RD} \\ S_{RC,RA} & S_{RC,RB} & S_{RC}^2 & S_{RC,RD} \\ S_{RD,RA} & S_{RD,RB} & S_{RD,RC} & S_{RD}^2 \end{vmatrix}$$

Observe os elementos da matriz.

Risco de uma carteira de ações (Modelo de Markowitz)

O Risco da Carteira é dado pelo valor esperado da variância.

O retorno médio de cada investimento pode ser agrupado numa matriz de retornos:

$$\tilde{R} = |R_1 \quad R_2 \quad R_3 \quad R_4|$$

Da mesma forma as participações em cada investimento podem ser escritas como uma matriz:

$$\tilde{P} = |p_1 \quad p_2 \quad p_3 \quad p_4|$$

Exercício no Excel

5.3.2 Considere a matriz de covariância obtida em 5.3.1 e determine o Retorno Esperado definido como

$$R(E) = \tilde{R}\,\tilde{P}^T$$

e a Variância da Carteira, definida por:

$$VarCart = \tilde{P}\ \tilde{COV}\ \tilde{P}^T$$

Altere os valores da matriz de participações e observe o valor da variância da carteira.
Função Excel: =MATRIZ.MULT(matriz_A;matriz_B)

Para completar o comando de Matriz no Excel, selecione as células para obter a resposta (matriz inteira) e digite:
[F2] e depois [ctrl]+[Shift]+[Enter]

5.4 Análise de Regressão

A análise de regressão possibilita desenvolver um modelo para prever valores de uma variável numérica com base em valores de outras variáveis.

Na regressão a variável que se deseja prever é chamada variável dependente. As variáveis utilizadas para fazer a previsão são chamadas variáveis independentes.

❼ Arquivo de recuperação → ***EstatisticaExcel_07.xlsm***

Regressão Linear Simples

Exercícios Excel (*EstatisticaExcel_07.xlsm; AnaliseMultivariada*)

5.4.1 Examine a relação entre tamanho de uma loja e as vendas anuais com base numa amostra de 14 lojas filiais da Sunflowers Roupas. Faça um diagrama de dispersão para analisar visualmente uma possível relação linear entre os dados. Calcule o coeficiente de correlação de Pearson. Desenvolva um modelo de regressão linear simples, utilize os conceitos a seguir.

Equação da Reta

O valor previsto de Y é igual ao intercepto de Y somado à inclinação vezes o valor de X :

$$\hat{Y} = b_0 + b_1 X \ ,$$

b_0 é o intercepto e b_1 é a inclinação.

A previsão

A forma mais rápida de se obter um valor previsto é utilizando a função PREVISÃO(X; DadosY;DadosX)

Coeficiente linear, Coeficiente Angular (beta)

O intercepto b_0 é o coeficiente linear, no Excel é obtido através da função INTERCEPÇÃO:

b_0 = INTERCEPÇÃO(valores_y; valores_x)

A inclinação b_1 é o coeficiente angular (beta). No Excel é obtido através da função INCLINAÇÃO:

b_1 = INCLINAÇÃO(valores_y; valores_x)

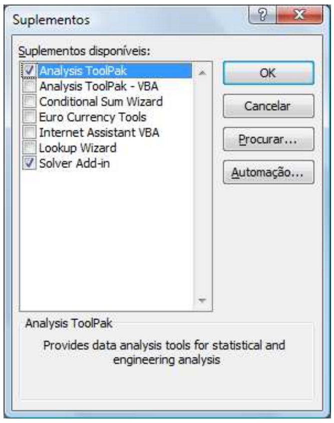
Coeficiente de determinação R^2

Uma medida que informa a qualidade do modelo é o coeficiente de determinação, o R^2. É numericamente igual ao quadrado do coeficiente de correlação de Pearson. Quanto mais próximo de 1 melhor é o modelo.

Ferramentas de Análise de Dados no Excel:

Entre as ferramentas estatísticas disponíveis no Excel temos a análise de regressão. Lembre, para utilizá-las devemos habilitar os suplementos de Ferramentas de Análise. Para o Excel 2007:

- Clique no Botão do Office (localizado na parte superior esquerda da planilha);
- Clique em Opções do Excel;
- Clique em suplementos;
- Clique em [ir] no campo Gerenciar [suplementos do Excel]
- Selecione a ferramenta de análise.



A aba **Dados** da faixa de opções apresenta a categoria de **Análise**, com os suplementos selecionados.

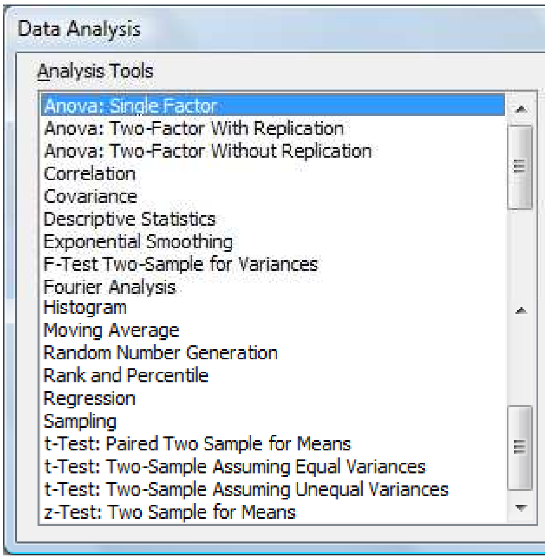
Clique em **Análise de Dados** para acessar as ferramentas de análise disponibilizadas.

A seguir são indicadas as ferramentas estatísticas que podem ser utilizadas para o conteúdo visto até este ponto.

Descriptive Statistics: Centro, Variação, Intervalo de Confiança
Histogram (Histograma): Distribuição.
Rank and Percentile (Rank e Percentil): Medidas de posição relativa.

Random Number Generator (Gerador de números aleatórios): Simulação, Amostragem com reposição.
Sampling (Amostragem): Amostragem com reposição.
t-Test: Two-Sample Assuming Equal Variances: Teste t de variância agrupada para a diferença entre duas médias (populações independentes)
t-Test: Paired Two-Sample for Means: Teste t em pares (populações dependentes)
Correlation (correlação): Matriz de correlação de Pearson.
Covariance (covariância): Matriz de covariância.
Regression (regressão): Regressão linear simples e múltipla.

O quadro a seguir apresenta as ferramentas de análise disponíveis.



Exercícios Excel (*EstatisticaExcel_07.xlsm; AnaliseMultivariada*)
5.4.2 Considere, para um determinado produto, os dados de preços e quantidades vendidas em vários anos conforme a tabela abaixo.

Anos	Quantidade (Q) (1.000 un.)	Preço (P) ($ 1.000)
1	2	4
2	1	6
3	3	3
4	1	5
5	4	1
6	3	2

Encontre a função linear que melhor se ajusta aos dados ou seja os valores de b_0 e de b_1 para o caso 5.4.2.

Faça uma previsão para quantidade (Q) quando o preço (P) for igual a 3,5.

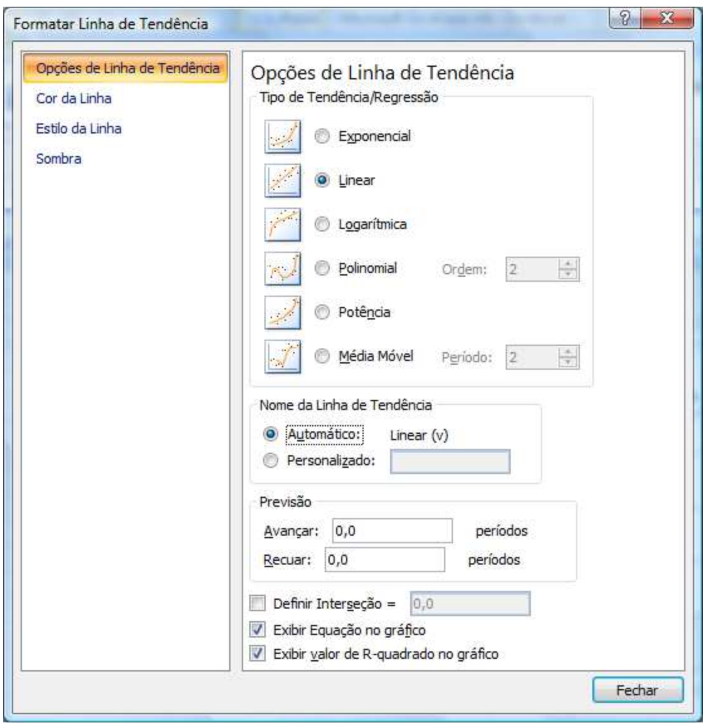
Regressão não Linear

Através de um diagrama de dispersão podemos visualizar se a relação entre duas variáveis pode ser ajustada por um modelo linear ($\hat{Y} = b_0 + b_1 X$) ou se existe indícios de não-linearidade, ou seja, outro tipo de função diferente da linear.

Considere os dados de volume de venda (v) e gastos com propaganda (p) registrados nos últimos anos, informados na tabela abaixo.

Anos	Gastos com propaganda (p) ($ 1.000,00)	Volume de vendas (v) (1.000 un.)
1	7,0	7,0
2	6,0	5,0
3	4,5	3,0
4	3,0	1,5
5	2,0	1,0
6	1,0	0,5
7	8,0	7,0
8	8,0	9,0

Exercícios Excel (*EstatisticaExcel_07.xlsm; AnaliseMultivariada*)
5.4.3 Faça um diagrama de dispersão para os dados informados na tabela acima. Através do ajuste de "linha de tendência" obtenha uma regressão linear e informe o coeficiente de determinação R^2. Explore outras formas funcionais disponibilizadas no Excel conforme a figura abaixo. Para cada tipo de tendência informe o coeficiente de determinação R^2.



Regressão Linear Múltipla

Quando o problema apresentado tem por objetivo prever uma variável dependente a partir do conhecimento de mais de uma variável independente, a técnica estatística é denominada Regressão Múltipla.

Considerando n variáveis independentes, a variável dependente é escrita como uma função linear das outras variáveis da seguinte forma:

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + \ldots + b_n X_n$$

A análise de regressão múltipla realiza estimativas das constantes b_i 's.

Como visto na página anterior, existe no Excel a ferramenta **Regressão**, nas opções de ferramentas de análise de dados, utilize esta ferramenta de análise para o exercício que segue.

Exercícios Excel (*EstatisticaExcel_07.xlsm; AnaliseMultivariada*)

5.4.4 Uma base de dados extraída da revista EXAME – 500 Melhores & Maiores de 2001 contém indicadores financeiros de referentes a empresas brasileiras de diversos setores econômicos. Considere na planilha *Indicadores* as seguintes variáveis:

- RENTAT (Rentabilidade do Ativo)
- RENTPL (Rentabilidade do Patrimônio Líquido)
- ALOPER (Alavancagem Operacional)
- MARVEN (Margem Líquida de Vendas)
- ALFIN (Alavancagem Financeira)
- LUPRE (Lucro = 1 ou Prejuízo = 0)

O exercício consiste em testar estatisticamente as relações de análise de rentabilidade. A amostra contém 297 observações válidas.

- **Correlação**: Analise a matriz de correlação linear de Pearson.
- **Regressão Linear Múltipla**: Considere a RENTPL como sendo a variável dependente e escreva a equação que relaciona RENTPL com as outras variáveis.

Comece pela planilha “Indicadores”:

Analise a correlação entre os dados

Obtenha a matriz de correlação na célula B119 da planilha “AnaliseMultivariada”

Formate a matriz obtida. Utilize recursos da categoria “Estilo”, na aba “Página Inicial”, Ex. Formatação condicional.

Obtenha um modelo de Regressão Linear para explicar RENTPL

- Selecione a aba “Dados” e na categoria “Análise de Dados”
 Escolha: Regressão
- Para Intervalo Y: Escolha a variável “RENTPL”
- Para Intervalo X: Escolha as variáveis:
 ALOPER; ALFIN; LUPRE; MARVEN; RENTAT
(Escolha a partir da linha com rótulos)
- Marque [x] Rótulos na primeira linha
- Intervalo de saída, na planilha “AnaliseMultivariada”, escolha a célula B128.

Como resultado o Excel Fornece todos os coeficientes da função linear que relaciona as variáveis ALOPER, ALFIN, LUPRE, MARVEN e RENTAT com a variável dependente RENTPL:

No Resumo dos resultados temos o quadro abaixo, no qual pode ser visto que o modelo tem um R^2 ajustado igual a 0,8005. Ou seja, variações de ALOPER, ALFIN, LUPRE, MARVEN e RENTAT conseguem explicar 80,05% das variações de RENTPL.

Estatística de regressão	
R múltiplo	0,8966
R-Quadrado	0,8039
R-quadrado ajustado	0,8005
Erro padrão	8,4374
Observações	297

Ainda, no quadro abaixo obtemos os coeficientes (betas) da regressão. As variáveis ALOPER e ALFIN apresentaram baixo poder de explicação (o valor-P é alto), não devem ser consideradas no modelo.

	Coeficientes	*Erro padrão*	*Stat t*	*valor-P*
Interseção	1,63928	0,975	1,682	0,094
ALOPER	-0,01153	0,046	-0,252	0,801
ALFIN	-0,02955	0,089	-0,334	0,739
LUPRE	-3,66001	1,079	-3,391	0,001
MARVEN	-5,79801	0,669	-8,668	0,000
RENTAT	28,42287	0,887	32,028	0,000

8 Arquivo de recuperação → ***EstatisticaExcel_08.xlsm***

Módulo 3

6 Análise de Séries Temporais

Séries temporais são formadas por dados registrados ao longo do tempo.

Através da análise de séries temporais estudamos tendências, variações sazonais, variações cíclicas, ... Nesta revisão vamos estudar diferentes modelos de previsão; analisar como escolher o modelo de previsão mais apropriado; estudar índices de preços.

Existem muitos modelos matemáticos para estudo de séries temporais. Talvez o mais fundamental seja o modelo clássico multiplicativo que identifica fatores de tendência, variações sazonais, cíclicas e irregulares.

Componentes de uma série temporal:

Tendência (T): Movimento geral, ascendente ou descendente, de longo prazo.

Componente Cíclico (C): Oscilações periódicas ao longo da série, geralmente de 2 a 10 anos. Correlacionado ao ciclo do negócio.

Componente Irregular (I): Dados que não seguem a tendência modificada pelo componente cíclico são considerados aleatórios ou irregulares.

Componente Sazonal (S): Quando os dados são mensais ou trimestrais as oscilações têm período mais curto de acordo com data no ano.

Modelo Clássico multiplicativo

$$Y_i = T_i \times S_i \times C_i \times I_i$$

6.1 Ajuste de série temporal

O primeiro passo em uma análise de séries temporais é **desenhar o gráfico para os dados** e observar quaisquer padrões que ocorram ao longo do tempo.

Ajuste de uma série temporal anual

Considere os valores de receita operacionais líquidas anuais (em bilhões de dólares) da Coca-Cola Company de 1975 a 2005 (LEVINE et al. 2008) apresentados na Tabela 6.1. Os anos foram codificados de forma que t = 0 corresponde ao ano 1975. O gráfico de série temporal pode ser obtido no Excel pelo gráfico de dispersão de dados com linhas retas e marcadores. O eixo x é o tempo (que pode ser codificado ou não).

Uma análise gráfica (veja Figura 6.1) revela que os valores apresentam uma tendência ascendente. Também apresentam variações cíclicas e variações irregulares. Vamos estudar alguns modelos de **ajustes** de tendências: Médias móveis, Tendências Linear, Quadrática, Exponencial. Pretende-se descrever a série através de um modelo matemático, onde, a partir de valores observados Y_i fazemos estimativas (ou ajustes) $\hat{Y}_i$.

Médias móveis

Médias móveis para um período L consistem em uma série de médias calculadas ao longo do tempo para uma sequência de L valores observados. São representadas pelo símbolo $MM(L)$. Por exemplo, para uma série como a da Tabela 6.1 com 31 anos, o ajuste $MM(5)$ ou estimativas, são dadas por

$$\hat{Y}_3 = \frac{Y_1 + Y_2 + Y_3 + Y_4 + Y_5}{5}$$

$$\hat{Y}_4 = \frac{Y_2 + Y_3 + Y_4 + Y_5 + Y_6}{5}$$

$$\hat{Y}_5 = \frac{Y_3 + Y_4 + Y_5 + Y_6 + Y_7}{5}$$

$$\vdots$$

$$\hat{Y}_{28} = \frac{Y_{26} + Y_{27} + Y_{28} + Y_{29} + Y_{30}}{5}$$

Cada estimativa é escrita como $\hat{Y}_i$, sendo i um índice que representa o período para o qual a estimativa está sendo realizada. Na forma como descrito anteriormente, obtemos uma série que **suaviza** as variações aleatórias bruscas permitindo observar variações cíclicas e tendência. Mas o ajuste de MM pode ser útil para realizar **comparações**, basta considerar que o cálculo representa estimativa para o período atual.

Ano	t	Receitas $bi	Ano	t	Receitas $bi
1975	0	2,9			
1976	1	3,1	1991	16	11,6
1977	2	3,6	1992	17	13,0
1978	3	4,3	1993	18	14,0
1979	4	4,5	1994	19	16,2
1980	5	5,3	1995	20	18,0
1981	6	5,5	1996	21	18,5
1982	7	5,9	1997	22	18,9
1983	8	6,6	1998	23	18,8
1984	9	7,2	1999	24	19,8
1985	10	7,9	2000	25	20,5
1986	11	7,0	2001	26	20,1
1987	12	7,7	2002	27	19,6
1988	13	8,3	2003	28	21,0
1989	14	9,0	2004	29	21,9
1990	15	10,2	2005	30	23,1

Tabela 6.1 Coca-Cola Company: Receitas Operacionais Líquidas anuais em bilhões de dólares.

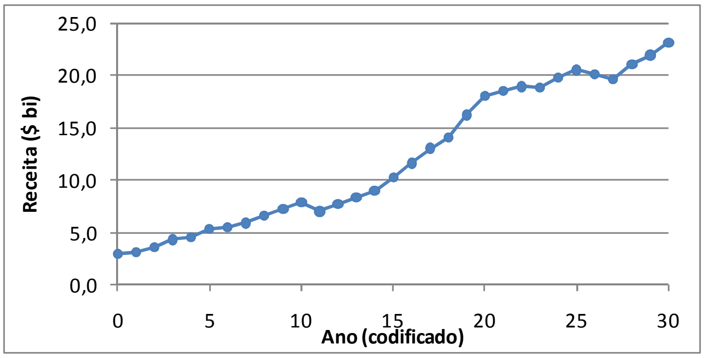


Figura 6.1 Coca-Cola Company: Receitas Operacionais Líquidas anuais em bilhões de dólares.

Assim, ao invés de chamar a estimativa final de $\hat{Y}_{28}$ a chamamos de $\hat{Y}_{30}$, mas o valor é o mesmo. Podemos assim verificar se o valor atual (real da série) está acima ou abaixo da média dos últimos períodos, ou ainda, podemos analisar datas nas quais ocorrem inversões.

Exercícios Excel (*EstatisticaExcel_08.xlsm; A3Series*)

6.1.1 Considere os valores da Tabela 6.1

(a) faça um ajuste de média móvel $MM(5)$.

(b) faça um ajuste de média móvel $MM(9)$.

(c) faça gráficos com os resultados.

Ajuste Exponencial

A equação desenvolvida para ajustar exponencialmente uma série em qualquer período de tempo, i, é baseada somente em três termos – o valor corrente na série temporal, Y_i; o valor exponencialmente ajustado calculado anteriormente, E_{i-1}; e um peso atribuído, ou coeficiente de ajuste, W. O ajuste exponencial é chamado $E(W)$ e o valor da série ajustada, para um tempo *i*, é dado por:

$$E_i = WY_i + (1-W)E_{i-1} \qquad i = 2, 3, 4, \ldots$$
$$E_1 = Y_1 .$$

O coeficiente de ajuste W assume valores entre 0 e 1. Para W pequeno, o ajuste considera peso maior para observações passadas e para W próximo a 1 as observações recentes recebem maior peso.

Exercícios Excel (*EstatisticaExcel_08.xlsm; A3Series*)

6.1.2 Faça um ajuste exponencial na série de receitas da empresa Cabot Corporation (Figura 6.2). Veja arquivo Excel *EstatisticaExcel.xlsx*, planilha *A3Series*, linha 38.

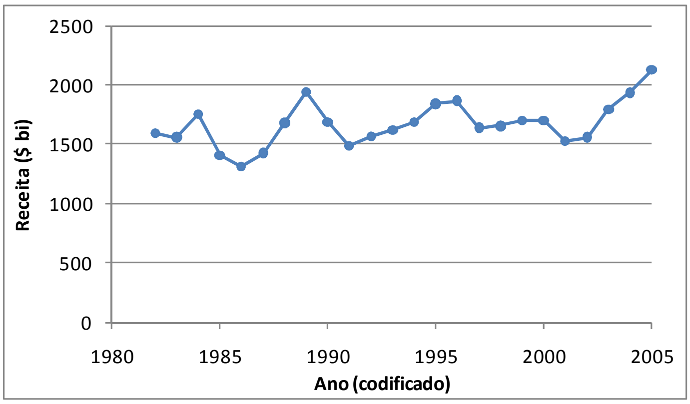


Figura 6.2 Cabot Corporation: Receitas em bilhões de dólares.

Modelo de tendência Linear, Quadrática, Exponencial

O fator mais frequentemente estudado em uma série temporal é a **tendência**. O intuito é realizar projeções de médio e longo prazos. O método mais aplicado é o de mínimos quadrados. O Excel possui uma ferramenta com os principais modelos.

- Modelo de tendência linear: $\hat{Y}_i = \beta_0 + \beta_1 X_i$
- Modelo de tendência quadrática: $\hat{Y}_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2$
- Modelo de tendência exponencial: $\hat{Y}_i = \beta_0 \ \beta_1^{X_i}$

  Neste modelo exponencial $(\beta_1 - 1)\times 100\%$ é a taxa de crescimento anual composta (em %).

Exercícios Excel (*EstatisticaExcel_08.xlsm; A3Series*)

6.1.3 Desenvolva **modelos de tendência** (linear, quadrática e exponencial) para a série de receitas da Coca-Cola Company (Tabela 6.1). Utilize a linha de tendência do Excel, e verifique qual modelo se ajusta melhor através do valor de R^2.

Resíduos e Desvio Médio Absoluto

Os resíduos são diferenças entre os valores observados e os valores previstos. Obtenha resíduos e analise os seus gráficos. O melhor modelo é aquele no qual os resíduos são aleatórios.

A escolha quantitativa do modelo de previsão apropriado pode ser feita com base no desvio médio absoluto, dado por:

$$DMA = \frac{\sum_{i=1}^{n} \left| Y_i - \hat{Y}_i \right|}{n}$$

Obtenha esta grandeza para as séries apresentadas na planilha em comparação com os ajustes e modelos.

Observe as séries temporais na planilha *A3Series* do arquivo *EstatisticaExcel.xlsx*. Verifique a análise de componentes da série a partir da linha 66.

Componente Tendência: Altere o valor de F67 e observe o efeito da tendência. O valor corresponde à inclinação. Por exemplo, digite 2.

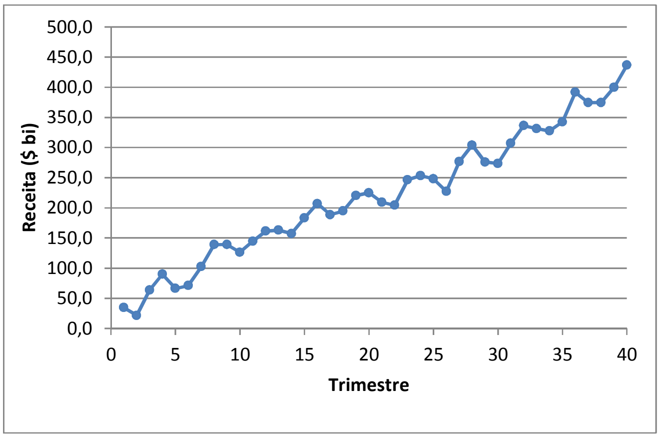
Componente Cíclica: Altere o valor de F69 e observe a variação Cíclica dada por uma função seno. O período é dado em F70.

Componente Sazonal: Tem a mesma forma da cíclica, mas com períodos menores.

Componente Irregular: Adicione uma amplitude aleatória alterando o valor em F75.

Sugestão de séries:

Célula	Início	Série1	Série2	Série3
F67	0	2	10	5
F69	0	20	20	50
F70	40	40	40	40
F72	0	10	20	20
F73	4	4	4	4
F75	0	15	30	100



6.2 Análise de séries para dados sazonais

Inúmeras séries temporais são coletadas trimestralmente ou mensalmente. Ou ainda semanalmente, diariamente e de hora em hora. Um modelo para extrair informações de tendência e sazonalidade pode ser obtido considerando variáveis dummy para sazonalidade.

Modelo Exponencial com dados trimestrais:

$$\hat{Y}_i = \beta_0 \beta_1^{X_i} \beta_2^{T_1} \beta_3^{T_2} \beta_4^{T_3}$$

Para resolver utiliza-se uma equação transformada:

$$\hat{y}_i = b_0 + b_1 X_i + b_2 T_1 + b_3 T_2 + b_4 T_3$$

sendo:

$\hat{y}_i = \log(\hat{Y}_i)$, ou seja, uma vez obtido $\hat{y}_i$, a estimativa da receita é dada por $\hat{Y}_i = 10^{\hat{y}_i}$;

$b_k = \log(\beta_k)$ e assim $\beta_k = 10^{b_k}$;

Interpretação:

$\beta_0 =$ é o ajuste para o primeiro trimestre da série.

$(\beta_1 - 1) \times 100\% =$ é a taxa de crescimento trimestral estimada composta das receitas.

β_i corresponde ao multiplicador sazonal para o trimestre T_j.

Observe que para obter os β's é necessário resolver um regressão linear múltipla. Esta regressão pode ser feita pela **ferramenta de análise de dados do Excel** mas pode também ser obtida programando a expressão para os β's na planilha de acordo com a expressão:

$$\beta = (X'X)^{-1}(X'Y)$$

Sendo:

- X a matriz formadas pelas variáveis independentes em cada coluna, porém a primeira coluna tem os elementos iguais a 1;
- Y a matriz coluna formada pela variável dependente;
- A notação X' significa transposta de X;
- O menos 1 significa inversa da matriz.

Este desenvolvimento está apresentado na planilha *RegressMatriz*. Em resumo, as matrizes são tais que o modelo seja descrito por:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1r} \\ 1 & x_{21} & x_{22} & \cdots & x_{2r} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nr} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_r \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

No exercício abaixo utilizaremos a análise de dados do Excel.

Exercícios Excel (*EstatisticaExcel_08.xlsm; A3Series*)

6.2.1 Considere os valores de Receitas trimestrais do Wal-Mart na planilha *A3Series* A110 (gráfico abaixo). Desenvolva uma equação de previsão de tendência exponencial com componentes trimestrais. Faça uma previsão para os próximos 4 trimestres.

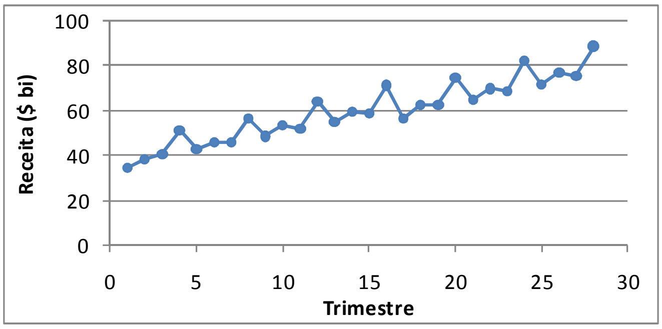


Figura 6.3 Wal-Mart: Receitas trimestrais.

❾ Arquivo de recuperação → ***EstatisticaExcel_09.xlsm***

6.3 Números-índice

Números-índice medem o valor de um item (ou grupo de itens) de um ponto específico no tempo sob a forma de uma percentagem do valor de um item (ou grupo de itens) em outro ponto no tempo.

Índice de Preços Simples

Um índice de preços simples rastreia o preço de uma única mercadoria num determinado período de tempo com o preço dessa mesma mercadoria em um determinado ponto no tempo no passado.

O período-base corresponde ao ponto no tempo no passado em relação ao qual estão sendo feitas todas as comparações. Assim, se possível deve-se escolher um ponto no qual exista certa estabilidade econômica.

$$I_i = \frac{P_i}{P_{base}} \times 100$$

$I_i =$ índice de preços para o ano *i*;

$P_i =$ preço para o ano *i*;

$P_{base} =$ preço para o ano-base.

Deslocando a base: $I_{novo} = \dfrac{I_{antiga}}{I_{nova\ base}} \times 100$

$I_{novo} =$ novo índice de preços;

$I_{antiga} =$ antigo índice de preços;

$I_{nova\ base} =$ valor do antigo índice de preços no novo ano-base.

Índice de Preços Agregados Ponderados

Um índice de preços agregados rastreia o preço de um grupo de mercadorias (cesta de mercadorias) num determinado período de tempo com o preço pago por esse mesmo grupo de mercadorias em um determinado ponto no tempo no passado.

Índices de preços agregados ponderados consideram diferentes ponderações. Por exemplo, o **índice de preço de Laspeyres** (I_L) (pronúncia Laspér) utiliza quantidades consumidas associadas ao ano-base no cálculo de todos os índices de preço na série:

$$I_L^{(t)} = \frac{\sum_{i=1}^{n} P_i^{(t)} Q_i^{(0)}}{\sum_{i=1}^{n} P_i^{(0)} Q_i^{(0)}} \times 100$$

em que

$t =$ período de tempo (0, 1, 2, ...)

$i =$ item (1, 2, ..., *n*)

$n =$ número total de itens que são considerados

$Q_i^0 =$ quantidade do item *i* no período de tempo 0.

$P_i^{(t)} =$ preço pago pela mercadoria *i* no período de tempo *t*;

$P_i^{(0)} =$ preço pago pela mercadoria *i* no período de tempo 0.

Exercícios Excel (*EstatisticaExcel_09.xlsm; Cesta*)

6.3.1 A planilha *Cesta* contém o preço de uma cesta de itens de alimentação, de 1992 a 2006. Estão incluídos os preços, em

dólares, de uma bisnaga com peso de uma libra (454 g) de pão branco, uma libra de carne moída, uma dúzia de ovos grandes tipo A e uma libra de alface.
(a) Calcule os índices de preços simples ao longo do período 1992-2006 para pão, carne, ovos e alface. Utilize o ano de 1992 como ano-base.
(b) Recalcule os índices utilizando o ano de 1996 como ano-base.
(c) Calcule o índice de preços de Laspeyres de 2006 para a cesta desses quatro itens de alimentação para uma família que tenha consumido em 1992, 50 bisnagas de pão, 22 libras de carne, 24 dúzias de ovos e 28 libras de alface.

❿ Arquivo de recuperação → ***EstatisticaExcel_10.xlsm***

7 Controle Estatístico de Processos

A partir dos anos 80, tem ocorrido nos Estados Unidos, um interesse renovado pela qualidade e pela produtividade. Alguns métodos para melhoria contínua de produtos têm sido desenvolvidos com ênfase cada vez maior na estatística, na melhoria de processos e na otimização do sistema como um todo.

7.1 Gestão da qualidade Seis Sigma

A metodologia de gestão Seis Sigma representa o sistema de melhoria da qualidade originalmente desenvolvido pela Motorola, em meados da década de 1980 (LEVINE et al.2008).

A metodologia foi projetada para criar processos que resultem em não mais do que 3,4 defeitos por milhão (ROTONDARO, 2002). A abordagem dos Seis Sigma pressupõe que o processo pode se desviar no máximo 1,5 desvios-padrão no longo prazo. Seis desvios-padrão (6 σ) menos uma variação de 1,5 produzem uma meta de 4,5 desvios-padrão. A área sob a curva normal fora de 4,5 desvios é de 3,398E-06. Assim, a metodologia é projetada para criar processos que resultem em não mais do que 3,4 defeitos por milhão.

A implementação da metodologia de gestão Seis Sigma requer uma abordagem orientada para dados fortemente baseada na utilização de ferramentas estatísticas tais como gráficos de controle e experimentos projetados. Envolve o treinamento das pessoas da organização nas etapas do programa Seis Sigma.

7.2 Gráficos de Controle

Um Gráfico de Controle monitora a variação em uma característica de um produto ou serviço ao longo do tempo. Pode ser utilizado para estudar o desempenho passado, avaliar as condições presentes ou para prever resultados no futuro.

Diferentes tipos e gráficos de controle permitem que se analisem tipos diferentes de variáveis **Críticas Para a Qualidade** (CPQ): variáveis categóricas (proporção de itens em não-conformidade) e variáveis contínuas (tempo de operação, dinheiro, peso, ...).

Os gráficos de controle, além de fornecer uma exposição visual dos dados que representam os processos, tem o objetivo de separar as causas de variação especiais das causas de variação comuns.
- **Causas de variação especiais**: Grandes oscilações ou padrões que não são inerentes ao processo, geralmente causadas por problemas a serem solucionados ou oportunidades a serem exploradas. Podem ser chamadas **causas de variações identificáveis**.
- **Causas de variação comuns**: Variabilidade inerente que existe num processo. Podem ser chamadas **causas de variação do acaso**.

Vários tipos de gráficos de controle são utilizados para monitorar processos. Nesta seção vamos estudar os seguintes gráficos:
- Gráfico p, utilizado para variáveis categóricas.
- Gráfico R, controle da amplitude.
- Gráfico $\bar{X}$, gráfico de controle para a média.

Construção de Limites de Controle
A forma mais comum estabelece que os limites de controle para uma medida num processo estão a uma distância de ±3 desvios-padrão da média estatística de interesse, veja figura abaixo.

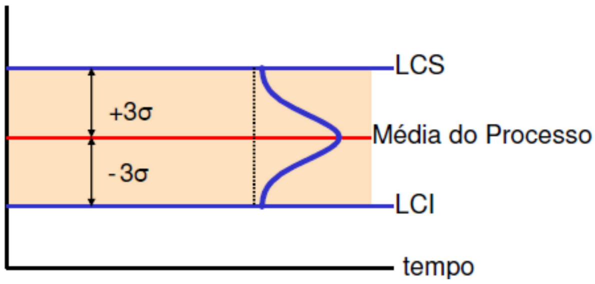


Figura 7.1 Ilustração de um gráfico de controle.

Devemos assim determinar os Limites de controle:

Limite de controle
Média do Processo ± 3 desvios-padrão
sendo: Limite de controle superior (**LCS**) = Média + 3 desvios-padrão Limite de controle inferior (**LCI**) = Média - 3 desvios-padrão

Gráficos de Controle específicos

O Gráfico *p*
O gráfico *p* é utilizado para variáveis categóricas. O gráfico *p* apresenta a proporção de itens em uma amostra.

Limites de controle para o Gráfico *p*:

$$LCS = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{\bar{n}}} \quad \text{e} \quad LCI = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{\bar{n}}}$$

sendo

$$\bar{n} = \frac{\sum_{i=1}^{k} n_i}{k} \quad \text{e} \quad \bar{p} = \frac{\sum_{i=1}^{k} X_i}{\sum_{i=1}^{k} n_i},$$

em que

X_i = número de itens não-conformes na amostra i.

n_i = tamanho da amostra (subgrupo) i.

k = número de amostras selecionadas.

$\bar{n}$ = média do tamanho das amostras.

$\bar{p}$ = média da proporção de itens não-conformes

Se LCI for negativo então LCI não existe.

Exercícios Excel (*EstatisticaExcel_10.xlsm;* ControleProcesso, *CP1e2*)

7.2.1 Durante a fase de análise do modelo dos Seis Sigma, foram coletados dados de não-conformidade de uma amostra de 200 quartos de hotel. A Tabela 7.1 apresenta a lista do número e de quartos não-conformes para cada dia no período de quatro semanas. Construa um **Gráfico *p*** de controle com base nesses dados. (Veja na página 26 como fazer gráficos de controle).

Tabela 7.1 Número de quartos não-conformes (n_i = 200).

Dia	Quartos não-conformes	Dia	Quartos não-conformes
1	16	15	18
2	7	16	13
3	21	17	15
4	17	18	10
5	25	19	14
6	19	20	25
7	16	21	19
8	15	22	12
9	11	23	6
10	12	24	12
11	22	25	18
12	20	26	15
13	17	27	20
14	26	28	22

O Gráfico *R*

O Gráfico R é o mais simples para monitorar a variabilidade ao lidar com dados numéricos (ao invés de proporções de categorias). Monitora a amplitude. Para desenvolver os limites é necessário ter uma estimativa para a média e para o desvio-padrão da amplitude.

Estes limites dependem de duas constantes que estão tabeladas, d_2 e d_3. Os valores de d_2 e d_3 dependem do número de observações na amostra. A Tabela 7.2 apresenta valores até 25 observações na amostra (também na planilha **TabelaControle**).

Tabela 7.2 Fatores para gráficos de controle.

Observações	d_2	d_3	D_3	D_4	A_2
2	1,128	0,853	0	3,267	1,880
3	1,693	0,888	0	2,575	1,023
4	2,059	0,880	0	2,282	0,729
5	2,326	0,864	0	2,114	0,577
6	2,534	0,848	0	2,004	0,483
7	2,704	0,833	0,076	1,924	0,419
8	2,847	0,820	0,136	1,864	0,373
9	2,970	0,808	0,184	1,816	0,337
10	3,078	0,797	0,223	1,777	0,308
11	3,173	0,787	0,256	1,744	0,285
12	3,258	0,778	0,283	1,717	0,266
13	3,336	0,770	0,307	1,693	0,249
14	3,407	0,763	0,328	1,672	0,235
15	3,472	0,756	0,347	1,653	0,223
16	3,532	0,750	0,363	1,637	0,212
17	3,588	0,744	0,378	1,622	0,203
18	3,640	0,739	0,391	1,609	0,194
19	3,689	0,733	0,404	1,596	0,187
20	3,735	0,729	0,415	1,585	0,180
21	3,778	0,724	0,425	1,575	0,173
22	3,819	0,720	0,435	1,565	0,167
23	3,858	0,716	0,443	1,557	0,162
24	3,895	0,712	0,452	1,548	0,157
25	3,931	0,708	0,459	1,541	0,153

Limites de controle para o Gráfico *R*:

$$LCS = \bar{R} + 3\bar{R}\frac{d_3}{d_2} \quad \text{e} \quad LCI = \bar{R} - 3\bar{R}\frac{d_3}{d_2}$$

sendo $$\bar{R} = \frac{\sum_{i=1}^{k} R_i}{k},$$

onde R_i é a amplitude da amostra i. Os cálculos podem ser simplificados utilizando as constantes D_3 e D_4 da Tabela 7.2:

$$LCS = D_4\bar{R} \quad \text{e} \quad LCI = D_3\bar{R}.$$

Exercícios Excel (*EstatisticaExcel_10.xlsm;* ControleProcesso, *CP1e2*)

7.2.2 Qualidade de serviço de um hotel. O tempo de entrega de bagagens foi definido como o tempo desde o momento que o hóspede completa os procedimentos de entrada até o momento em que a bagagem chega ao quarto do hóspede. As amostras foram de tamanho cinco em cada dia, durante 28 dias. Faça um Gráfico R a partir dos valores da Tabela 7.3.

Observações = 5.

O Gráfico $\bar{X}$

Como a amplitude está sob controle, agora será examinado o gráfico de controle para a média do processo, o Gráfico $\bar{X}$. Para controlar os limites de controle para a média, calcula-se a média aritmética de cada amostra ($\bar{X}_i$) e depois a média das médias, denominada $\bar{\bar{X}}$.

Limites de controle para o Gráfico $\bar{X}$:

$$LCS = \bar{\bar{X}} + 3\frac{\bar{R}}{d_2\sqrt{n}} \quad \text{e} \quad LCI = \bar{\bar{X}} - 3\frac{\bar{R}}{d_2\sqrt{n}}$$

em que $\bar{\bar{X}}$ é a média das médias de amostras e $\bar{R} = \frac{\sum_{i=1}^{k} R_i}{k}$,

onde R_i é a amplitude da amostra i e k, o número de amostras. Os limites podem ser obtidos utilizando o fator A_2 da Tabela 7.2:

$$LCS = \bar{\bar{X}} + A_2\bar{R} \quad \text{e} \quad LCI = \bar{\bar{X}} - A_2\bar{R}.$$

Tabela 7.3 Tempos de entrega de bagagens.

Dia	Tempos de Entrega em Minutos					Média	Amplitude
1	6,7	11,7	9,7	7,5	7,8	8,68	5,0
2	7,6	11,4	9,0	8,4	9,2	9,12	3,8
3	9,5	8,9	9,9	8,7	10,7	9,54	2,0
4	9,8	13,2	6,9	9,3	9,4	9,72	6,3
5	11,0	9,9	11,3	11,6	8,5	10,46	3,1
6	8,3	8,4	9,7	9,8	7,1	8,66	2,7
7	9,4	9,3	8,2	7,1	6,1	8,02	3,3
8	11,2	9,8	10,5	9,0	9,7	10,04	2,2
9	10,0	10,7	9,0	8,2	11,0	9,78	2,8
10	8,6	5,8	8,7	9,5	11,4	8,80	5,6
11	10,7	8,6	9,1	10,9	8,6	9,58	2,3
12	10,8	8,3	10,6	10,3	10,0	10,00	2,5
13	9,5	10,5	7,0	8,6	10,1	9,14	3,5
14	12,9	8,9	8,1	9,0	7,6	9,30	5,3
15	7,8	9,0	12,2	9,1	11,7	9,96	4,4
16	11,1	9,9	8,8	5,5	9,5	8,96	5,6
17	9,2	9,7	12,3	8,1	8,5	9,56	4,2
18	9,0	8,1	10,2	9,7	8,4	9,08	2,1
19	9,9	10,1	8,9	9,6	7,1	9,12	3,0
20	10,7	9,8	10,2	8,0	10,2	9,78	2,7
21	9,0	10,0	9,6	10,6	9,0	9,64	1,6
22	10,7	9,8	9,4	7,0	8,9	9,16	3,7
23	10,2	10,5	9,5	12,2	9,1	10,30	3,1
24	10,0	11,1	9,5	8,8	9,9	9,86	2,3
25	9,6	8,8	11,4	12,2	9,3	10,26	3,4
26	8,2	7,9	8,4	9,5	9,2	8,64	1,6
27	7,1	11,1	10,8	11,0	10,2	10,04	4,0
28	11,1	6,6	12,0	11,5	9,7	10,18	5,4

Exercício Excel (*EstatisticaExcel_10.xlsm;* ControleProcesso, *CP1e2*)

7.2.3 Construa um Gráfico $\bar{X}$ para o processo de entrega de bagagens. Verifique através do Gráfico R e do Gráfico $\bar{X}$ se o processo está sob controle estatístico.

Exercício Excel (*EstatisticaExcel_10.xlsm;* ControleProcesso, *CP3*)

7.2.4 Um artigo apresenta uma análise sobre uma operação de envasamento de água mineral. Uma das características de interesse é a quantidade de magnésio, medidas em partes por milhão (ppm), existente na água. Os dados a representam o nível de magnésio de 30 subgrupos de 4 garrafas (Observações = 4) coletadas ao longo de um período de 30 horas.

7.2.4.a Construa um gráfico de controle para a Amplitude

7.2.4.b Construa um gráfico de controle para a Média

7.2.4.c O processo está sob controle?

7.3 Variabilidade de Processos

Os pontos devem ser investigados de forma a observar se as variações existentes são de causa comum ou se existe algum ponto fora dos limites ou ainda se uma série de pontos está abaixo ou acima da média. Na figura 7.2 abaixo o processo está sob controle.

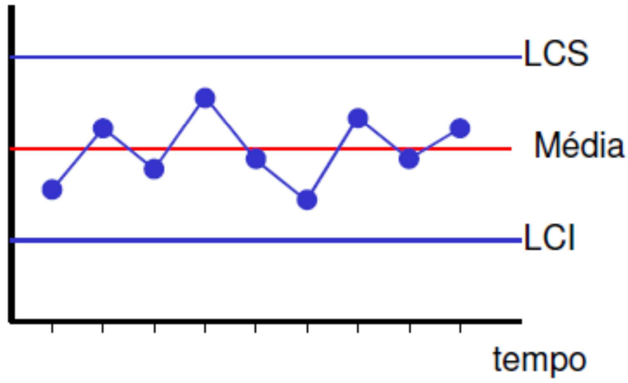


Figura 7.2 Processo sob controle: Variação comum.

As figuras 7.3 a 7.5 são exemplos de processos fora do controle. Na Figura 7.3 um valor ultrapassou o limite de controle superior, a variação não é comum. Na Figura 7.4, embora todos os valores estejam entre os limites, ocorre uma série de pontos abaixo da média e uma série de pontos acima da média sendo apresentada uma forte tendência. Na Figura 7.5 ocorrem pontos consecutivos abaixo da média

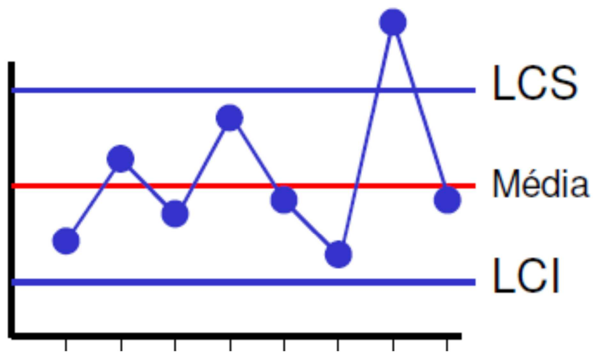


Figura 7.3 Processo fora de controle: variação de causa especial.

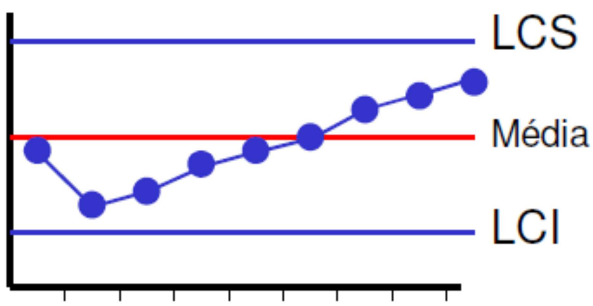


Figura 7.4 Processo fora de controle: variação de causa especial - Pontos consecutivos, tendência.

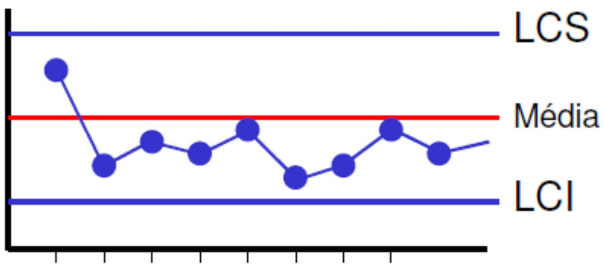


Figura 7.5 Processo fora de controle: Variação de causa especial - pontos consecutivos abaixo da média.

Gráfico de controle no Excel:

No Excel não existe um modelo específico para Gráficos de Controle, mas eles podem ser feitos com gráficos de **dispersão com pontos de dados conectados por linhas**.

Exercício Excel (*EstatisticaExcel_10.xlsm;* ControleProcesso, *CP4*)

7.4 A gerente de uma filial de uma corretora de valores está preocupada com o número de operações não-desejadas (com erro no recibo da operação), realizadas pela equipe de vendas. As operações com erros são canceladas e refeitas. Na Tabela 7.4 são apresentados dados coletados ao longo de 30 dias

Tabela 7.4 Registro de operações não desejadas.

Dia	Op. não Desejadas	Total	Dia	Op. não Desejadas	Total
1	2	74	16	3	54
2	12	85	17	12	74
3	13	114	18	11	103
4	33	136	19	11	100
5	5	97	20	14	88
6	20	115	21	4	58
7	17	108	22	10	69
8	10	76	23	19	135
9	8	69	24	1	67
10	18	98	25	11	77
11	3	104	26	12	88
12	12	98	27	4	66
13	15	105	28	11	72
14	6	98	29	13	118
15	21	204	30	15	138

A gerente deseja saber se a proporção de operações não desejadas está em um estado de controle estatístico, de modo que ela possa planejar a etapa subsequente em um processo de melhoria de qualidade.

7.4.1 Construa um gráfico de controle para esses dados

7.4.2 O processo está sob controle?

7.4.3 Com base nos resultados, o que deveria ser feito para melhorar o processo?

Gráfico de controle Excel:

- Gráfico de dispersão com pontos de dados conectados por linha.
- Incluir LCS, LCI. x: menor dia, maior dia
  y: LCS ou LCI
- Incluir p barra. x: menor dia, maior dia
  y: p barra
- Incluir os dados x: de 1 a 30
  y: p_i

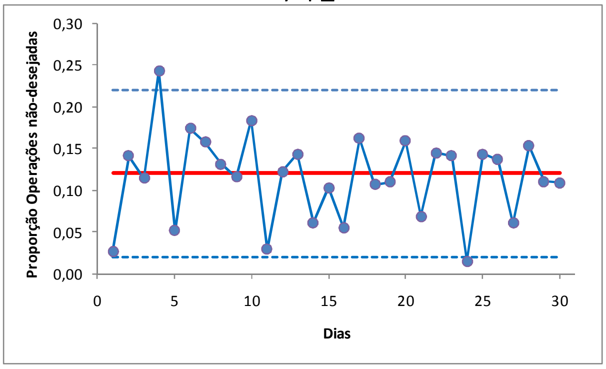


Figura 7.6 Resultado para gráfico de controle no Excel

Causas especiais de variação: O dia 4 está acima do LCS, o dia 24 está abaixo do LCI, portanto o processo está fora de controle.

As causas especiais devem ser investigadas para melhorar o processo.

⓫ Arquivo completo → ***EstatisticaExcel_11.xlsm***

Tabela Excel2. Funções Excel, parte 2.

Função	Descrição
MÉDIASE(interv;crit;intev. méd**)**	Média condicionada
SUBTOTAL(Função; Referencia**)**	Para dados com filtro
INT.CONFIANÇA.T(alfa;desvp;n**)**	Erro estatístico p/ IC
INT.CONFIANÇA.NORM(a;dp;n**)**	Erro estatístico p/ IC
DISTT(x; graus de lib.; caudas**)**	Distribuição t-student
INVT(P; graus de liberdade**)**	Distribuição t-student
VARP(v_1**)**	Variância população
TESTE.Z(Amostra;x;sigma**)**	Teste z, média 1 amostra
TESTE.T(Am1;Am2;caudas;tipo**)**	Teste t, par, indep, ...
COVAR(v_1; v_2**)**	Covariância
PEARSON(v_1; v_2**)**	Correlação linear
CORREL(v_1; v_2**)**	Correlação linear
INTERCEPÇÃO(v_y; v_x**)**	Coeficiente linear da reta
INCLINAÇÃO(v_y; v_x**)**	Coeficiente angular da reta
PREVISÃO(x; v_y; v_x**)**	Previsão linear a partir de x
MATRIZ.MULT(matr_A; matr_B**)**	Multiplicação de Matrizes
MATRIZ.INVERSO(MatrizA**)**	Inverte a MatrizA
TRANSPOR(MatrizA_nxm**)**	Obtém transposta mxn

BIBLIOGRAFIA

BRUNI A.L. **Estatística aplicada à gestão empresarial**, 2a ed. São Paulo: Atlas, 2008.

CORRAR L.J., PAULO E, DIAS FILHO J.M. **Análise multivariada para cursos de administração, ciências contábeis e economia**, São Paulo: Atlas, 2007.

CORRAR L.J., THEÓPHILO C.R. **Pesquisa operacional para decisão em contabilidade e administração: contabilometria**, 2a ed. São Paulo: Atlas, 2008.

FÁVERO L.P. et. al. **Análise de dados: modelagem multivariada para tomada de decisão**, Rio de Janeiro: Elsevier, 2009.

LEVINE, STEPHAN, KREHBIEL, BERENSON. **Estatística: teoria e aplicações usando o Microsoft® Excel**, 5a ed. Rio de Janeiro: LTC, 2008.

ROTONDARO, R.G (coordenador) **Seis Sigma: Estratégia gerencial para a melhoria de processos, produtos e serviços**, 1a ed. São Paulo: Atlas, 2002.

TRIOLA, M.F. **Introdução à Estatística**, Rio de Janeiro: LTC, 2005.