

Avaliação de Desempenho de Sistemas

Paulo Maciel

Centro de Informática - UFPE

Objetivo

- É o estudo, fixação e aplicação de métodos e modelos para Avaliação de Desempenho.

Slides:

https://dl.dropboxusercontent.com/u/5147622/ADS/AvalDes_Measuring_INDT.pdf

<https://dl.dropboxusercontent.com/u/5147622/ADS/SPN1.pdf>

Programa

- ❑ Conceitos Básicos
- ❑ Leis Operacionais
- ❑ Técnicas de Medição e Ferramentas
- ❑ Tópico em Inferência e Estatística Descritiva
- ❑ Cadeias de Markov e Filas
- ❑ Redes de Petri Estocásticas
- ❑ Simulação Estocástica

Ementa (Desempenho: medição)

- Breve Introdução à avaliação de desempenho
- Leis Operacionais
- Técnicas de Medição
- Monitoração no Windows (PowerShell)
- Monitoração no Linux
- Profiling
- Análise exploratória de dados
 - Métodos gráficos
 - Resumos estatísticos
 - Apresentação de resultados
- Métodos para estimativa com confiança
 - Estimativa de tendências centrais
 - Estimativa de variabilidade
 - Comparação de sistemas
 - Estimativa de percentuais

Bibliografia básica (primeira parte)

1. Measuring Computer Performance: A Practitioner's Guide, David J. Lilja , Cambridge University Press, 2000.
2. Art of Computer Systems Performance Analysis Techniques For Experimental Design Measurements Simulation And Modeling, *Raj Jain*, Wiley Computer Publishing, John Wiley & Sons, Inc,1991.
3. <http://www.itl.nist.gov/div898/handbook/>
4. <http://www.bipm.org/en/bipm/tai/tai.html>

Bibliografía básica

- **Measuring Computer Performance: A Practitioner's Guide.** David J. Lilja , Cambridge University Press, 2000.
- **Art of Computer Systems Performance Analysis Techniques For Experimental Design Measurements Simulation And Modelingm.** Raj Jain, Wiley Computer Publishing, John Wiley & Sons, Inc,1991.
- **Probability and Statistics with Reliability, Queuing, and Computer Science Applications.** K. Trivedi, John Wiley and Sons, New York, 2001. ISBN number 0-471-33341-7

Metodologia

- Aulas expositivas
- Aulas práticas

Avaliação

- Resolução de listas de exercícios.
(Não serão permitidos atrasos.)

Avaliação de Desempenho

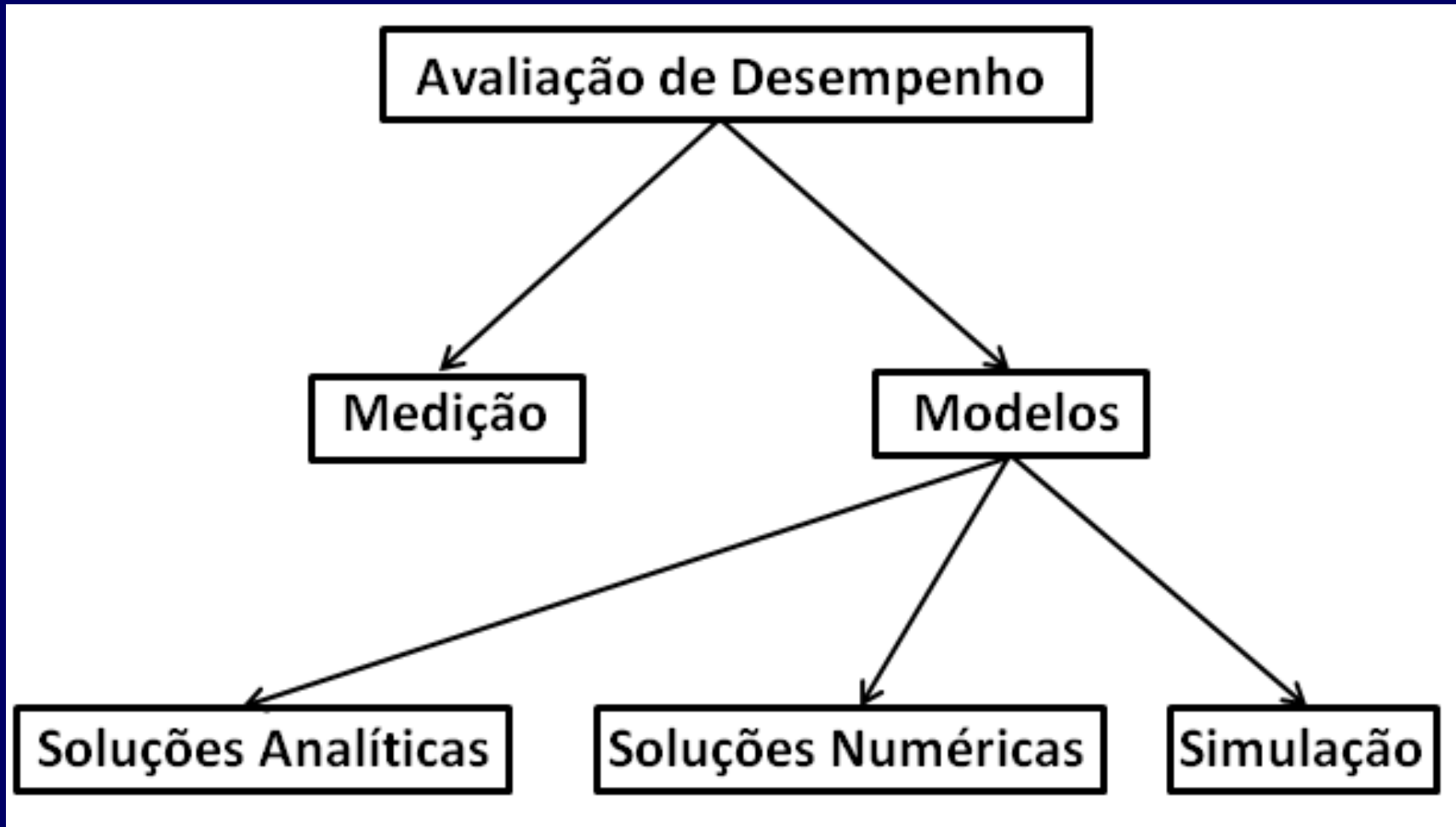
- ❑ Refere-se a um conjunto de métodos que possibilita investigar o comportamento temporal de sistemas.
- ❑ Possui longa tradição no estudo e dimensionamento de sistemas de comunicação, sistemas de manufatura, pesquisa operacional.
- ❑ Avaliação de desempenho de sistemas computacionais
- ❑ Sistemas de recursos compartilhados × Sistemas de tempo real.

Avaliação de Desempenho

- Sistema de Tempo Real
 - Exemplos
 - Controle de processo
 - Rôbos
 - Sistemas de controle de aeronaves.
 - Objetivos
 - Corretude
 - Tolerância a falha
 - Propriedades de interesse
 - *Liveness, safety*
 - Tempo determinístico
 - Modelos
 - TA, EFSM, TPN, RTPA

- Sistema de Recursos Compartilhados
 - Exemplos
 - *Time-sharing computers.*
 - Arquiteturas cliente-servidor
 - Sistemas de telefonia, comunicação,
 - Linhas de produção.
 - Objetivos
 - Uso econômico de recursos
 - Tolerância à falhas
 - Propriedades de interesse
 - *Throughput*, utilização, retardo, probabilidade de perda.
 - Tempo estocástico
 - Modelos
 - QN, SPN, SPA

Classificação da Técnicas de Avaliação de Desempenho



Processo de Avaliação de Desempenho

Macro-Atividades de um Processo de Avaliação de Desempenho (com modelagem)

1. **Compreensão geral do problema/sistema a ser avaliado.**
2. **Definição inicial dos critérios de desempenho a serem avaliados.**
3. **Identificação dos componentes.**
4. **Refinamento dos critérios de avaliação**
5. **Geração do modelo abstrato.**
6. **Planejamento da medição.**
7. **Coleta dos dados.**
8. **Análise dos dados coletados** associados aos componentes (influentes) do sistema/problema.
9. **Geração do modelo refinado.**
10. **Definição e mapeamento das métricas no modelo refinado.**
11. **Escolha dos métodos de avaliação dos modelos.**
12. **Desagregação do modelo refinado.**
13. **Avaliação.**
14. **Agregação.**
15. **Análise dos resultados e recomendações.**

Informações do Documento do Processo

Para cada Atividade constam:

- Objetivo
- Responsável
- Pré-condições
- Entradas
- Ações
- Saídas
- Pós-Condições

***Checklist* para evitar erros comuns
em um Projeto de Avaliação de
Desempenho**

***Checklist* para evitar erros comuns em um Projeto de Avaliação de Desempenho**

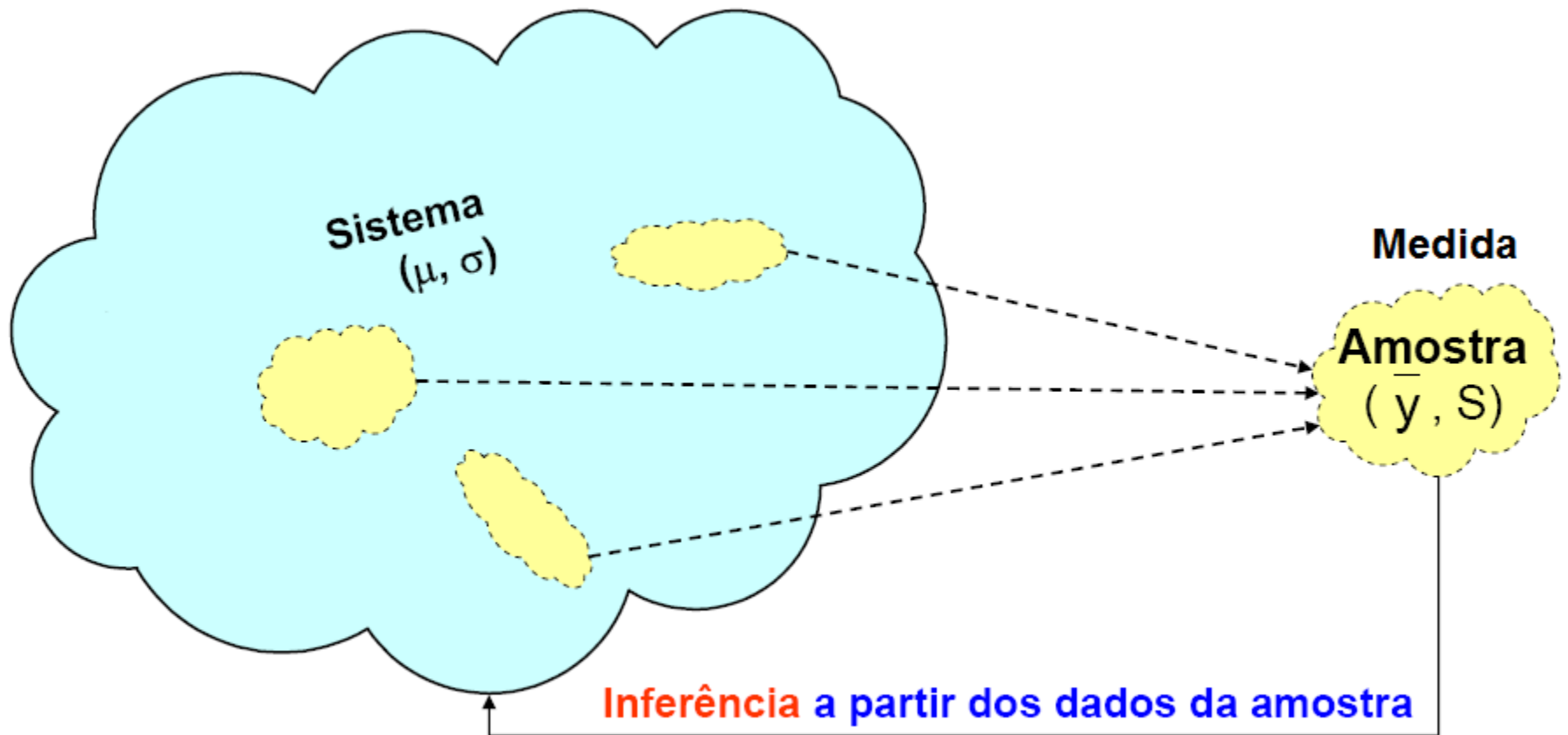
1. O sistema/problema a ser avaliado deve ser claramente definido e compreendido, assim como os critérios da avaliação.
2. Os critérios de avaliação devem ser definidos de maneira clara, objetiva e de forma não viesada.
3. Apresente os objetivos e estratégia de de forma clara e precisa.
4. Envolver a alta-gerência para seja dada a devida prioridade ao projeto.
5. Defina um plano de ação (metodologia), ressaltando as etapas, pré-condições, insumos, produtos, evidências (pós-condições), funções e responsáveis.
6. Solucione disputas (internas e externas).
7. Verifique se etapas da avaliação foram seguidas de maneira sistemática.
8. Avalie se as métricas definidas são relevantes para a avaliação.
9. Certifique-se que o conjunto de parâmetros que afeta o desempenho do sistema está devidamente definido.
10. Certifique-se que a carga considerada é adequada.
11. Certifique-se que a técnica de avaliação é adequada.
12. Certifique-se que o nível de abstração é apropriado.

***Checklist* para evitar erros comuns em um Projeto de Avaliação de Desempenho**

13. Defina os parâmetros a serem variados.
14. Defina se será realizada análise de sensibilidade.
14. Assegure-se de que adotou métodos para prover resultados estatisticamente confiáveis.
15. Verifique se os erros das “entradas” podem causar erros significantes nos resultados.
16. Avalie se os *Outliers* devem ser retirados da análise dos dados.
17. Avalie se serão consideradas alterações futuras na carga do sistema.
18. Verifique se os resultados obtidos são fáceis de explicar e os apresente com as devidas interpretações e com o auxílio de gráficos.
19. Certifique-se que a forma de apresentação dos resultados é adequada para o público, audiência ou cliente.

Conceitos Básicos sobre Medição

Conceitos Básicos



Conceitos Básicos

- **População:** conjunto de todas as medidas de interesse de um sistema.
- **Amostra:** porção da população através da qual informações do sistema são obtidas.
- **Inferência estatística** auxilia na estimativa de parâmetros do sistema através da escolha adequada de uma **amostra**.

Conceitos Básicos

□ Exatidão (*accuracy*),

- proximidade entre o resultado de uma medição e o valor referência correspondente.
- Ausência de Viés e de Precisão

□ Precisão

- É uma medida da dispersão do conjunto de dados obtidos na medição. Esta relacionada a repetibilidade.

□ Resolução

- Corresponde a menor alteração que pode ser detectada por ferramental de medição.

Conceitos Básicos

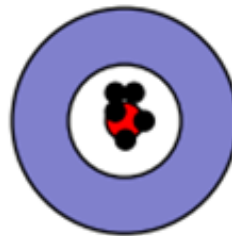
Qual você prefere?

Exatidão

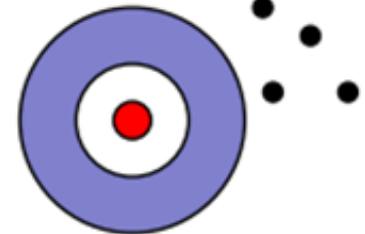
Ausência de
Viés

Viés

Precisão



Imprecisão



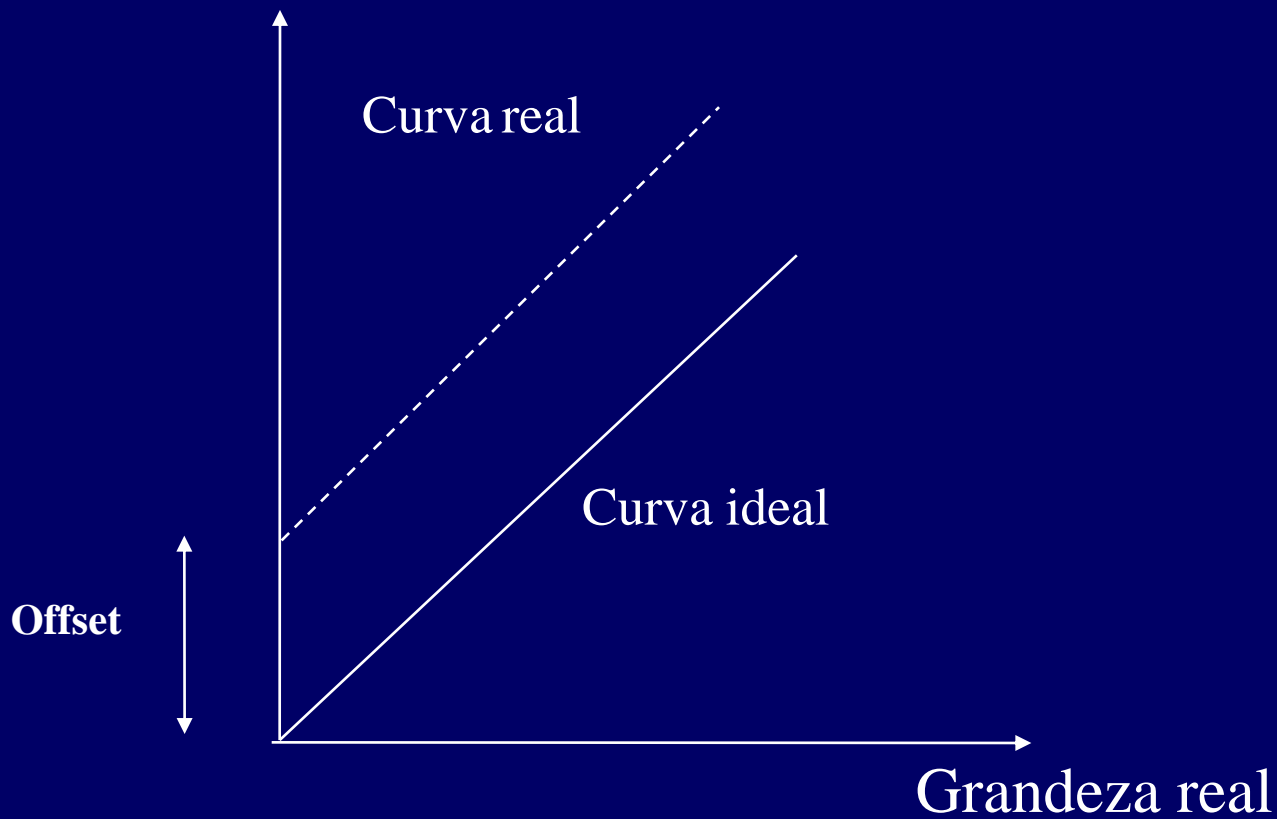
Conceitos Básicos

Erro de
offset

Erro de Offset – **provoca viesamento**

- **É constante**

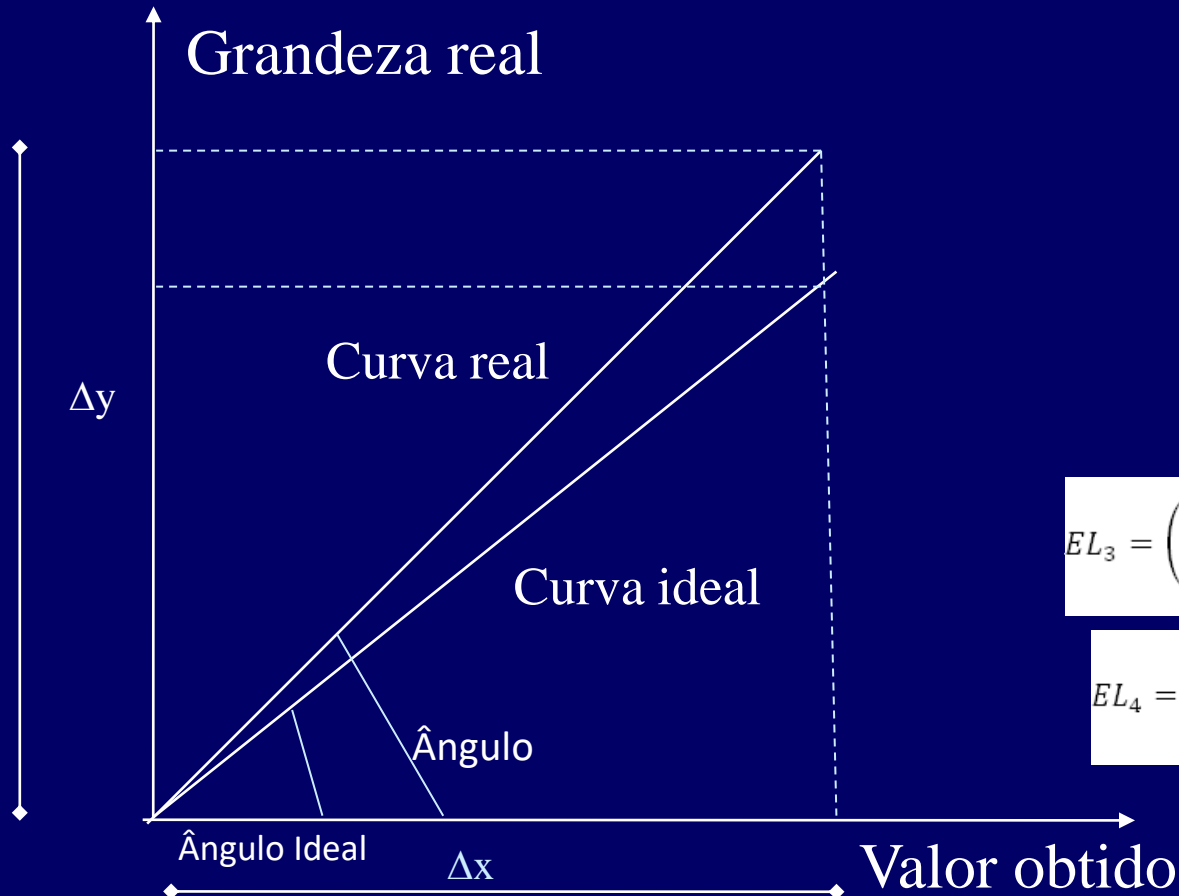
Valor obtido



Conceitos Básicos

Erro de Linearidade - Dependente do valor de entrada.

Provoca viesamento



$$EL_1 = \left(\frac{\left(\frac{\Delta y}{\Delta x} - 1 \right)}{2} \right) \times 100$$

$$EL_2 = \left(\frac{\left(\sum_{j=1}^{N-1} \left(\frac{\Delta y}{\Delta x} \right) \right) - 1}{2} \right) \times 100$$

$$\Delta y = MMP_j - MMP_{j-1}$$

$$\Delta x = TRP_j - TRP_{j-1}$$

$$EL_3 = \left(\frac{(\theta_{ideal} - \theta_{EL_1})}{2 \times \pi} \right) \times 100 = \left(\frac{\left(\frac{\pi}{4} - \theta_{EL_1} \right)}{2 \times \pi} \right) \times 100$$

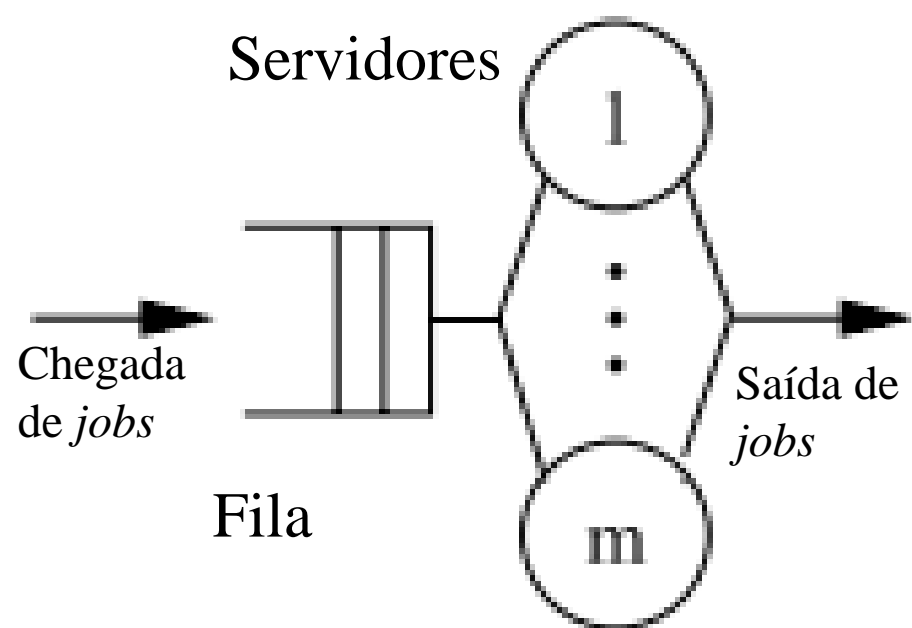
$$EL_4 = \left(\frac{(\theta_{ideal} - \theta_{EL_2})}{2 \times \pi} \right) \times 100 = \left(\frac{\left(\frac{\pi}{4} - \theta_{EL_2} \right)}{2 \times \pi} \right) \times 100$$

Notação de Kendall para Sistemas de Fila

Notação de Kendall para Sistemas de Fila

□ $A/B/m/K$

- A – distribuição do tempo entre chegadas.
- B – distribuição do tempo de serviço.
- m – número de servidores.
- K = capacidade de armazenamento.



$A, B = \{M, D, G, E, H\}$

- M - *Markovian*,
- D - *Determinística*,
- G - *General*
- E_r - *Erlangian*
- H_r - *Hiper-exponencial*

Notação de Kendall para Sistemas de Fila

□ A/B/m/K

- A – distribuição do tempo entre chegadas.
- B – distribuição do tempo de serviço.
- m – número de servidores.
- K = capacidade de armazenamento.
- Muitas vezes quando K e m são ∞ , estes termos são omitidos ou usa-se / /

– Exemplos:

- M/M/1
- M/M/1/K
- M/G/2

Análise Operacional

□ Variáveis operacionais

T : Período de observação

K : Número de recursos do sistema

A_i : Número total de solicitações (ex:.chegadas) do recurso i no período T .

A_0 : Número total de solicitações (ex:.chegadas) ao sistema no período T .

C_i : Número total de serviços finalizados pelo recurso i no período T .

C_0 : Número total de serviços finalizados pelo sistema no período T .

B_i : Tempo de ocupação do recurso i no período T

Métricas derivadas (*derived measures*)

S_i : Tempo médio de serviço por finalização relativa ao recurso i ; $S_i = B_i/C_i$

U_i : Utilização média do recurso i ; $U_i = B_i/T$

X_i : *throughput* (ex.: finalizações por unidade de tempo) do recurso i ; $X_i = C_i/T$

λ_i : taxa de chegada (ex.: chegadas por unidade de tempo) ao recurso i ; $\lambda_i = A_i/T$

X_0 : *throughput* do sistema; $X_0 = C_0/T$

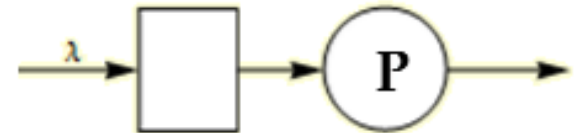
V_i : Número médio de visitas ao recurso i por solicitação; $V_i = C_i/C_0$

Análise Operacional

OL

□ Exemplo1

Suponha que ao se monitorar um processador por um período de 1 min, verificou-se que o recurso esteve ocupado por 36s. O número total de transações que chegaram ao sistema é 1800. O sistema também finalizou a execução de 1800 transações no mesmo período.



Qual a taxa de chegada ao sistema (λ_0)?

Qual é o throughput do sistema (X_0)?

Qual é a utilização da CPU (U_{CPU})?

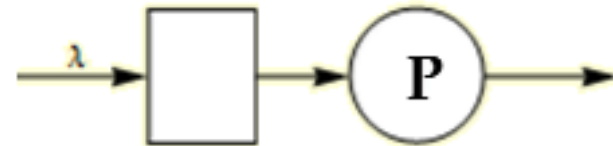
Qual é o tempo médio por transações finalizadas pelo sistema (S_0)?

Análise Operacional

OL

Exemplo1

É importante salientar que o único recurso do sistema é a CPU, portanto as métricas associadas à CPU serão as mesmas associadas ao sistema.



$$T = 1 \text{ min} = 60s$$

$$A_0 = 1800 \text{ transactions}$$

$$C_0 = 1800 \text{ transactions}$$

$$B_{cpu} = 36s$$

$$U_{cpu} = \frac{36s}{60s} = 60\%$$

$$X_0 = \frac{C_0}{T} = \frac{1800 \text{ transactions}}{60} = 30.00 \text{ tps}$$

$$S_{cpu} = \frac{B_{cpu}}{C_0} = \frac{36s}{1800 \text{ transactions}} = 0.02s$$

Análise Operacional



OL

Utilization Law

$$U_i = \frac{B_i}{T} = \frac{B_i}{T} \times \frac{C_i}{C_i} = \frac{B_i}{C_i} \times \frac{C_i}{T} = S_i \times X_i$$

Relacionamento da utilização de um dispositivo com o seu throughput.

Análise Operacional



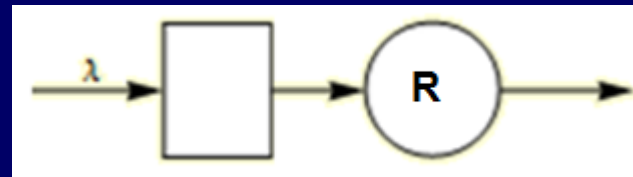
OL

Utilization Law

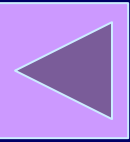
$$U_i = S_i \times X_i$$

Exemplo: Considere que 125 pacotes por segundo chegam a um roteador e que o roteador leva em média 2 milisegundos para tratar o pacote. Portanto:

$$U_i = 0,002 \times 125 = 25\%$$

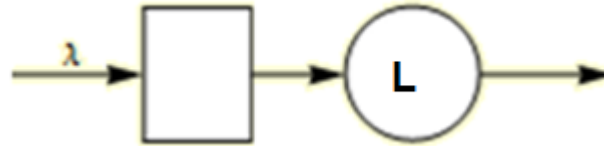


Análise Operacional



OL

Exemplo2



A banda passante de um *link* de comunicação é 56000 bps. Pacotes de 1500 bytes são transmitidos ao *link* a uma taxa de 3 pacotes por segundo

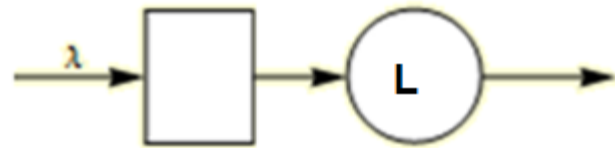
Qual é a utilização do link?

Análise Operacional



OL

Exemplo2



$$U = S \times \lambda_0$$

$$\text{Packet Size} = 1500 \text{ Bytes} = 1500 \text{ Bytes} \times 8 \text{ bits} = 12000 \text{ bits}$$

$$\text{Band Width} = 56000 \text{ bps} = \frac{56000 \text{ bps}}{8 \text{ bits}} = 7000 \text{ Bps} = \frac{7000 \text{ Bps}}{1500 \text{ Bytes}} = 4,66667 \text{ pps}$$

$$S (\text{time to send a packet}) = \frac{1}{4,66667 \text{ pps}} = 0,214285714 \text{ s}$$

$$\lambda_0 = 3 \text{ pps}$$

$$U = S \times \lambda_0 = 0,214285714 \text{ s} \times 3 \text{ pps} = 0,642857 = 64,2857\%$$

Análise Operacional



OL

Forced Flow Law

$$X_i = \frac{C_i}{T} = \frac{C_i}{T} \times \frac{C_0}{C_0} = \frac{C_i}{C_0} \times \frac{C_0}{T} = V_i \times X_0$$

Uma maneira interessante de relacionar o throughput do sistema ao throughput dos recursos.

Análise Operacional

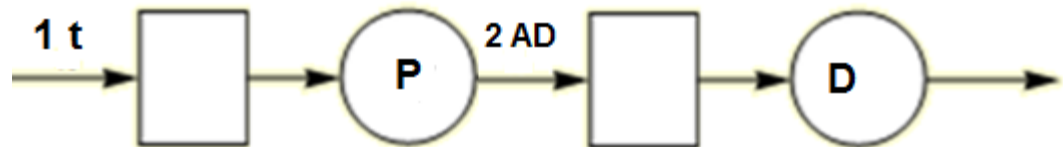


OL

Forced Flow Law

$$X_i = V_i \times X_0$$

Exemplo: suponha que quando se executa uma (1) transação em um servidor (e consequentemente no processador), faz-se dois (2) acessos a unidade de disco. Se 5,6 transações são finalizadas no servidor por segundo, quantos acessos foram feitos à unidade de disco?



$$X_i = 2 \times 5,6 = 11,2 \text{ acessos por segundo}$$

Análise Operacional



OL

Service Demand Law

- *Service demand de um recursos* é o tempo médio total que uma transação passa em no recurso.

Da *Utilization Law*, tem-se:

$$U_i = X_i \times S_i$$

Da *Forced Flow Law*, tem-se:

$$X_i = V_i \times X_0$$

Portanto:

$$U_i = V_i \times X_0 \times S_i = D_i \times X_0$$



OL

Análise Operacional

Service Demand Law

$$U_i = V_i \times X_0 \times S_i = D_i \times X_0$$

Portanto:

$$D_i = \frac{U_i}{X_0}$$

Observe que a utilização U_i do dispositivo i é diretamente proporcional à demanda D_i (*service demand*), portanto o dispositivo com mais alta demanda $\max_i \{D_i\}$ tem a mais alta utilização e é o “gargalo” do sistema.

Análise Operacional

OL

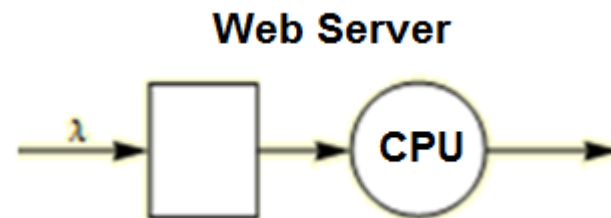
Exemplo3

Considere que um *Web Server* foi monitorado por 10 min e que a CPU (durante este período) teve ocupação média de 90,2348%. O *log* do *Web Server* registrou 30.000 solicitações processadas. Qual é a *CPU Service Demand* (D_{CPU}) relativa as solicitações ao *Web Server*?

$$T = 10 \times 60s = 600s$$

$$X_0 = 30.000/600 = 50 \text{ solicitações por segundo.}$$

$$D_{CPU} = U_{CPU}/X_0 = 0,902348/50 = 0,018047 \text{ s/solicitação}$$



Análise Operacional

OL

Example 5.3.2. A system composed of five servers was monitored for four hours ($T = 4 \times 60 \times 60s = 14400s$) under operational conditions. In this period, the log registered $C_0 = 28978$ transactions processed. The servers' utilizations were obtained over the period every $30s$. Hence a sample of 480 utilizations for each server was recorded. The average utilizations over the four hours period of each server were $\overline{U}_{s_1} = 0.3996$, $\overline{U}_{s_2} = 0.2389$, $\overline{U}_{s_3} = 0.2774$, $\overline{U}_{s_4} = 0.5253$, and $\overline{U}_{s_5} = 0.2598$, respectively. The mean time demanded of a typical transaction in each server may be estimated through

$$D_{s_i} = \frac{\overline{U}_{s_i}}{X_0},$$

since $X_0 = C_0/T = 28978 \text{ trans.} / 14400s = 2.0124 \text{ tps}$. Hence

Table 5.1: Demands

Server	Demand (s)
<i>Server 1</i>	0.1986
<i>Server 2</i>	0.1187
<i>Server 3</i>	0.1378
<i>Server 4</i>	0.2610
<i>Server 5</i>	0.1291

Análise Operacional



OL

Therefore, each typical transaction demanded the respective times of each specific server. Now, assume that a considerable demand increase is forecasted. It is expected that $C'_0 = 60000$ transactions would be requested in the same four hours period. Thus, the expected throughput would be $X'_0 = 60000/14400s = 4.1667tps$. The foreseen utilization of each server may be estimated through

$$U'_{si} = [D_{si} \times X'_0] .$$

since the maximal utilization is 1. Therefore

Table 5.2: Utilization

Server	U'_{si}
<i>Server 1</i>	0.8274
<i>Server 2</i>	0.4947
<i>Server 3</i>	0.5743
<i>Server 4</i>	1
<i>Server 5</i>	0.5378

and the Server 4 would reach 80% of utilization for a throughput of $X_0 = 3.0649tps$.

Análise Operacional



OL

Example 5.3.3. A server was monitored for $T = 2h$ considering a specific workload. In this period, the average processor utilization was $U_{cpu} = 0.38$, and $C_0 = 374,356$ transactions were processed. Each transaction, on average, reads/writes 18,427.44 bytes from/to the disk. The average time to read or write one sector (512 bytes) from/to the disk is 0.26 ms. The transaction throughput is

$$X_0 = \frac{C_0}{T} = \frac{374,356}{2 \times 60 \times 60 s} = 51.99389 tps.$$

The number of sectors read/written per transaction

$$\frac{18,427.44 bytes}{512 bytes} = 35.9911 sectors.$$

As the time to read/write one sector (512 bytes) to the disk is $0.00026 s$, the time to read/write 35.9911 sectors is

$$D_{disk} = 0.00026 s \times 35.9911 sectors = 0.009358 s = 9.358 ms.$$

Therefore,

$$U_{disk} = D_{disk} \times X_0 = 0.009358 s \times 51.99389 tps = 0.486542.$$

And as $U_{cpu} = 0.38$, the bottleneck is the disk.

Análise Operacional



OL

□ **Exemplo4** Suponha um departamento composto por quatro recursos (pessoas: R1, R2, R3 e R4). Esse departamento foi monitorado por um período de 6 horas. Verificou-se que R1 esteve ocupado por 4h25min, R2 por 4h5min, R3 por 5h15min e R4 por 3h56min. O número total de transações que chegaram ao departamento foi 96. O sistema também finalizou a execução de 96 transações no mesmo período. O número total de chegadas a cada recurso e as respectivas finalizações são $A_1 = C_1 = 60$, $A_2 = C_2 = 110$, $A_3 = C_3 = 100$ e $A_4 = C_4 = 55$.

Qual a taxa de chegada ao sistema (λ_0)?

Qual é o *throughput* do sistema (X_0)?

Qual é a utilização de cada recurso (U_i)?

Qual é o tempo médio por transações finalizadas por cada recurso do sistema (S_i)?

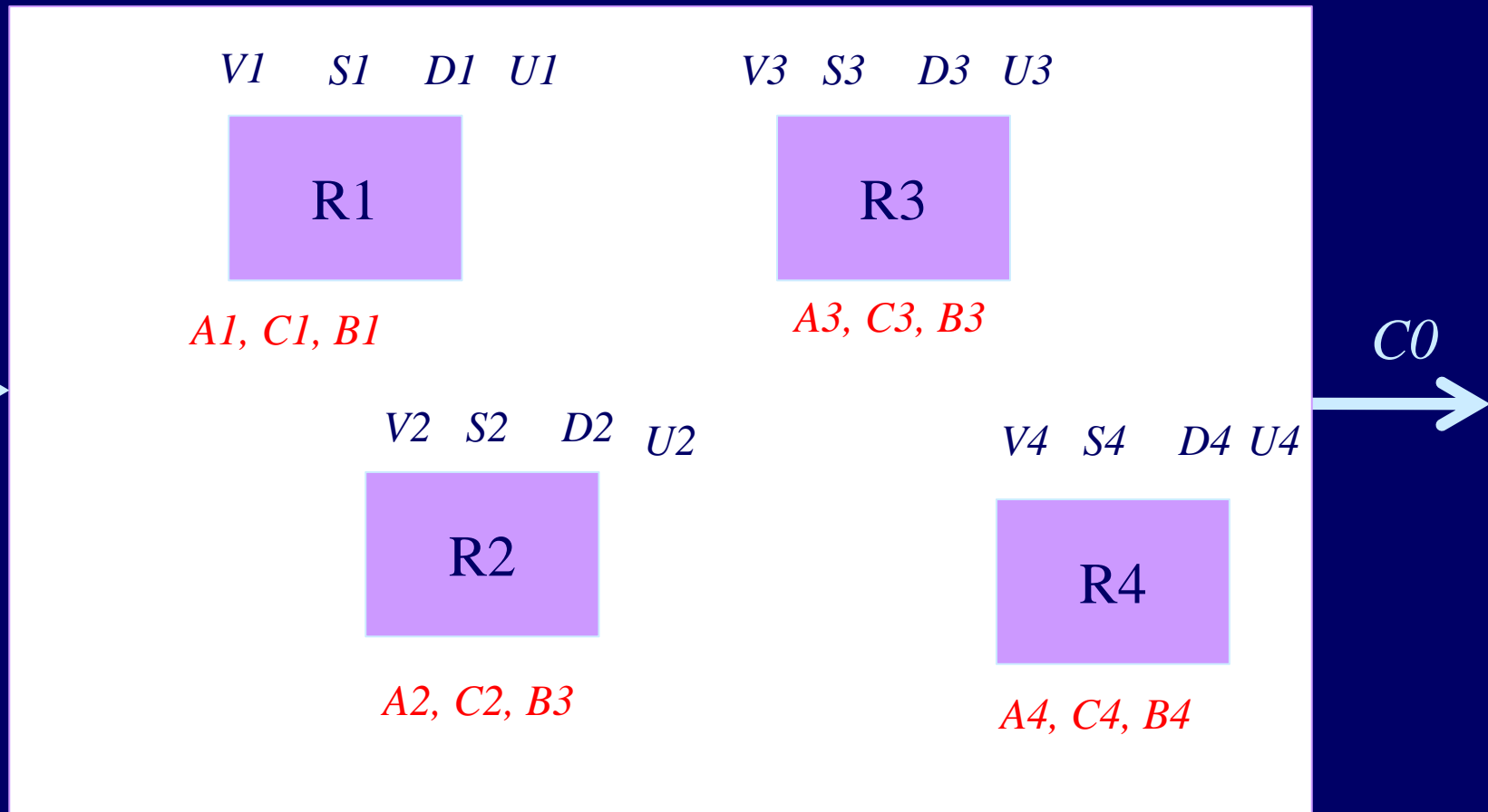
Qual é o número médio de visitas por recurso (V_i)?

Qual é tempo médio de uma transação qualquer (não necessariamente a que visitou o recurso i) no recurso i (D_i)?

Análise Operacional

OL

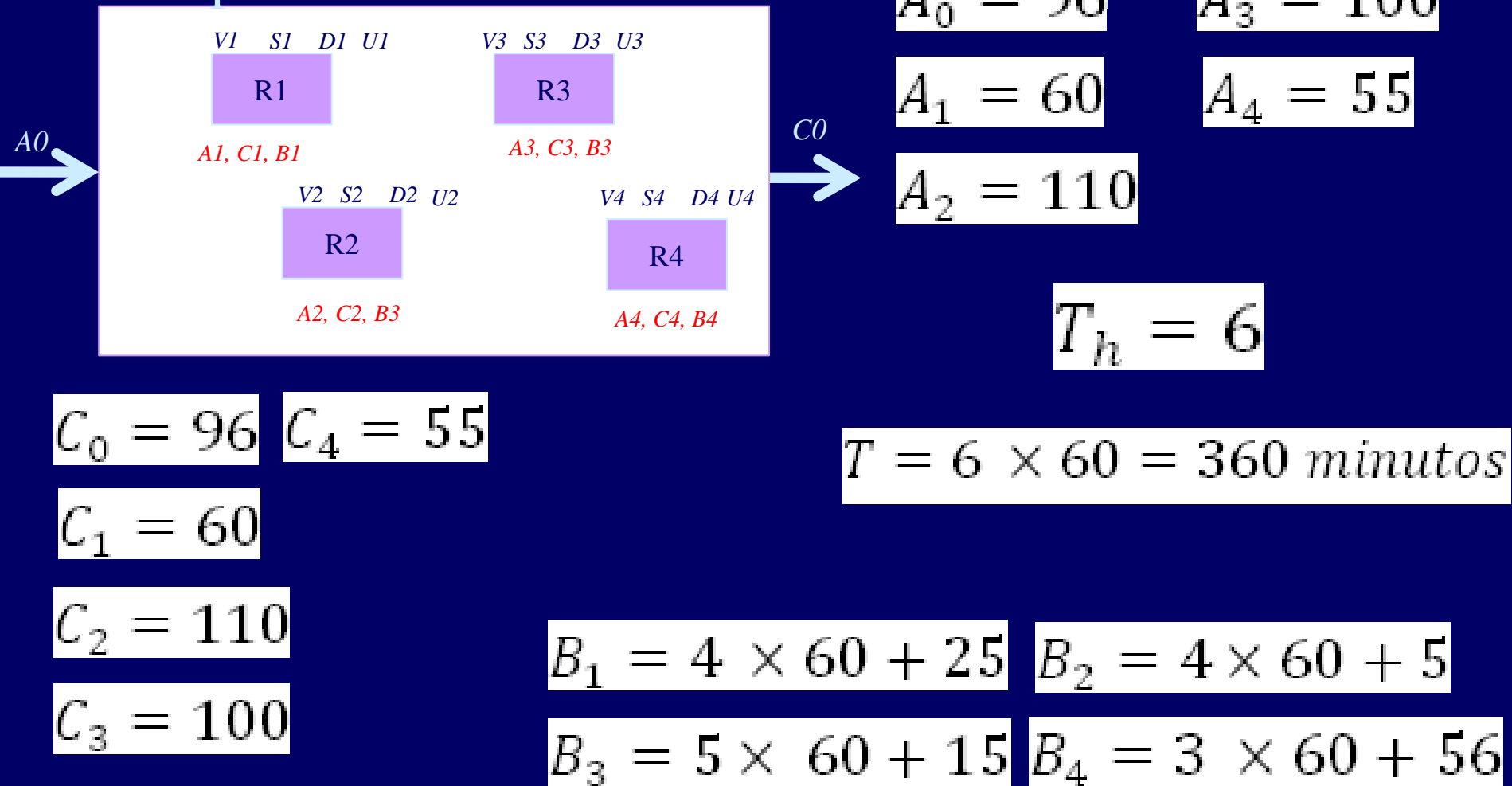
Exemplo 4



Análise Operacional

OL

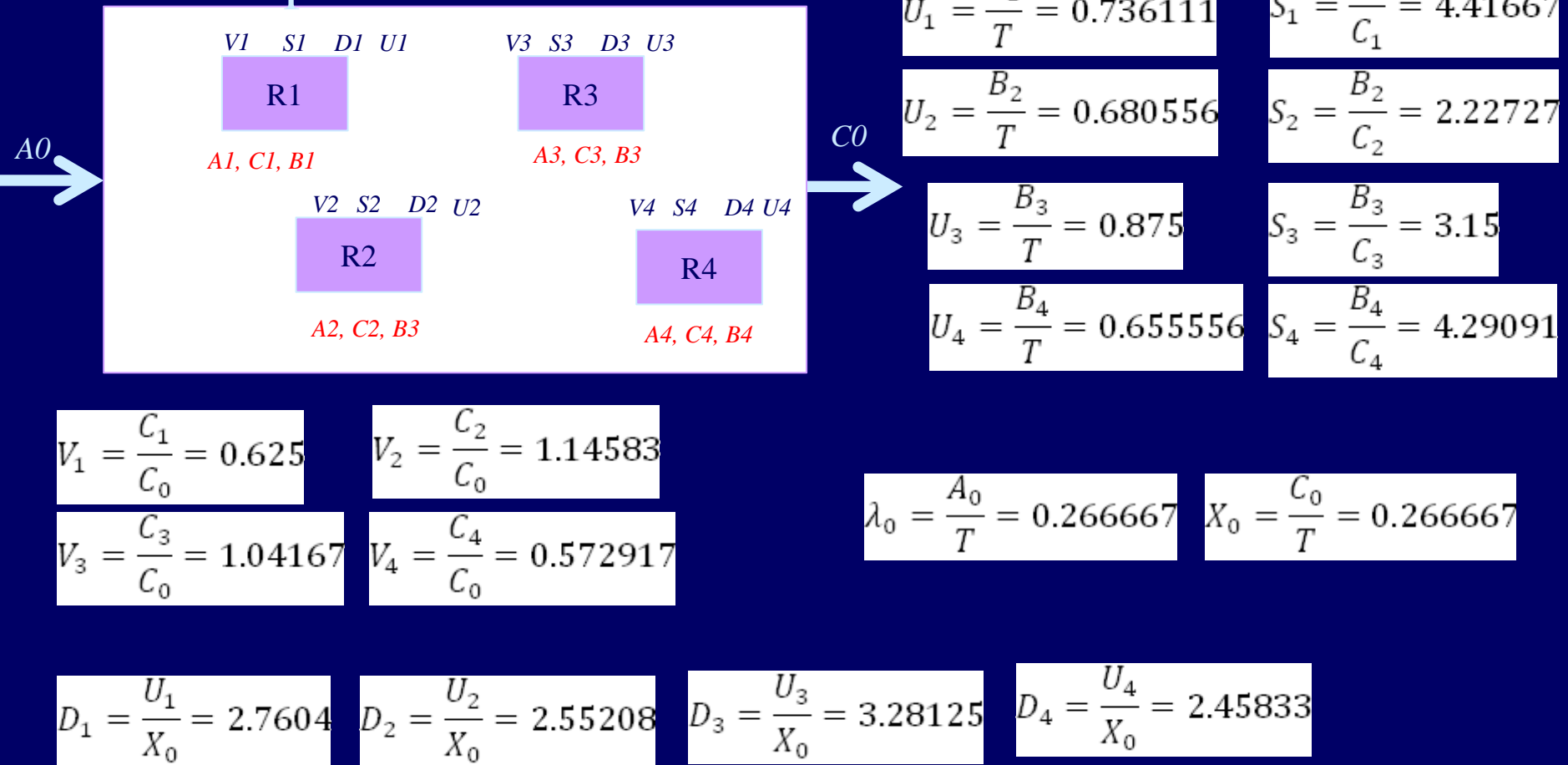
Exemplo 4



Análise Operacional

OL

Exemplo 4

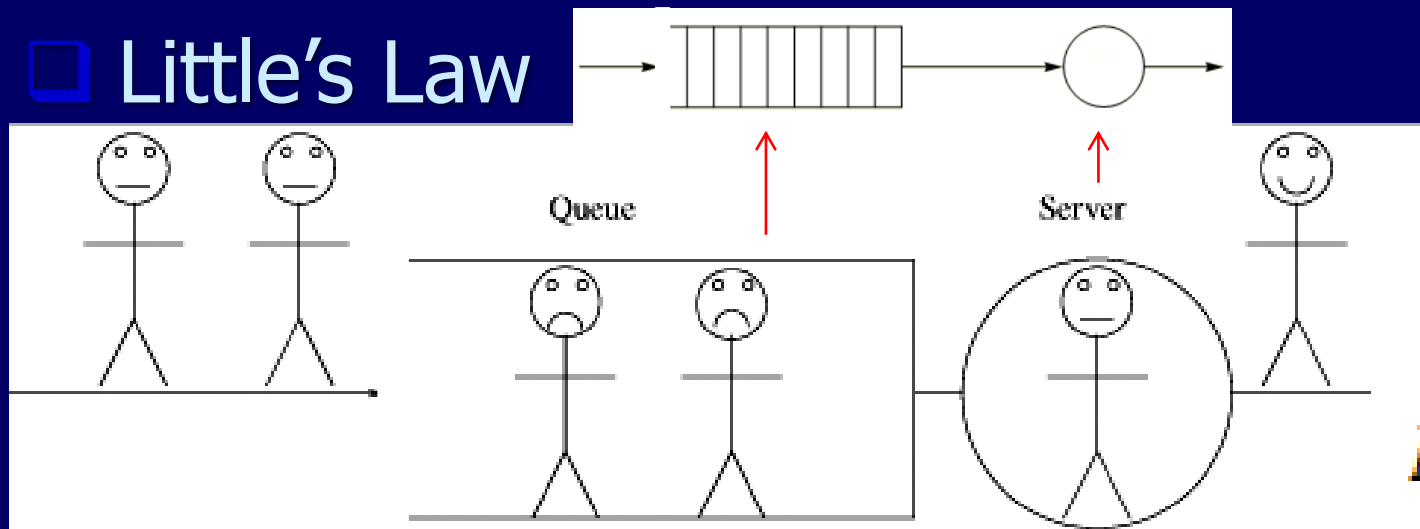




OL

Análise Operacional

Little's Law



$$N_i = \lambda_i \times R_i$$

A lei de Little também é uma lei operacional, pois utiliza apenas informações mensuráveis. Adotamos essa lei para relacionar o tamanho da fila N_i de um dispositivo i ao tempo de resposta deste dispositivo R_i , em função do número de chegadas (A_i) observadas no período (T). $\lambda_i = \frac{A_i}{T}$

R_i – Response time

W_i – Waiting time

S_i – Service time

N_i – Número de clientes no sistema

X_i – Throughput (vazão)

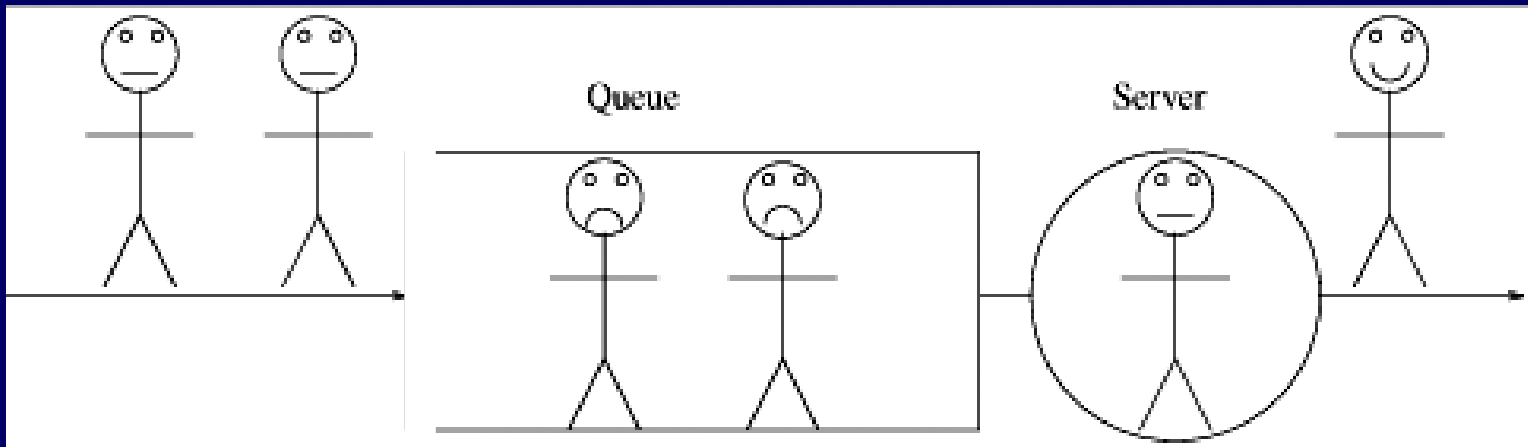
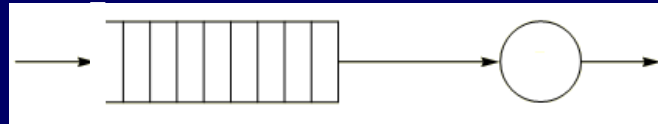
λ_i - taxa de chegada

Análise Operacional



OL

□ Little's Law



Se o sistema é balanceado, a taxa de chegada é igual ao *throughput*, portanto:

$$N_i = \lambda_i \times R_i = X_i \times R_i$$

Quando não há fila e se considera apenas um servidor, a Little's law corresponde a Utilization law:

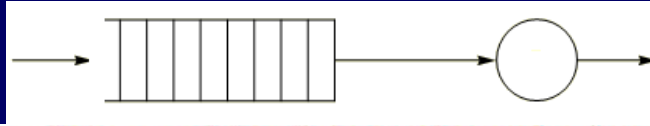
e $R_i = S_i$

Análise Operacional



OL

Exemplo



Considere um sistema de atendimento a usuários foi monitorado durante o período de 8h. A cada 30 minutos foi feita uma contagem do número de clientes que estavam aguardando para serem atendidos.



A amostra é esta:

11	7	13	17	17	19	3	11
16	8	0	15	5	18	6	14

O número médio de clientes que aguardaram para serem atendidos foi, portanto:

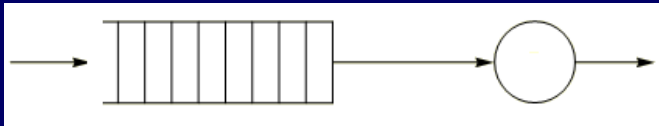
$$N = 11,2 \text{ clientes}$$

Análise Operacional



OL

Exemplo



Durante o mesmo período de 8h, verificou-se o número de clientes que foram atendidos. Este número foi $C_0 = 324$ clientes.

Qual é o tempo médio de permanência (*residence time, response time*) de cada cliente no sistema?

Sabemos que $N = \lambda \times R$ e que se o sistema é balanceado $N = X \times R$. Portanto:

$$R = \frac{N}{X}$$

Sabemos também que $X_0 = \frac{C_0}{T} = \frac{324 \text{ cliente}}{8 \times 60 \text{ minutos}} = 0,675 \text{ clientes por minuto (cpm)}$. Desta forma:

$$R = \frac{N}{X_0} = \frac{11,2 \text{ clientes}}{0,675 \text{ cpm}} = 16,67 \text{ minutos}$$

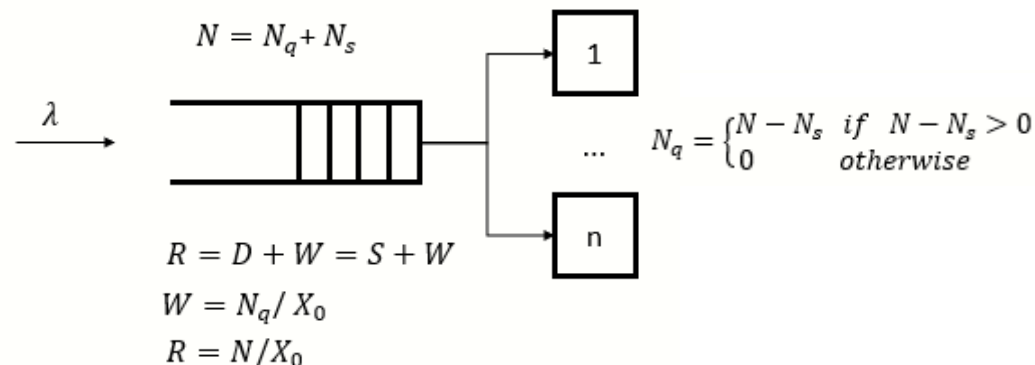


Análise Operacional

OL

Exemplo

Ni	N_average	
10	11.5	
8	T (min)	480
13	C0 (# clients that concludes their transactions)	120
10	X0 (Throughput)	0.25
18	R - Response time	46
8	Nq=N-Ns - Queue Size	10.5
9	Ns - Number of servers	1
11	W=R-D=R-S=Nq/X0 - Waiting time	42
16	D=S=R-W	4



Análise Operacional



OL

General Response Time Law

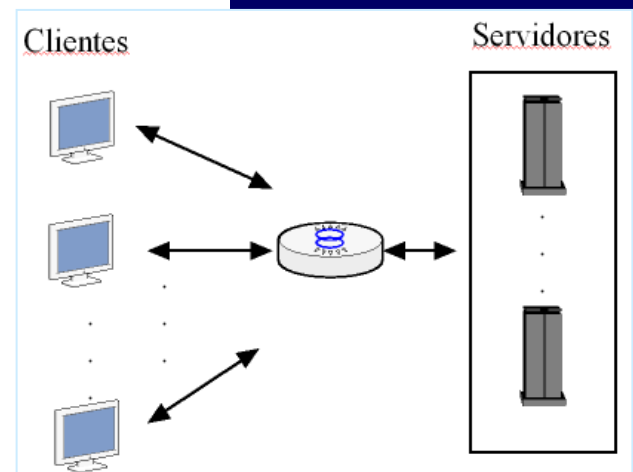
A Little's law pode ser aplicada a qualquer parte do sistema, basta apenas que o fluxo esteja “balanceado”. Portanto, pode-se aplicá-la a parte central do sistema (servidores) e ao sistema periférico (clientes).

N é o número total de transações no sistema, R é o *response time*, e X é o *throughput* do sistema.

$$N = X \times R$$

Dado que N_i é o número de transações em cada dispositivo, N pode ser calculado:

$$N = N_1 + N_2 + \dots + N_M$$



Análise Operacional



OL

General Response Time Law

$$N = N_1 + N_2 + \dots + N_M$$

$$X \times R = X_1 \times R_1 + X_2 \times R_2 + \dots + X_M \times R_M$$

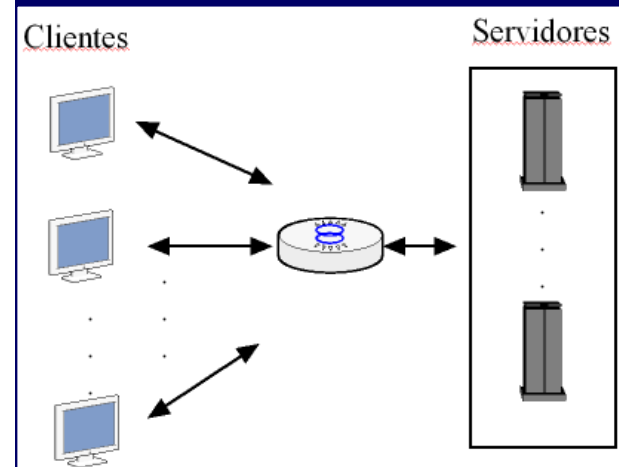
Dividindo-se por X , tem-se:

$$\frac{X \times R}{X} = \frac{X_1 \times R_1 + X_2 \times R_2 + \dots + X_M \times R_M}{X}$$

$$\frac{X \times R}{X} = \frac{X_1 \times R_1}{X} + \frac{X_2 \times R_2}{X} + \dots + \frac{X_M \times R_M}{X}$$

$$R = V_1 \times R_1 + V_2 \times R_2 + \dots + V_M \times R_M$$

$$R = \sum_i^M V_i \times R_i$$



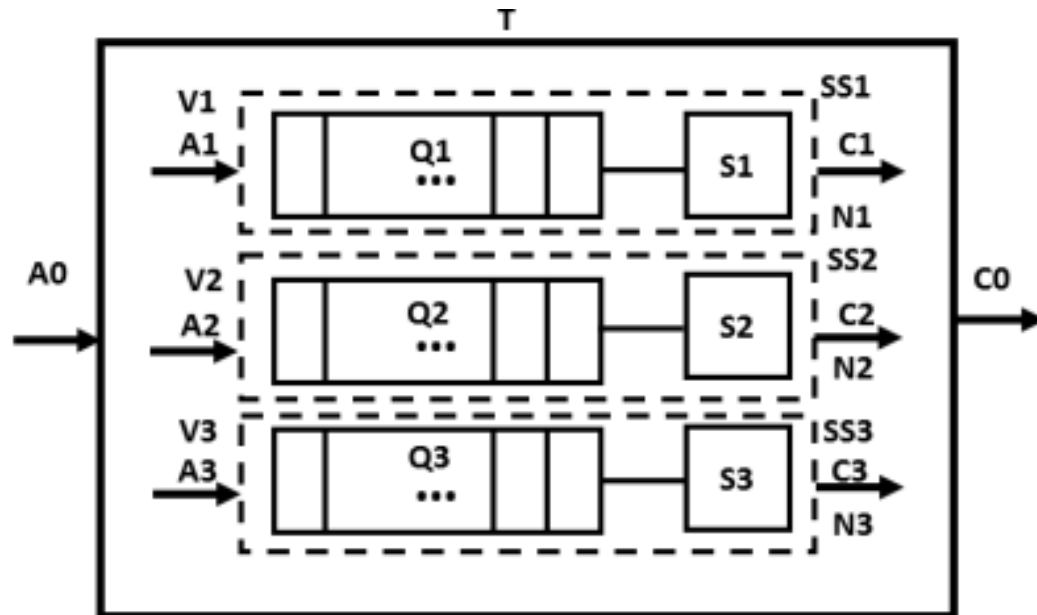


OL

Análise Operacional

General Response Time Law

Example 6.5.1. Consider a balanced system composed of three subsystems (ss_1 , ss_2 , and ss_3). Each of these subsystems consists of a machine (S_i) and an input queue (Q_i). Samples of the row sizes of Q_1 , Q_2 and Q_3 were collected every 18s for a period $T = 30min$ (1800s). We obtained, therefore, three (3) samples with one hundred (100) measures each. A queue representation of such system is depicted in Figure 6.6. The three samples were recorded as in Table 6.1. The average sizes of the respective queues were $\bar{N}_1 = 20.862$, $\bar{N}_2 = 8.123$ and $\bar{N}_3 = 12.237$.

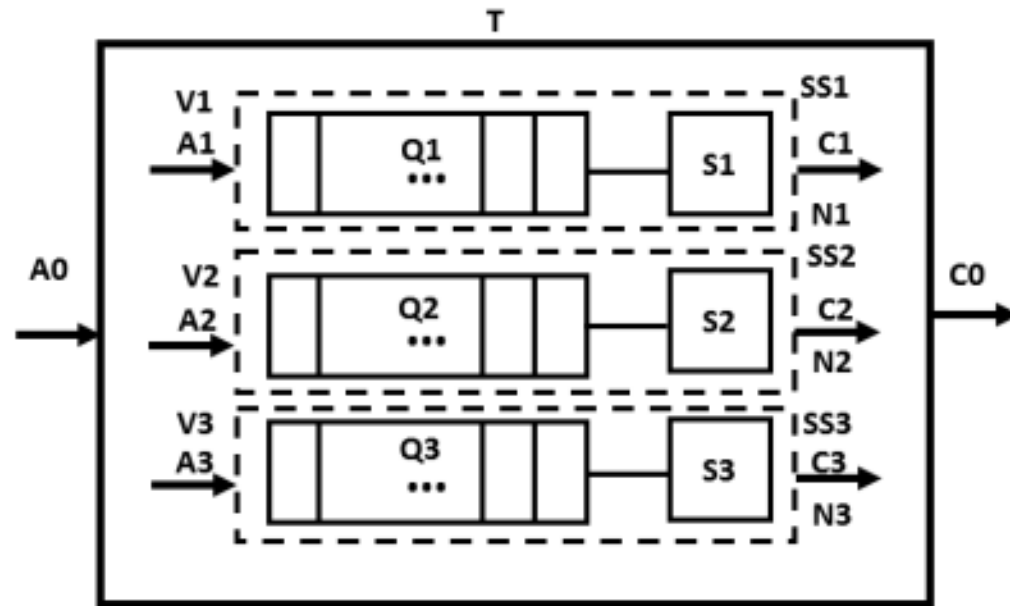




OL

Análise Operacional

General Response Time Law



The number of transactions completed by (ss_1 , ss_2 , and ss_3) and by the system were $C_1 = 120$, $C_2 = 90$, $C_3 = 70$ and $C_0 = 210$, respectively. Therefore, the respective throughputs were:

$$X_1 = \frac{120}{1800} = 0.0667 \text{ tps}, X_2 = \frac{90}{1800} = 0.05 \text{ tps}, X_3 = \frac{70}{1800} = 0.0389 \text{ tps} \text{ and } X_0 = \frac{210}{1800} = 0.1167 \text{ tps}.$$



OL

Análise Operacional

General Response Time Law

$$R_1 = \frac{\overline{N}_1}{X_1} = 312.936s$$

$$R_2 = \frac{\overline{N}_2}{X_2} = 162.451s$$

$$R_3 = \frac{\overline{N}_3}{X_3} = 314.653s$$

$$\text{As } R = \sum_{i=1}^k V_i \times R_i,$$

and $k = 3$, we have:

$$R = V_1 \times R_1 + V_2 \times R_2 + V_3 \times R_3,$$

$$\text{hence, } R = 0.5714 \times 312.9368s + 0.4286 \times 162.451s + 0.3333 \times 314.6533s.$$

$$R = 353.3273s.$$

We also know that $V_1 = \frac{X_1}{X_0}$, $V_2 = \frac{X_2}{X_0}$, and $V_3 = \frac{X_3}{X_0}$.

Thus:

$$V_1 = \frac{0.0667}{\frac{0.1167}{82}} = 0.5714,$$

$$V_2 = \frac{0.05}{0.1167} = 0.4286, \text{ and}$$

$$V_3 = \frac{0.0389}{0.1167} = 0.3333.$$

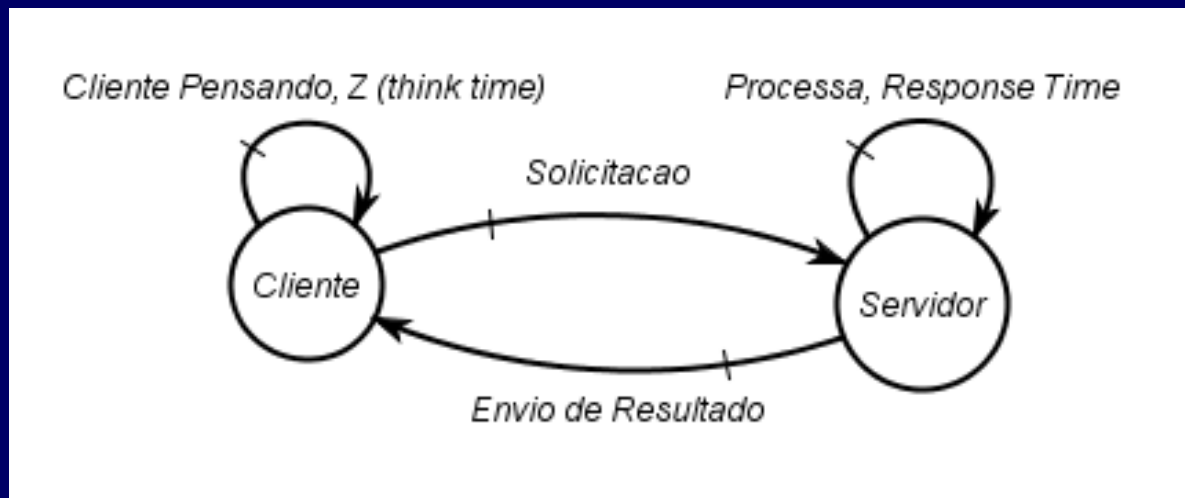
Análise Operacional



OL

Interactive Response Time Law

Em um sistema interativo, os clientes fazem uma solicitação a um sistema servidor, o sistema servidor processa essa solicitação e devolve um resultado ao cliente. Após um período de espera (*think time*) Z , o cliente faz uma nova solicitação. Se o *system response time* é R , o tempo total desse ciclo é $R + Z$.





Análise Operacional

□ Interactive Response Time Law

Se considerarmos um período T , cada cliente gerará:

$$\frac{T}{R + Z} \text{ solicitações no período } T.$$

Se considerarmos N clientes, teremos:

$$\frac{N \times T}{R + Z} \text{ solicitações no período } T.$$

Portanto, o *throughput* do sistema é:

$$X = \frac{N \times T}{R + Z} \frac{1}{T}$$

$$X = \frac{N}{R + Z} \quad \text{and} \quad R = \frac{N}{X} - Z$$

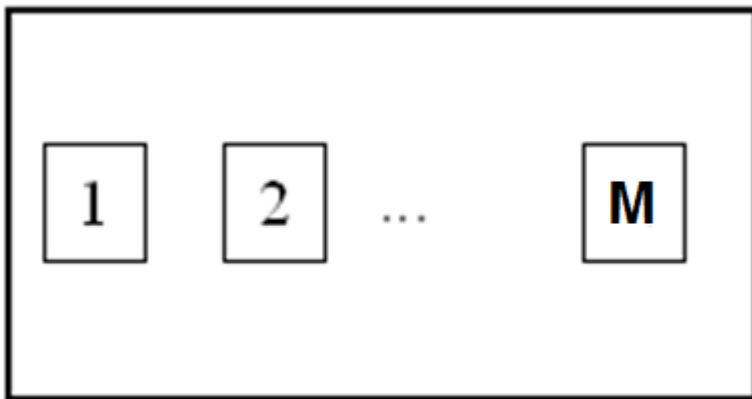
Análise Operacional



OL

□ *Bottleneck Analysis*

Observe que a utilização U_i do dispositivo i é diretamente proporcional à demanda D_i (*service demand*), portanto o dispositivo com mais alta demanda $\max_i \{D_i\}$ tem a mais alta utilização e é o “gargalo” do sistema.



Sistema com M componentes em paralelo

$$X_0 \leq \frac{1}{D_{max}}$$

$$R \geq N \times D_{max} - Z.$$

Todas atividades começam no mesmo momento, mas a tarefa “maior” só finaliza quando todas as atividades finalizarem.



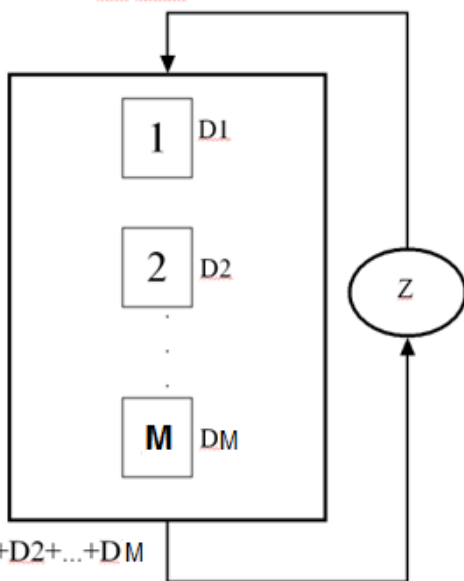
OL

Análise Operacional

□ Bottleneck Analysis

Considere agora outra situação limite: um sistema composto por M componentes em série e que o clientes tenham um *think time* Z .

Sistema com M componentes em série



$$R = V_1 \times R_1 + V_2 \times R_2 + \dots + V_M \times R_M$$

$$\text{Dado que } R_i = W_i + S_i$$

$$R \geq v_1 \times S_1 + v_2 \times S_2.$$

$$D_i = v_i \times S_i,$$

$$D = \sum_{i=1}^M D_i.$$

$$R \geq D = \sum_{i=1}^M D_i$$

Sabemos que

$$X = \frac{N}{R + Z} \leq \frac{N}{D + Z}$$

Se considerarmos

$Z = 0$, temos

$$X \leq \frac{N}{D}$$

Portanto

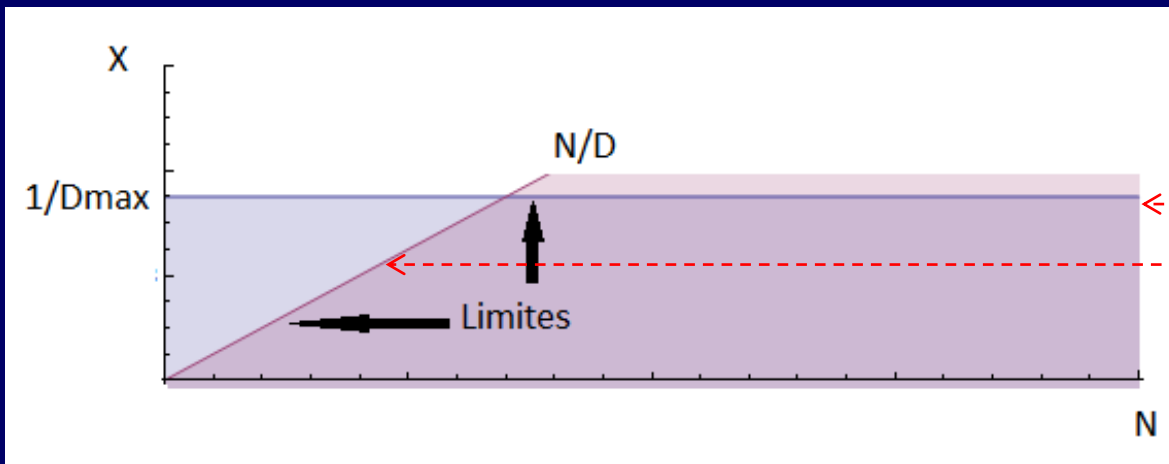
$$X \leq \min \left\{ \frac{1}{D_{\max}}, \frac{N}{D} \right\}$$

Análise Operacional



OL

□ Bottleneck Analysis



Portanto, $R = D$

E como sabemos que

$$X = \frac{N}{R + Z}$$

Temos:

$$X = \frac{N}{D + Z}$$

Portanto:

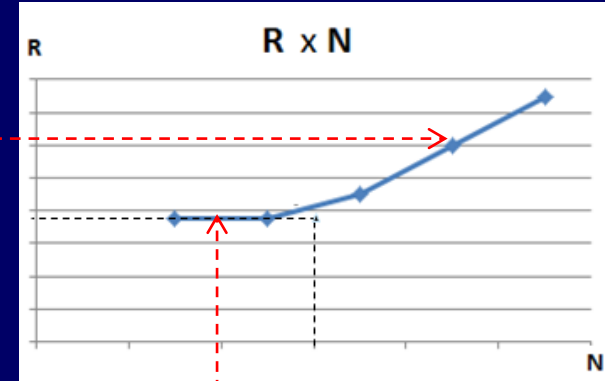
$$X \leq \frac{N}{D}$$

Consequentemente:

$$X \leq \min \left\{ \frac{1}{D_{max}}, \frac{N}{D} \right\}$$

Análise Operacional

□ Bottleneck Analysis



Sabemos que: $R = \frac{N}{X} - Z$

Como $X_0 \leq \frac{1}{D_{max}}$

Temos:

$$R = \frac{N}{X_0} - Z \geq N \times D_{max} - Z$$

$$R \geq N \times D_{max} - Z$$

Sabemos que:

$$R = \sum_i^M V_i \times R_i$$

$$R_i = W_i + S_i$$

Para um sistema com M components, temos:

$$R = \sum_i^M V_i(W_i + S_i) \geq \sum_i^M V_i S_i = \sum_i^M D_i = D$$

Portanto: $R \geq D$

Desta forma:

$$R \geq \max\{D, N \times D_{max} - Z\}$$

Se $Z = 0$, temos:

$$R \geq \max\{D, N \times D_{max}\}$$

Análise Operacional

□ Bottleneck Analysis

Knee

Quando as duas assíntotas se encontram, temos:

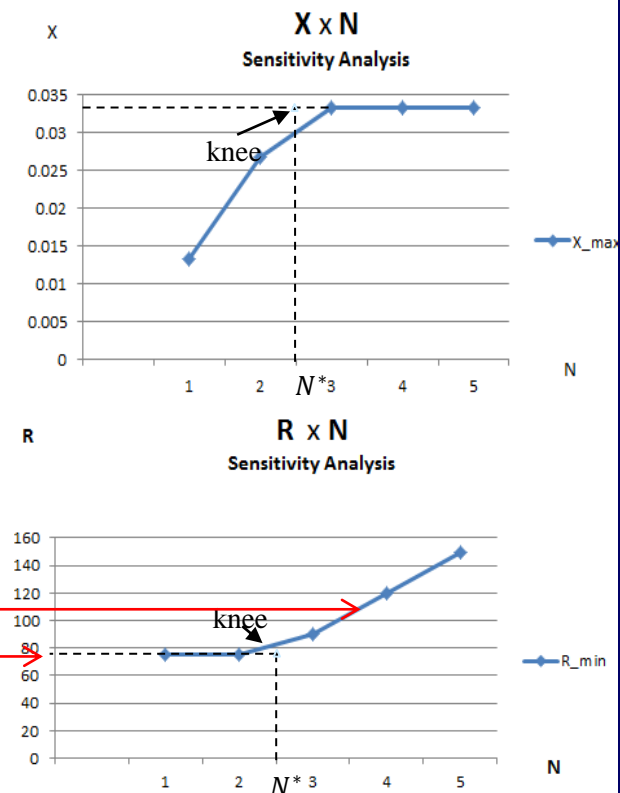
$$D = N \times D_{max} - Z$$

Portanto:

$$N^* = \frac{D}{D_{max}} + Z$$

Se $Z = 0$, temos:

$$N^* = \frac{D}{D_{max}}$$



Análise Operacional



OL

□ *Bottleneck Analysis*

Problema:

Considere um sistema composto por um servidor de aplicação (SA), um servidor de banco de dados (SBD), uma unidade de disco (D) e um servidor de autenticação (SAut).

Clientes se “logam” no sistema, procuram documentos (textos) e fazem downloads dos documentos de seu interesse. Considere situações em que se tem até 5 clientes.

Análise Operacional



OL

□ *Bottleneck Analysis*

Problema:

Esse sistema e seus componentes foram monitorados por 4h. Observaram-se, nesse período, a conclusão de 400.000 transações. A utilização média medida de cada recurso foi:

$$U_{SA} = 0.555556, U_{SBD} = 0.833333, U_D = 0.416667, \text{ e } U_{SAut} = 0.277778.$$

Qual é a vazão máxima do sistema?



OL

Análise Operacional

□ Bottleneck Analysis

Problema:

Sabemos que:

$$X_0 = \frac{400.000}{4 \times 60 \text{ min} \times 60 \text{ s} \times 1000 \text{ ms}} = 0.027778 \text{ ms}$$

Dado que $D_i = \frac{U_i}{X_0}$, temos:

$$D_{SA} = \frac{U_{SA}}{X_0} = \frac{0.555556}{0.027778 \text{ ms}} = 20 \text{ ms}$$

$$D_{SBD} = \frac{U_{SA}}{X_0} = \frac{0.833333}{0.027778 \text{ ms}} = 30 \text{ ms}$$

$$D_D = \frac{U_D}{X_0} = \frac{0.416667}{0.027778 \text{ ms}} = 15 \text{ ms}$$

$$D_{DSA_{ut}} = \frac{U_{SA_{ut}}}{X_0} = \frac{0.277778}{0.027778 \text{ ms}} = 10 \text{ ms}$$



OL

Análise Operacional

□ *Bottleneck Analysis*

Problema: Sabemos, portanto, que:

$$\begin{aligned} D_{max} &= \max \{D_{SA}, D_{SBD}, D_D, D_{SAut}\} \\ &= 30ms \end{aligned}$$

E que

$$D = \sum_{i \in \{D_{SA}, D_{SBD}, D_D, D_{SAut}\}} D_i = 75 \text{ ms}$$

Como

$$X = \min_{N \in \{1,2,3,4,5\}} \left\{ \frac{1}{D_{max}}, \frac{N}{D} \right\}$$

e $N = \{1,2,3,4,5\}$,



Análise Operacional

OL - res

□ Bottleneck Analysis Problema:

temos:

$$N = 1$$
$$X = \min_{N=1} \left\{ \frac{1}{D_{max}}, \frac{N}{D} \right\} = \min \left\{ \frac{1}{30ms}, \frac{1}{75ms} \right\}$$
$$= 0.013333$$

$$N = 2$$
$$X = \min_{N=2} \left\{ \frac{1}{D_{max}}, \frac{N}{D} \right\} = \min \left\{ \frac{1}{30ms}, \frac{2}{75ms} \right\}$$
$$= 0.026667$$

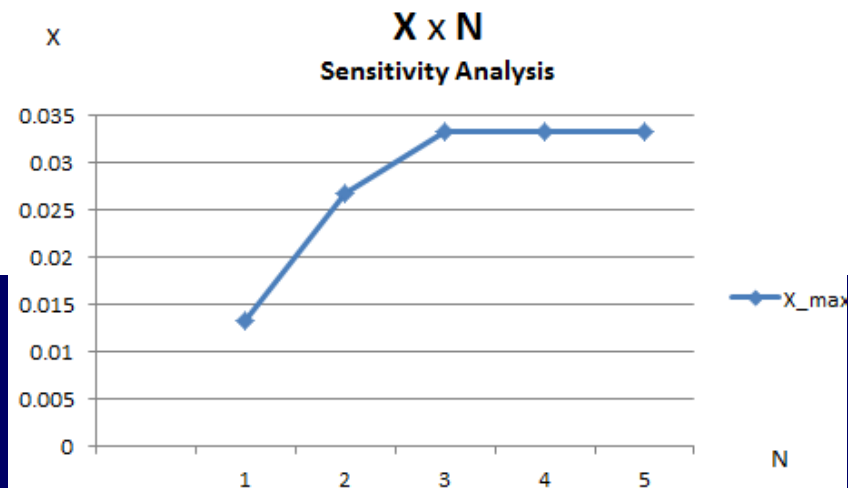
$$N = 3$$
$$X = \min_{N=3} \left\{ \frac{1}{D_{max}}, \frac{N}{D} \right\} = \min \left\{ \frac{1}{30ms}, \frac{3}{75ms} \right\}$$
$$= 0.033333$$

$$N = 4$$

$$X = \min_{N=4} \left\{ \frac{1}{D_{max}}, \frac{N}{D} \right\} = \min \left\{ \frac{1}{30ms}, \frac{4}{75ms} \right\}$$
$$= 0.033333$$

$$N = 5$$

$$X = \min_{N=5} \left\{ \frac{1}{D_{max}}, \frac{N}{D} \right\} = \min \left\{ \frac{1}{30ms}, \frac{5}{75ms} \right\}$$
$$= 0.033333$$

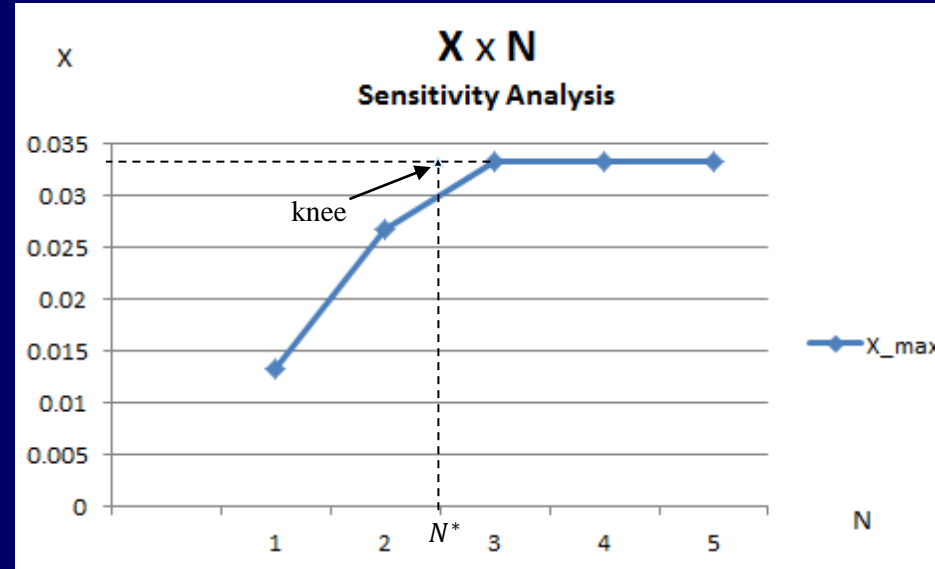
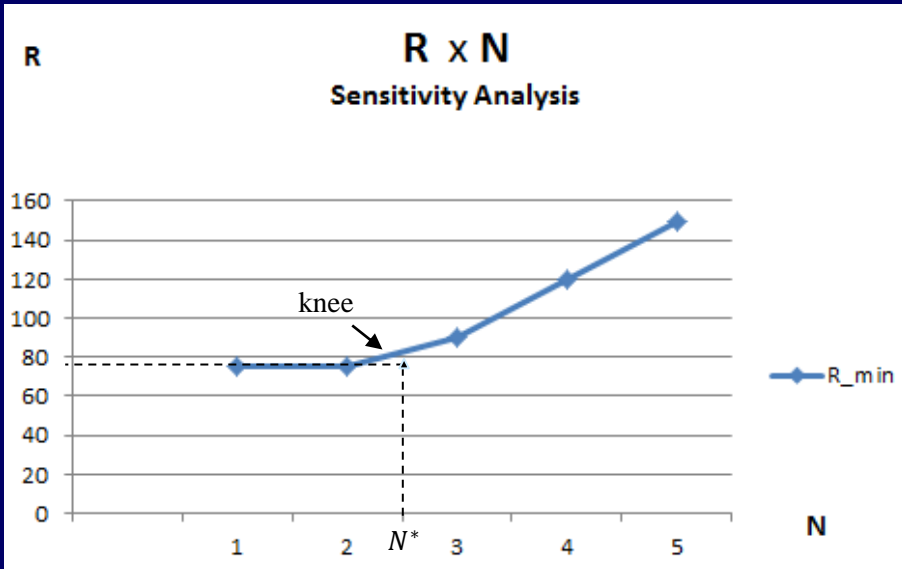




Análise Operacional

OL - res

□ *Bottleneck Analysis* Problema:



O número de transações (clientes, jobs, operações) N^* no “joelho” (*knee*) é

$$\text{estimado por: } N^* = \frac{D+Z}{D_{max}} = 2,5$$



Análise Operacional

□ *Bottleneck Analysis*

Problema:

Considere um sistema composto por um servidor (S) e duas unidades de disco (D1 e D2). Uma carga de trabalho, associada a um cliente, foi aplicada ao sistema durante 6000s e os recursos (S, D1 e D2) foram monitorados. Neste período, 300000 transações foram concluídas. As utilizações do servidor e dos discos D1 e D2 foram 35%, 40% e 28%, respectivamente.

- Trace as assíntotas que delimitam a vazão máxima do sistema. Considere que até sete usuários (N) podem fazer uso do sistema simultaneamente.
- Trace as assíntotas que delimitam a vazão máxima do sistema, considerando agora que a unidade de disco D1 foi substituída por outra unidade (NewD1) e que o respectivo *time demand* passou a ser (D_{NewD1}) 4ms. Considere também que até sete usuários (N) podem fazer uso do sistema simultaneamente.
- Trace também as assíntotas de R para as situações de carga descritas na letra a.
- Considerando as situações de carga descritas na letra b, obtenha as assíntotas de R.



Análise Operacional

Bottleneck Analysis

$T = 6000s$, $C_0 = 300000 \text{ trans.}$, $U_S = 35\%$,
 $U_{D1} = 40\%$, $U_{D2} = 28\%$.

$$X_0 = \frac{300000 \text{ trans.}}{6000s} = 50 \text{ tps}$$

$$D_S = \frac{U_S}{X_0} = \frac{35\%}{50 \text{ tps}} = 0.0070s = 7ms$$

$$D_{D1} = \frac{U_{D1}}{X_0} = \frac{40\%}{50 \text{ tps}} = 0.0080s = 8ms$$

$$D_{D2} = \frac{U_{D2}}{X_0} = \frac{28\%}{50 \text{ tps}} = 0.0056s = 5,6ms$$

$$D_{max} = \max_i \{D_S, D_{D1}, D_{D2}\} = 8ms$$

$$D = D_S + D_{D1} + D_{D2} = 0,0206s = 20,6ms$$

$$\text{Bottleneck} = \arg \left(i, \max_i \{D_S, D_{D1}, D_{D2}\} \right) = D1$$

a)

$$\text{Knee} = \frac{D + Z}{D_{max}} = \frac{D}{D_{max}} = \frac{20,6ms}{8ms} = 2,575$$

Sabemos que:

$$X_{max} = \min_i \left\{ \frac{1}{D_{max}}, \frac{N}{D} - Z \right\}$$

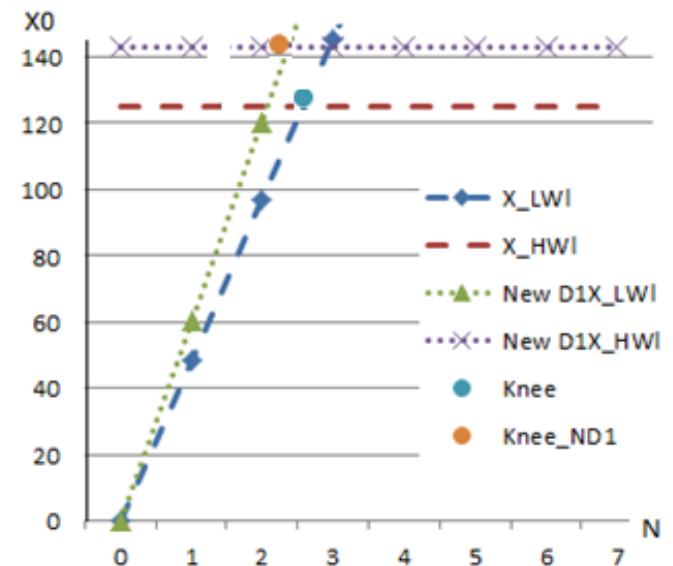
Consideraremos que o *think time* é

$Z = 0$, portanto:

$$X_{max} = \min_i \left\{ \frac{1}{D_{max}}, \frac{N}{D} \right\}$$

$$\frac{1}{D_{max}} = \frac{1}{8ms} = 125 \text{ tps}$$

N	$X_{LWI} = \frac{N}{D}$ (tps)	$X_{HWI} = \frac{1}{D_{max}}$ (tps)
0	0	125
1	48.54369	125
2	97.08738	125
3	145.6311	125
4	194.1748	125
5	242.7184	125
6	291.2621	125
7	339.8058	125





Análise Operacional

b)

$$New_{D_{D1}} = \frac{8ms}{2} = 4ms$$

$$D_S = \frac{U_S}{X_0} = \frac{35\%}{50tps} = 0.0070s = 7ms$$

$$D_{D2} = \frac{U_{D2}}{X_0} = \frac{28\%}{50tps} = 0.0056s = 5,6ms$$

$$D_{max} = \max_i \{D_S, D_{D1}, D_{D2}\} = 7ms$$

$$D = D_S + D_{D1} + D_{D2} = 16,6ms$$

$$Bottleneck = \arg \left(i, \max_i \{D_S, D_{New_D1}, D_{D2}\} \right) = S$$

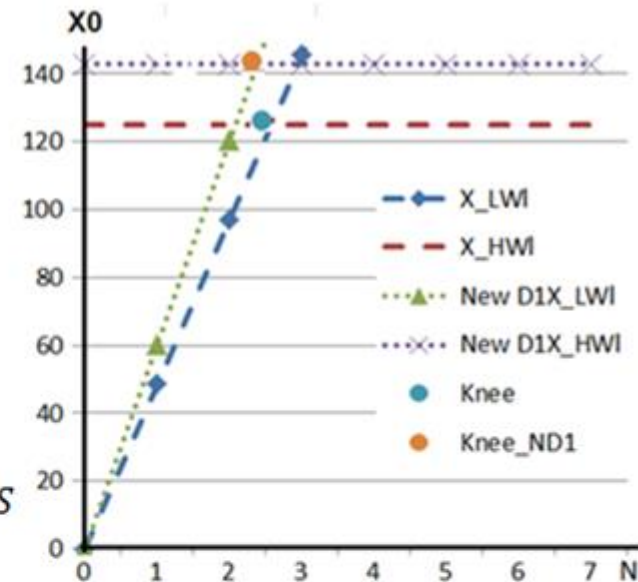
$$X_{max} = \min \left\{ \frac{1}{D_{max}}, \frac{N}{D} - Z \right\}$$

Consideraremos que o think time
é $Z = 0$, portanto:

$$X_{max} = \min \left\{ \frac{1}{D_{max}}, \frac{N}{D} \right\} \quad \frac{1}{D_{max}} = \frac{1}{7ms} = 142,86tps$$

$$X_{max} = 142,86tps$$

$$Knee = \frac{D + Z}{D_{max}} = \frac{D}{D_{max}} = \frac{16,6ms}{7ms} = 2,371$$



N	New D1X_LWI
0	0
1	60.2409639
2	120.4819277
3	180.7228916
4	240.9638554
5	301.2048193
6	361.4457831
7	421.6867470



Análise Operacional

c)

$T = 6000s, C_0 = 300000 \text{ trans.}, U_S = 35\%,$
 $U_{D1} = 40\%, U_{D2} = 28\%.$

$$\lambda_0 = \frac{300000 \text{ trans.}}{6000s} = 50 \text{ tps}$$

$$D_S = \frac{U_S}{\lambda_0} = \frac{35\%}{50 \text{ tps}} = 0.0070s = 7ms$$

$$D_{D1} = \frac{U_{D1}}{\lambda_0} = \frac{40\%}{50 \text{ tps}} = 0.0080s = 8ms$$

$$D_{D2} = \frac{U_{D2}}{\lambda_0} = \frac{28\%}{50 \text{ tps}} = 0.0056s = 5,6ms$$

$$D_{\max} = \max_i \{D_S, D_{D1}, D_{D2}\} = 8ms$$

$$D = D_S + D_{D1} + D_{D2} = 0,0206s = 20,6ms$$

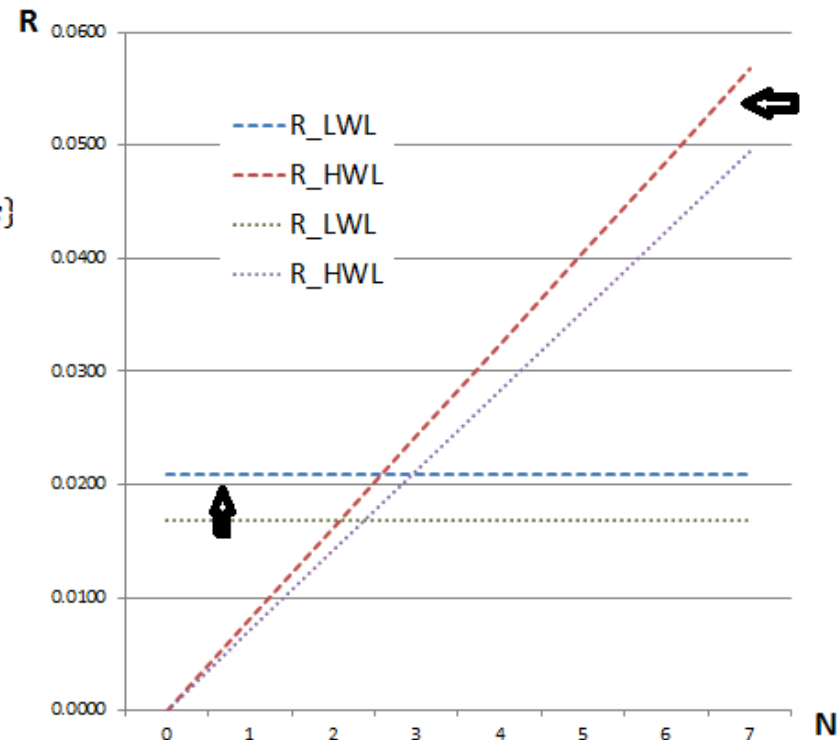
$$\text{Bottleneck} = \arg(i, \max_i \{D_S, D_{D1}, D_{D2}\}) = D1$$

$$R = \frac{N}{X} - Z \quad Z = 0$$

$$R \geq \max\{D, N \times D_{\max}\}$$

$$R \geq \max\{20,6ms, N \times 8ms\}$$

N	R_LWL	R_HWL
0	0.0208	0.0000
1	0.0208	0.0081
2	0.0208	0.0162
3	0.0208	0.0243
4	0.0208	0.0324
5	0.0208	0.0405
6	0.0208	0.0486
7	0.0208	0.0567





Análise Operacional

b)

d)

$$New_{D_{D1}} = \frac{8ms}{2} = 4ms$$

$$D_S = \frac{U_S}{X_0} = \frac{35\%}{50tps} = 0.0070s = 7ms$$

$$D_{D2} = \frac{U_{D2}}{X_0} = \frac{28\%}{50tps} = 0.0056s = 5,6ms$$

$$D_{max} = \max_i \{D_S, D_{D1}, D_{D2}\} = 7ms$$

$$D = D_S + D_{D1} + D_{D2} = 16,6ms$$

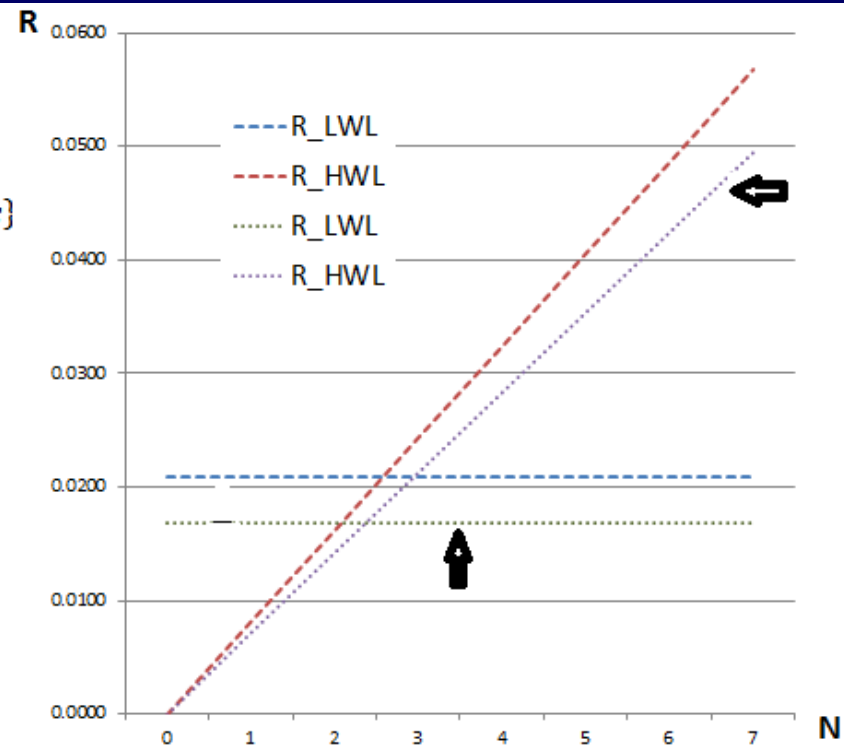
$$Bottleneck = \arg \left(i, \max_i \{D_S, D_{New_D1}, D_{D2}\} \right) = S$$

$$R = \frac{N}{X} - Z \quad Z = 0$$

$$R \geq \max\{D, N \times D_{max}\}$$

$$R \geq \max\{20,6ms, N \times 8ms\}$$

N	R_LWL	R_HWL
0	0.0168	0.0000
1	0.0168	0.0071
2	0.0168	0.0141
3	0.0168	0.0212
4	0.0168	0.0283
5	0.0168	0.0354
6	0.0168	0.0424
7	0.0168	0.0495





Análise Operacional

Leis Operacionais (*derived measures*)

Utilization Law:

$$U_i = X_i \times S_i = \lambda_i \times S_i$$

Forced Flow Law:

$$X_i = V_i \times X_0$$

Service Demand Law:

$$D_i = V_i \times S_i = U_i / X_0$$

Little's Law:

$$N = X \times R$$

Interactive Response Time Law

$$R = \frac{N}{X} - Z$$

Análise Operacional



OL

Exemplo

A server was monitored for $T=2h$ considering a specific workload. In this period, the average processor utilization was $U_{CPU} = 0.38$, and $C0 = 3,743,566$ transactions were processed. Each transaction, on average, reads/writes 842,744.2 *bytes* from/to the disk. The average time to read or write 1 sector (512 bytes) from/to the disk is 0.26 ms. Find out the bottleneck (CPU or Disk).



Análise Operacional

OL

Resolution:

The transaction throughput:

$$X_0 = \frac{3,743,566 \text{ trans.}}{2 \times 60 \times 60 \text{ s}} = 1.308546 \text{ tps}$$

The number of sectors read/written per transaction =

$$\frac{842,744.2 \text{ bytes}}{512 \text{ bytes}} = 1645.985 \text{ sectors}$$

The time to read/write one sector (512 bytes) to the disk = 0.00026 s.

$$\begin{aligned} \text{Time to read/write } 1645.985 \text{ sectors} &= D_{\text{disk}} = \\ 0.00026 \text{ s} \times 1645.985 \text{ sectors} &= 0.427956 \text{ s} \end{aligned}$$

$$U_{\text{disk}} = D_{\text{disk}} \times X_0 = 0.427956 \text{ s} \times 1.308546 \text{ tps} = 0.56$$

As $U_{\text{CPU}} = 0.38$, the bottleneck is the disk.

Análise Operacional



OL

Exemplo

Example 6.3.2. A system composed of five servers was monitored for four hours ($T = 4 \times 60 \times 60s = 14400s$) under operational conditions. In this period, the log registered $C_0 = 28978$ transactions processed. The servers' utilizations were obtained over the period every $30s$. Hence a sample of 480 utilizations for each server was recorded. The average utilizations over the four hours period of each server were $\overline{U}_{s_1} = 0.3996$, $\overline{U}_{s_2} = 0.2389$, $\overline{U}_{s_3} = 0.2774$, $\overline{U}_{s_4} = 0.5253$, and $\overline{U}_{s_5} = 0.2598$, respectively.

What is the demand of each server?



Análise Operacional

OL

$$X_0 = C_0/T = 28978 \text{ trans.}/14400 \text{ s} = 2.0124 \text{ tps.}$$

$$\overline{U}_{s_1} = 0.3996, \quad \overline{U}_{s_2} = 0.2389,$$

$$\overline{U}_{s_3} = 0.2774, \quad \overline{U}_{s_4} = 0.5253, \quad \text{and} \quad \overline{U}_{s_5} = 0.2598$$

$$D_{s_i} = \frac{\overline{U}_{s_i}}{X_0}$$

Server	Demand (s)
<i>Server 1</i>	0.1986
<i>Server 2</i>	0.1187
<i>Server 3</i>	0.1378
<i>Server 4</i>	0.2610
<i>Server 5</i>	0.1291

Therefore, each typical transaction demanded the respective times presented above of each specific server. Now, assume that a considerable demand increase is forecasted. It is expected that $C'_0 = 60000$ transactions would be requested in the same four hours period.

What would be the foreseen utilization of each server?



OL

Análise Operacional

$$U'_{s_i} = \lceil D_{s_i} \times X'_0 \rceil .$$

since the maximal utilization is 1.

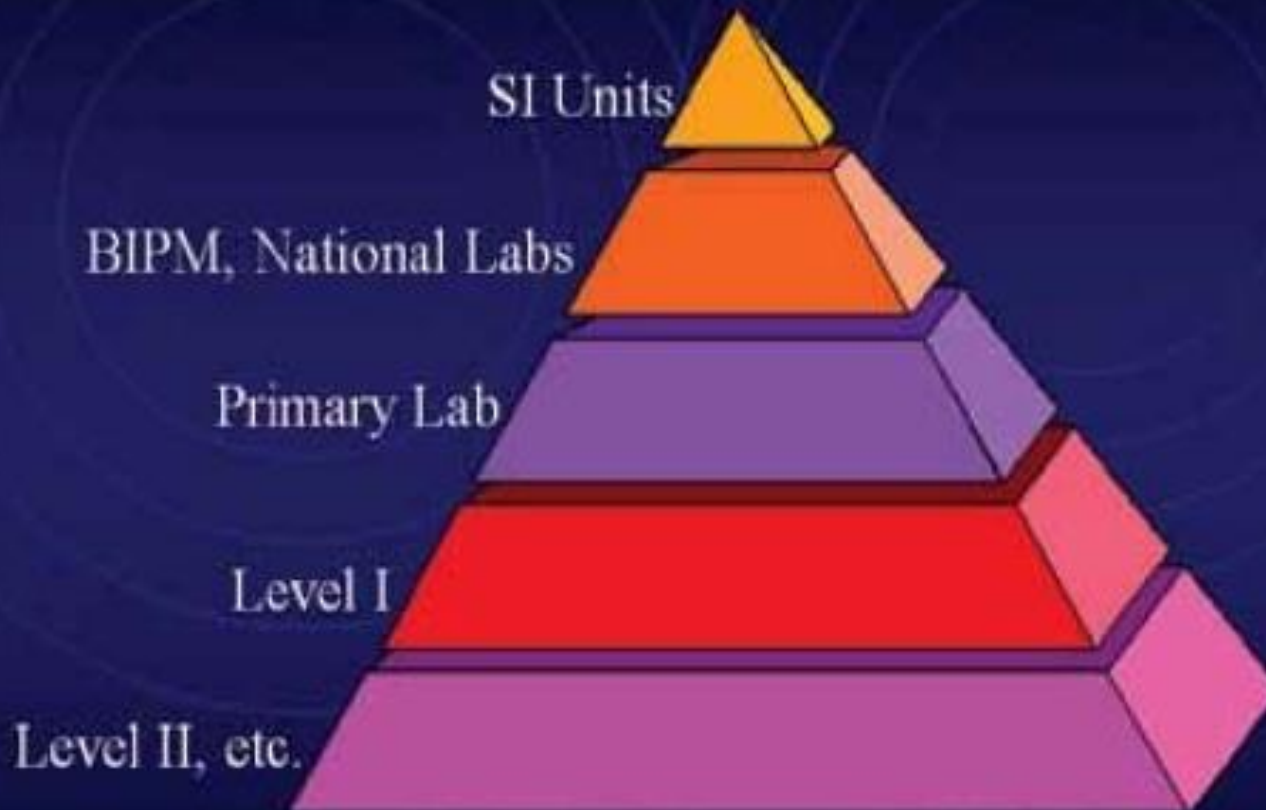
Therefore

Server	U'_{s_i}
<i>Server 1</i>	0.8274
<i>Server 2</i>	0.4947
<i>Server 3</i>	0.5743
<i>Server 4</i>	1
<i>Server 5</i>	0.5378

Sincronização de Relógios de Computadores

Sincronização de Relógios de Computadores

Typical Calibration Hierarchy



Sincronização de Relógios de Computadores

Sistema Internanacional de medidas (SI)

Meter

Original (1793): 1/10000000 of the meridian through Paris between the North Pole and the Equator. A set of 30 prototypes of the meter prototypes made of a 90% platinum-10% iridium

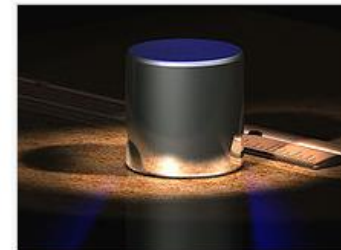


Metre
Historical International Prototype Metre bar, made of an alloy of platinum and iridium, that was the standard from 1889 to 1960.

Present definition: the meter is the length of the path travelled by light in vacuum during a time interval of 1/299,792,458 of a second.

kilogram

The **kilogram (kg)** is the base unit of mass in the International System of Units (SI) and is defined as being equal to the mass of the *International Prototype of the Kilogram (IPK)*. A set of 40 prototypes of the kilogram made of a 90% platinum-10% iridium.



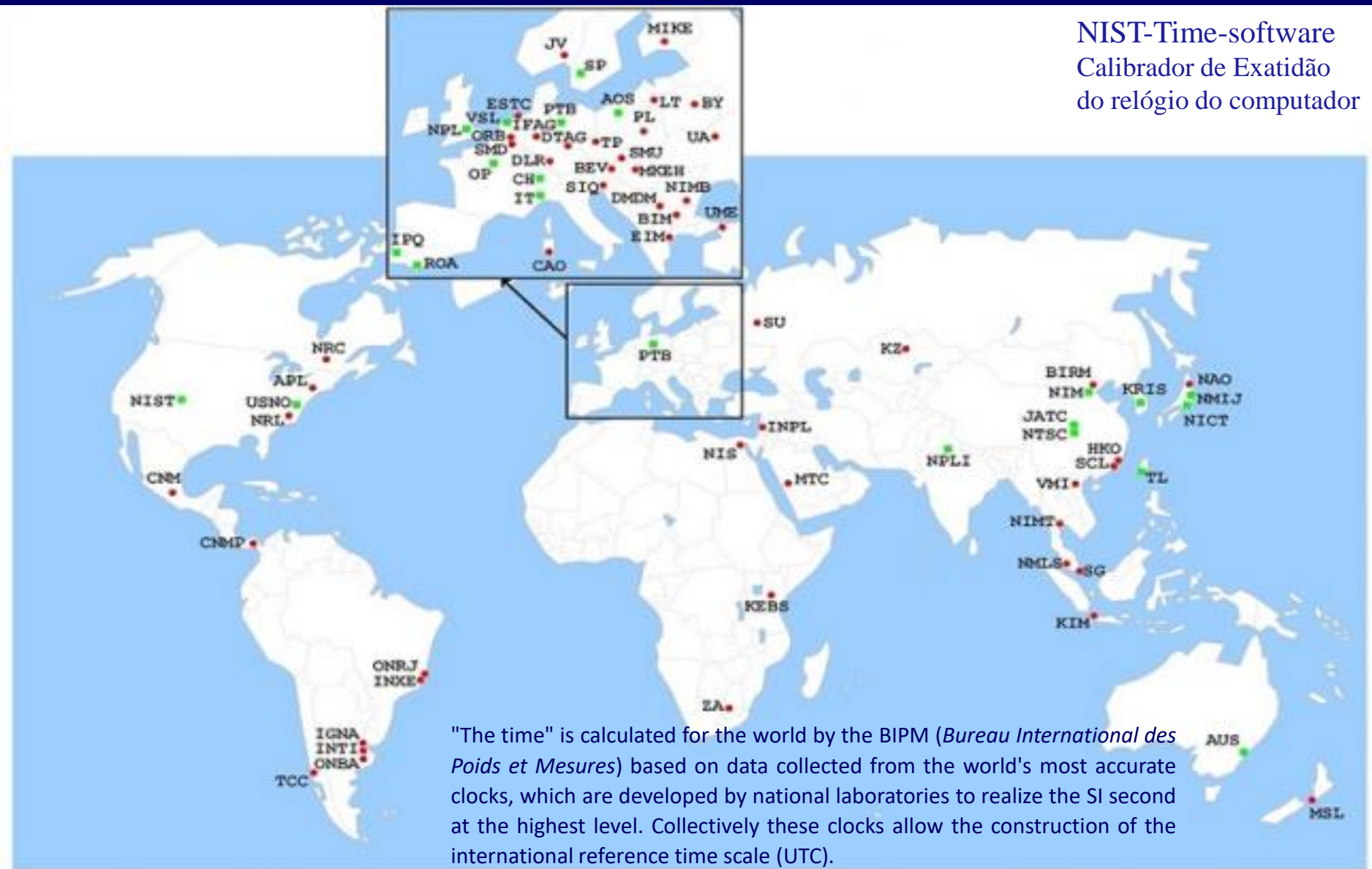
A CGI of the international prototype kilogram (the inch ruler is for scale). The prototype is manufactured from a platinum-iridium alloy and is 39.17 mm in both diameter and height, its edges have a four-angle (22.5°, 45°, 67.5° and 79°) chamfer to minimize wear.

Second

Current (1967): The duration of 9192631770 periods of the radiation corresponding to the transition between the two hyperfine levels of the ground state of the caesium 133 atom.

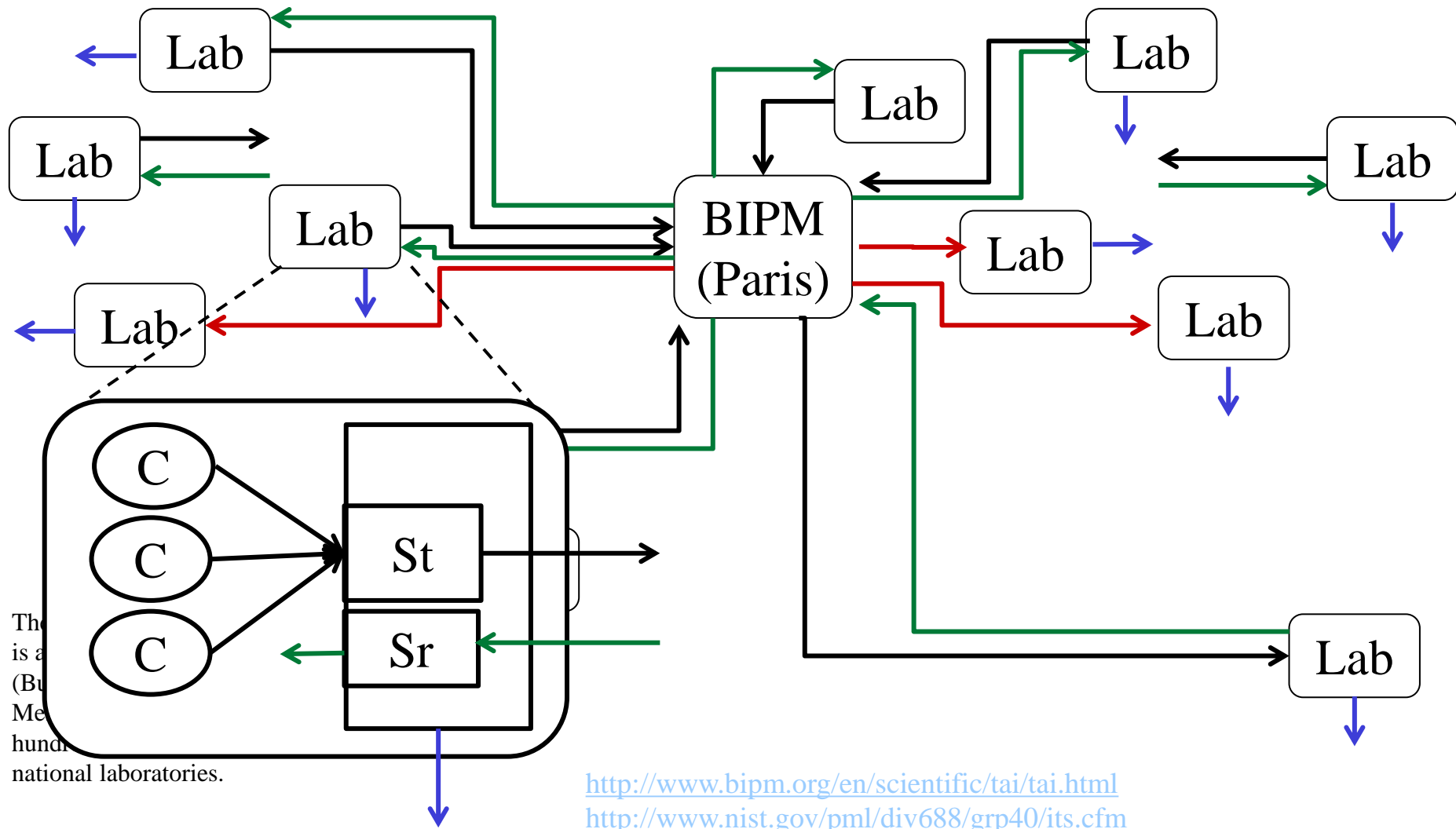
Sincronização de Relógios de Computadores

NIST-Time-software
Calibrador de Exatidão
do relógio do computador



"The time" is calculated for the world by the BIPM (*Bureau International des Poids et Mesures*) based on data collected from the world's most accurate clocks, which are developed by national laboratories to realize the SI second at the highest level. Collectively these clocks allow the construction of the international reference time scale (UTC).

Sincronização de Relógios de Computadores



Estratégias de Medição

Algumas métricas

- Contagem
- Duração de evento
- Instante de ocorrência de evento
- Tamanho
- Frequência de clock
- MIPS
- MFLPOS
- Vazão (*throughput*)
- Utilização
- Aceleração (*speed up*)
- Probabilidade de descarte
- ...

Estratégias para medição

Baseadas em detecção de evento

Baseadas em amostragem

Direta

Indireta

Métodos baseados em detecção de eventos

Medição direta

Usado quando a métrica desejada é coletada através da monitoração de um evento diretamente observável.

Medição indireta

Usada quando a métrica desejada não pode ser obtida diretamente,

Mede-se algo diretamente e deriva-se ou se deduz a métrica desejada,

Depende da habilidade e criatividade do avaliador

Métodos baseados em detecção de eventos

Eventos

□ **Eventos** são mudanças predefinidas que ocorrem no estado do sistema.

- Referência a memória,
- Acesso a disco,
- Mudança do valor do registrador de estado,
- Mensagem na rede,
- Interrupção do processador
- Uso de um recurso
- Tempo de utilização de um recurso
- ...

Medição Indireta

□ Métricas baseadas em *eventos secundários*

- Registra-se um valor quando sempre que ocorre um evento,
- Registra o tamanho do bloco para cada operação de I/O,
- Contagem do número de operações,
- Encontra-se a tamanho médio da informações de I/O transferida.

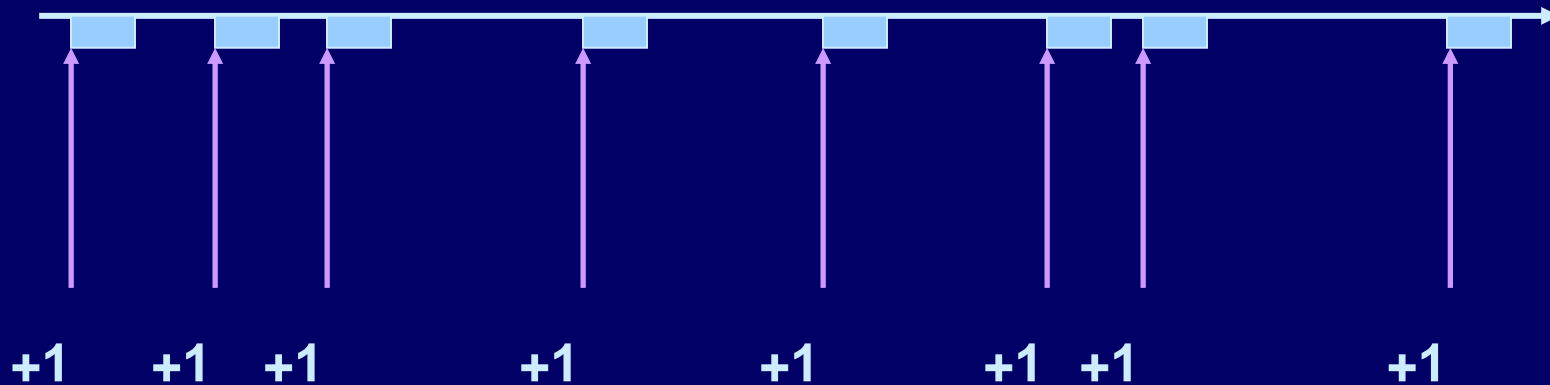
Ferramentas Básicas

- Contador

- Temporizador (*timer*)

Contador

- Conta o número de ocorrência de eventos.

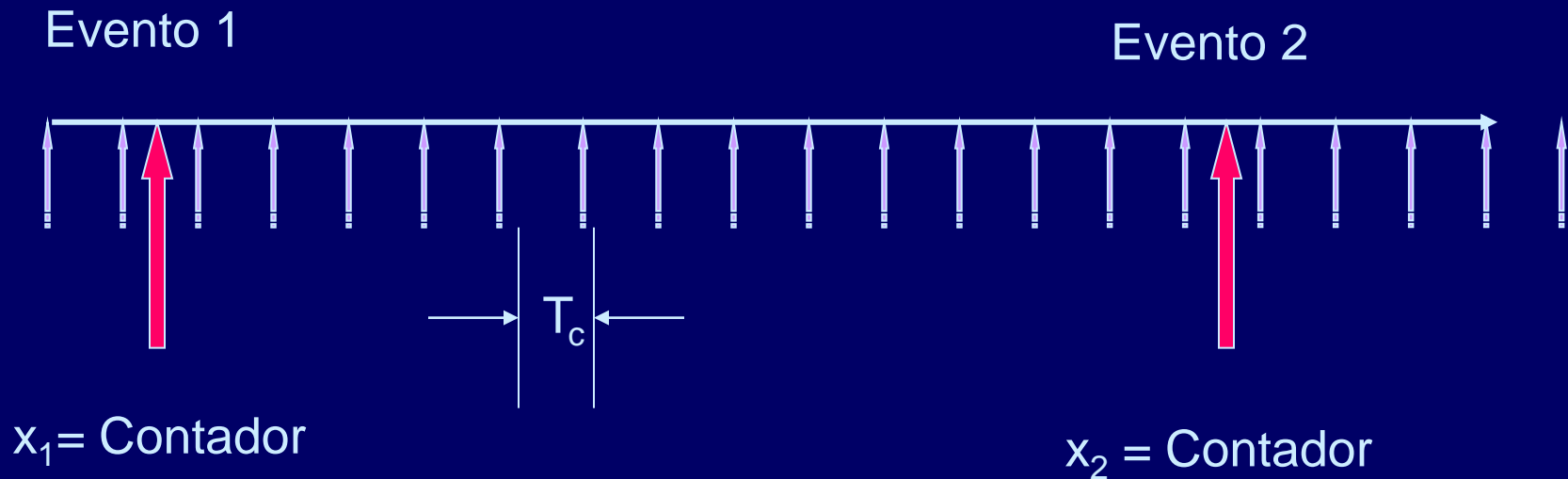


Conta exatamente 8 eventos

Timers

- Temporizadores (Timer)

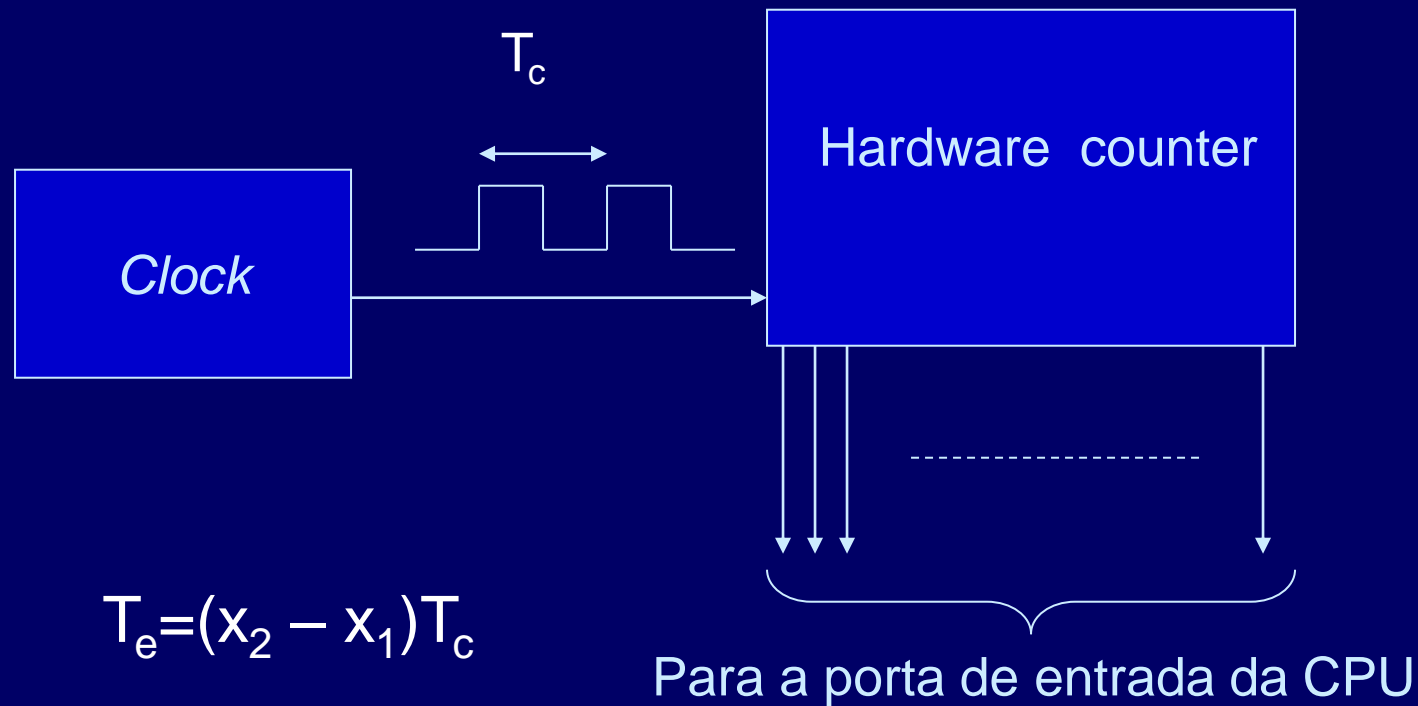
- Conta o número de pulsos entre dois eventos.



$$T_e = (x_2 - x_1) T_c$$

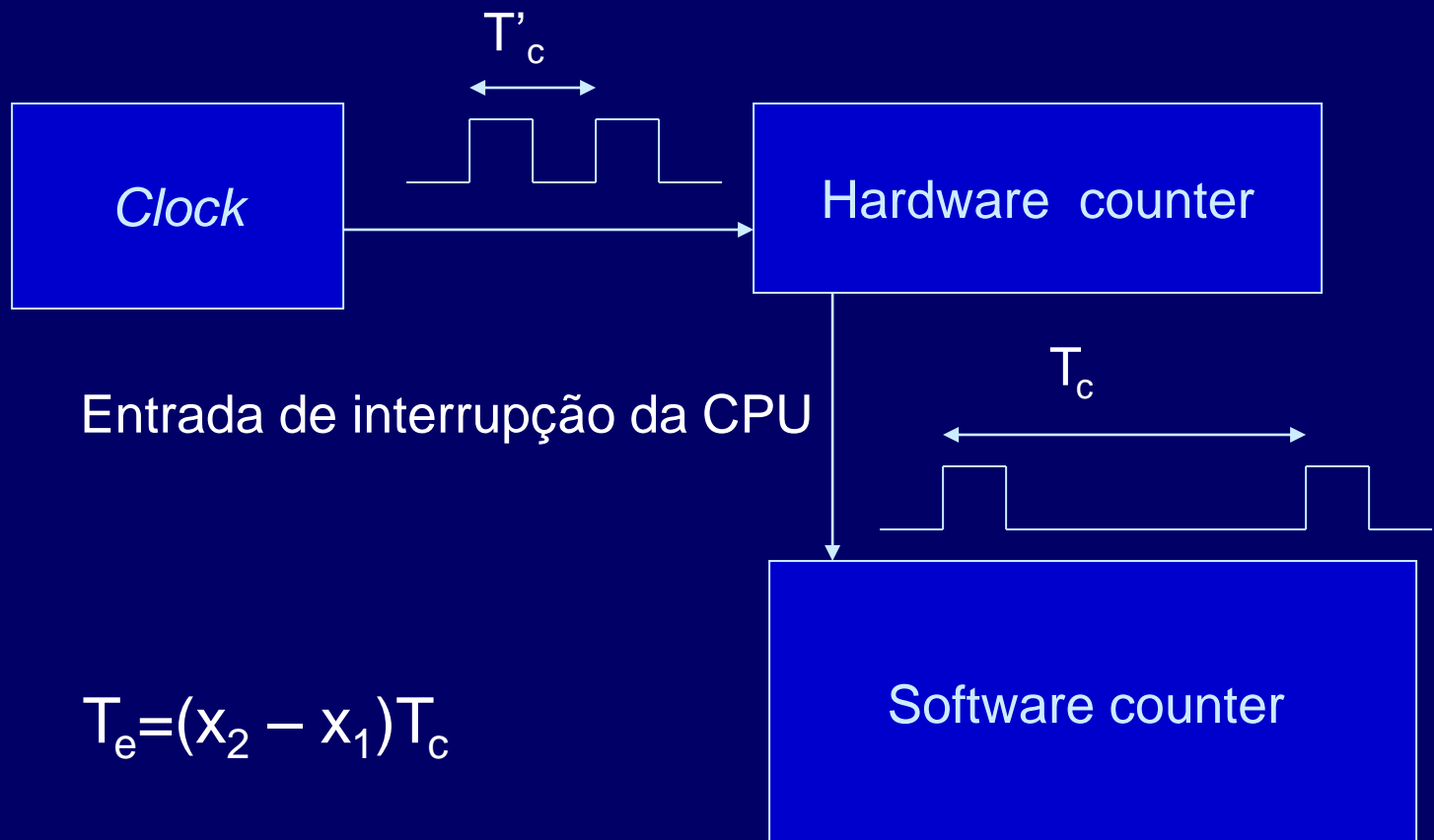
Timers

■ Hardware Timer



Timers

■ Hardware/Software Timer system



Obs.: fale sobre a Interrupção INT 08h, INT 1Ch (processadores X86).

Timers

□ SoftwareTimer

```
Start_count = read_timer();
```

porção do programa a ser medido

```
Stop_count = read_timer();
```

```
Elapsed_time = (stop_count – start_count)*  
clock_period;
```

Timers

```
import time
import pstats
#-----
def fast():
    """
    start = time.time()
    print("I run fast!")
    end = time.time()
    print("fast() function takes", end-start, "seconds")

#-----
def slow():
    """
    start = time.time()
    time.sleep(3)
    print("I run slow!")
    end = time.time()
    print("slow() function takes", end-start, "seconds")
```

```
#-----
def medium():
    """
    start = time.time()
    time.sleep(0.5)
    print("I run a little slowly...")
    end = time.time()
    print("medium() function takes", end-start, "seconds")
#-----
fast()
slow()
medium()
```

```
I run fast!
fast() function takes 0.00016307830810546875 seconds
I run slow!
slow() function takes 3.0038845539093018 seconds
I run a little slowly..
medium() function takes 0.5014247894287109 seconds
```

Timers

```
#include <stdio.h>
#include <time.h> // for time()
#include <unistd.h> // for sleep()

int main()
{
    time_t begin = time(NULL);

    // do some stuff here
    sleep(3);

    time_t end = time(NULL);

    printf("Time elapsed is %ld seconds \n", (end - begin));

    return 0;
}
```

Time elapsed is 3 seconds

Timers

```
#include <stdio.h>
#include <time.h>
```

```
// A function that terminates when enter key is
pressed
```

```
void fun()
{
    printf("fun() starts \n");
    printf("Press enter to stop fun \n");
    while(1)
    {
        if (getchar())
            break;
    }
    printf("fun() ends \n");
}
```

```
// The main program calls fun() and measures time taken
by fun()
```

```
int main()
{
    // Calculate the time taken by fun()
    time_t begin;
    time_t end;
    time_t t;

    begin = time(NULL);
    fun();
    end = time(NULL);
    t = end - begin;

    printf("fun() took %ld seconds to execute \n", t);
    return 0;
}
```

fun() starts
Press enter to stop fun

fun() ends
fun() took 7 seconds to execute

Restrições

- *Timer Rollover* (provoca erro relacionado à resolução)
 - Contador de n-bit
 - *contagem* = $[0, 2^n - 1]$
 - *Rollover* \equiv transição de $(2^n - 1) \rightarrow 0$
 - Se ocorre *rollover* entre eventos *start/stop*
 - então *contador* = $(x_2 - x_1) < 0$
 - Verifique se *contador* < 0
 - Medir outra vez
 - Some 2^n a contagem

Restrições

- *Overhead* no sistema
 - Apenas quando ocorre o evento de interesse,
 - Eventos pouco frequentes → **pouca perturbação**
 - Eventos muito frequentes → **forte perturbação.**
- O comportamento do sistema continua sendo típico?
 - A perturbação altera o sistema em medição

Timer Overhead

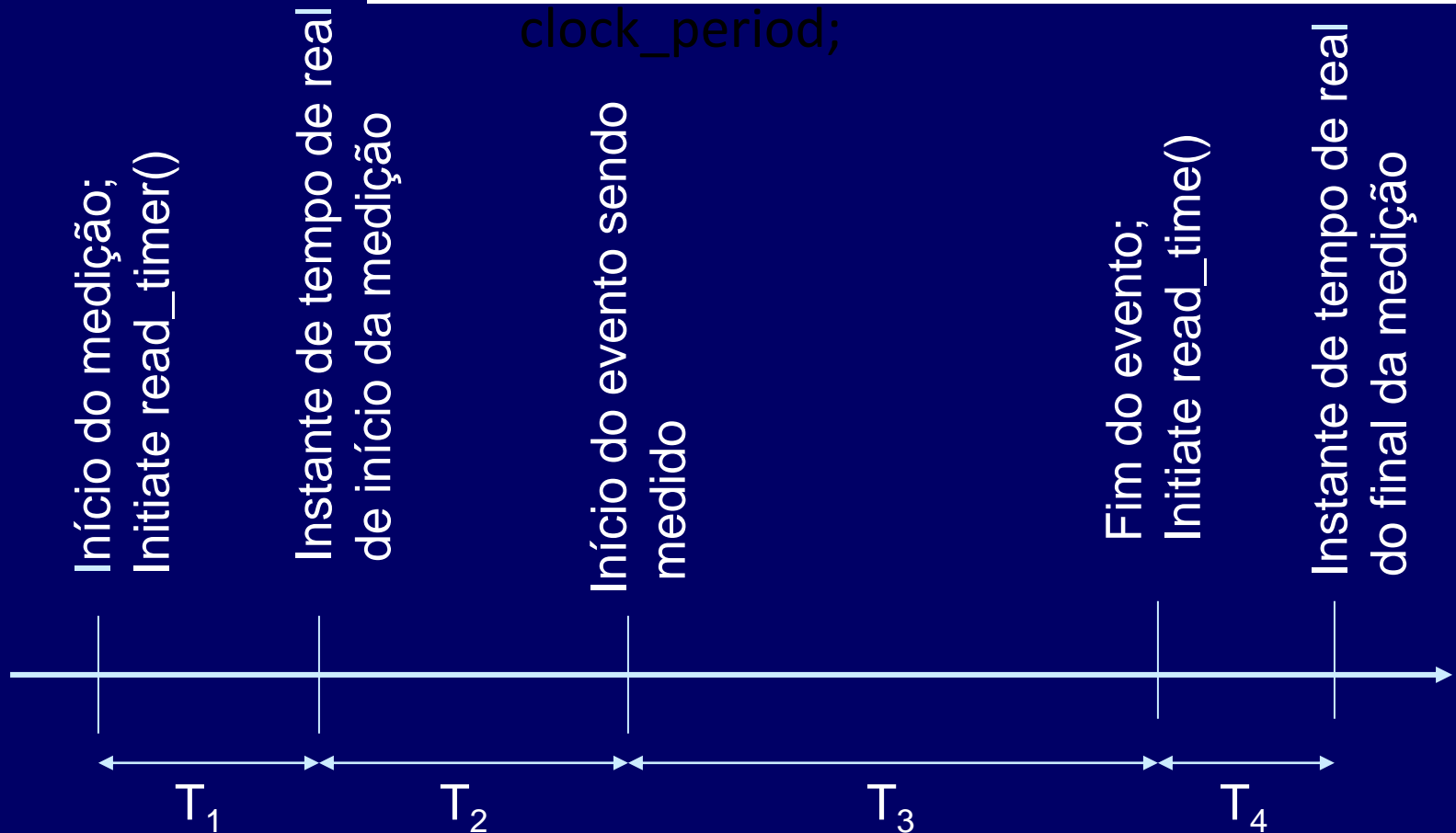
```
start_count = read_timer();
```

porção do programa a ser medido

```
stop_count = read_timer();
```

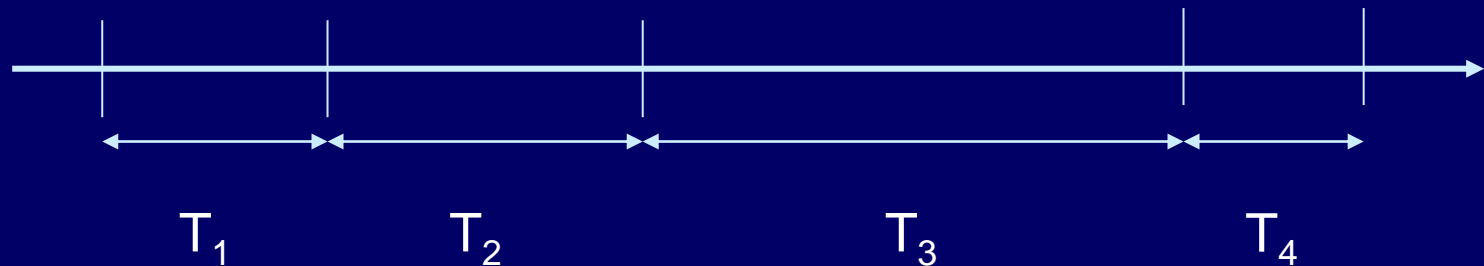
```
elapsed_time = (stop_count - start_count)*
```

```
clock_period;
```



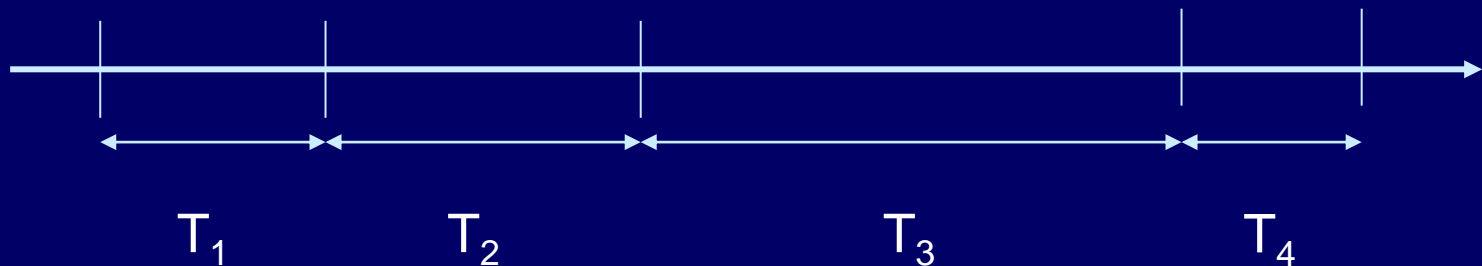
Timer Overhead

- T_1 = tempo para ler o contador,
- T_2 = tempo para armazenar o valor do contador,
- T_3 = intervalo de tempo do evento em medição,
- T_4 = tempo para ler o contador
 - $T_4 = T_1$



Timer Overhead

- T_e = intervalo de tempo do evento = T_3
- No entanto, o tempo medido é:
 - $T_m = T_2 + T_3 + T_4$
- $T_e = T_m - (T_2 + T_4) = T_m - (T_2 + T_1)$
- *Timer overhead* = $T_{ovhd} = (T_1 + T_2)$



Timer Overhead

- Se $T_e \gg T_{ovhd}$
 - Ignore o *overhead*
- Se $T_e \approx T_{ovhd}$
 - Medição será altamente suspeita
- T_{ovhd} pode variar substancialmente,
- *Good rule of thumb*
 - T_e deve ser $100-1000x > T_{ovhd}$

Sobre a medição baseada na detecção de eventos

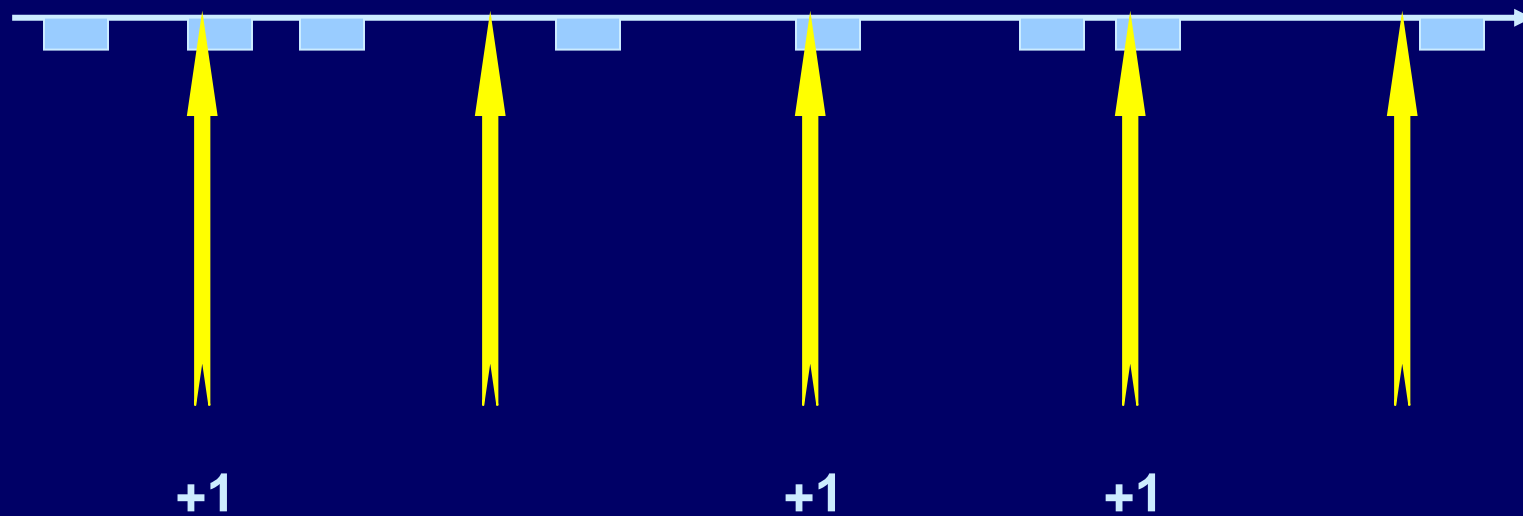
- ❑ Depende de quando o evento ocorre,
- ❑ Pode não ser fácil se estimar perturbações,
- ❑ Por quanto tempo medir?
- ❑ Pode ser uma boa alternativa quando os eventos inseridos têm baixa frequência de execução.

Métodos baseados em Amostragem

Características dos métodos baseados em Amostragem

- Registra o estado necessário em intervalos de tempo
- *Overhead*
 - Independente da frequência específica do evento,
 - Depende da frequência de amostragem
- Perde alguns eventos
- Produz resumo estatístico
 - Pode perder eventos raros,
 - Cada replicação produzirá resultados diferentes.

Amostragem



□ Conta 3 eventos em 5 amostras

Comparação

	Contagem de eventos	Amostragem
Resolução	Cont. exata	Resumo estatístico
<i>Overhead e</i> Perturbação	\sim #eventos	Constante

Contagem de eventos:

Melhor para eventos de baixa frequência,
Necessário quando uma contagem exata é exigida.

Amostragem:

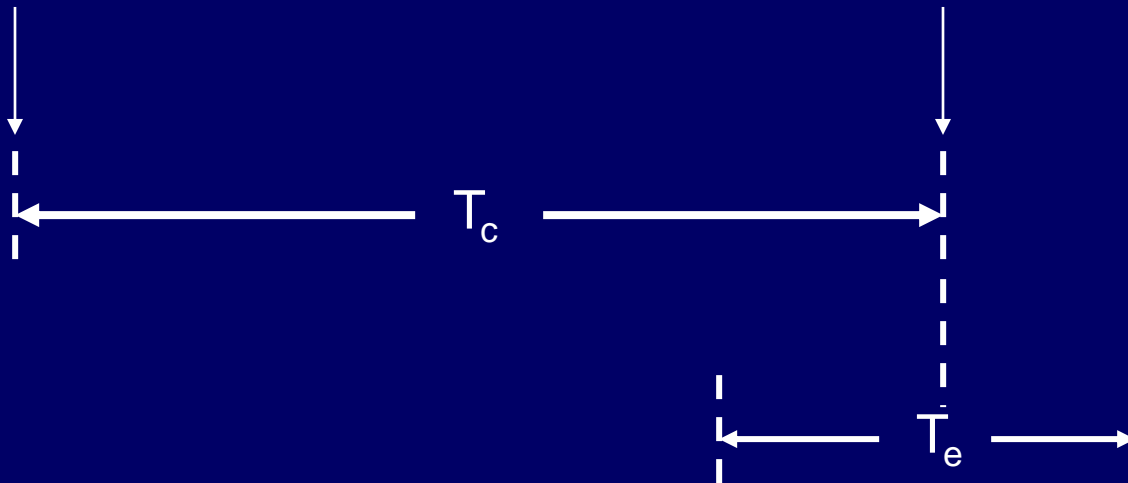
Melhor alternativa quando a frequência dos eventos é alta,
Se um resumo estatístico é adequado.

Medidas Aproximadas de Intervalos de Tempo Curtos

Medidas Aproximadas de Intervalos de Tempo Curtos

- ❑ Como medir um evento, cuja duração é menor que a resolução do ferramental de medição?
- ❑ Não é possível medir diretamente um evento cuja duração é menor que o tempo de resolução ($T_e < T_c$).
- ❑ Mesmo quando $T_e > nT_c$, mas o n é pequeno, o *overhead* torna difícil medir esses eventos.

Medidas Aproximadas de Intervalos de Tempo Curtos



Caso 1:
Contador+1



Caso 2:
Contador+0

Medidas Aproximadas de Intervalos de Tempo Curtos



Math

Binomial

Binomial
Excel

Applet

❑ Experimento de Bernoulli

- Resultado = +1 com probabilidade p
- Resultado = +0 com probabilidade $(1-p)$
- Equivalente ao lançamento de uma moeda (com viés – se $p \neq 0,5$),

❑ Repita n vezes

- Aproxima-se de uma distribuição binomial
- Apenas aproxima, pois não há garantia de que cada medição seja independente.
 - ❑ Na prática, normalmente é próximo.

Medidas Aproximadas de Intervalos de Tempo Curtos

- m = número de ocorrência do Caso 1 Contador+1
- n = Número total de medidas,
- A proporção média é a razão m/n
- Use intervalo de confiança para proporção.

$$T_e = \frac{m}{n} T_c$$



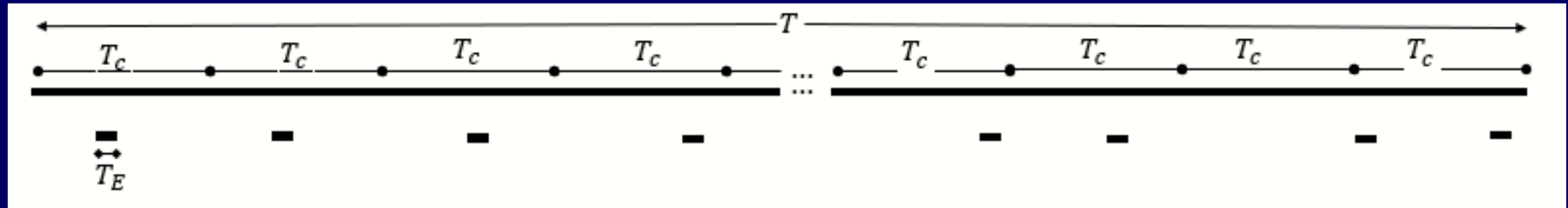
Math

Binomial

Binomial
Excel

Applet

Medidas Aproximadas de Intervalos de Tempo Curtos



$$T_c = 100 \mu s$$

$$T = n \times T_c = 876\,400 \mu s$$

$$n = 8764$$

$$m = 467$$

$$p = \frac{m}{n} = \frac{467}{8764} = 0.053286$$

$$T_c \times p = 100 \mu s \times 0.053286 = 5.3286 \mu s$$

Medidas Aproximadas de Intervalos de Tempo Curtos



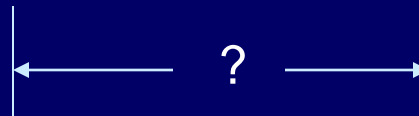
Math

Binomial

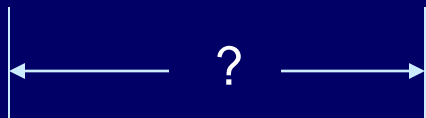
Binomial
Excel

Applet

- Resolução do Clock = 100 us
- $n = 8764$ medições
- $m = 467$ ticks de clock ticks contados
- 95% confidence interval



Caso 1:
467



Caso 2:
8297

Medidas Aproximadas de Intervalos de Tempo Curtos



Math

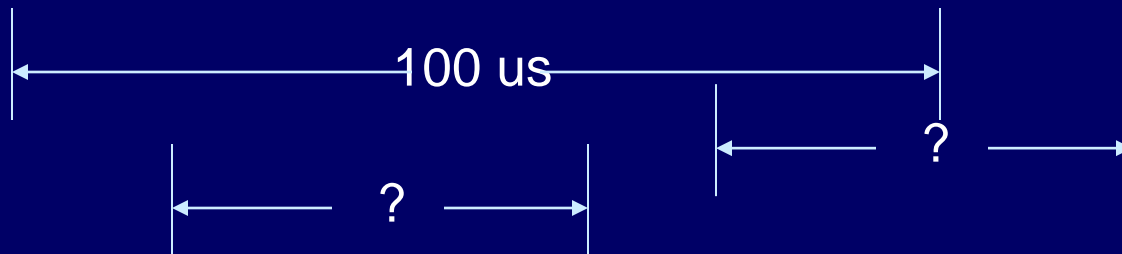
Binomial

Binomial
Excel

Applet

- $T = 876400 \text{ us}$
- $T_c = 100 \text{ us}$
- $n = 8764$ medições
- $m = 467$ ticks de clock ticks contados
- 95% confidence interval

$$T_e = \frac{m}{n} T_c$$

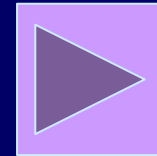


Caso 1:
467

Caso 2:
8297

$$T_e = \frac{m}{n} \times T_c = \frac{467}{8764} \times 100 \mu s = 5,32$$

Exemplo



Use Excel,
then use

Minitab

Applet

$$(c_1, c_2) = \frac{467}{8764} \mp 1.96 \sqrt{\frac{\frac{467}{8764} \left(1 - \frac{467}{8764}\right)}{8764}}$$
$$= (0.0486, 0.0580)$$

- Escalar pelo período de clock = 100 us
- Probabilidade de 95% do tempo do evento estar no intervalo (4.87, 5.82).

Test and CI for One Proportion

Sample	X	N	Sample p	95% CI
1	467	8764	0,053286	(0,048676; 0,058196)

← Obtido através do
MINITAB

Medição - Ferramentas

□ *Windows monitoring tools*

- Perfmon ([http://technet.microsoft.com/en-us/library/cc771692\(WS.10\).aspx](http://technet.microsoft.com/en-us/library/cc771692(WS.10).aspx))
- Powershell

□ *Linux measuring syscalls and commands*

- top, vmstat, procinfo, mpstat, sar,oprofile,meninfo,slabinfo,slabtop,vmstat,iostat,iptraf
- Sysstat :
- SystemTap

Alguns links:

- Lucifredi_Lecture_Feb07.pdf,
- http://www.volny.cz/linux_monitor/
- <http://syscalls.kernelgrok.com/>
- <http://sebastien.godard.pagesperso-orange.fr/documentation.html>

□ *Profiling tools*

- gprof (http://www.cs.utah.edu/dept/old/texinfo/as/gprof_toc.html)
- jprof (<http://perfinsp.sourceforge.net/jprof.html>)
- Valgrind (<http://valgrind.org/>)
- Tau (<http://www.cs.uoregon.edu/Research/tau/home.php>)

Geração de carga e medição com o Powershell

About Powershell Scripts

For running scripts (name notation: "name.ps1")

First:

Execute "powershell" in command line "as administrator".

After:

Set-ExecutionPolicy RemoteSigned

Then:

Example 1:

Execute "Measure-Command {echo hi}"

Geração de carga e medição com o Powershell

Other examples:

```
Measure-Command {echo dir}  
Measure-Command {echo cd}
```

Example 2:

```
Set-ExecutionPolicy RemoteSigned  
$i = 1  
for ($i=1; $i -le 100; $i++)  
{Measure-Command {echo dir} | Select-Object -Property  
TotalMilliseconds | convertto-csv >> output4.txt}
```

Geração de carga e medição com o Powershell

If you are benchmarking an .exe in the current directory, use this: `Measure-Command { .\your.exe }`.

Example 3:

```
Set-ExecutionPolicy RemoteSigned
$i = 1
for ($i=1; $i -le 100; $i++)
{Measure-Command {echo dhry1nnt.exe} | Select-Object -
Property TotalMilliseconds | convertto-csv >> output5.txt}
```

Look at:

<http://technet.microsoft.com/en-us/library/ee176949.aspx>

Exemplo 1

Script

Estimar o tempo de execução do comando "dir" padrão de um diretório especificado.

```
Set-ExecutionPolicy RemoteSigned
$i = 1
for ($i=1; $i -le 100; $i++)
{Measure-Command {echo dir} | Select-Object -Property
TotalMilliseconds | convertto-csv >> output3.txt}
```

Abrir os arquivos com o Excel

Usar o Excel, o Minitab e o Statdisk

Raw data
output3

Data analysis
output3

Exemplo 2

Estimar o tempo de execução do benchmark Dhrystone.

Script

```
Set-ExecutionPolicy RemoteSigned
$i = 1
for ($i=1; $i -le 100; $i++)
{Measure-Command {echo dhry1nnt.exe} | Select-Object -
Property TotalMilliseconds | convertto-csv >> output5.txt}
.
```

Abrir os arquivos com o Excel

Usar o Excel, o Minitab e o Stadisk

Benchmarks

Roy Longbottom's PC Benchmark Collections

<http://www.roylongbottom.org.uk/dhrystone%20results.htm>

CPU Utilization –
Benchmark Dhrystone –
10000 runs

Raw data
output5

Medição com Perfmon

Sobre o Perfmon

Criar o “User Defined” “Data Collector Set”.

Alterar as propriedades do respectivo “Performance Counter” para que o “log format” seja “comma separated”.

Clicar sobre o respectivo “User Defined” “Data Collector Set” e alterar suas propriedades definindo a “Stop Condition”.

Clicar sobre o respectivo “User Defined” “Data Collector Set” e “Run”.

About Perfmon

<http://searchitchannel.techtarget.com/feature/Using-Windows-7-performance-monitor-to-view-data>

Exemplo 3

Executar o benchmark Dhrystone 10000 vezes consecutivas através do script t4_ps.ps1. Estimar a utilização da CPU durante a execução do script.

Script

```
Set-ExecutionPolicy RemoteSigned
$i = 1
for ($i=1; $i -le 100; $i++)
{Measure-Command {echo dhrylnnt.exe} | Select-Object -
Property TotalMilliseconds | convertto-csv >> output5.txt}
```

Abrir os arquivos com o Excel

Usar o Excel, o Minitab e o Stadisk

About Perfmon

<http://searchitchannel.techtarget.com/feature/Using-Windows-7-performance-monitor-to-view-data>

CPU Utilization (perfmon) –
Benchmark Dhrystone –
10000 runs

Raw data
Output5 - script

Profiling

Profiling

□ Características básicas

- Caracterização do comportamento global,
- Fornece uma visão global da aplicação,
- Obtém-se o tempo de “permanência” em cada funcionalidade.

Profiling

- ❑ Fornece um visão global do tempo de execução da aplicação,
- ❑ Calcula-se a proporção do que se permanece em determinados estados em relação ao tempo total.
 - Fração de tempo em cada rotina,
 - Fração de tempo no núcleo do SO,
 - Fração do tempo em operações de I/O,
- ❑ Encontra-se gargalos e *hot-spots*
 - Otimize estas parte primeiro.

Profiling

- Normalmente utilizados para:
 - encontrar *mixes* de instruções utilizadas,
 - estatísticas de execução de funcionalidades,
 - estatísticas de uso de registradores,
 - estatísticas de desvios.
- Comumente aceitam:
 - um programa executável como entrada,
 - decodificam e analisam as instruções do executável.
- Adicionam código (*probes*) à aplicação a ser monitorada. alguns adicionam o código (*probes*) durante a compilação
- ou obtém amostras do contador de programa.

Profiling

Estratégias:

- *Basic Block counting*
- *Program Counter sampling*

Profiling

□ Contagem de Bloco Básico

– Bloco Básico

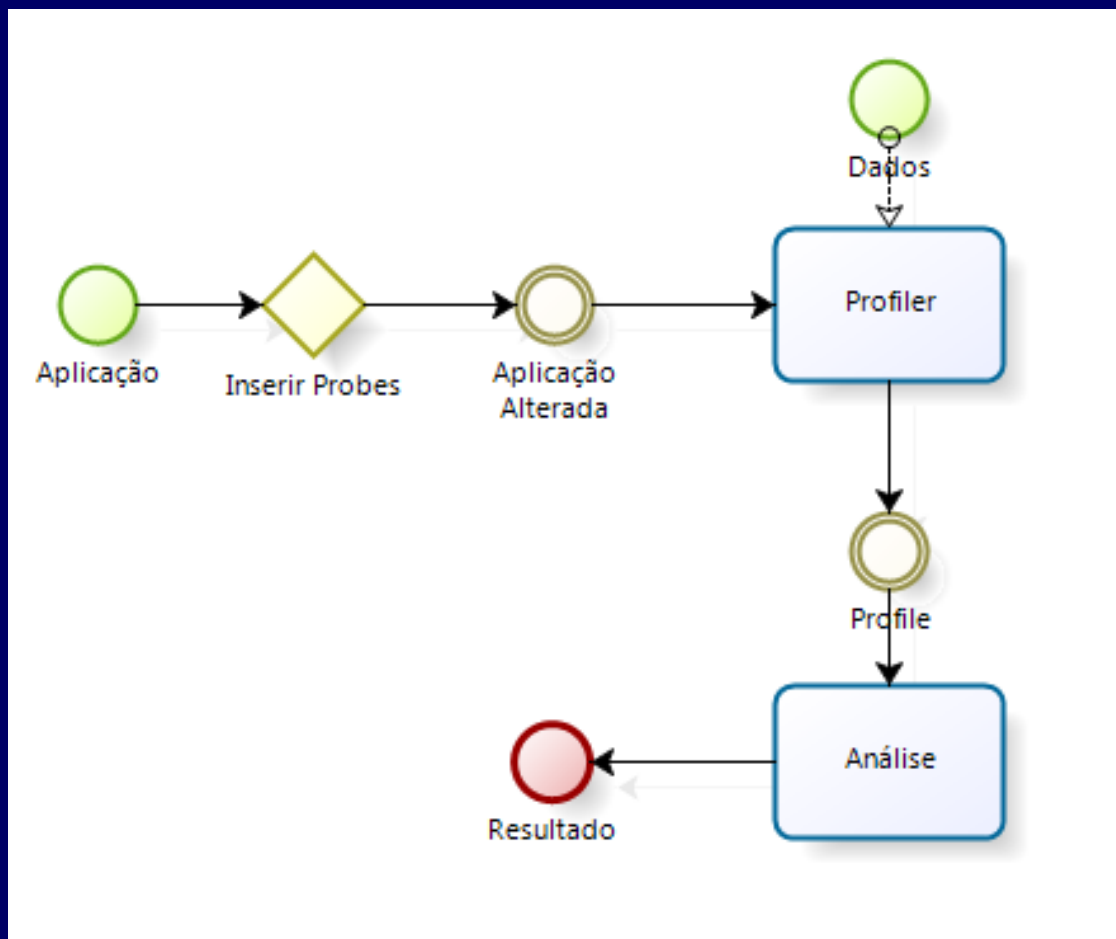
□ Sequência de instruções sem desvios de fluxo,

□ Quando a primeira instrução é executada, todas as demais do mesmo bloco serão executadas,

□ Entrada única e saída única.

Profiling

□ Contagem de Bloco Básico



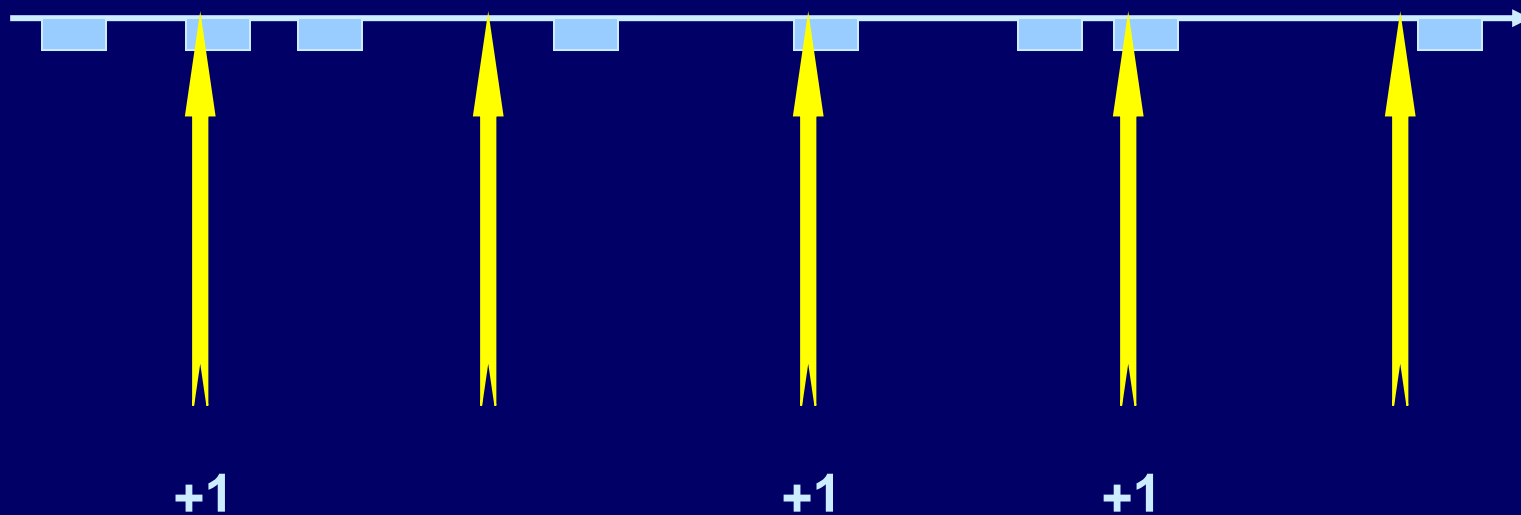
Profiling

□ Contagem de Bloco Básico

- Gere um programa *profile* inserindo instruções adicionais em cada bloco:
 - Incremente um único contador em todas as vezes que se entra no bloco.
- Gere um histograma da execução do programa.
- Pode realizar um pós-processamento para encontrar a frequência de execução das instruções (dos blocos).

Profiling

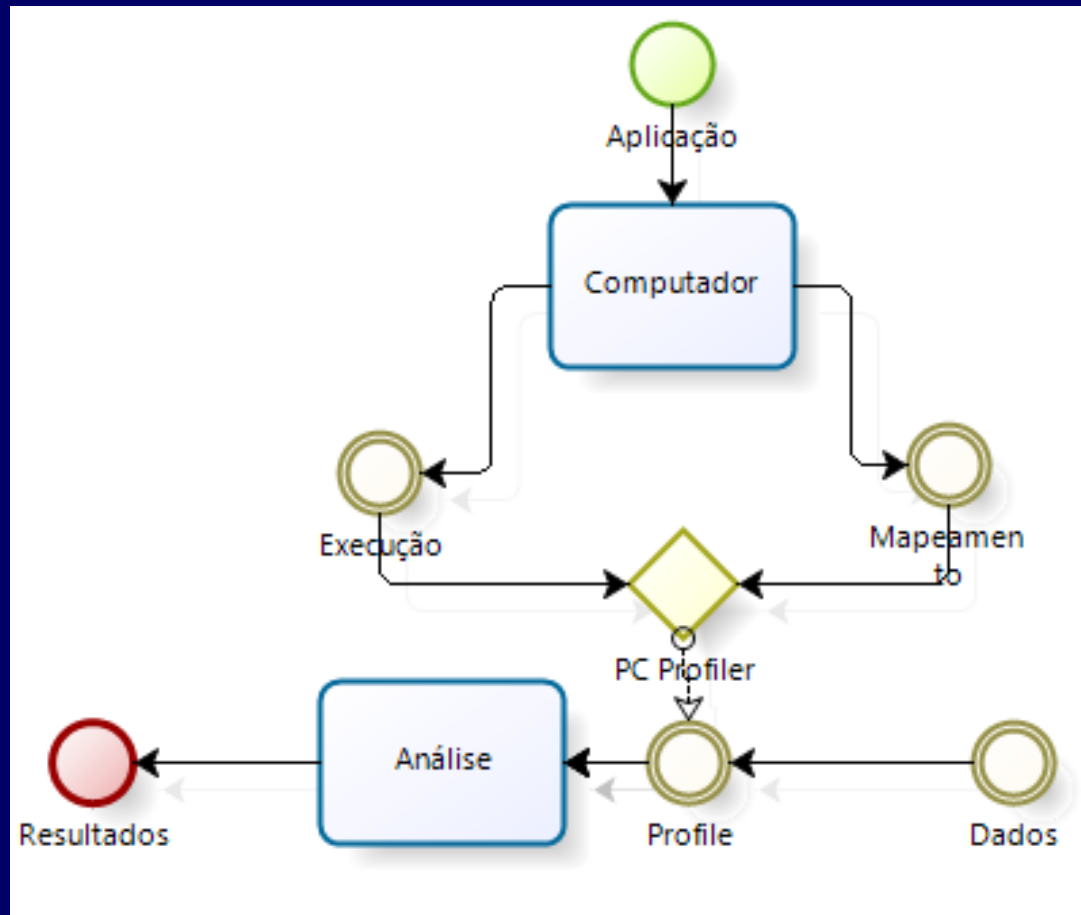
▪ Amostragem do Contador de Programa



- ❑ Periodicamente interrompe a aplicação em intervalos de tempo fixos,
- ❑ Registra-se a informação do estado no serviço de interrupção,
- ❑ Após a finalização, obtém-se um *profile* global

Profiling

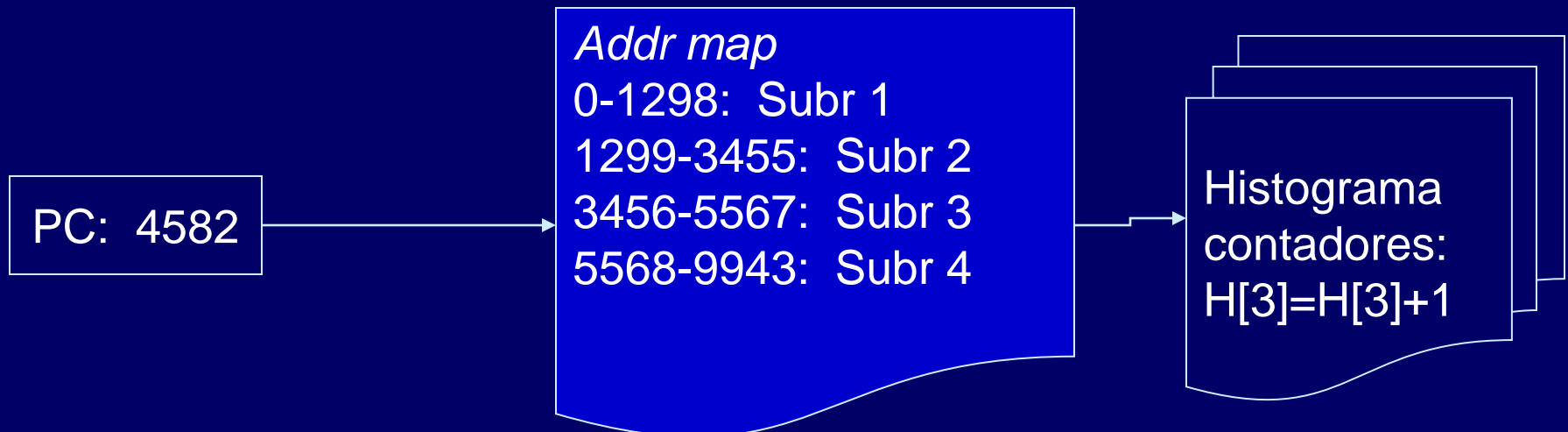
▪ Amostragem do Contador de Programa



Profiling

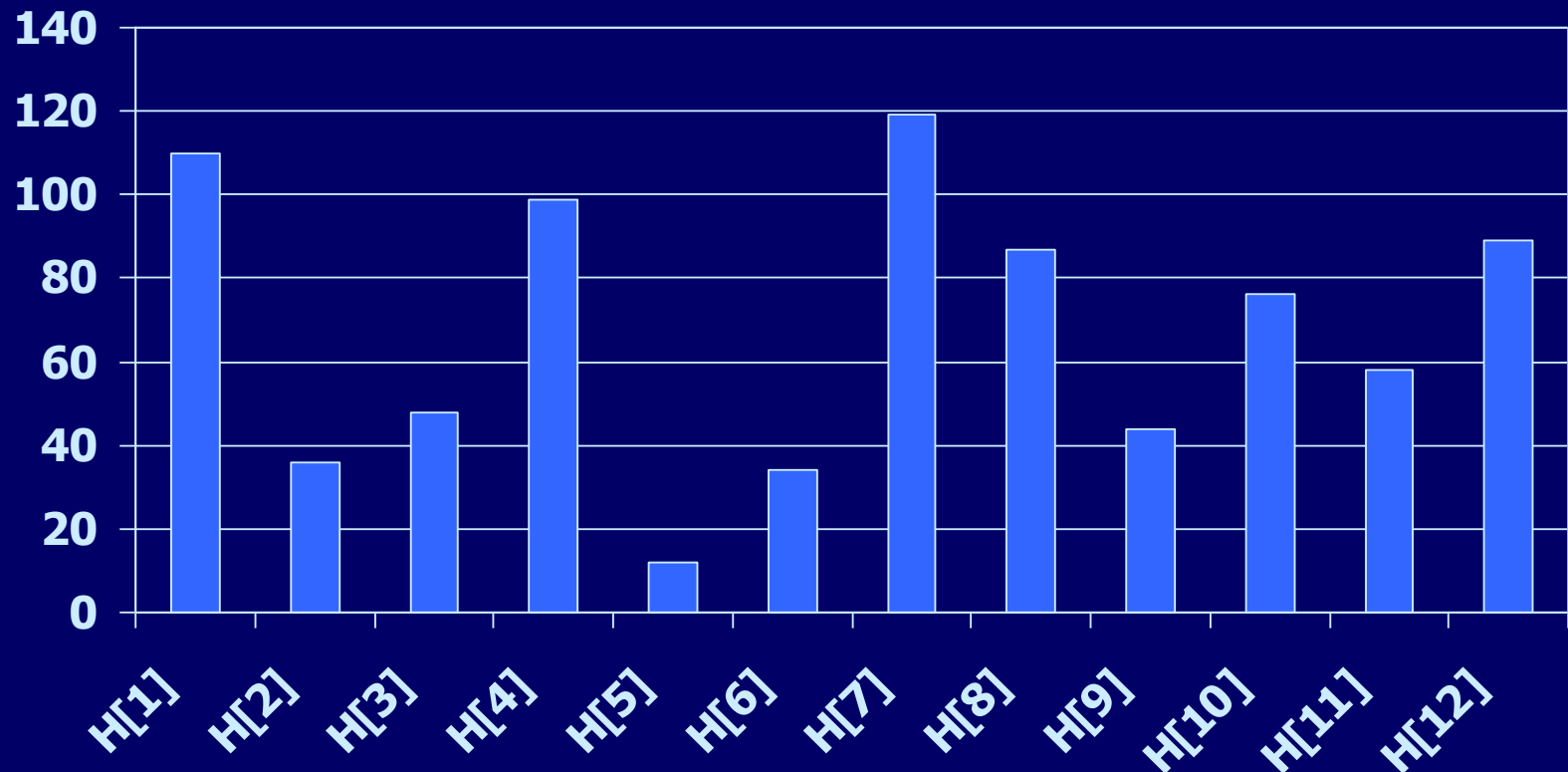
□ Em cada interrupção

- Examina-se o PC no endereço de retorno armazenado na pilha,
- Usa-se o mapa de endereço para se referenciar a sub-rotina *i*.
- Incrementa-se o elemento (*i*) do *array* $H[i]$



Profiling

▪ Amostragem do Contador de Programa



Profiling

Amostragem do Contador de Programa

- n = total de interrupções,
- Pós-processamento
 - $H[i]/n$ = proporção de execução na sub-rotina i
 - $(H[i]/n) \times (\text{Tempo de observação}) = \text{Tempo em cada sub-rotina } i$

Profiling

▪ Amostragem do Contador de Programa

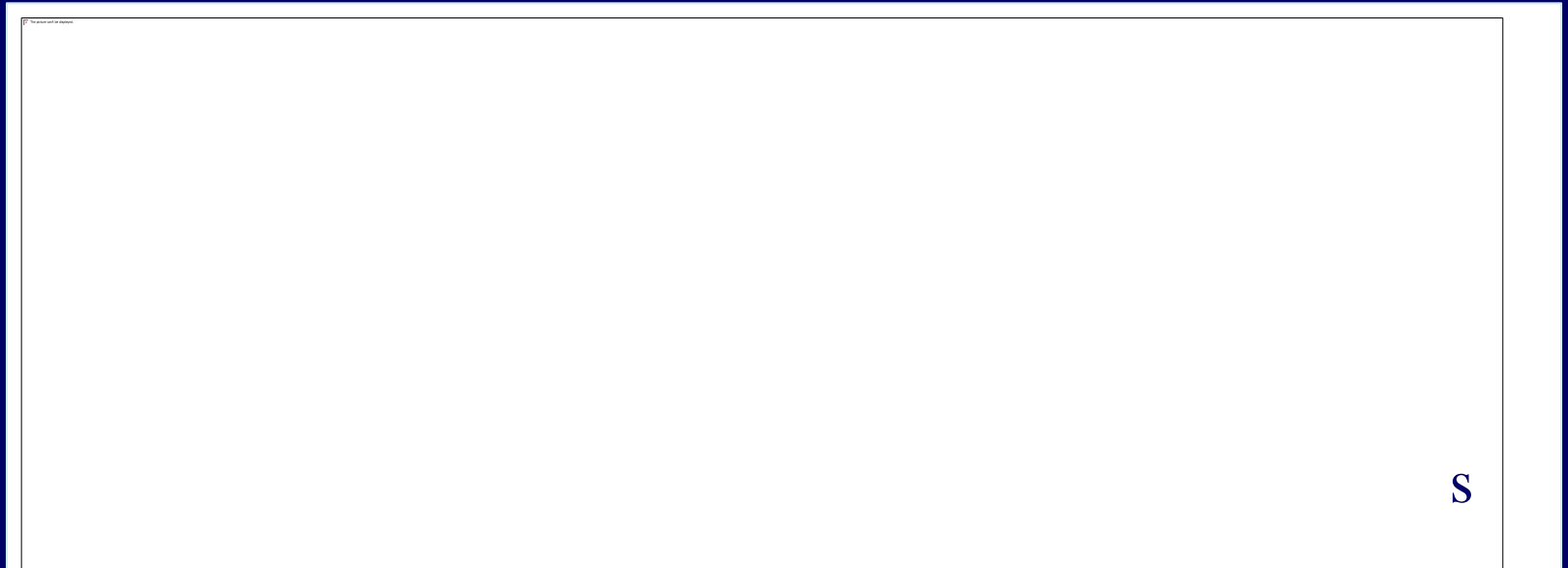
- Neste processo
 - Diferentes contagens são obtidas a cada vez que o experimento é realizado,
- Infere-se sobre o comportamento do programa a partir de uma amostra.
- Deve ser utilizado intervalo de confiança para quantificar a precisão dos resultados.

Profiling

▪ **Exemplo1** (Amostragem do Contador de Programa)

- Tempo entre interrupções = 10 ms
- $H[A] = 12$ interrupções na sub-rotina A,
- $n = 800$ amostras
 - Programa executou por 8 s
- Tempo na sub-rotina A?
 - Intervalo de confiança de 99%
- $p = H[A] / n = 12 / 800 = 0,015$

Profiling



S

- ❑ 99% de probabilidade de que o programa permaneça entre 0,39-2,61% (proporção) do seu tempo de execução na sub-rotina A
- ❑ Intervalo largo.
- ❑ No entanto, em menos de 3% do tempo de execução se esteve dentro da sub-rotina A,
- ❑ Inicie otimizando em outro local.

Profiling

▪ Exemplo 2

(Amostragem do Contador de Programa)

- Tempo entre interrupções = 40 us
- 36128 interrupções na sub-rotina A ($H[A]$)
- Programa executou por 10 s
- Qual o tempo da sub-rotina?
 - intervalo de confiança de 90%
- $H[A] = 36128$ (interrupções na sub-rotina A)
- $n = 10 \text{ s} / 40 \text{ us} = 250\,000$ (número de total de interrupções)
- $p = H[A] / n = 0,144$ (proporção de execução na sub-rotina A)

Profiling

$$\begin{aligned}(c_1, c_2) &= 0.144512 \mp 1.645 \sqrt{\frac{0.144512(0.855488)}{250000}} \\ &= (0.144, 0.146) \% \\ &= (10 \times 0.144, 10 \times 0.146) \\ &= (1.44, 1.46)\text{s}\end{aligned}$$

- 90% de probabilidade de que o programa permaneça entre 14.4-14.6% (proporção) do seu tempo de execução na sub-rotina A.
- O tempo que o processador ficou executando o programa está entre 1.44 e 1.46 s.

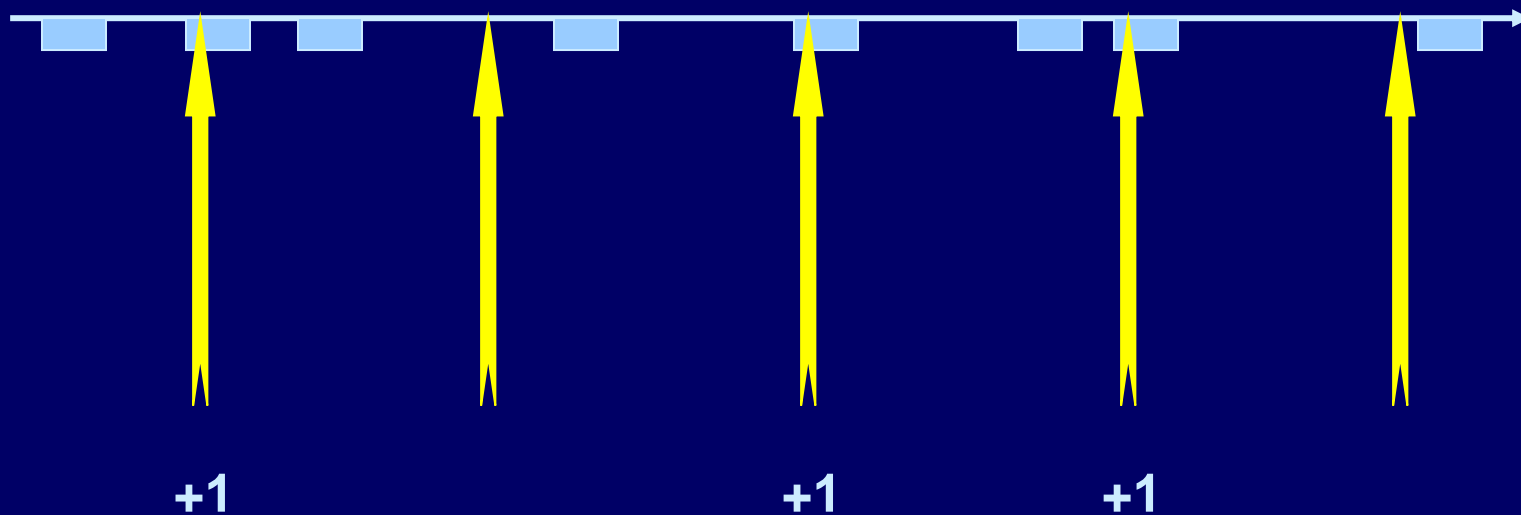
Profiling

■ Amostragem do Contador de Programa

- Redução do Tamanho do Intervalo
 - Adote um intervalo de confiança menor,
 - Colete mais amostras
- Execute o programa por mais tempo
 - Pode não ser possível.
- Aumente a taxa de amostragem
 - Pode ser fixada (limitada) pelo sistema ou ferramental.
 - Aumenta o *overhead* e/ou perturbação
- Execute múltiplas vezes e some as amostras de cada execução.

Profiling

▪ Amostragem do Contador de Programa



- ❑ Amostragem periódica × Amostragem aleatória
- ❑ Interrupções devem ocorrer assíncronamente para qualquer evento do sistema (evitar viés, correlação)
 - As amostras devem ser independentes entre si,
 - Caso contrário, a correlação entre os eventos pode ser forte.

Comparação

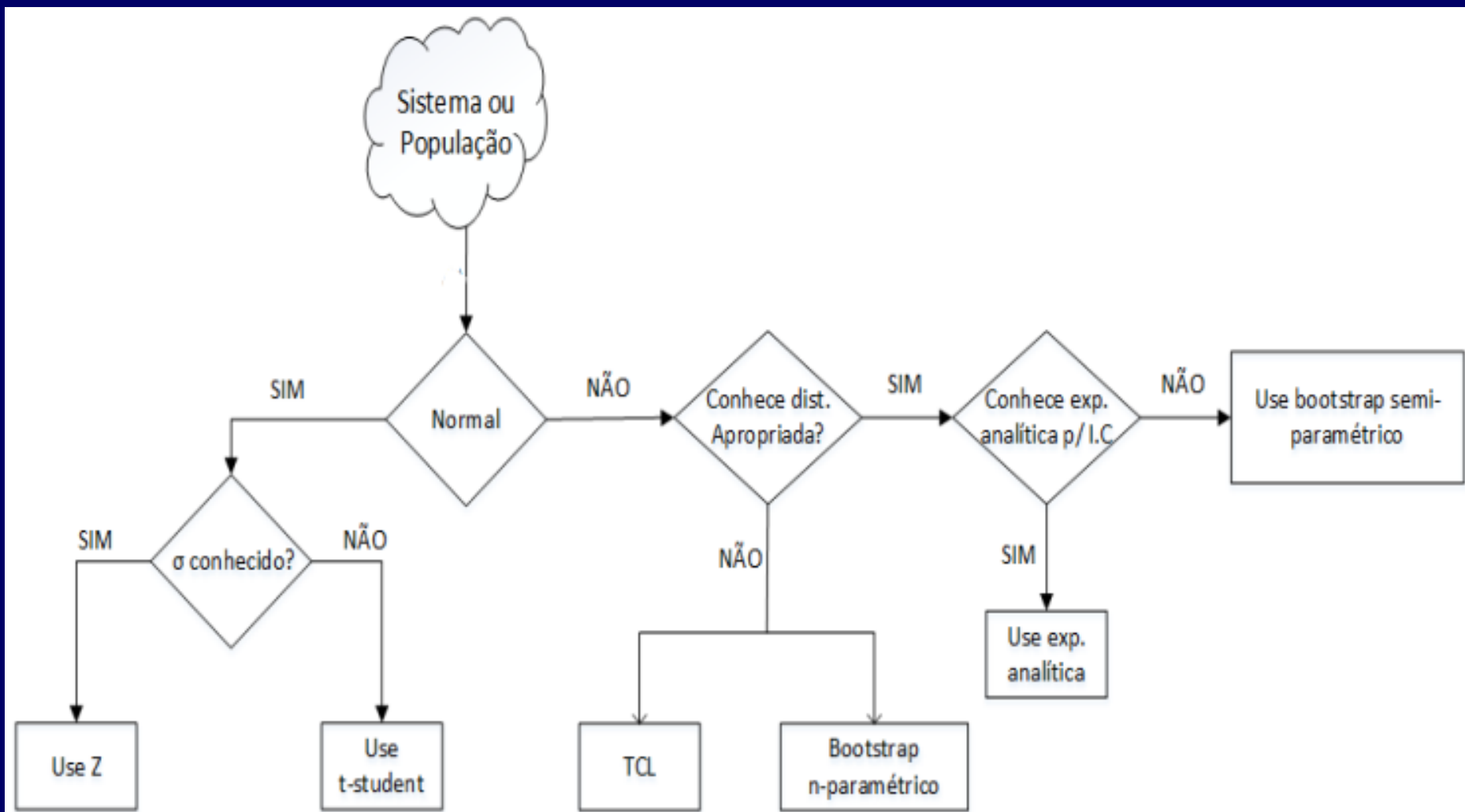
	Amostragem do Contador de Programa	Contagem de Bloco Básico
Saída	Estimativa estatística	Contagem exata
<i>Overhead</i>	Rotina de serviço de interrupção	Intruções extra em cada bloco
Perturbação	Aleatoriamente distribuída	Alta
Repetitibilidade	Dentro de variância estatística	Perfeita

Análise Exploratória de Dados e Inferência

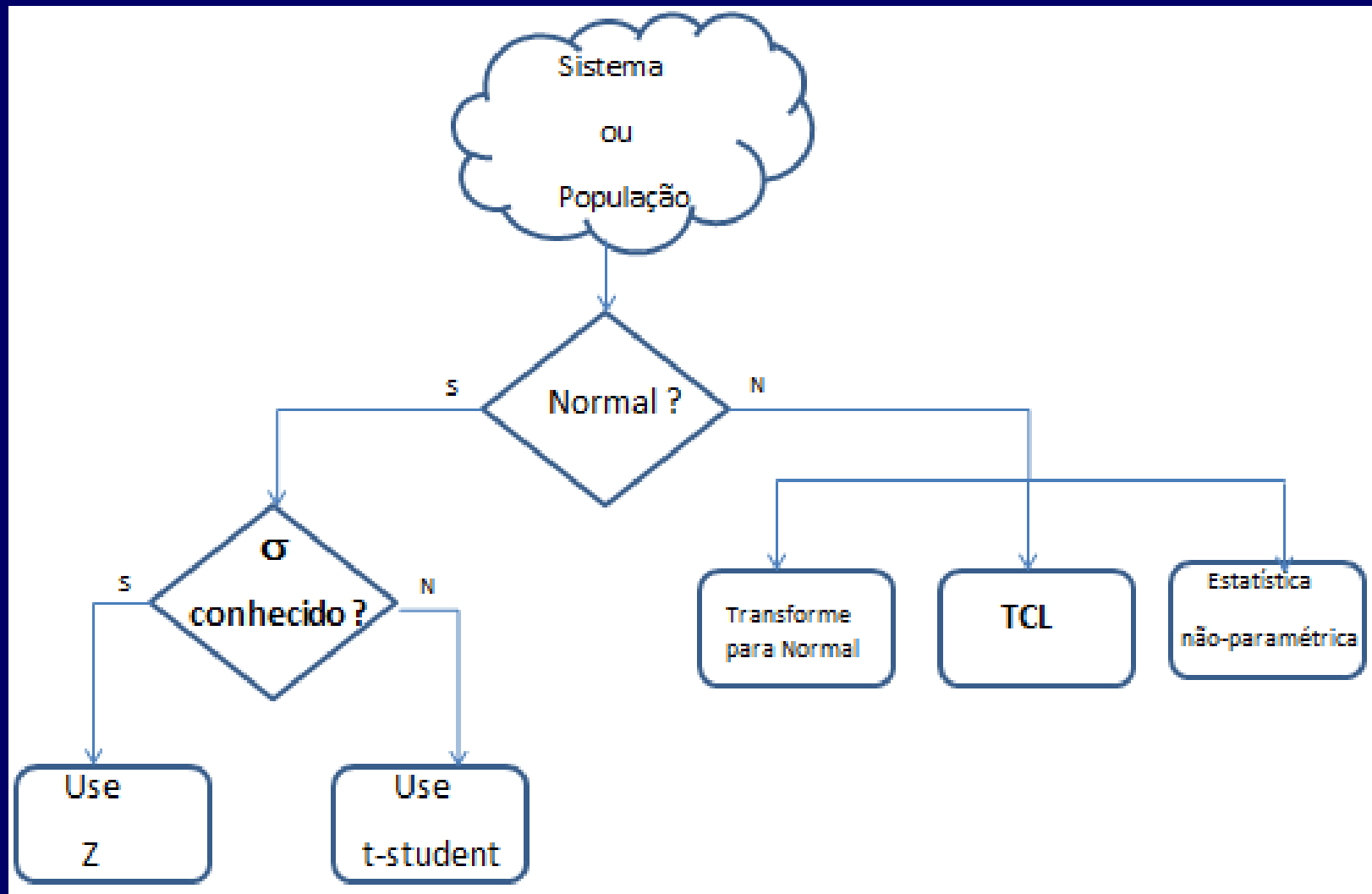
Orientação para Inferência

- ❑ Qual é o nível de mensuração dos dados?
 - (Classe ou Contínuo)
- ❑ Qual é o parâmetro relevante?
 - Média, variância/desvio-padrão, proporção.
- ❑ Há razões para supor que a população seja normalmente distribuída?
- ❑ O desvio populacional é conhecido?

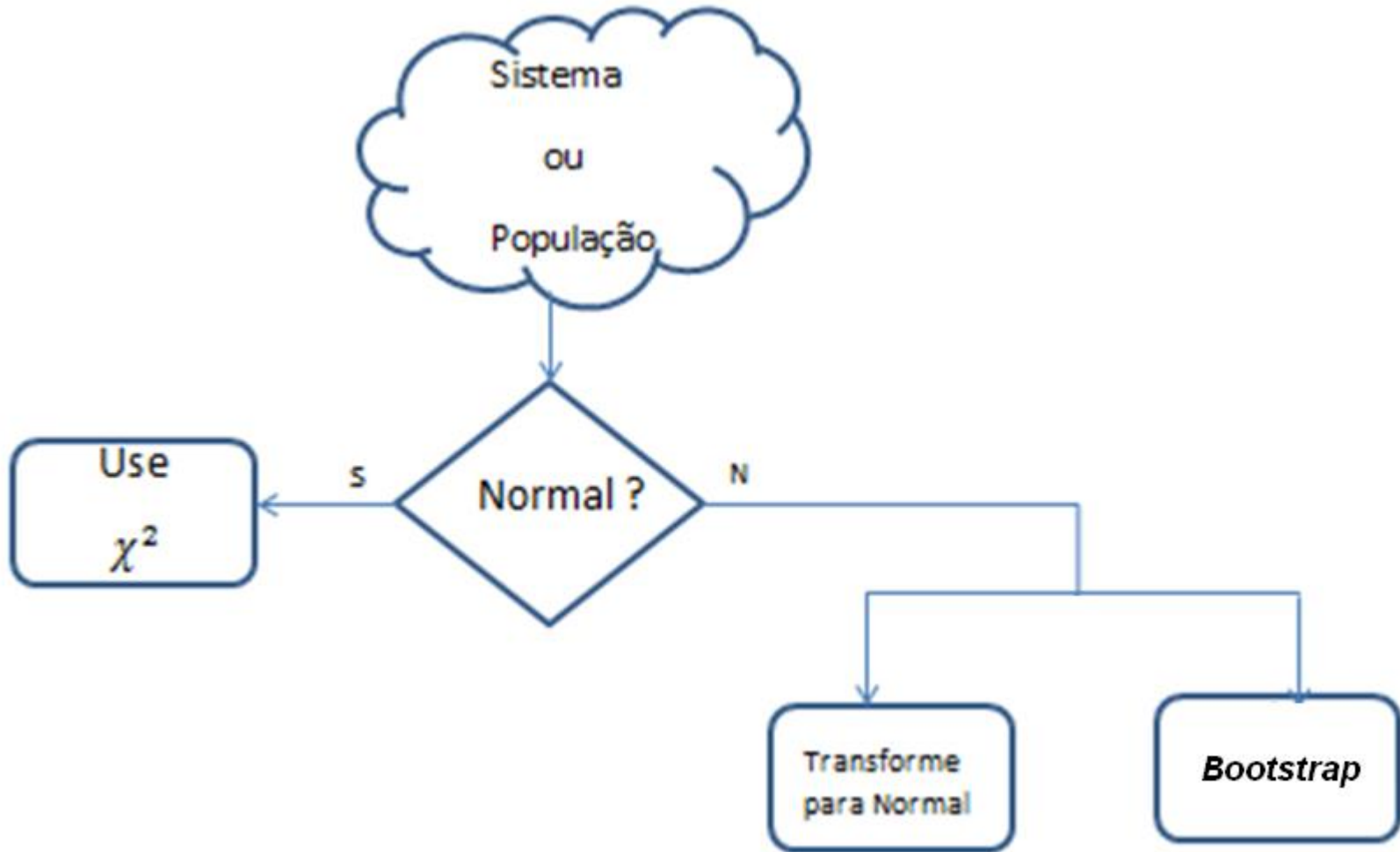
Orientação para Inferência – Média



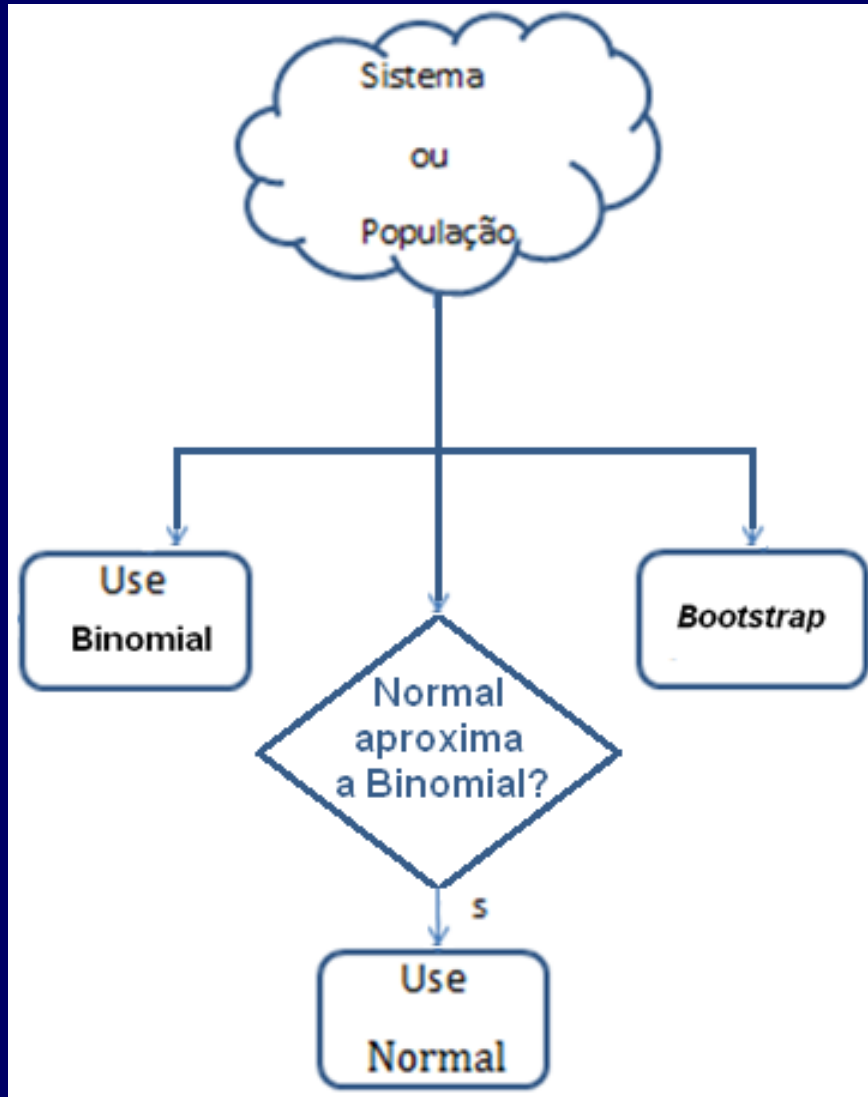
Orientação para Inferência – Média



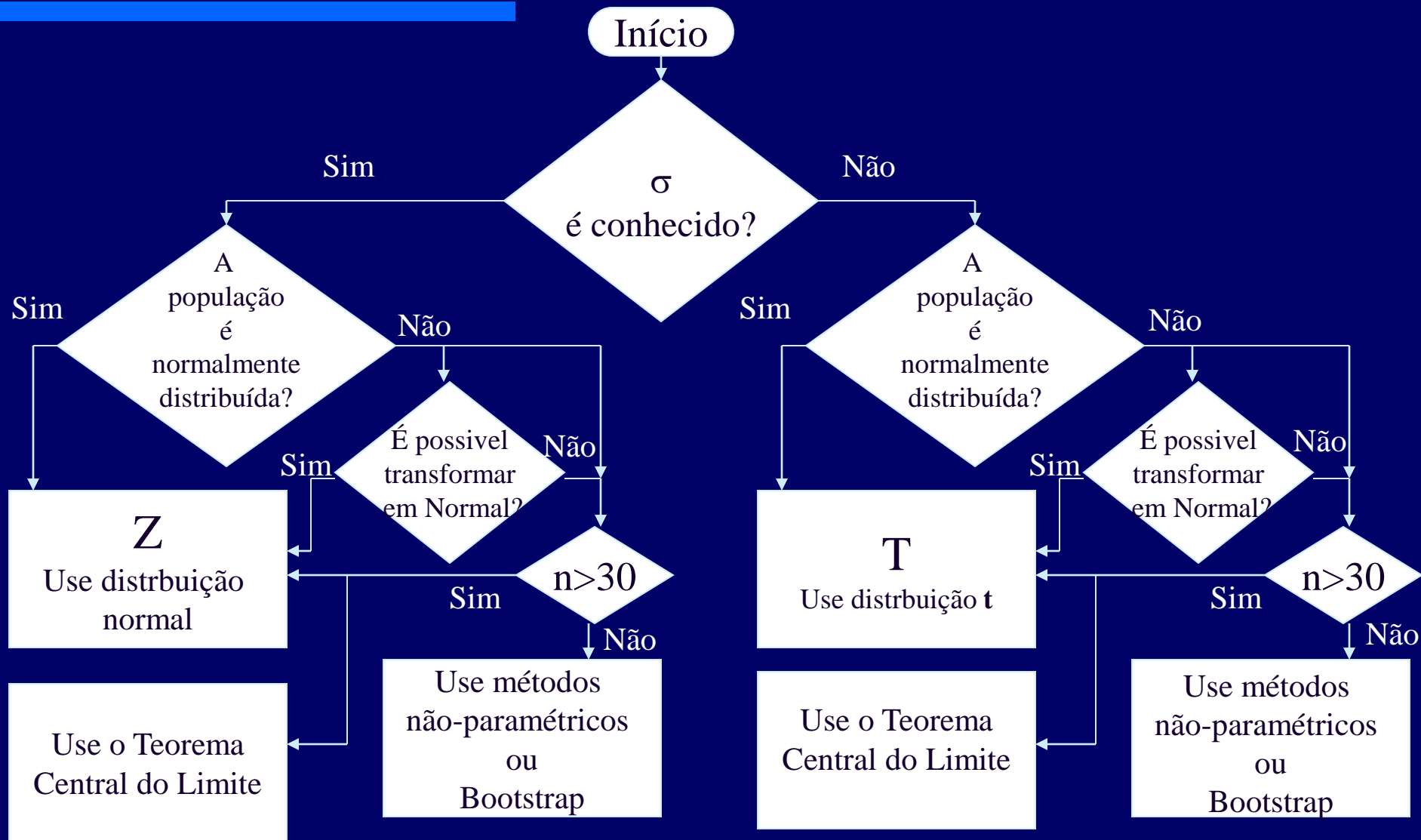
Orientação para Inferência - Variância



Orientação para Inferência - Proporção



Orientação para Inferência



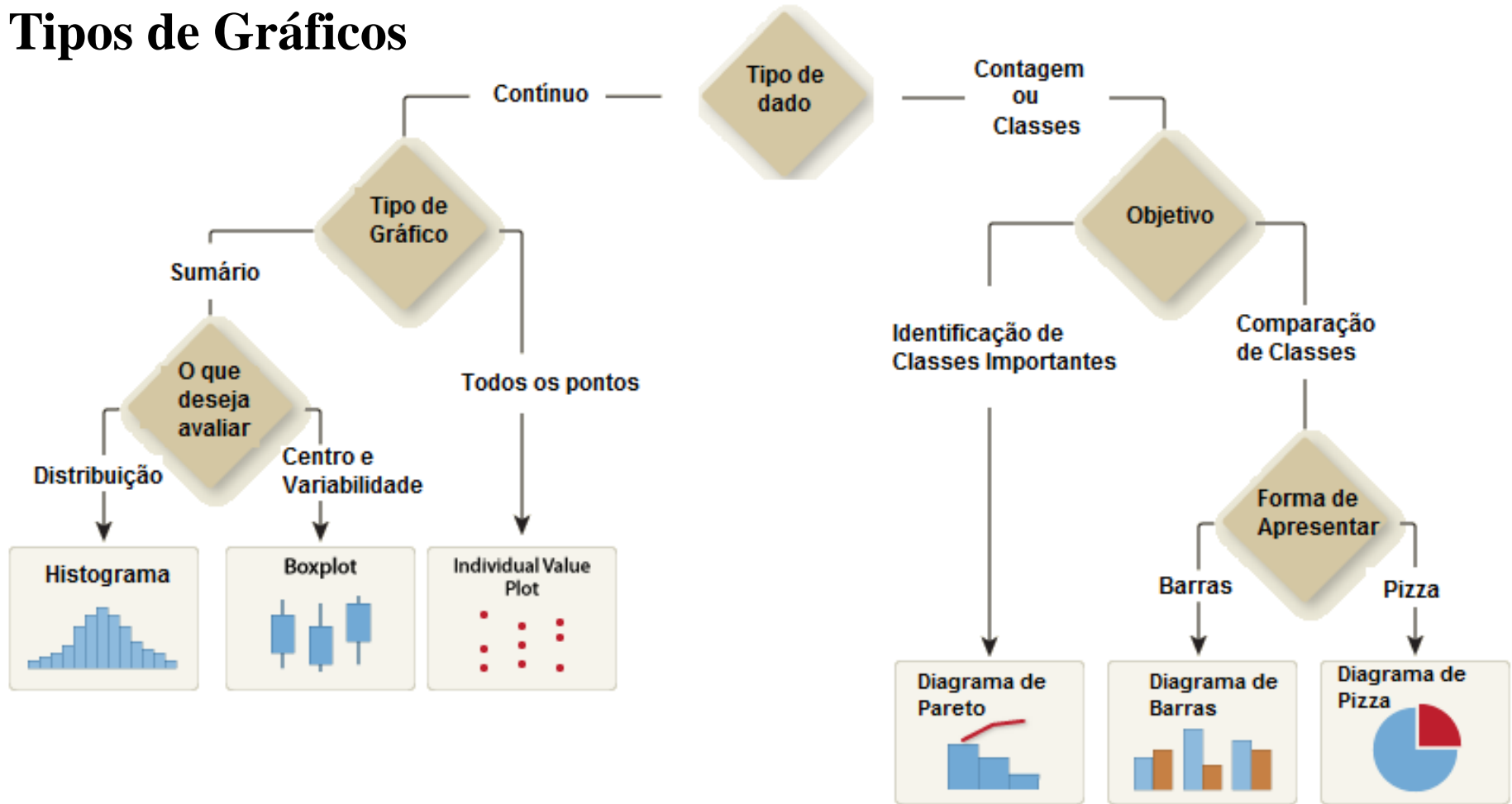
Análise dos Dados

- Uma variável categórica situa um indivíduo numa determinada classe.
- Uma variável quantitativa assume valores numéricos
- A distribuição de uma variável nos fornece os valores que esta assume e sua frequência.

ANÁLISE EXPLORATÓRIA

Análise dos Dados

Tipos de Gráficos



- Alguns gráficos para variáveis categóricas relacionam a categoria a uma contagem ou percentagem.
 - Gráfico de Barras
 - Gráfico de Pareto
 - Gráfico de Setores

Análise dos Dados

Variáveis Categóricas

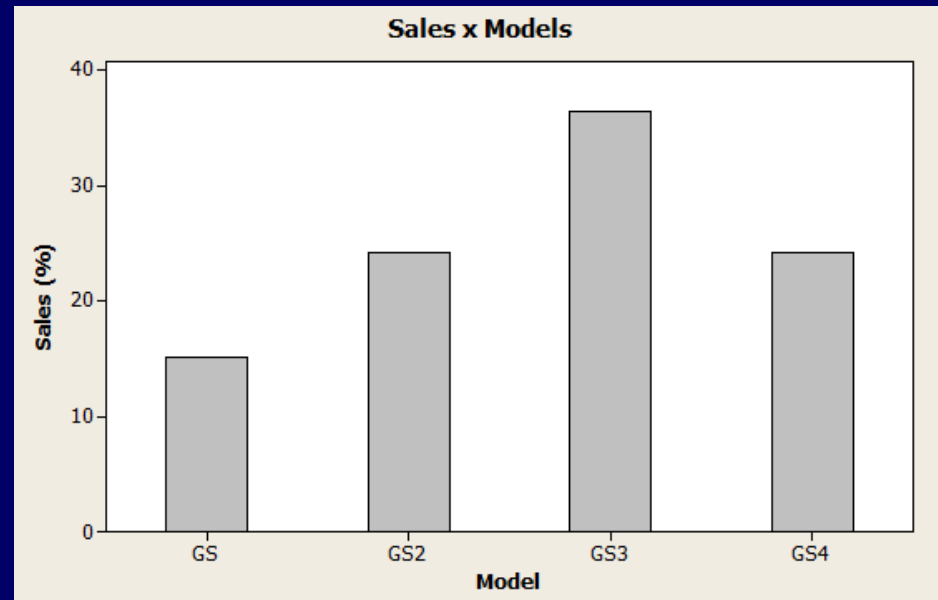
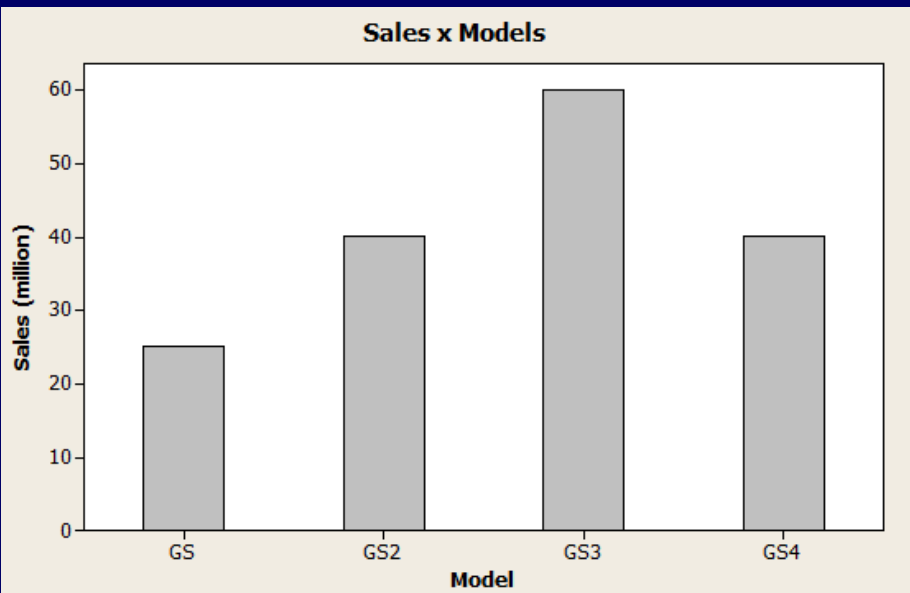
Sales	Numbers in million	%
GS	25	15.2
GS2	40	24.2
GS3	60	36.4
GS4	40	24.2

Análise dos Dados

Variáveis Categóricas

Gráfico de Barras Simples

Sales	Numbers in million	%
GS	25	15.2
GS2	40	24.2
GS3	60	36.4
GS4	40	24.2



O eixo vertical pode ser uma contagem, frequência, percentual ou uma função.

Análise dos Dados

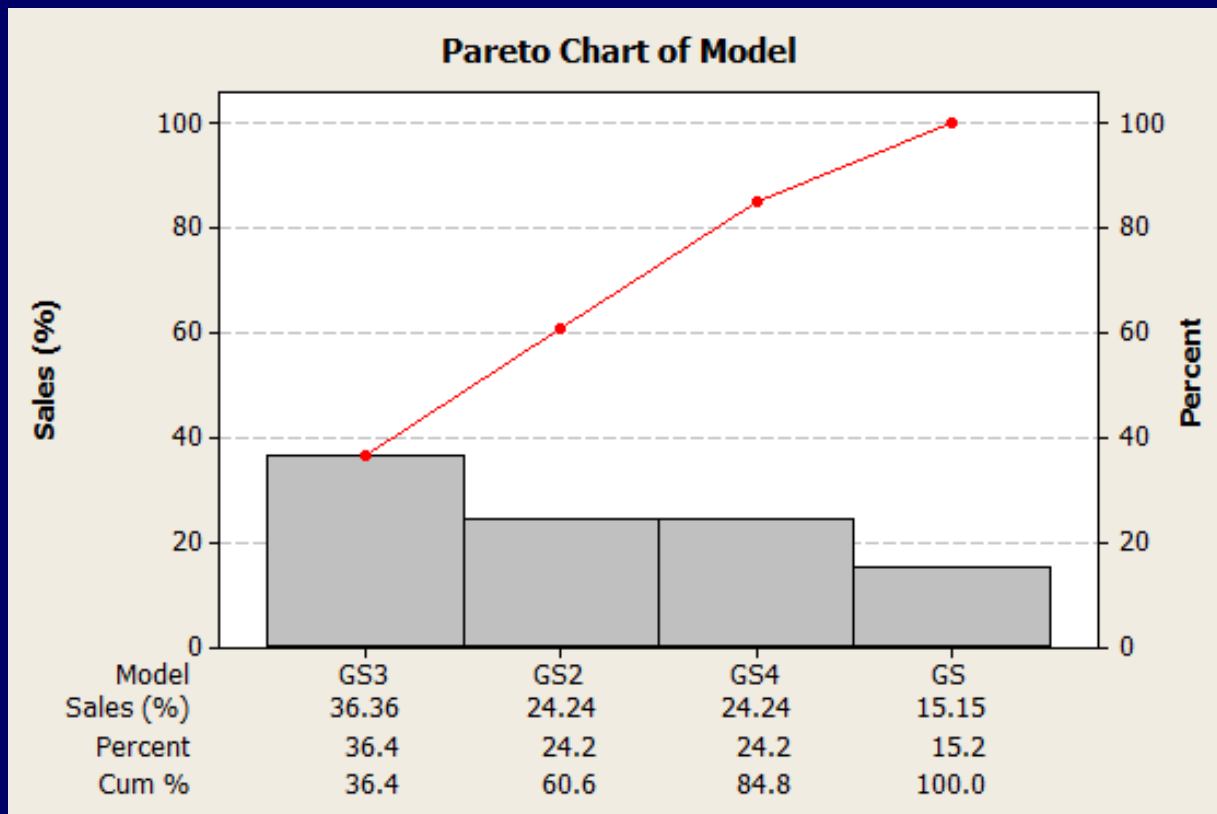
Variáveis Categóricas

C:\Paulo Maciel\Tools\Minitab 14\Exemplos\QSBCToolBo
Files & Templates for Six Sigma and MINITAB - V14.0\
PARETO-POSTAGE.mpj

Gráfico de Pareto

- Opcionalmente pode-se ter uma linha apresentando frequência acumulada

- O eixo vertical pode ser uma contagem, frequência, percentual ou uma função.

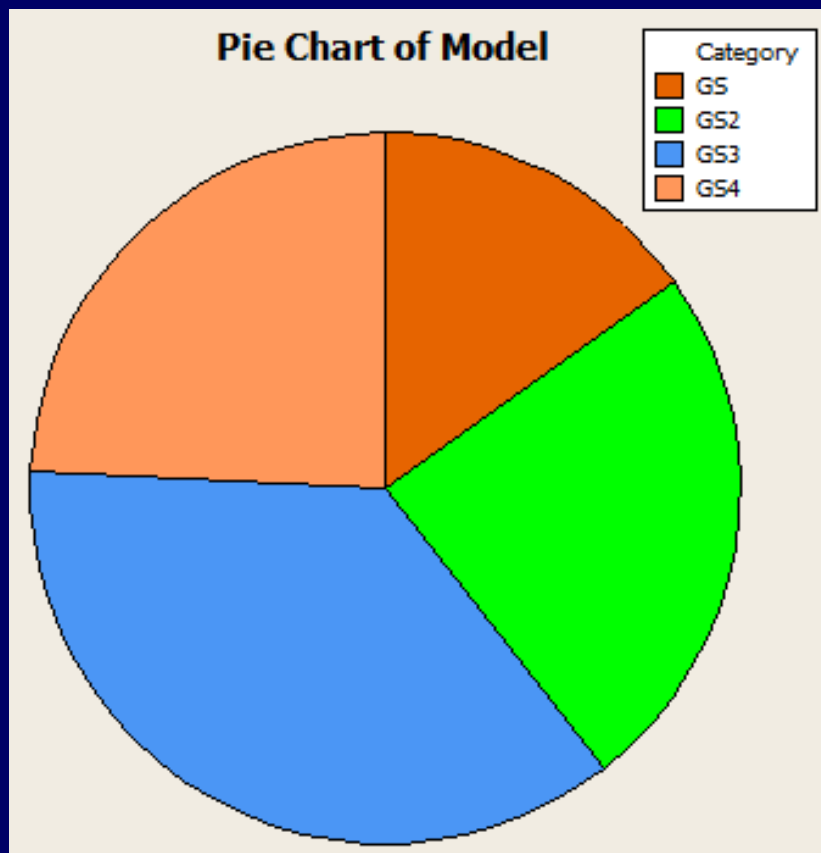


- O gráfico de Pareto é um Gráfico de Barras para onde o arranjo das barras se dá em função da frequência. as barras com maior frequência são ficam mais a esquerda que as barras de menor frequência.

Análise dos Dados

Variáveis Categóricas

Gráfico de Setores



Sales	Numbers in million	%
GS	25	15.2
GS2	40	24.2
GS3	60	36.4
GS4	40	24.2

- Os setores podem ser rotulados com contagens, frequências, percentuais ou funções.

Análise dos Dados

Minitab

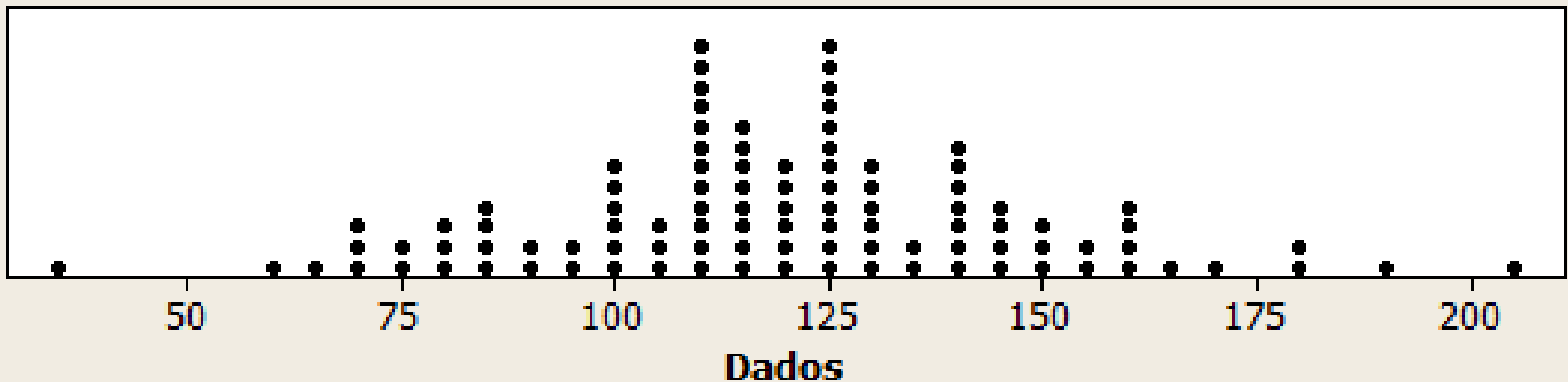
- Distribuição de uma variável quantitativa registra seus valores numéricos e a frequência de ocorrência de cada valor.
- Alguns gráficos para representação de distribuição.
 - Gráfico de Pontos (dotplot)
 - Gráfico de Ramos-e-folhas
 - Impróprio para representação de grandes conjuntos de dados.
 - Histograma
 - Divide o intervalo de valores de uma variável em intervalos e apresenta o número ou percentagem que se enquadra em cada intervalo.
 - Grafo de Distribuição Empírica
 - Diagrama de caixa

Análise dos Dados

C:\Users\Paulo Maciel\Dropbox\Models\Minitab\dotplot.MPJ

□ Dot Plot

Dotplot of Dados



O eixo x é dividido em intervalos. Os dados que estiverem dentro de cada intervalo são representados por pontos (dots).

Análise dos Dados

□ Histograma

- Dividir o intervalo dos dados em classes de igual amplitude. Na prática, quando o número de observações é grande, normalmente se considera o Número de Classes

$$NC = (\text{Número de observações})^{1/2}$$

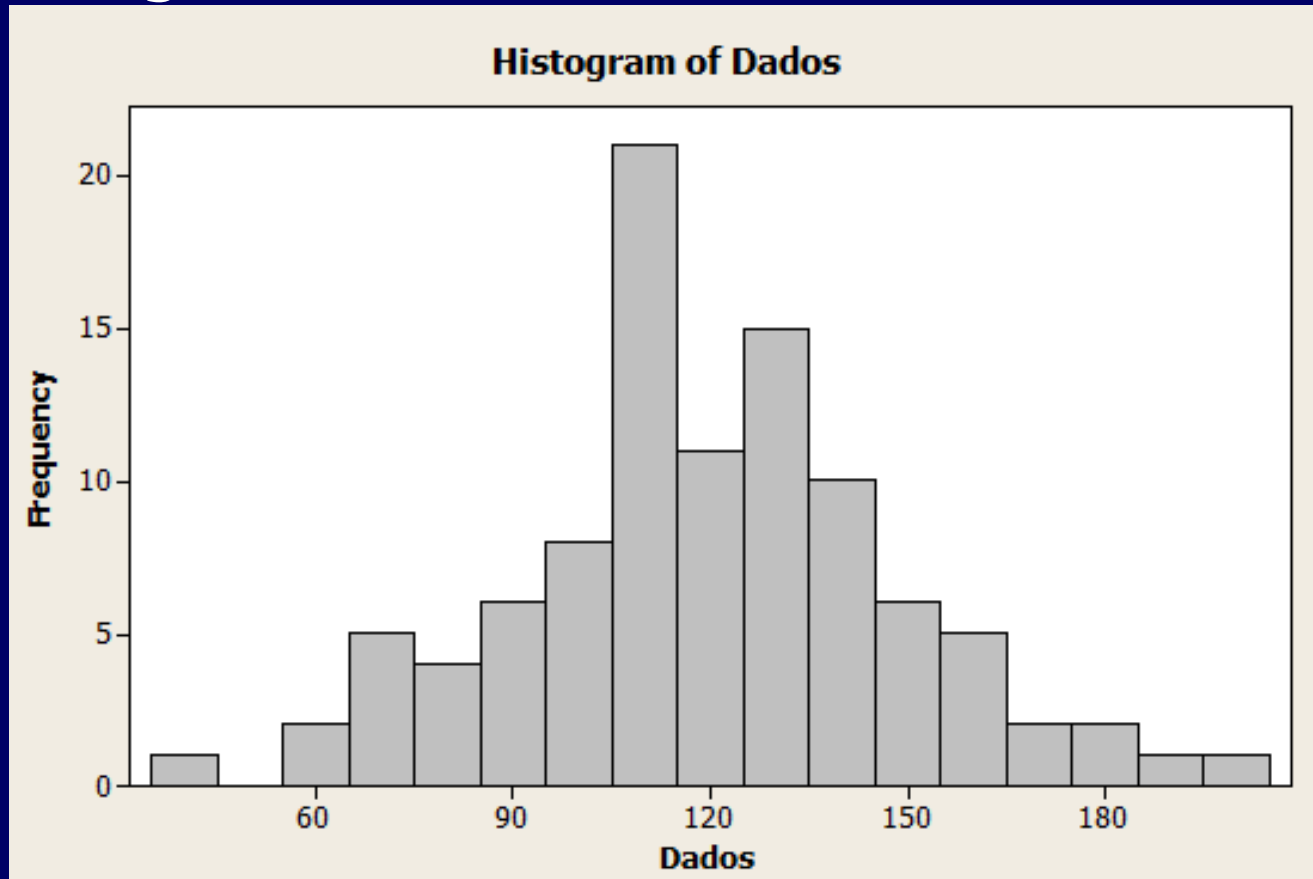
$$NC = \lceil 1 + n \times \log_{10} SS \rceil$$

- Contar o número de observações em cada classe (tabela de frequência).
- Traçar o histograma. As classes são colocadas na horizontal e as frequências na vertical. Não há espaçamento entre as classes. Cada classe é representada por uma barra de altura igual a frequência.

Análise dos Dados

C:\Paulo Maciel\Tools\Minitab 14\Exemplos\QSBCToolBo
Files & Templates for Six Sigma and MINITAB - V14.0\
TIME TO DELIVER.MPJ

Histograma

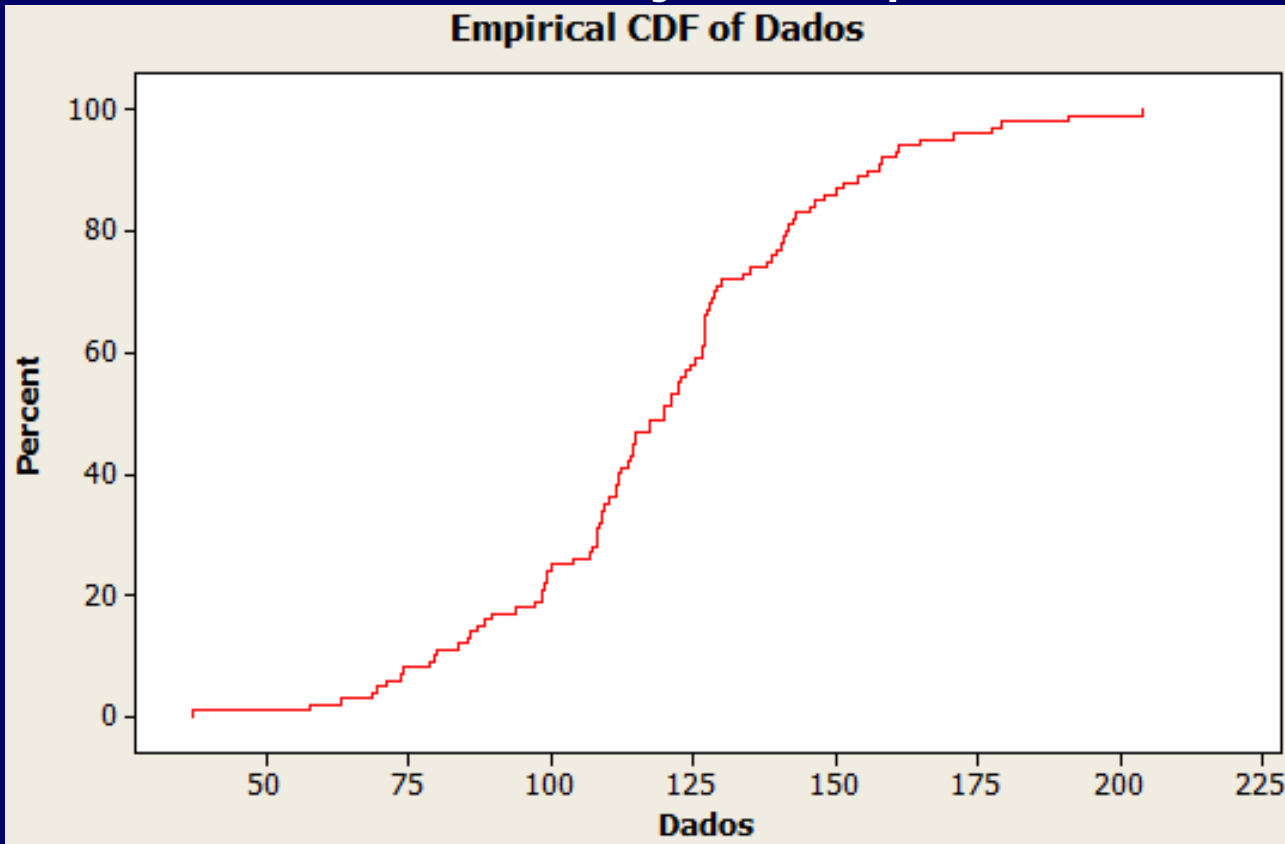


O eixo x é dividido em intervalos. As barras representam o número (ou frequência) de observações em cada intervalo.

Análise dos Dados

C:\Paulo Maciel\Tools\Minitab 14\Exemplos\QSBCToolBo
Files & Templates for Six Sigma and MINITAB - V14.0\
TIME TO DELIVER.MPJ

Gráfico de Distribuição Empírica



O gráfico de distribuição empírica grafa o valor de cada observação, considerando todos os valores menores ou iguais ao valor em avaliação, em relação ao percentual total dos valores da amostra.

Análise dos Dados

- Exame de uma distribuição
 - Padrão geral e desvio acentuados.
 - Padrão geral:
 - Forma
 - Centro
 - Dispersão
 - Desvios acentuados
 - *Outliers*

Análise dos Dados

□ Medidas de Centro

- A Média \bar{x} de um conjunto de observações é obtida somando os valores das observações e dividindo pelo número de observações.

$$- \quad \bar{x} \text{ or } \mu = \frac{1}{N} \sum_{i=1}^N x_i$$

- A média não imune à influência de observações extremas.
- Não é resistente

Análise dos Dados

□ Medidas de Centro

- A Média Aritmética Ponderada.

$$\bar{x} \text{ or } \mu = \frac{\sum_{i=1}^N x_i f_i}{\sum_{i=1}^N f_i}$$

• Exemplo

Suponha que a utilização de uma CPU foi medida em 5 intervalos de tempo. Qual é a utilização média da CPU?

Measurement Duration (s)	CPU Utilization
1	45,00%
1	45,00%
1	45,00%
1	45,00%
100	20,00%
Mean CPU Utilization	20,96%

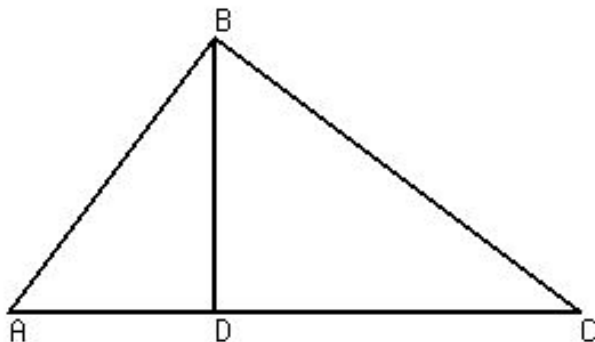
Análise dos Dados

- Medidas de Centro
 - A Média Geométrica.

$$\bar{X} = \sqrt[n]{x_1 x_2 \cdots x_n}$$

O termo média geométrica é originado da Geometria.

Se traçarmos uma reta perpendicular do ângulo reto de um triângulo retângulo até a hipotenusa, a altura do triângulo é exatamente igual a média geométrica das duas partes da hipotenusa, ou seja: $BD = \sqrt{AD \times DC}$



Análise dos Dados

A Média Geométrica

A media aritmética é relevante quando quantidades são SOMADAS para produzir o total.

Da mesma forma, a média geométrica é relevante sempre que quantidades são MULTPLICADAS para produzir o total.

Por exemplo, considere que você investiu numa aplicação e recebeu 10% no primeiro ano, 50% no segundo ano e 30% no terceiro ano. Qual é a taxa média de retorno?

A média aritmética NÃO é apropriada, porque o seu investimento no primeiro ano foi MULTIPLICADO por 1,1. No segundo ano, o novo valor total é MULTIPLICADO por 1,5, e, finalmente, no terceiro ano o valor obtido ao final do segundo ano é MULTIPLICADO por 1,3. Desta forma, a média relevante é a média geométrica.

$$\bar{X} = \prod_{i=1}^n x_i^{1/n}$$

Análise dos Dados

A Média Geométrica

$$\bar{X} = \prod_{i=1}^n x_i^{1/n}$$

A média geométrica é:

$$\bar{x} = \sqrt[3]{1,1 \times 1,5 \times 1,3} = 1,289662$$

Portanto, a taxa média de retorno é $1,289662 - 1 = 0,289662 = 28,9662\%$

Análise dos Dados

A Média Geométrica

Exemplo:

Considere um sistema de gerenciamento de banco de dados (SGBD) instalado em uma infraestrutura composta por um servidor (S) que executa um sistema operacional (OS).

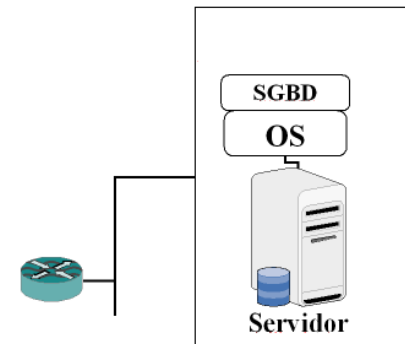
A vazão (*throughput*) deste SGBD foi monitorada em condições de *stress* e o valor máximo obtido para transações do tipo “ T_1 ” foi de 80 TPS.

Uma série de ajustes foi executada no sistema. Foram feitos ajustes no hardware do servidor e otimizações no sistema operacional. Cada um desses ajustes, respectivamente, provocou um aumento da vazão do SGBD.

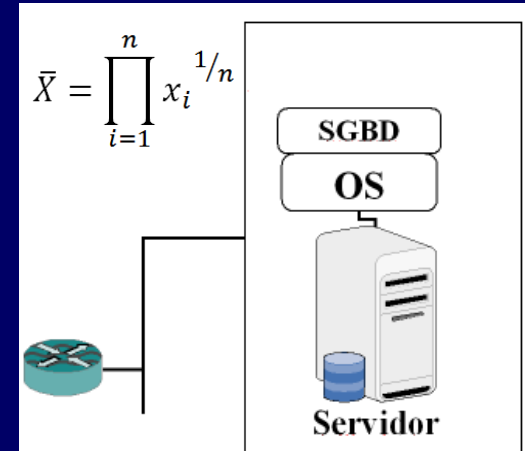
Os ajustes no HW do servidor aumentaram a vazão média em 22%. Os ajustes relativos ao SO provocaram um aumento médio de 8% na vazão do SGBD.

Além desses ajustes, o SGBD foi reconfigurado. Este ajuste, em especial, aumentou a vazão média em 17%.

$$\bar{X} = \sqrt[n]{\prod_{i=1}^n x_i}$$



Análise dos Dados



A Média Geométrica

Exemplo:

- a) Qual é o percentual médio de otimização?
- b) Qual é a vazão média da plataforma otimizada (considerando os três ajustes efetuados)?

$$a) \sqrt[3]{1,22 \times 1,08 \times 1,17} = 1,1552 = 115,52\%$$

$$115,52\% - 100\% = 15,52\%$$

$$b) TP' = TP \times 115,52\% = 80 \text{ tps} \times 115,52\% = 92.42 \text{ tps}$$

Componente i		Otimização devido ao Componente i	TPS	
S	Baseline	22%	97.60	Throughput (SGBD) após a otimização do HW
OS	Baseline	8%	86.40	Throughput (SGBD) após a otimização do SO (TPS)
SGBD	Baseline	17%	93.60	Throughput (SGBD) após a reconfiguração do SGBD (TPS)
Throughput (TPS)	80.00	115.520%	172.42	Throughput (SGBD) considerando todas as otimizações (TPS)
% de otimização médio		15.520%		

Análise dos Dados

– Média Harmônica

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

Considere que através de um *link* de comunicação trafegam pacotes. Considere que:

- 6400 pacotes foram transferidos a uma taxa 320 pps,
- 6400 pacotes foram transferidos a uma taxa 480 pps,
- 6400 pacotes foram transferidos a uma taxa 520 pps, e que
- 6400 pacotes foram transferidos a uma taxa 280 pps.

Qual foi a taxa média de transferência de pacotes?

$$\overline{tx} = \frac{N}{\frac{1}{tx_1} + \frac{1}{tx_2} + \frac{1}{tx_3} + \frac{1}{tx_4}} = \frac{4}{\frac{1}{320} + \frac{1}{480} + \frac{1}{520} + \frac{1}{280}} = 373,7326 \text{ pps}$$

Análise dos Dados

– Média Harmônica

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

Ou usando a média aritmética, desta forma:

Tempo necessário para 6400 pacotes a uma taxa 320 pps $T_1 = \frac{6400 \text{ p}}{320 \text{ pps}} = 20 \text{ s}$

Tempo necessário para 6400 pacotes a uma taxa 480 pps $T_2 = \frac{6400 \text{ p}}{480 \text{ pps}} = 13,333333 \dots \text{ s}$

Tempo necessário para 6400 pacotes a uma taxa 520 pps $T_3 = \frac{6400 \text{ p}}{520 \text{ pps}} = 12,30769231 \text{ s}$

Tempo necessário para 6400 pacotes a uma taxa 280 pps $T_4 = \frac{6400 \text{ p}}{280 \text{ pps}} = 22,85714286 \text{ s}$

Tempo total necessário para transferir os $4 \times 6400 = 25600 \text{ pacotes}$ é

$$\sum_{i=1}^4 T_i = 68.4981685 \text{ s}$$

Portanto a taxa média de transferência é:

$$\overline{tx} = \frac{25600 \text{ pacotes}}{68.4981685} = 373,7326 \text{ pps}$$

Análise dos Dados

Média Harmônica

No entanto, considere que

- # $P1$ foram transferidos a uma taxa tx_1 durante um tempo T , que
- # $P2$ foram transferidos a uma taxa tx_2 durante um tempo T , que
- # $P3$ foram transferidos a uma taxa tx_3 durante um tempo T , e que
- # $P4$ foram transferidos a uma taxa tx_4 durante tempo T .

Se se deseja saber a taxa média de transferência, deve-se usar a **média aritmética**.

Veja o exemplo:

- # de pacotes transferidos no período T_1 (10s) a uma taxa tx_1 (320 pps) = 3200 p
- # de pacotes transferidos no período T_2 (10s) a uma taxa tx_2 (480 pps) = 4800 p
- # de pacotes transferidos no período T_3 (10s) a uma taxa tx_3 (520 pps) = 5200 p
- # de pacotes transferidos no período T_4 (10s) a uma taxa tx_4 (280 pps) = 2800 p

Total de pacotes transferidos em $(T = T_1 + T_2 + T_3 + T_4)$ 40s é 16000 pacotes.

Portanto a taxa média deve ser calculada através da média aritmética:

$$\overline{tx} = \frac{tx_1 + tx_2 + tx_3 + tx_4}{4}$$

$$\overline{tx} = \frac{320 \text{ pps} + 480 \text{ pps} + 520 \text{ pps} + 280 \text{ pps}}{4} = 400 \text{ pps}$$

O que equivale a:

$$\overline{tx} = \frac{\text{Total de pacotes transferidos em } T}{T}$$

$$\overline{tx} = \frac{16000}{40} = 400 \text{ pps}$$

Análise dos Dados

□ Medidas de Centro

- A Mediana de um conjunto de observações é o ponto médio de uma distribuição. É um número tal que metade das observações é inferior a ele e metade é superior.
- Disponha todas as observações em ordem de tamanho (da menor para a maior).
- Se o número de observações (n) é ímpar, a mediana é a observação central e localiza-se $(n+1)/2$ observações a partir da base.
- Se o número de observações for par, a mediana é a média das duas observações centrais. A localização é novamente $(n+1)/2$
- É resistente

Análise dos Dados

- Medidas de Centro

Midrange (Mr) é uma medida de centro.

$$Mr = \frac{\textit{Menor_Valor} + \textit{Maior_Valor}}{2}$$

Análise dos Dados

Mediana
E
Média

54

59

35

41

46

25

47

60

54

46

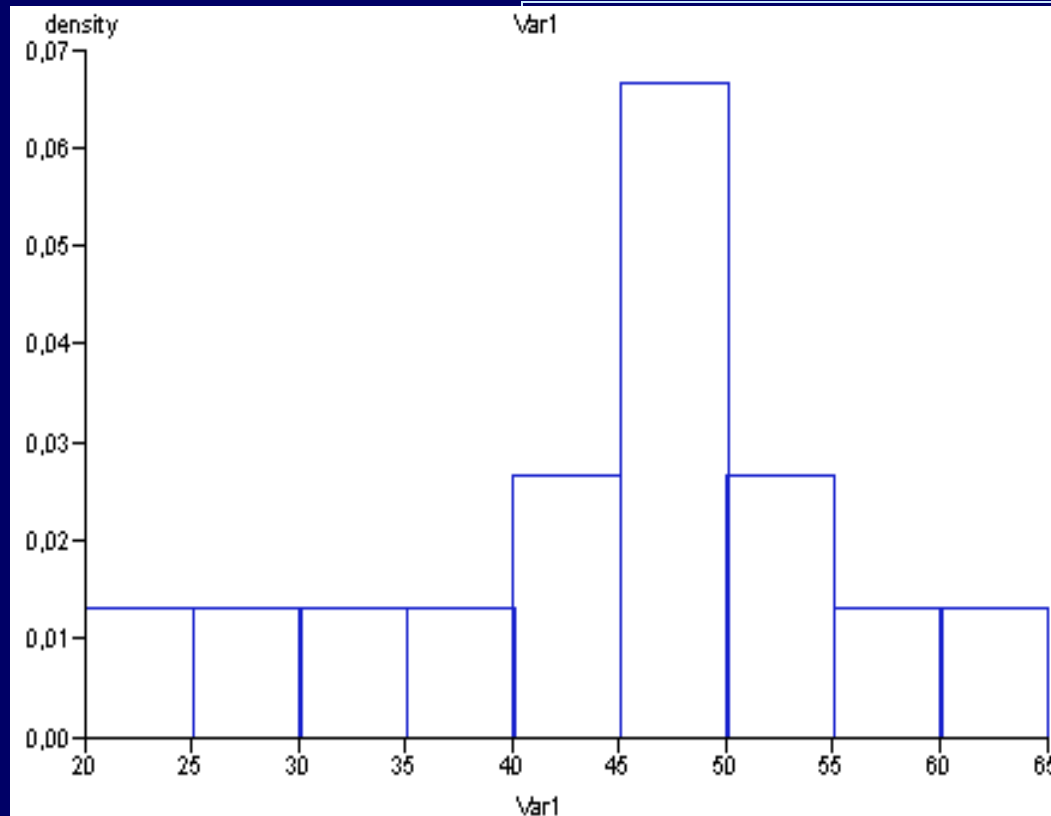
49

46

41

34

22



Mediana = 46

Máximo = 60

Mínimo = 22

1Q = 35

3Q = 54

IIQ = 19

\bar{X} = 43,9333

σ = 11,2470

σ^2 = 126,4952



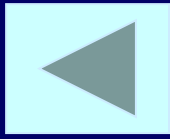
Análise dos Dados

□ Medidas de Dispersão

- **Amplitude** é a diferença entre o maior e o menor valor de um conjunto de dados.

- $A = \text{Maior valor} - \text{Menor Valor}$

Medida de dispersão simples, porém de fácil obtenção que provê informações importantes.



Análise dos Dados

□ Medidas de Dispersão

- A descrição numérica mais comum é a combinação da média (para medir o centro) e do desvio-padrão (s) para medir a dispersão.
- O desvio-padrão mede a dispersão considerando o quão afastadas da média estão as observações. (mesma unidade da média)
- A **variância** (s^2) de um conjunto de observações é a média do quadrado dos desvios destas (observações) em relação a média.

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

- Desvio-padrão:

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

Análise dos Dados

□ Medidas de Dispersão

– Propriedades do Desvio-padrão:

□ $s = 0$ indica que não há dispersão.

□ Quanto mais dispersas as observações maior o s .

□ s não é resistente. Alguns valores extremos (*outliers*) podem tornar s grande.

Análise dos Dados

□ Medidas de Dispersão

- Coeficiente de Variação descreve o desvio padrão em relação a média. Possibilita a comparar a variação para valores originados de diferentes populações.

$$\square CV = s / \bar{x}$$

Análise dos Dados

□ Quantil

- O quantil p de uma variável aleatória X é o valor x que soluciona

$$\square \quad p = P[X \leq x] \text{ ou } p = FX[x]$$

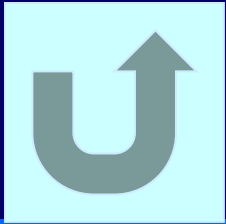
□ Percentis, Quartis e Mediana são quantis.

- O p -ésimo percentil de uma distribuição é o valor que tem p por cento das observações nele ou abaixo dele.
- O 50º percentil é a mediana (medida de centro).
- O 25º percentil é denominado 1º quartil.
- O 75º percentil é o terceiro quartil.

Análise dos Dados

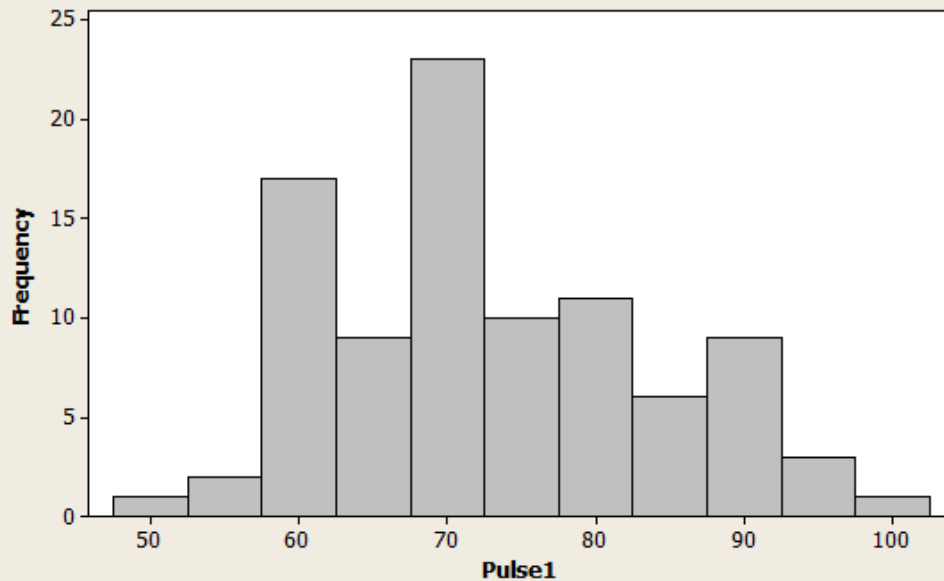
□ Medidas e Dispersão

- Podemos descrever a dispersão (variabilidade) de uma distribuição mediante os percentis.
- O 50º percentil é a mediana (medida de centro).
- O 25º percentil é denominado 1º quartil.
- O 75º percentil é o terceiro quartil.
- O 1º quartil pode ser obtido calculando-se a mediana dos dados que estão à esquerda (abaixo) da mediana global.
- O 3º quartil pode ser obtido calculando-se a mediana dos dados que estão à direita (acima) da mediana global.

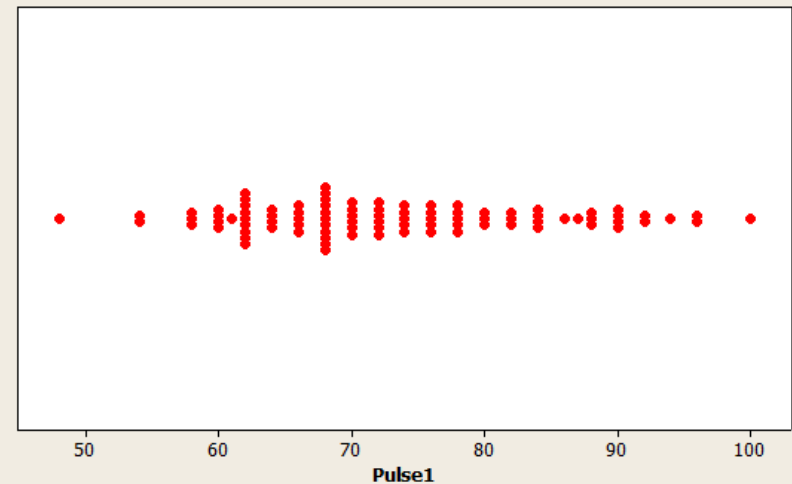


Análise dos Dados

Histogram of Pulse1



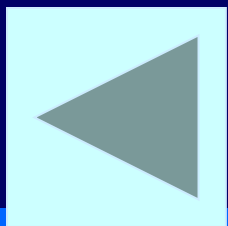
Individual Value Plot of Pulse1



Descriptive Statistics: Pulse1

Variable	Count	Mean	StDev	Variance	CoefVar	Minimum	Q1	Median	Q3
Pulse1	92	72,87	11,01	121,19	15,11	48,00	64,00	71,00	80,00

Maximum Range
100,00 52,00



Análise dos Dados

□ Medidas e Dispersão

- Intervalo interquartil: IIQ é a distância entre o primeiro e o terceiro quartil. $IIQ = Q_3 - Q_1$
- O critério $1,5 \times IIQ$ para definir *Outliers* suaves:
 - Dados que estão abaixo de $Q_1 - (1,5 \times IIQ)$ são *outliers* suaves.
 - Dados que estão acima de $Q_3 + (1,5 \times IIQ)$ são *outliers* suaves.
- Resumo dos cinco (5) números:
 - Mínimo Q_1 M Q_3 Máximo



Análise dos Dados

□ Medidas e Dispersão

- Intervalo interquartil: IIQ é a distância entre o primeiro e o terceiro quartil. $\text{IIQ} = Q_3 - Q_1$
- O critério $3,0 \times \text{IIQ}$ para definir *Outliers* extremos:
 - Dados que estão abaixo de $Q_1 - (3,0 \times \text{IIQ})$ são *outliers* extremos.
 - Dados que estão acima de $Q_3 + (3,0 \times \text{IIQ})$ são *outliers* extremos.
- Resumo dos cinco (5) números:

□ Mínimo	Q_1	M	Q_3	Máximo
----------	-------	---	-------	--------

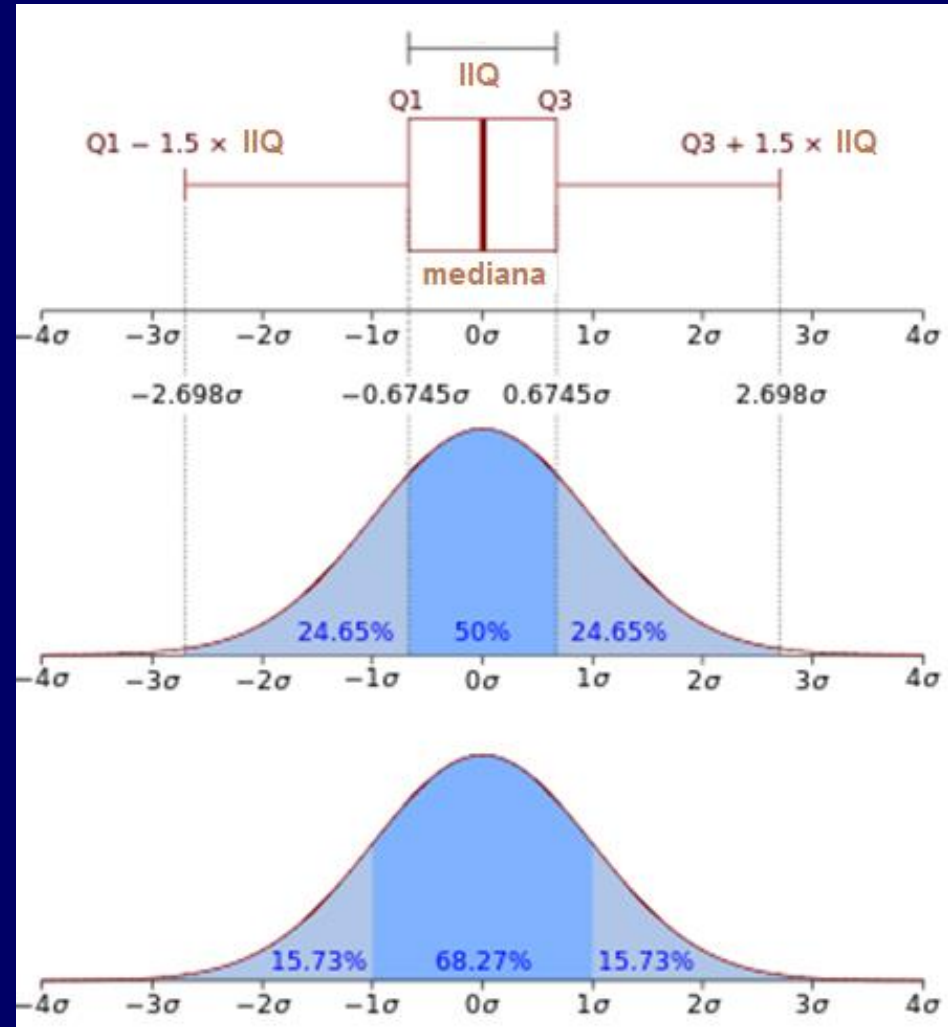
Análise dos Dados

Intervalo interquartil

Para dados Normalmente distribuídos, o $IIQ \cong \frac{4}{3} \times S$.

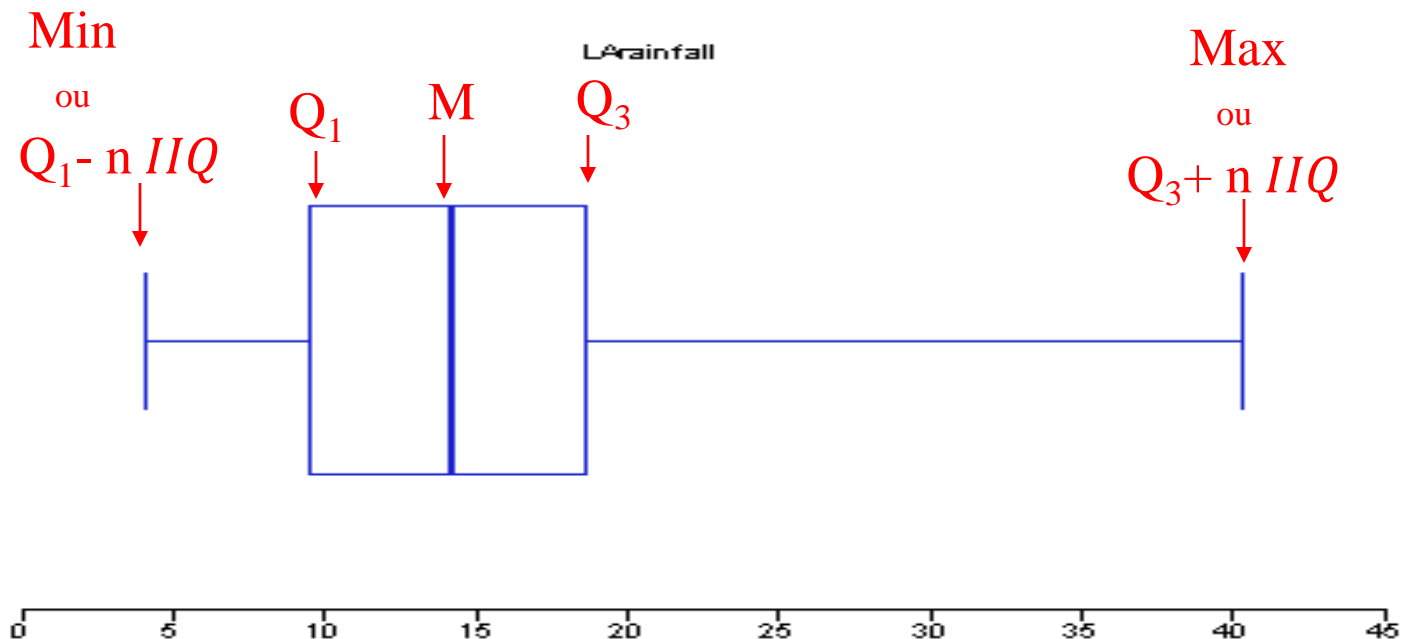
Portanto:

$$S \cong IIQ \times \frac{3}{4}$$



Análise dos Dados

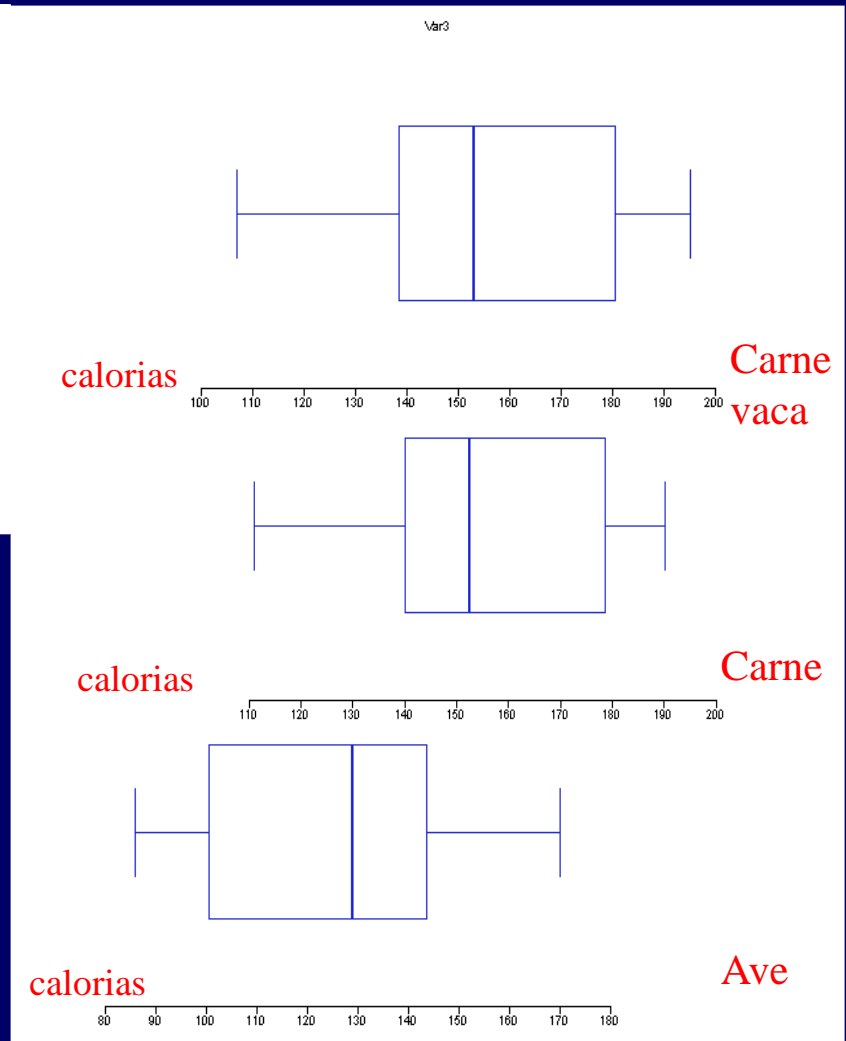
- **Diagrama de Caixa** é um diagrama do resumo dos cinco números, como os *outliers* suspeitos marcados individualmente.
 - Algumas ferramentas podem não marcar os *outliers* suspeitos, como também utilizar uma regra diferente do 1,5 IIQ e 3IIQ.

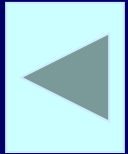


Análise dos Dados

□ Diagrama de Caixa é um diagrama do resumo dos cinco números, como os *outliers* suspeitos Marcados individualmente.

□ Comparando distribuições.





Análise dos Dados



- Escolha da medida de centro e de dispersão
 - O resumo dos cinco números é, em geral, melhor do que a média e o desvio-padrão quando as distribuições são assimétricas ou quando a distribuição tiver fortes *outliers*.
 - Quando a distribuição for razoavelmente simétricas e sem *outliers*, a média e o desvio-padrão são recomendados.

- Excel (C:\Paulo Maciel\Tools\Statistics\Excel\PLANILHA\pulse.XLS)
- Statistica (C:\Paulo Maciel\Tools\Statistics\Statistica\Examples\pulse.sta)
- Minitab (C:\Paulo Maciel\Tools\Minitab 14\Data\Data\pulse.MTW)

Análise dos Dados

□ Formas

Skewness (Assimétria) e *Kurtose* (Curtose) são estatísticas sem unidade.

São comumente normalizadas de forma que a Distribuição Normal tenha as respectivas estatísticas iguais a 0.

- Uma distribuição é assimétrica se uma das caudas for maior que a outra.
- Assimetria (*skewness*)

$$\text{Skew}[X] = \frac{E[(X - E[X])^3]}{(E[(X - E[X])^2])^{3/2}} = \frac{E[(X - E[X])^3]}{(V[X])^{3/2}}$$

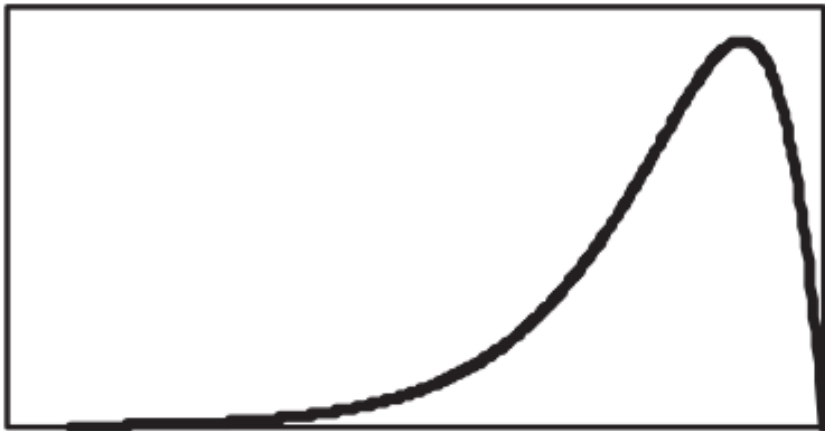
- Valores negativos ($\text{Skew}[X] < 0$) indicam que a cauda está a esquerda.
- Valores positivos ($\text{Skew}[X] > 0$), indicam cauda à direita.
- $\text{Skew}[X] = 0$ indica simetria.

Análise dos Dados

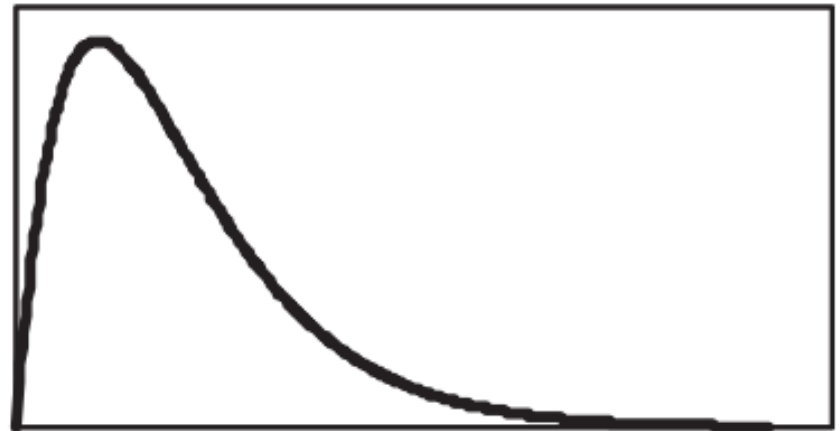
□ Formas

- Assimetria (*skewness*)

$$\text{Skew}(X) = -1.414$$



$$\text{Skew}(X) = +1.414$$



- Valores negativos ($\text{Skew}[X] < 0$) indicam que a cauda está a esquerda.
- Valores positivos ($\text{Skew}[X] > 0$), indicam cauda à direita.
- $\text{Skew}[X] = 0$ indica simetria.

■ Formas

Assimetria interquartil

$$SK_{IQ} = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1} = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}$$

$SK_{IQ} = 0$ - Dados em torno da mediana

$SK_{IQ} < 0$ - mediana mais próxima de Q_3 do que Q_1 (assimetria a esquerda)

$SK_{IQ} > 0$ - mediana mais próxima de Q_1 do que Q_3 (assimetria a direita)

Análise dos Dados

$$\text{Kurtosis} = \frac{\sum (x_i - \bar{X})^4}{ns^4}$$

□ Formas

- Curtose – medida de achatamento. Valores negativos indicam achatamento. Valores positivos indicam picos.

$$\text{Kurt}[X] = \frac{E[(X - E[X])^4]}{(E[(X - E[X])^2])^2} - 3 = \frac{E[(X - E[X])^4]}{(V[X])^2} - 3$$

- $\text{Kurt}[X] < 0$ indica achatamento no centro ou caudas truncadas (*platykurtic*),
- $\text{Kurt}[X] > 0$ indica pico no centro ou caudas longas (*leptokurtic*),
- $\text{Kurt}[X] = 0$ é denominada distribuição *mesokurtic*. A distribuição Normal é *mesokurtic*.

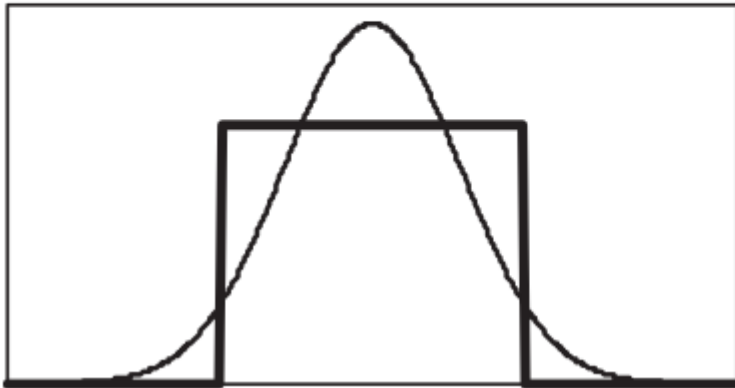
Análise dos Dados

Ver exemplo:
Time to Deliver
Em C:\Paulo
Maciel\Tools\Minitab 14
\Exemplos\QSBCToolBox
\Files & Templates for Six Sigma
and MINITAB - V14.0

Formas

- Curtose

Platykurtic
 $Kurt(X) = -1.2$



Leptokurtic
 $Kurt(X) = +3$



- $Kurt[X] < 0$ indica achatamento no centro ou caudas truncadas (*platykurtic*),
- $Kurt[X] > 0$ indica pico no centro ou caudas longas (*leptokurtic*),
- $Kurt[X] = 0$ é denominada distribuição *mesokurtic*.

REVISÃO DE CONCEITOS BÁSICOS DE PROBABILIDADE

Revisão de Probabilidade

Espaço de Probabilidade

- Um modelo probabilístico é uma representação matemática de um problema/sistema.
- Essa representação matemática é fundamentada através do conceito de espaço de probabilidade (Ω, S, P) .
- Um espaço de probabilidade envolve a realização de um experimento aleatório.

Revisão de Probabilidade

Espaço de Probabilidade

- Experimento aleatório é um experimento que em cada realização se pode gerar diferentes resultados (*outcomes* – evento simples), mesmo que as condições sejam as mesmas em cada realização.
- Um evento é qualquer conjunto de resultados (eventos simples) de um experimento.
- Evento simples é um resultado (*outcome*) que não pode ser decomposto em componentes mais simples.
- Os eventos simples de um espaço amostral devem ser mutuamente exclusivos.

Revisão de Probabilidade

■ Espaço de Probabilidade: $OS = (\Omega, S, P)$

- Ω é o conjunto de todos possíveis resultados de um experimento aleatório - espaço amostral
- S é o conjunto de todos sub-conjuntos formados com os eventos de Ω - conjunto de eventos
- P é uma função que atribui probabilidades aos eventos.
($P: S \rightarrow \mathbb{R}$)

Revisão de Probabilidade

- O conceito de experimento aleatório é abrangente. Um experimento pode ser o teste de um servidor, de dois servidores, de três servidores, por exemplo.
- Na formulação do modelo probabilístico temos apenas um experimento.
- O experimento deve ser definido de acordo com o objetivo do estudo.

Revisão de Probabilidade

□ Espaço de Probabilidade: $OS = (\Omega, S, P)$

Exemplo:

Suponha um experimento que consiste de uma única execução da atividade “ligar o condicionador de ar”. Os resultados possíveis são $\Omega = \{OK, Falha\}$.

O conjunto de eventos $S = \{\emptyset, \{OK\}, \{Falha\}, \{OK, Falha\}\}$

Se tivermos 95 por cento de chances de sucesso (ligar a chave e o condicionador funcionar), portanto:

$$P(\emptyset) = 0$$

$$P(\{OK\}) = 0.95$$

$$P(\{Falha\}) = 0.05$$

$$P(\{OK, Falha\}) = 1$$

Revisão de Probabilidade

□ Espaço de Probabilidade: $OS = (\Omega, S, P)$

Exemplo:

Suponha que quatro servidores foram testados por um período. Após o teste, cada servidor é aprovado (S) ou reprovado (F). Representemos os quatro servidores por um vetor $E_i = (SS_1, SS_2, SS_3, SS_4)$, em que $SS_i = \{S, F\}$ representa o status do servidor i . Este teste é um experimento aleatório, pois o resultado não é conhecido antes da realização do experimento.

$$|\Omega| = 16$$

$$\Omega = \{E_1 = (F, F, F, F), E_2 = (F, F, F, S), E_3 = (F, F, S, F), E_4 = (F, F, S, S), E_5 = (F, S, F, F), E_6 = (F, S, F, S), E_7 = (F, S, S, F), E_8 = (F, S, S, S), E_9 = (S, F, F, F), E_{10} = (S, F, F, S), E_{11} = (S, F, S, F), E_{12} = (S, F, S, S), E_{13} = (S, S, F, F), E_{14} = (S, S, F, S), E_{15} = (S, S, S, F), E_{16} = (S, S, S, S)\}$$

Revisão de Probabilidade

□ Espaço de Probabilidade: $OS = (\Omega, S, P)$

Exemplo:

- Considere um experimento em que se seleciona um conector de metal e se mede sua espessura. Os possíveis valores associados a espessura depende da resolução do mecanismo de medição. No entanto, pode ser conveniente definir o espaço amostral através do conjunto de reais positivos.

$$\Omega = R^+ = \{x \mid x > 0\}$$

Revisão de Probabilidade

□ Espaço de Probabilidade: $OS = (\Omega, S, P)$

Exemplo:

- Se se sabe que a espessura está entre 10 e 11 mm, o espaço amostral pode ser definido por:

$$\Omega = \{x \mid 10 < x < 11\}$$

- Se o objetivo da análise é considerar apenas se o conector tem espessura fina, média ou espessa, o espaço de amostral pode considerar apenas os três possíveis resultados:

$$\Omega = \{fina, média, espessa\}$$

Revisão de Probabilidade

□ Espaço de Probabilidade: $OS = (\Omega, S, P)$

Exemplo:

- Se o objetivo da análise é considerar apenas uma parte está em conformidade ou não com a especificação, o espaço amostral pode ser simplificado para:

$$\Omega = \{sim, não\}$$

Revisão de Probabilidade

□ Espaço de Probabilidade: $OS = (\Omega, S, P)$

Um espaço de probabilidade satisfaz as seguintes condições:

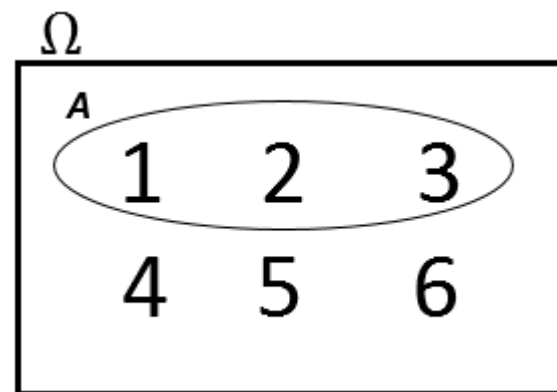
– Se $A \in S \Rightarrow A^c \in S$

O complementar de evento A é representado por A^c (ou \overline{A}) e consiste em todos os resultados em que A não ocorre.

$$A^c = \{4, 5, 6\}$$

$$\text{Se } P(A) = p = 1/2,$$

$$\text{então } P(A^c) = 1 - P(A) = 1 - p = 1/2$$



Revisão de Probabilidade

□ Espaço de Probabilidade: $OS = (\Omega, S, P)$

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

Sejam A e B dois eventos

$$A = \{1, 2, 3\}$$

$$B = \{1, 3, 5\}$$

União

$$A \cup B = \{1, 2, 3, 5\}$$

Interseção

$$A \cap B = \{3\}$$

Diferença

$$A - B = A \cap B^c = \{1, 2, 3\} \cap \{2, 4, 6\} = \{2\}$$

$$B - A = B \cap A^c = \{1, 3, 5\} \cap \{4, 5, 6\} = \{5\}$$

■ Espaço de Probabilidade: $OS = (\Omega, S, P)$

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

Sejam A e B dois eventos

$$A = \{1, 2, 3\}$$

$$B = \{1, 3, 5\}$$

União

$$A \cup B = \{1, 2, 3, 5\}$$

Interseção

$$A \cap B = \{3\}$$

Diferença

$$A - B = A \cap B^c = \{1, 2, 3\} \cap \{2, 4, 6\} = \{2\}$$

$$B - A = B \cap A^c = \{1, 3, 5\} \cap \{4, 5, 6\} = \{5\}$$

Revisão de Probabilidade

□ Sejam A e B dois eventos

$$A = \{1,2,3\}$$

$$B = \{1,3,5\}$$

Então

$$A \cup B = \{1,2,3,5\}$$

$$A \cap B = \{3\}$$

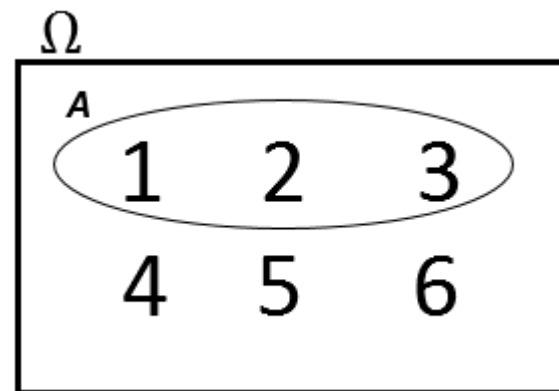
Se $P(A) = 1/2$, $P(B) = 1/2$ e $P(A \cap B) = \frac{1}{6}$

Como

$$P(A \cup B) = P(A) + P(B) - P(A \cap B),$$

então

$$P(A \cup B) = 1/2 + 1/2 - 1/6 = 0.833333$$



Revisão de Probabilidade

– Se A_1, A_2, \dots são todos os eventos então
 $\bigcup_{i=1}^{\infty} A_i \in S$

– Se A_1, A_2, \dots São disjuntos, então:

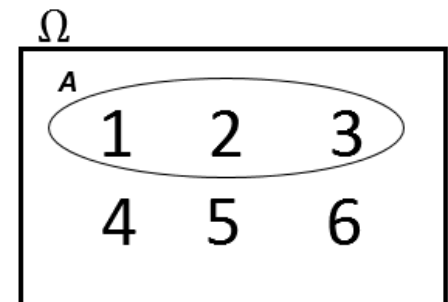
$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

Ex: $A = \{1, 2, 3\}$ e $C = \{4\}$

Se $P(A) = \frac{1}{2}$ e $P(C) = \frac{1}{6}$,

então $P(A \cup C) = \frac{1}{2} + \frac{1}{6} = 0.666667$,

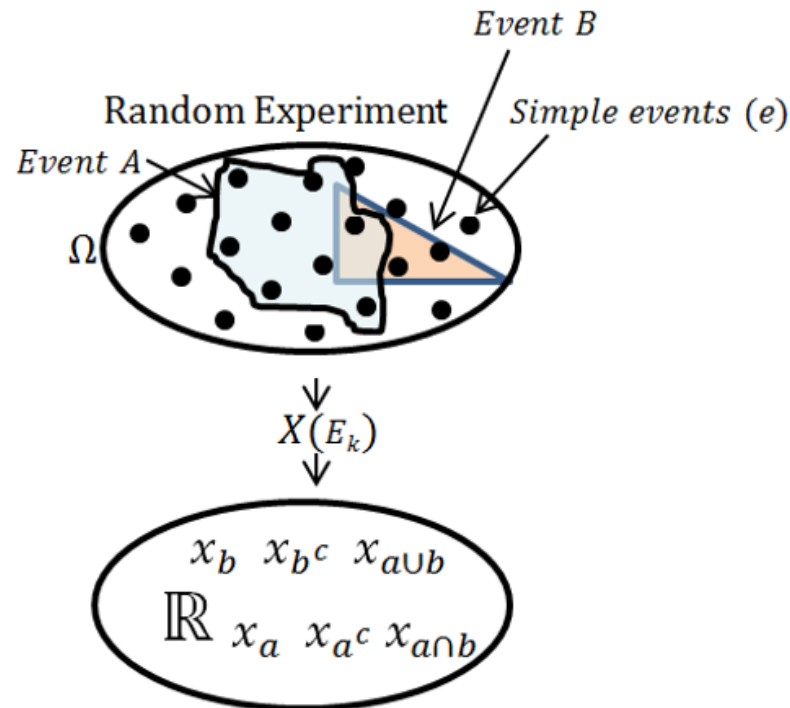
dado que $A \cap C = \emptyset$



Revisão de Probabilidade

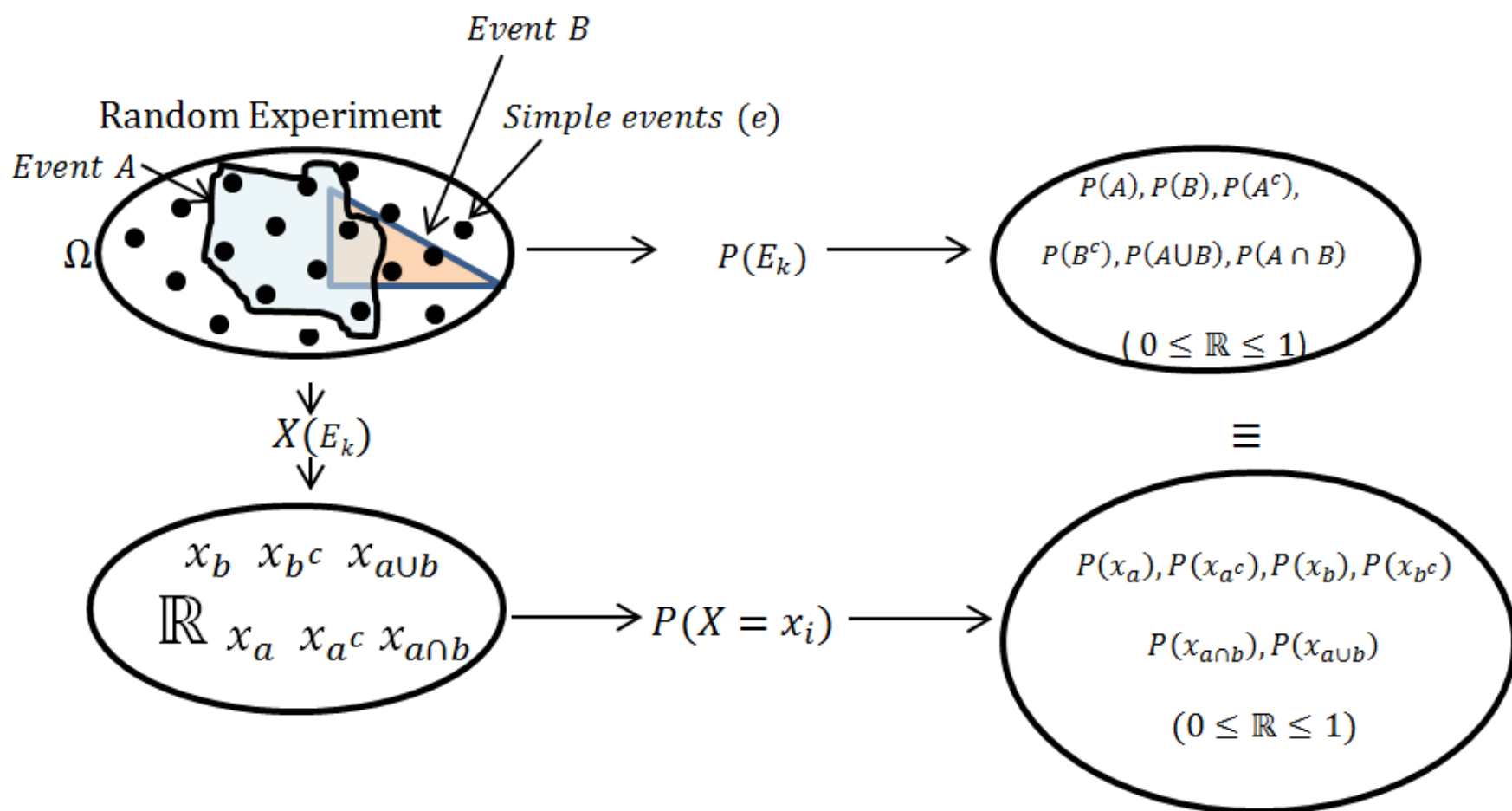
□ Variável Aleatória

É uma função que atribui um número real a resultados de um experimento aleatório. $X: \Omega \rightarrow \mathbb{R}$



Revisão de Probabilidade

□ Variável Aleatória

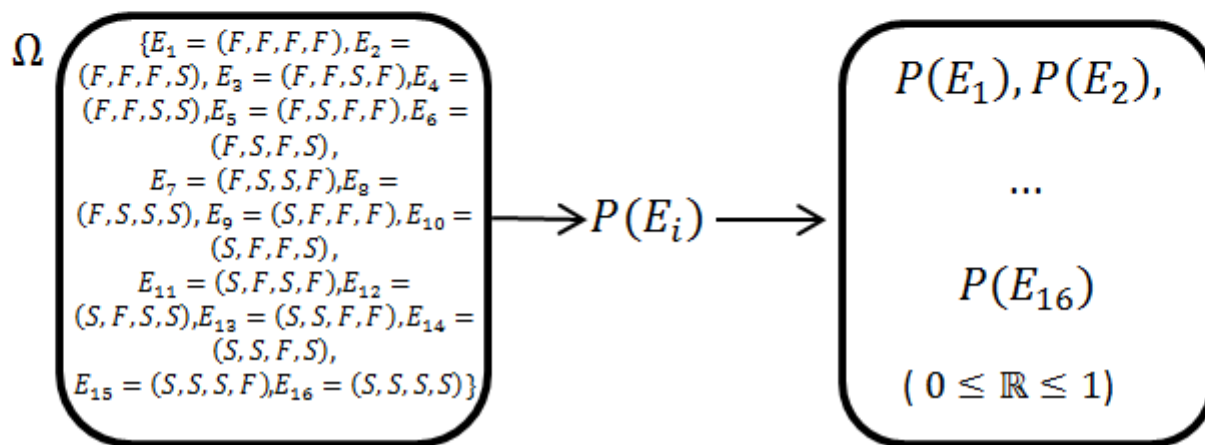


Revisão de Probabilidade

□ Variável Aleatória

Suponha que quatro servidores foram testados por um período. Após o teste, cada servidor é aprovado (S) ou reprovado (F). Representemos os quatro servidores por um vetor $E_i = (SS_1, SS_2, SS_3, SS_4)$, em que $SS_i = \{S, F\}$ representa o status do servidor i . Este teste é um experimento aleatório, pois o resultado não é conhecido antes da realização do experimento.

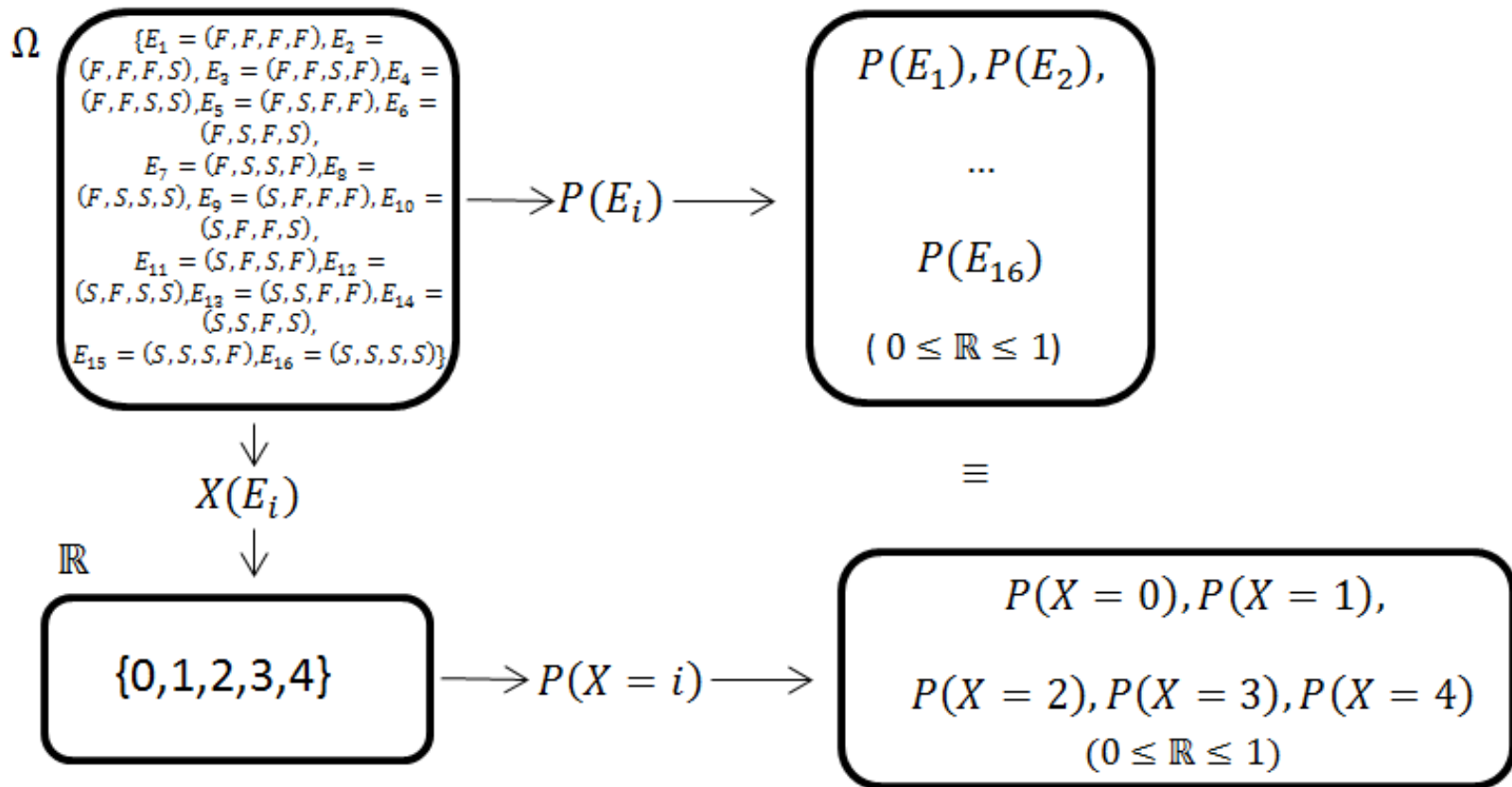
$$|\Omega| = 16$$



Revisão de Probabilidade

□ Variável Aleatória Discreta

Considere uma variável aleatória $X: \Omega \rightarrow \mathbb{R}$ que represente o número de servidores que passaram no teste.



Revisão de Probabilidade

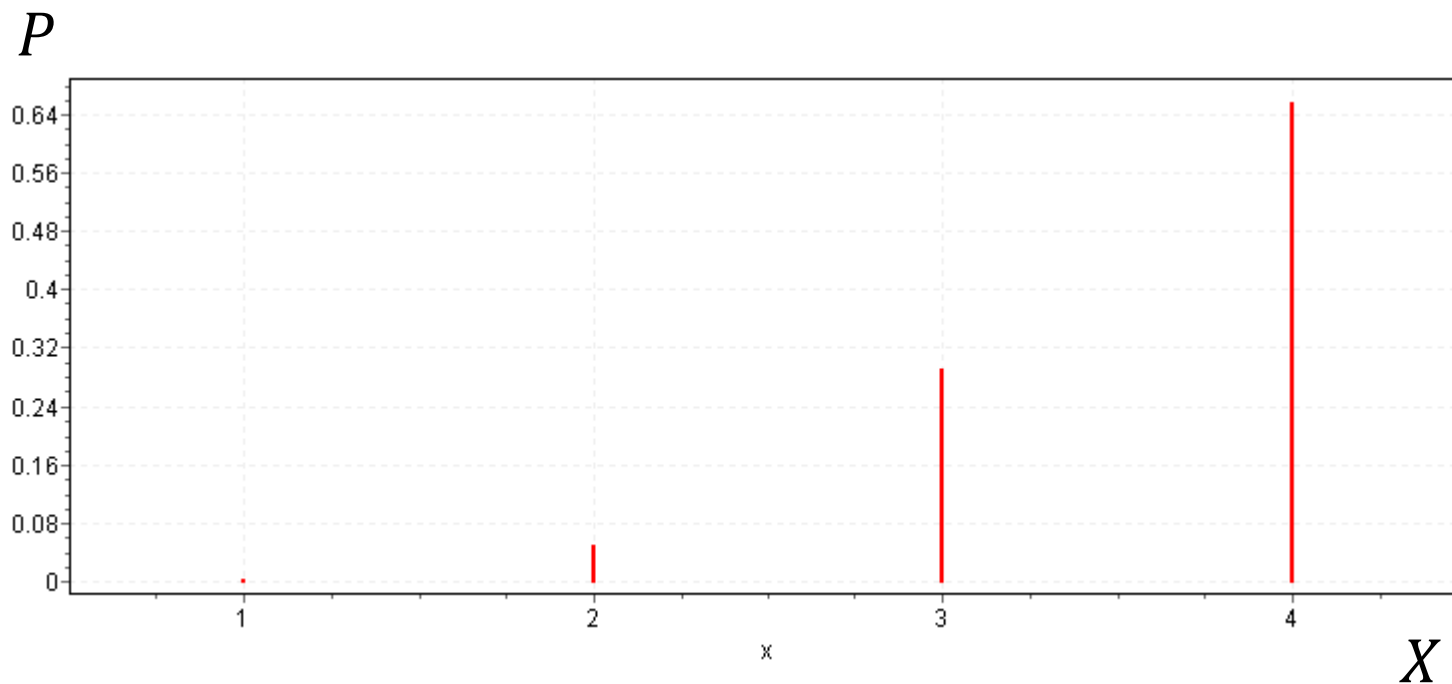
□ *Probability mass function (pmf)* – Seja Ω um espaço amostral discreto. $p(x)$ é uma função que atribui probabilidades aos valores (x) de uma variável aleatória X .

$$p(x) = P[X=x],$$

tal que x representa os valores atribuídos aos experimentos aleatórios de Ω .

Revisão de Probabilidade

□ *Probability mass function (pmf)*

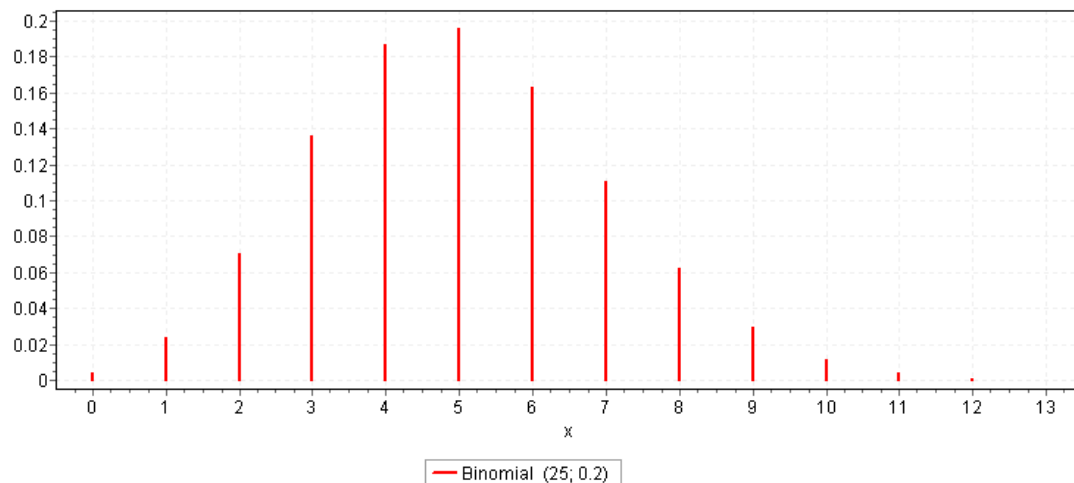
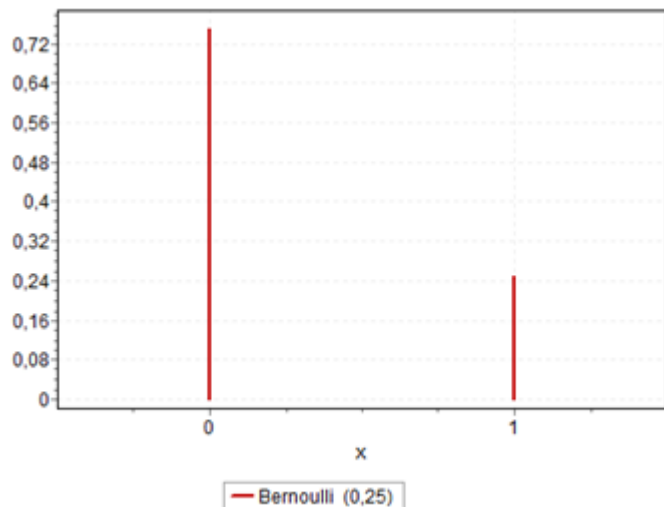


X — número de servidores
que passaram no teste

Revisão de Probabilidade

□ *Probability mass function (pmf)*

Exemplos de pmfs de V.A. discretas

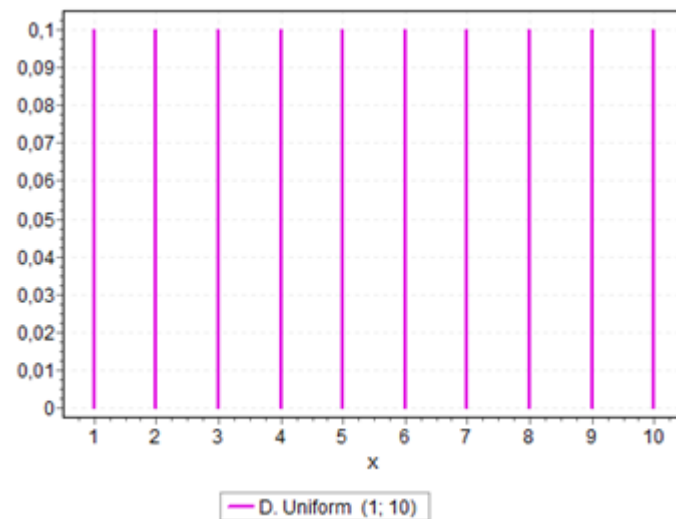
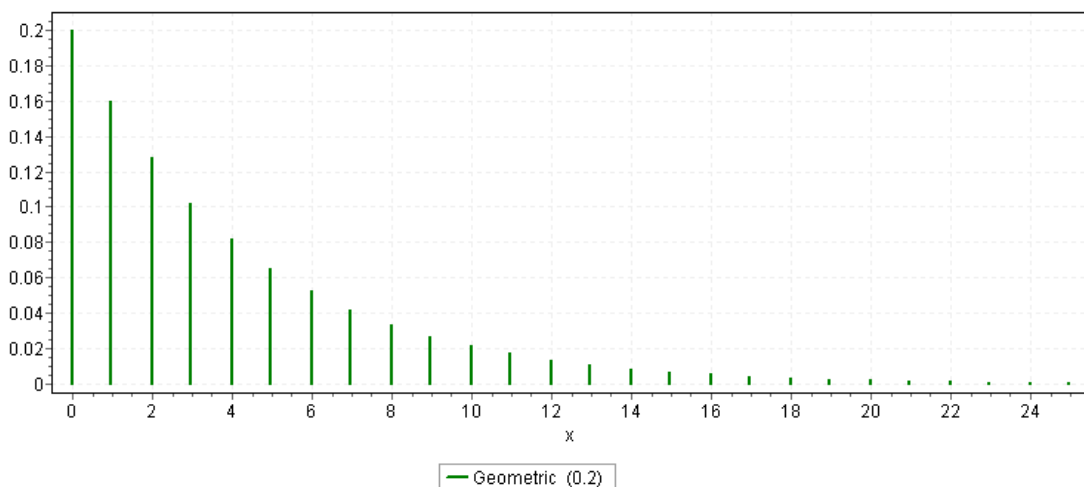


Essas distribuições serão apresentadas com detalhe posteriormente ao longo do curso.

Revisão de Probabilidade

□ *Probability mass function (pmf)*

Exemplos de pmfs de V.A discretas



Essas distribuições serão apresentadas com detalhe posteriormente ao longo do curso.

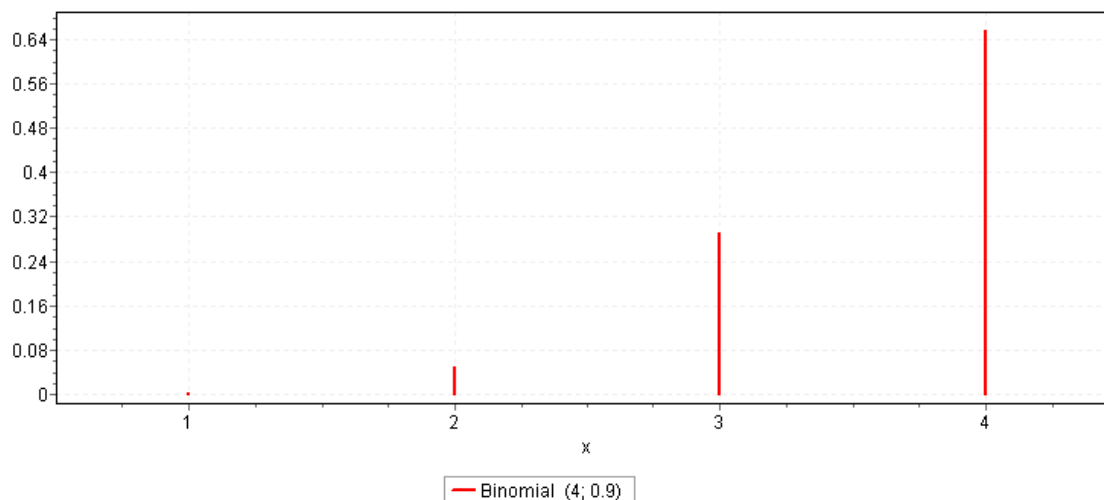
Revisão de Probabilidade

Portanto, se o estado de cada computador não interfere no estado dos demais e se considerarmos que $P(S) = p = 0.9$ é constante e igual para todos computadores, temos:

$$\begin{aligned} P(\text{de apenas 1 computador não falhar durante o teste}) &= P(E_2 \cup E_3 \cup E_5 \cup E_9) = \\ &= P(F, F, F, S) + P(F, F, S, F) + P(F, S, F, F) + P(S, F, F, F) \\ &= p(1-p)^3 + p(1-p)^3 + p(1-p)^3 + p(1-p)^3 = 4 \times p(1-p)^3 = 0.0036 \end{aligned}$$

\equiv

$$P(X = 1) = \binom{4}{1} p^1 (1-p)^3 = 4 \times p(1-p)^3 = 0.0036$$



Revisão de Probabilidade

□ Função de Distribuição de Probabilidade

Acumulativa (CDF) de uma variável aleatória X , denotada por $F(X)$, é definida por $F(X) = P[X \leq x] \forall x \in \mathbb{R}$

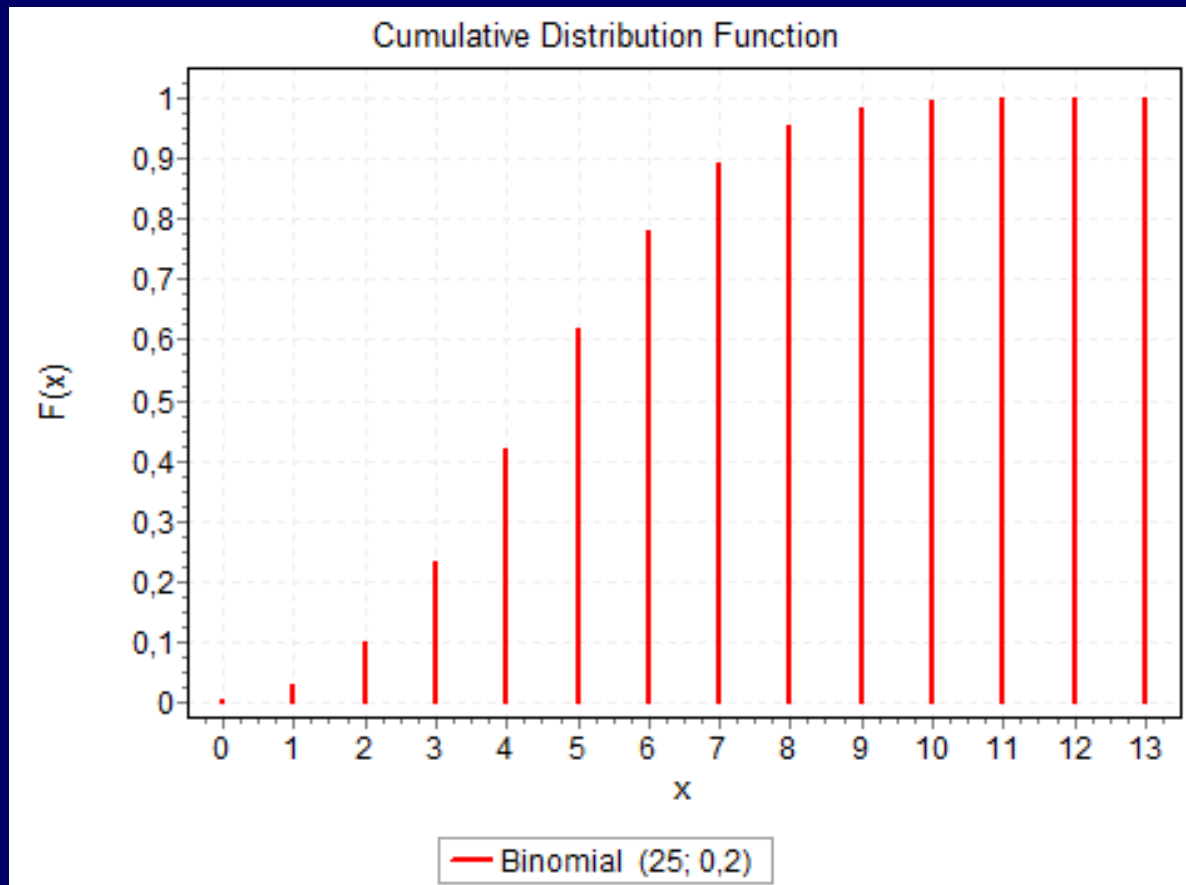
□ $F(X)$ é uma função monotônica não-decrescente tal que $0 \leq F(X) \leq 1$, onde $F(-\infty) = 0$ e $F(\infty) = 1$

□ $F(X) = \sum_{y \leq x} p(y) \Rightarrow F(\infty) = \sum_{\forall y} p(y) = 1$

Revisão de Probabilidade

EasyFit

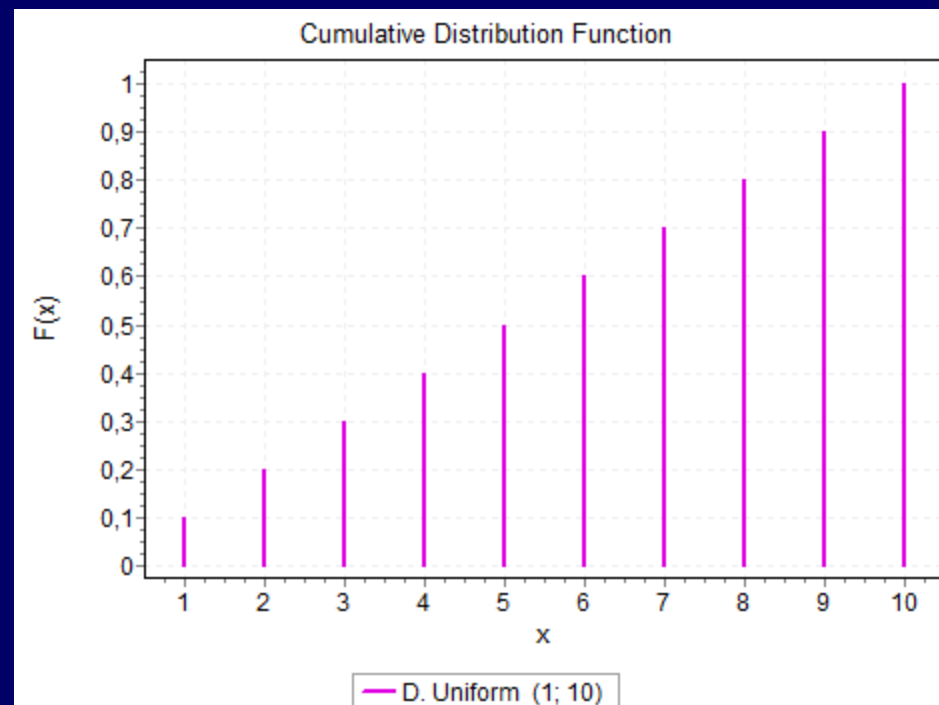
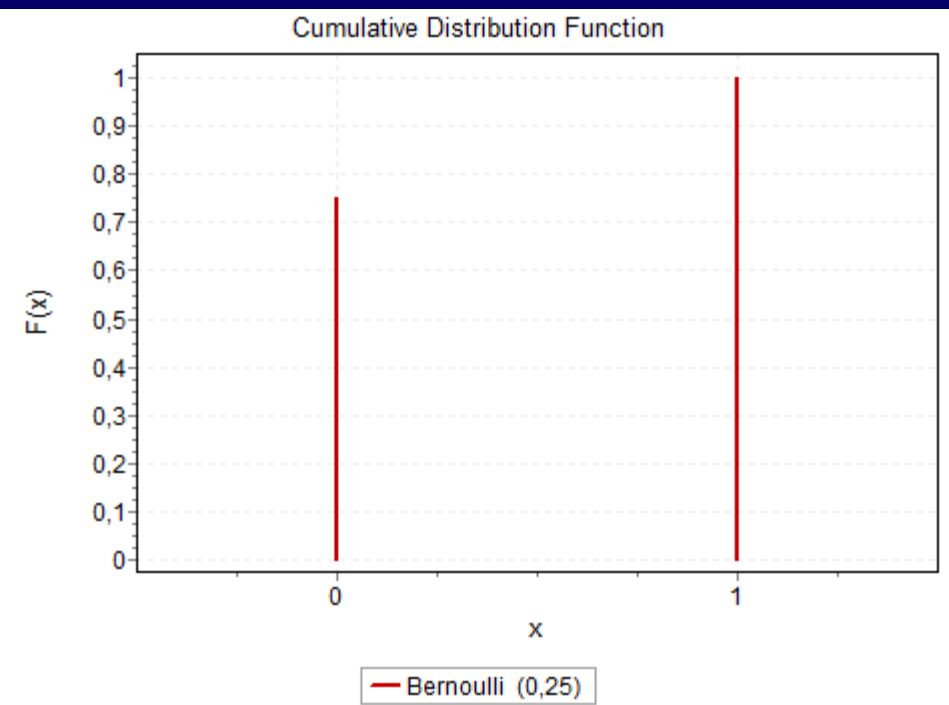
❑ Função de Distribuição de Probabilidade Acumulativa (CDF)



Revisão de Probabilidade

EasyFit

Função de Distribuição de Probabilidade Acumulativa (CDF)



Revisão de Probabilidade

□ Variável Aleatória Contínua

Exemplo: o tempo de resposta a solicitação de um serviço é uma variável aleatória contínua.

Como $F(x)$ não é decrescente,

$$F(x) = \int_{-\infty}^{x_1} f(x) dx$$

$$f(x) \geq 0$$

$$P[a < X \leq b] = F(b) - F(a) = \int_a^b f(x) dx$$

Para variáveis aleatórias contínuas, a

Função de Densidade de Probabilidade (pdf), $f(x)$, é definida por:

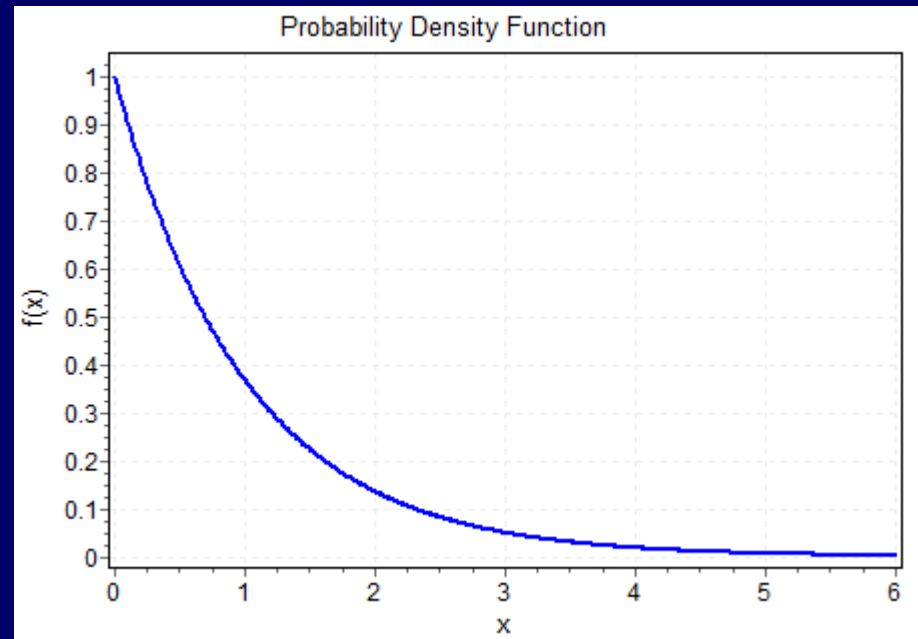
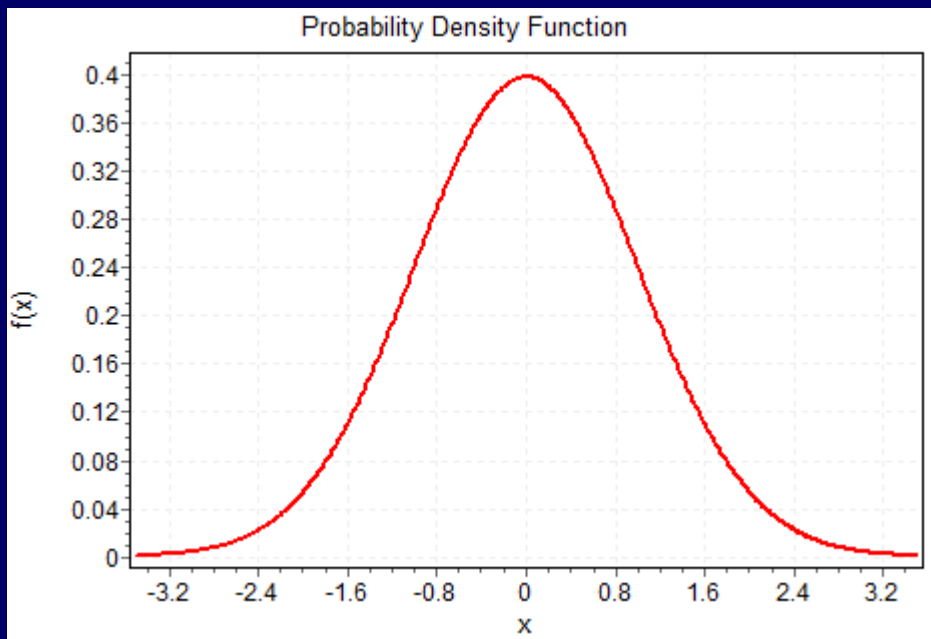
$$f(x) = dF(x)/dx, \quad \text{onde}$$

$F(x)$ é a Função de Distribuição de Probabilidade Acumulativa

Revisão de Probabilidade

EasyFit

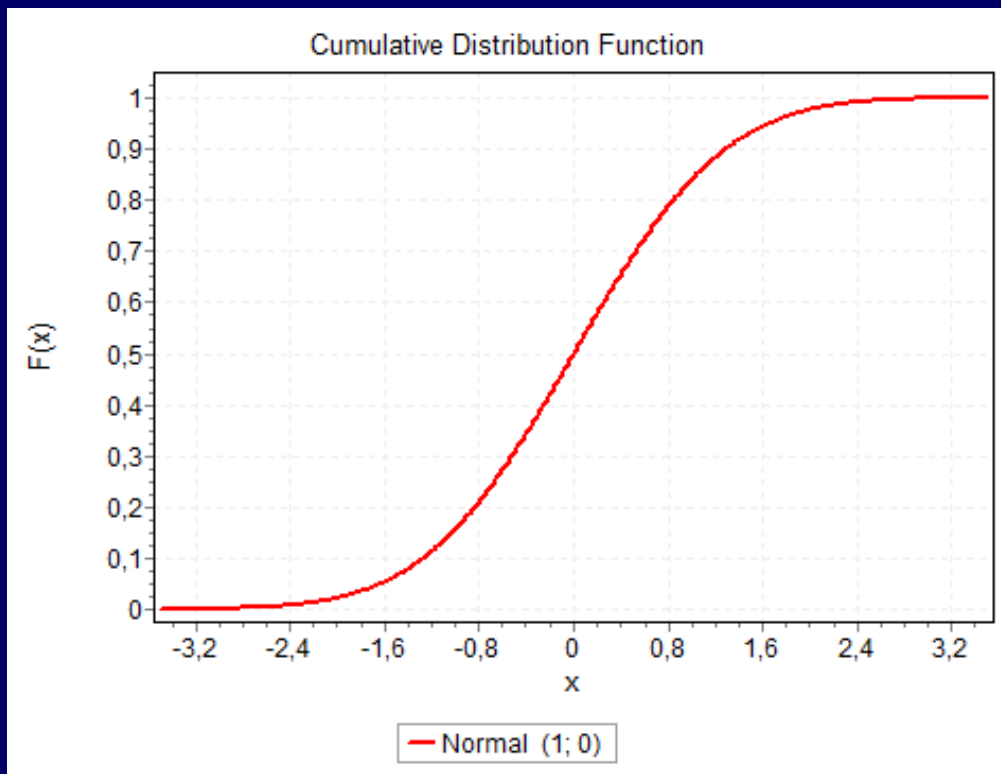
- Função de Densidade de Probabilidade $f(x)$, é definida por:
- $$f(x) = dF(x)/dx$$



Revisão de Probabilidade

□ Distribuição de Probabilidade Acumulativa

$$F(x) = \int_{-\infty}^{x_1} f(x)dx = P[a < X \leq b] = F(b) - F(a) = \int_a^b f(x)dx$$

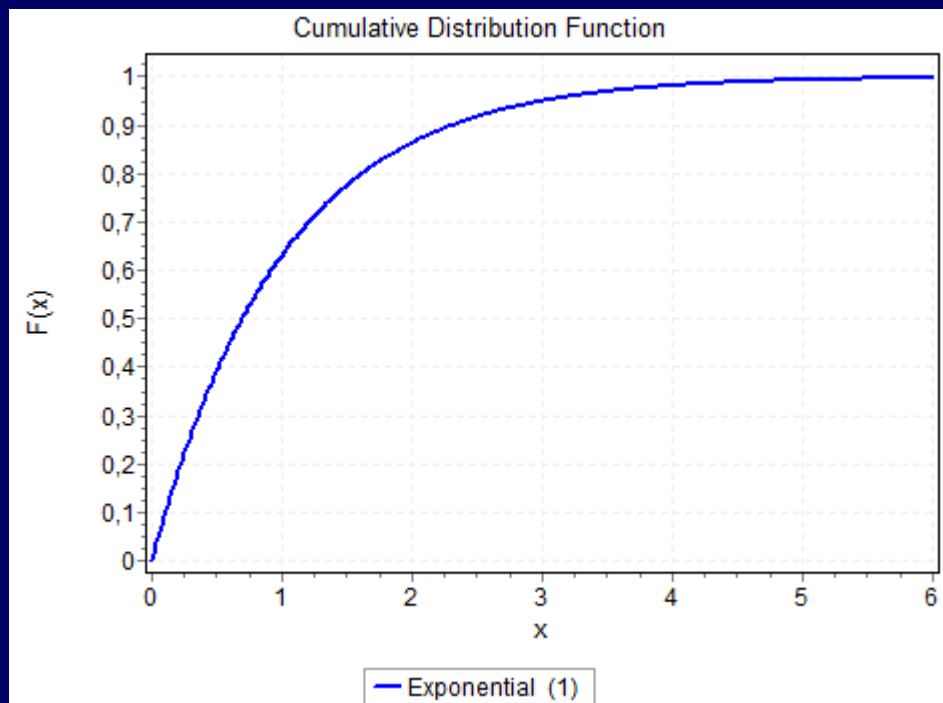
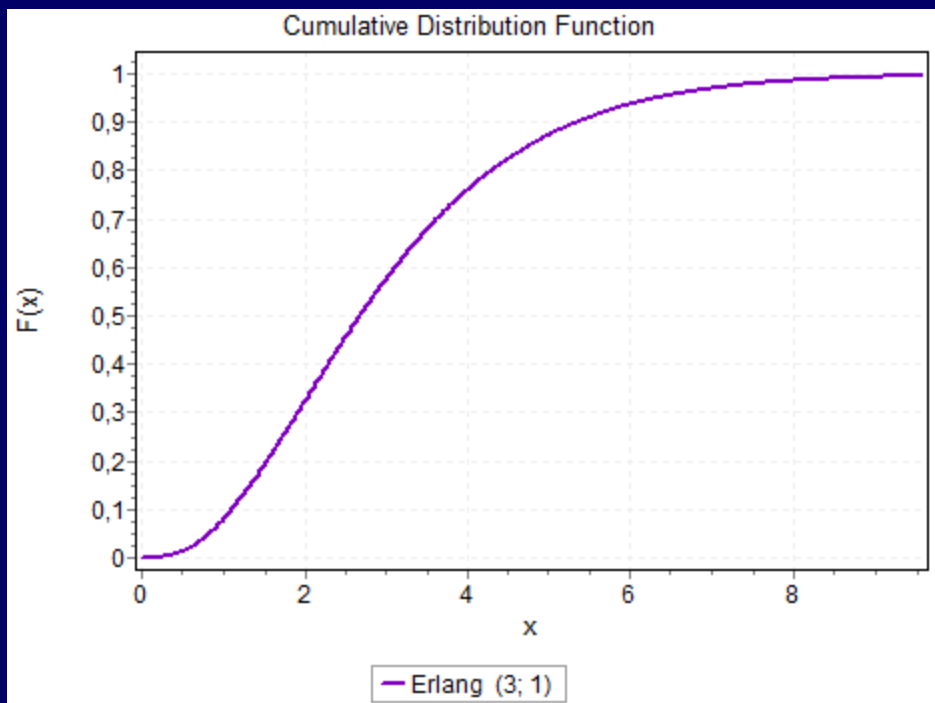


Revisão de Probabilidade

EasyFit

□ Distribuição de Probabilidade Acumulativa

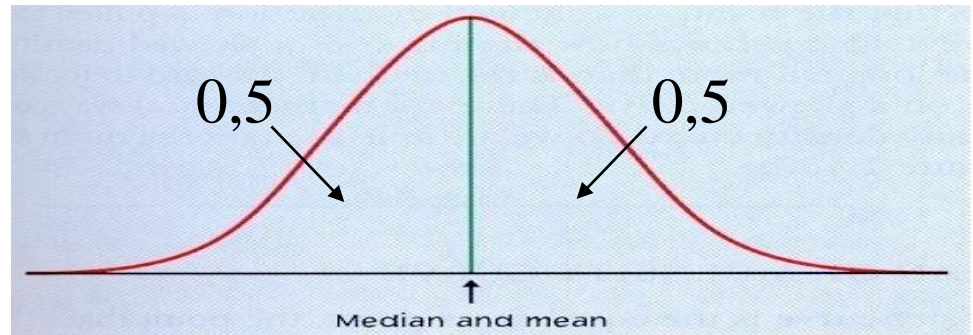
$$F(x) = \int_{-\infty}^{x_1} f(x)dx = P[a < X \leq b] = F(b) - F(a) = \int_a^b f(x)dx$$



Distribuições Normais

As curvas de densidade de probabilidade Normais são:

- simétricas,
- unimodais,
- em forma de sino,
- média, mediana e moda são iguais.
- $IIQ/\sigma \sim \frac{4}{3}$



Todas as distribuições normais têm a mesma forma global e são descritas pela média μ e o desvio-padrão σ .

Distribuições Normais

Função de Densidade das Distribuições Normais

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2 / 2\sigma^2}$$

onde:

x = cada valor da variável aleatória contínua que $-\infty < x < \infty$.

σ = desvio-padrão da população

e = **Base do logaritmo natural = 2.7183**

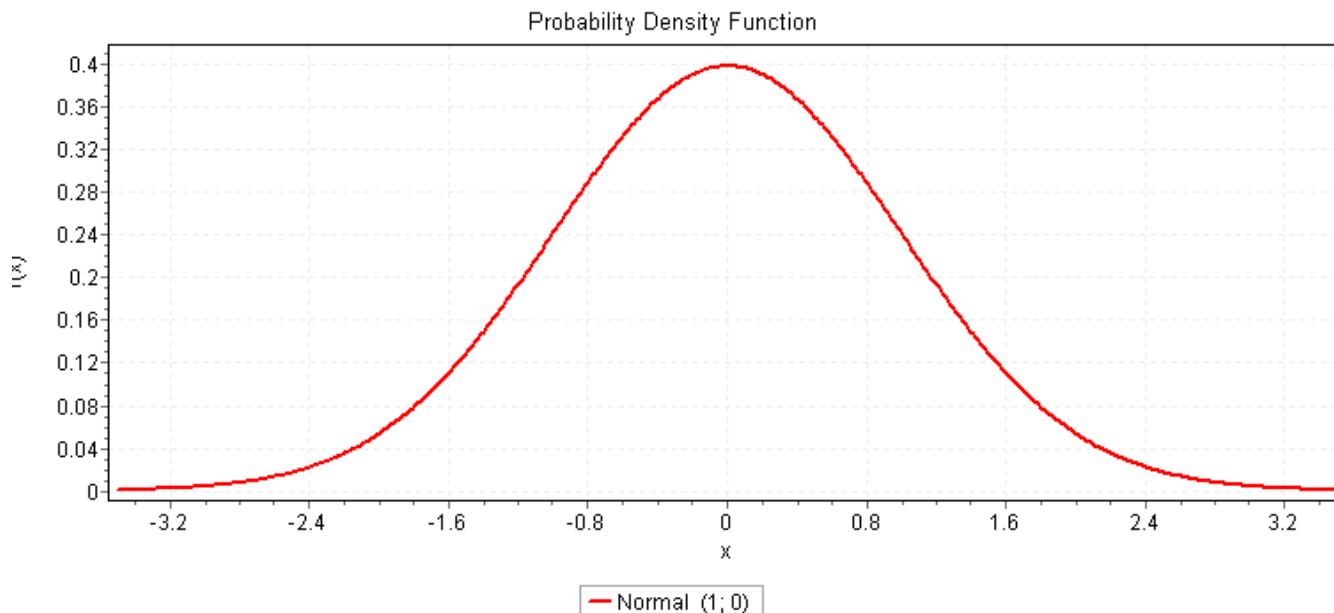
μ = média da população

Notação: $N(\mu, \sigma)$

Distribuição Normal

Distribuição Normal Padronizada

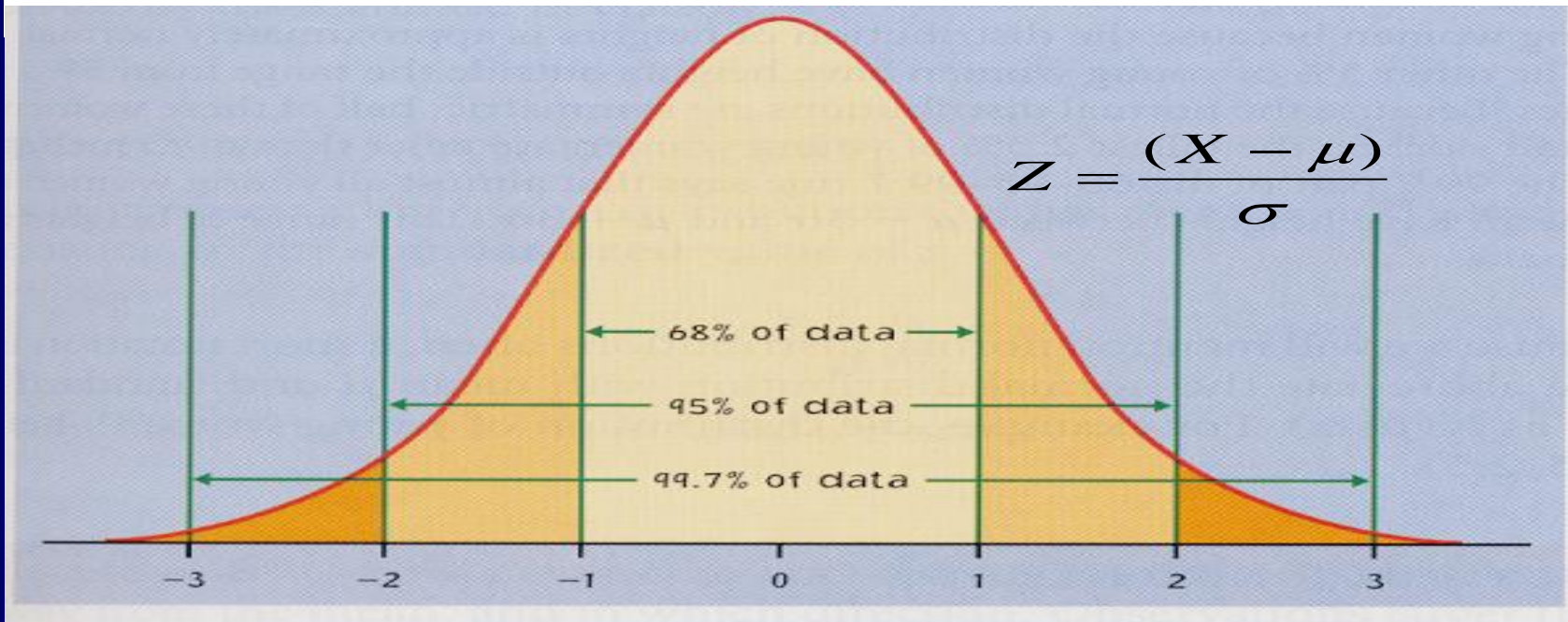
É uma distribuição normal com média 0 e desvio padrão 1.
 $N(0,1) = Z$



Distribuições Normais

Padronização

- Expressando valores em termos das distâncias da média.
- Distância medida em desvios-padrão.
- Valor padronizado (Z)



Distribuição Normal

Distribuição Normal Padronizada

Ex.:

- ❑ O tempo de execução de uma função A1 é descrito pela seguinte distribuição $N(64.5 \text{ s}, 2.5 \text{ s})$
- ❑ Qual a probabilidade da função ser executada em menos que 65,8 s?
- ❑ Qual é o valor padronizado de uma execução igual a 65,8 s?

$$Z = \frac{(X - \mu)}{\sigma}$$

Distribuição Normal

Distribuição Normal Padronizada

- Quando se padronizam, todas as distribuições normais têm

$$\mu = 0$$

$$\sigma = 1$$

- A padronização de uma variável com distribuição normal $N(\mu, \sigma)$ produz uma nova variável com distribuição normal padronizada $N(0, 1)$.

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

Área sob a Curva Normal

0,52

Aplicar no
Statistica e
calcular com
Z

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990

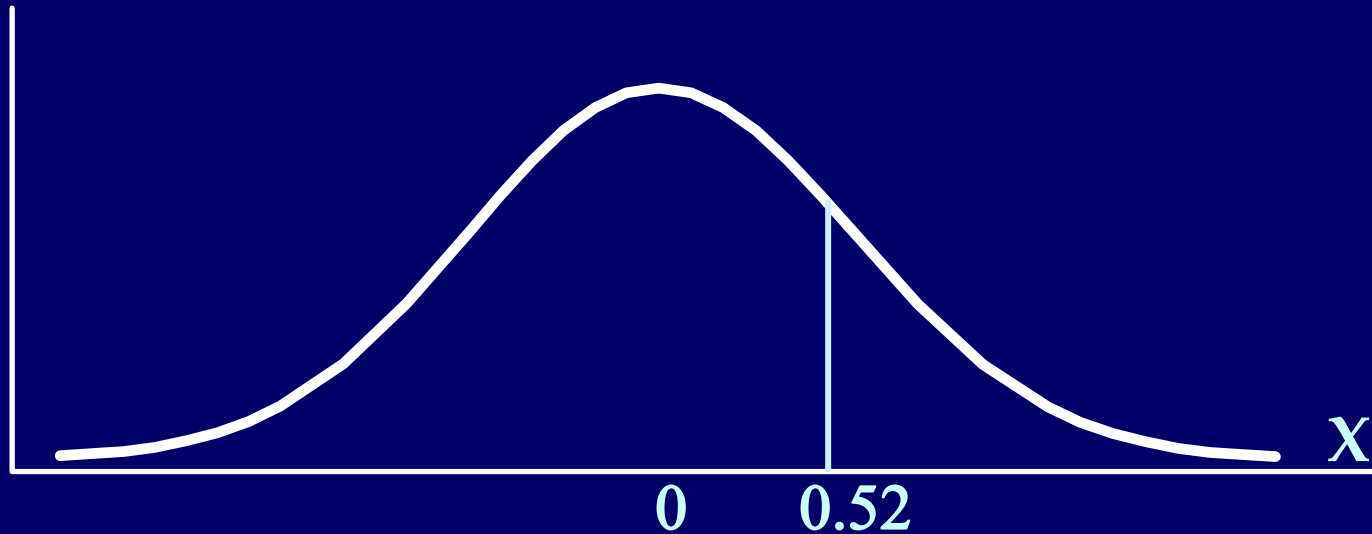
0,5 + valor =
0,6985

Distribuição Normal

Aplicar no
Statística e
calcular com
Z

Área sob a Curva Normal

0.6985



Exemplo:

$z = 0.52$ (ou -0.52)

$A(z) = 0.6985$ ou 69.85%

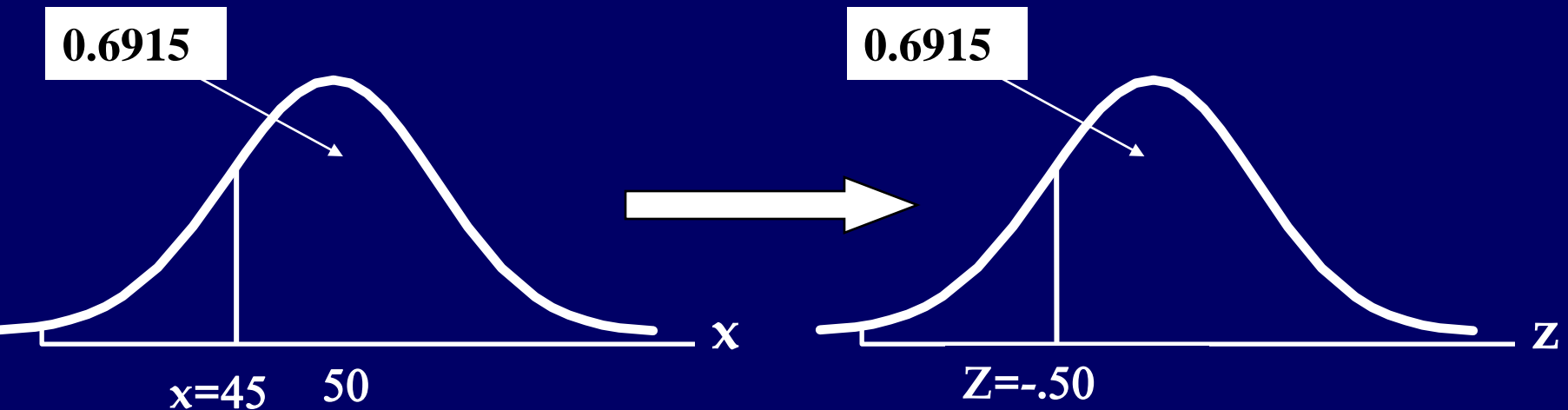
Distribuição Normal

Aplicar no
Statística e
calcular com
Z e X

Normal Padrão

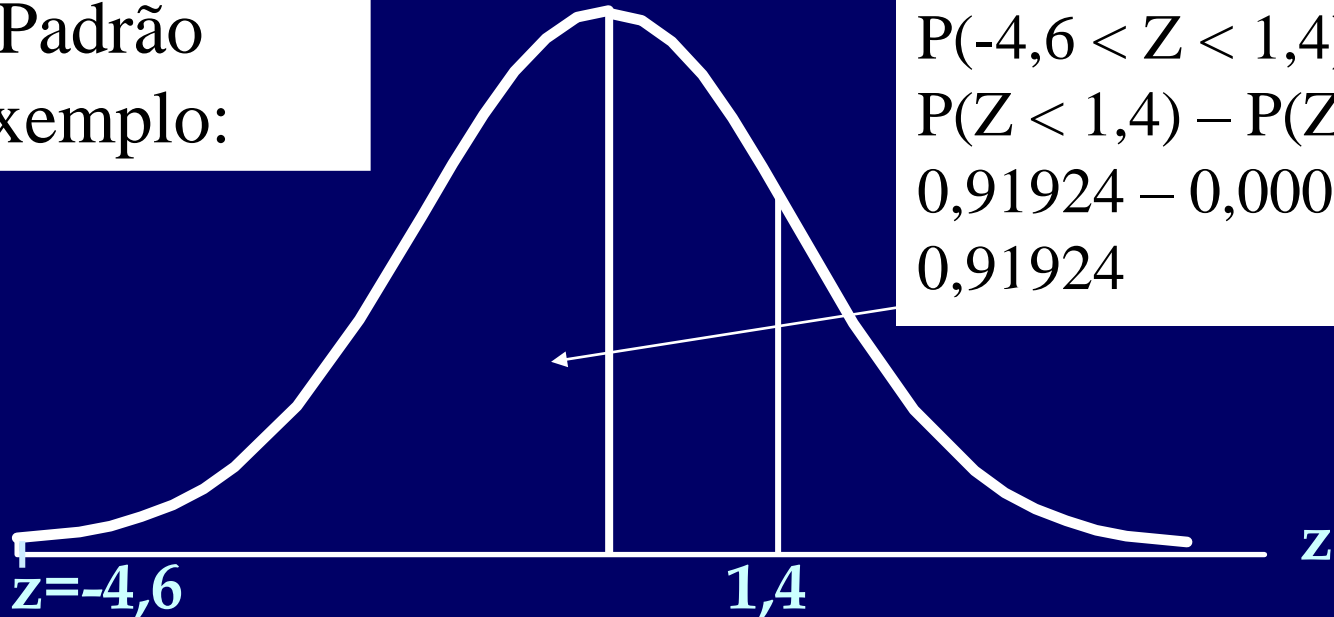
Exemplo:

$$z = \frac{x - \mu}{\sigma} = \frac{45 - 50}{10} = -0.50$$



Distribuição Normal

Normal Padrão
Exemplo:



$$\mu=0,2508$$

$$\sigma=0,0005$$

$$P(0,2485 < X < 0,2515) =$$

$$P(-4,6 < Z < 1,4) =$$

$$P(Z < 1,4) - P(Z < -4,6) =$$

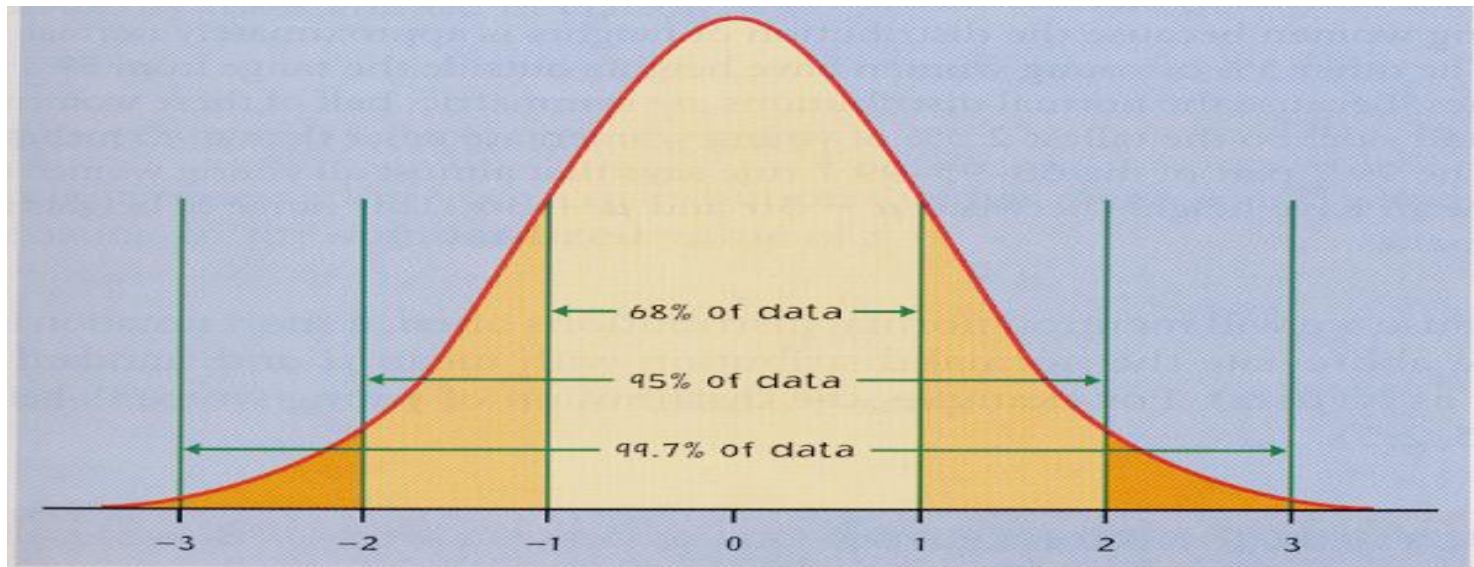
$$0,91924 - 0,0000 =$$

$$0,91924$$

Distribuições Normais

Regra 68-95-99.7

- Curva normal com média μ e desvio padrão σ
 - 68% das observações estão entre $\pm 1 \sigma$
 - 95% das observações estão entre $\pm 2 \sigma$
 - 99.7% das observações estão entre $\pm 3 \sigma$



Média Amostral e Variância da Média Amostral

- Considere X_1, X_2, \dots, X_n variáveis aleatórias mutuamente independentes e identicamente distribuídas.
- Seja (x_1, x_2, \dots, x_n) uma n-tupla de valores onde x_i é um valor específico de X_i .
- Também se diz que (x_1, x_2, \dots, x_n) é um experimento aleatório de tamanho n .
- A estatística amostral $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ é a média

Média Amostral e Variância da Média Amostral

Se a população tem média $E[X_i] = \mu$ e variância $Var[X_i] = \sigma^2$, a média de \bar{X} (média das médias) é

$$E[\bar{X}] = \frac{1}{n} \sum_{i=1}^n E[X_i] = \frac{1}{n} n E[X_i] = \frac{1}{n} n \mu = \mu$$

$$Var[\bar{X}] = E[(\bar{X} - \mu)^2] = E\left[\left(\frac{1}{n} \sum_{i=1}^n (\bar{X}_i - \mu)\right)^2\right] =$$

$$Var[\bar{X}] = \left[\frac{1}{n^2} \sum_{i=1}^n E(\bar{X}_i - \mu)^2 \right] =$$

$$Var[\bar{X}] = \frac{1}{n^2} (n\sigma^2) = \frac{\sigma^2}{n}$$

Média Amostral e Variância da Média Amostral

Quando se obtém amostras grandes, a média das amostras obedece (aproximadamente) a distribuição normal com média μ e desvio padrão da média amostral (erro padrão) igual a

$$\sigma_{\bar{x}} = \sigma / \sqrt{n}$$

Estimador

□ Estimador:

Qualquer estatística $\hat{\Theta} = \hat{\Theta}(X_1, X_2, \dots, X_n)$ usada para estimar o valor de um parâmetro θ de uma população é denominado estimador de θ .

Um estimador $\hat{\Theta} = \hat{\Theta}(X_1, X_2, \dots, X_n)$ é dito não-viesado de um parâmetro θ se $E[\hat{\Theta}(X_1, X_2, \dots, X_n)] = \theta$.

Média Amostral – Estimador não viesado da média populacional

- A média amostral \bar{X} é um estimador não-viesado da média da população μ (quando esta existe)

$$\begin{aligned} E[\bar{X}] &= E\left[\frac{1}{n} \sum_{i=1}^n X_i\right] \\ &= \frac{1}{n} \sum_{i=1}^n E[X_i] \\ &= \frac{1}{n} \sum_{i=1}^n E[X] \\ &= \frac{1}{n} n E[X] \\ &= E[X] \\ &= \mu. \end{aligned}$$

Variância da Amostra - Estimador não viesado da variância populacional



□ A variância amostral

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

é um estimador não-viesado da variância da σ^2 população
(quando esta existe)

$$\begin{aligned} S^2 &= \frac{1}{n-1} \sum_{i=1}^n (X_i^2 - 2X_i\bar{X} + \bar{X}^2) \\ &= \frac{1}{n-1} \left(\sum_{i=1}^n X_i^2 \right) - \frac{2n}{n-1} \left(\frac{1}{n} \sum_{i=1}^n X_i \right) \bar{X} + \frac{n\bar{X}^2}{n-1} \\ &= \frac{1}{n-1} \sum_{i=1}^n X_i^2 - \frac{n}{n-1} \bar{X}^2 \end{aligned}$$

Variância da Amostra - Estimador não viesado da variância populacional

Portanto

$$E[S^2] = \frac{1}{n-1} \sum_{i=1}^n E[X_i^2] - \frac{n}{n-1} E[\bar{X}^2].$$

No entanto

$$E[X_i^2] = \text{Var}[X_i] + (E[X_i])^2 = \sigma^2 + \mu^2$$

Expansão de Taylor

E

$$E[\bar{X}^2] = \text{Var}[\bar{X}] + (E[\bar{X}])^2 = \frac{\sigma^2}{n} + \mu^2.$$

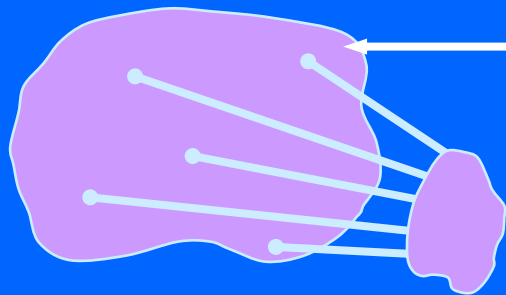
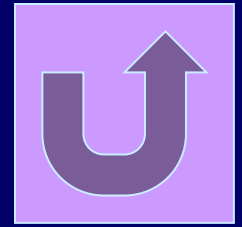
Variância da média amostral

Desta forma

$$\begin{aligned} E[S^2] &= \frac{1}{n-1} n(\sigma^2 + \mu^2) - \frac{n}{n-1} \left(\frac{\sigma^2}{n} + \mu^2 \right) \\ &= \sigma^2. \end{aligned}$$

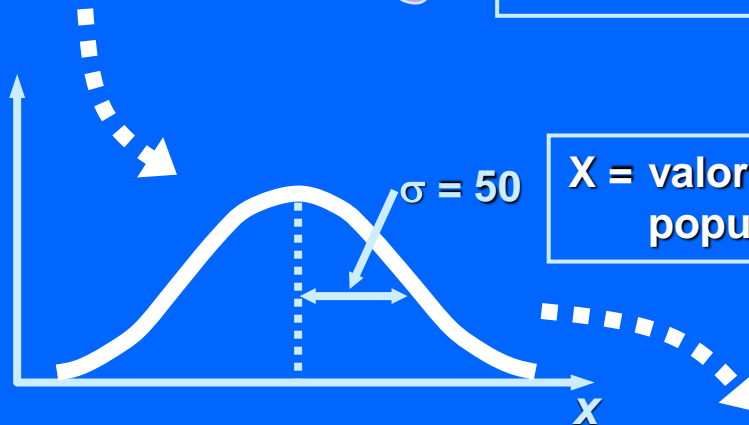
Teorema do Limite Central

(Distribuição de \bar{X})



População (média = μ ,
desvio padrão = σ)

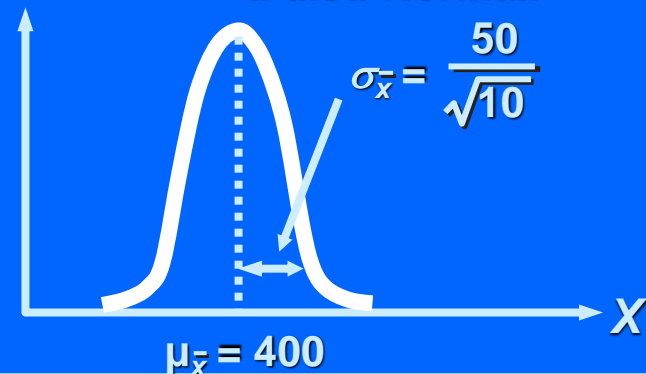
Amostra aleatória (média = \bar{X} ,
desvio padrão = s)



X = valor desta
população

Assume-se que as
observações
individuais obedecem
a dist. Normal.

\bar{X} segue a distribuição Normal, centrado
em μ com desvio padrão σ/\sqrt{n}



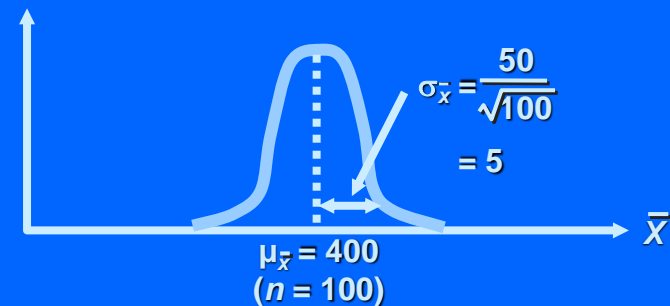
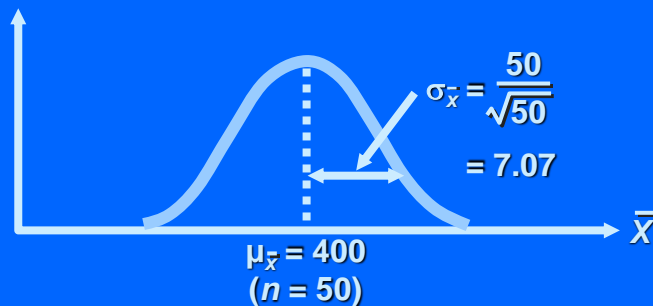
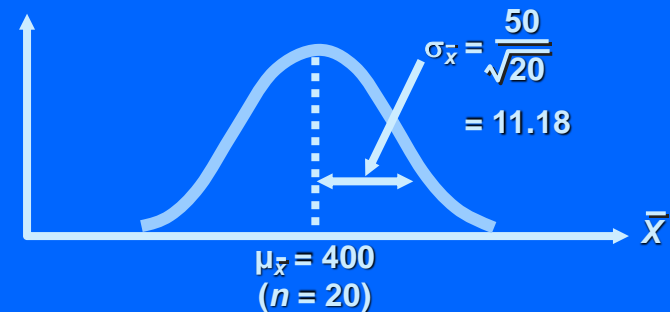
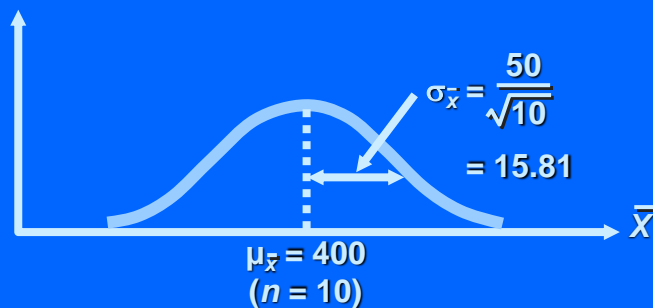
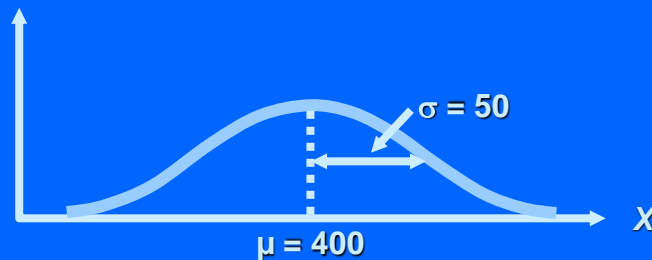
Teorema do Limite Central

(Distribuição de \bar{X})

Applet

[http://onlinestatbook.com/
stat_sim/
sampling_dist/index.html](http://onlinestatbook.com/stat_sim/sampling_dist/index.html)

População



Teorema do Limite Central

(Distribuição de \bar{X})

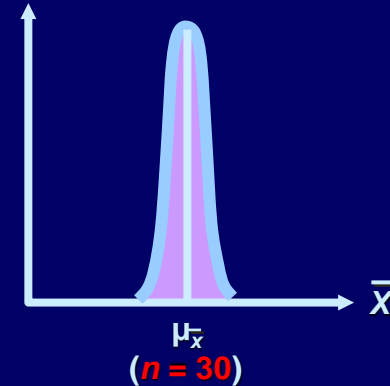
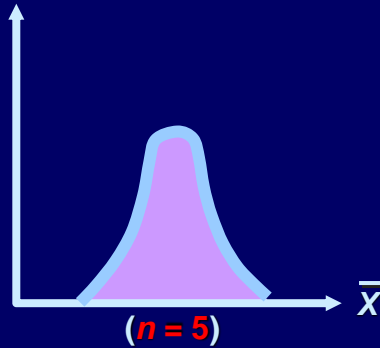
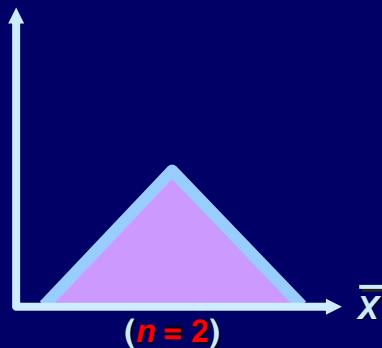
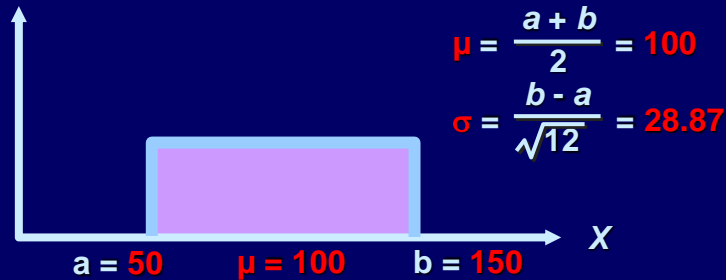
$$\text{Média} = \mu_{\bar{x}} = \mu$$

$$\begin{array}{l} \text{Desvio padrão} \\ \text{(erro padrão)} \end{array} = \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Teorema do Limite Central

(Distribuição de \bar{X})

População Uniforme



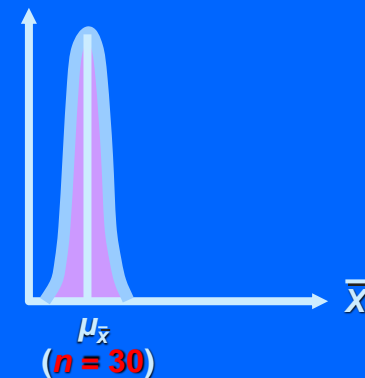
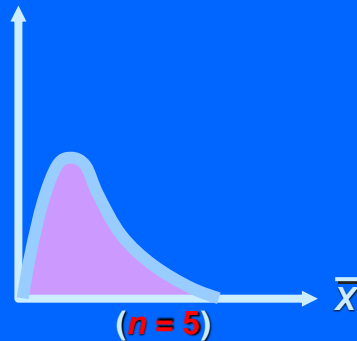
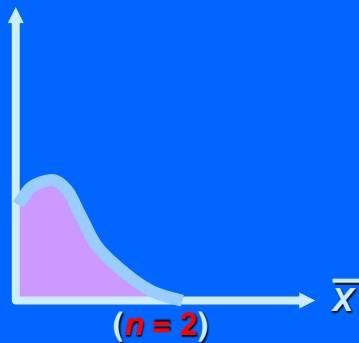
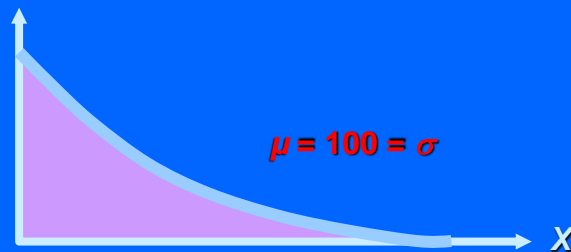
Pelo TCL, tem-se:

$$\mu_{\bar{X}} = \mu = 100$$
$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{28.87}{\sqrt{30}} = 5.27$$

Teorema do Limite Central

(Distribuição de \bar{X})

**População
Exponencial**



Pelo TCL, tem-se: $\mu_{\bar{X}} = \mu = 100$

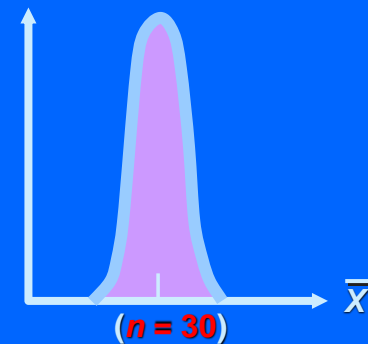
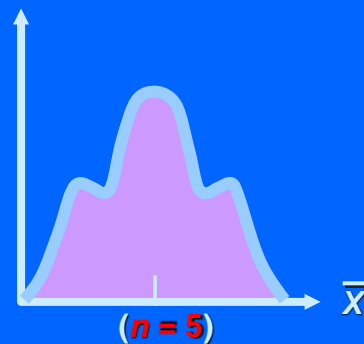
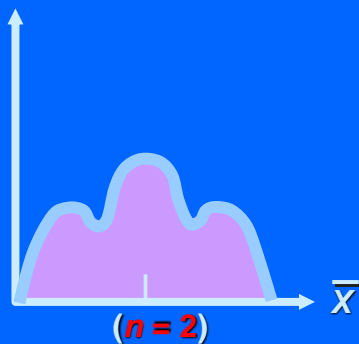
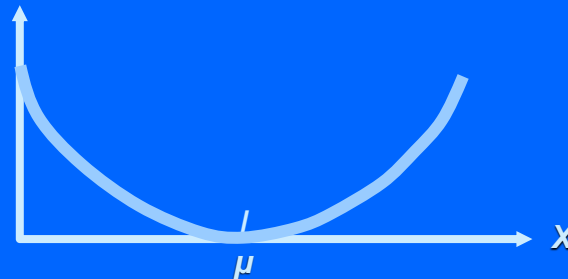
$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{100}{\sqrt{30}} = 18.26$$

Teorema do Limite Central

(Distribuição de \bar{X})

Demonstração
Minitab

**População
com dist.
em forma de U**



Inferência

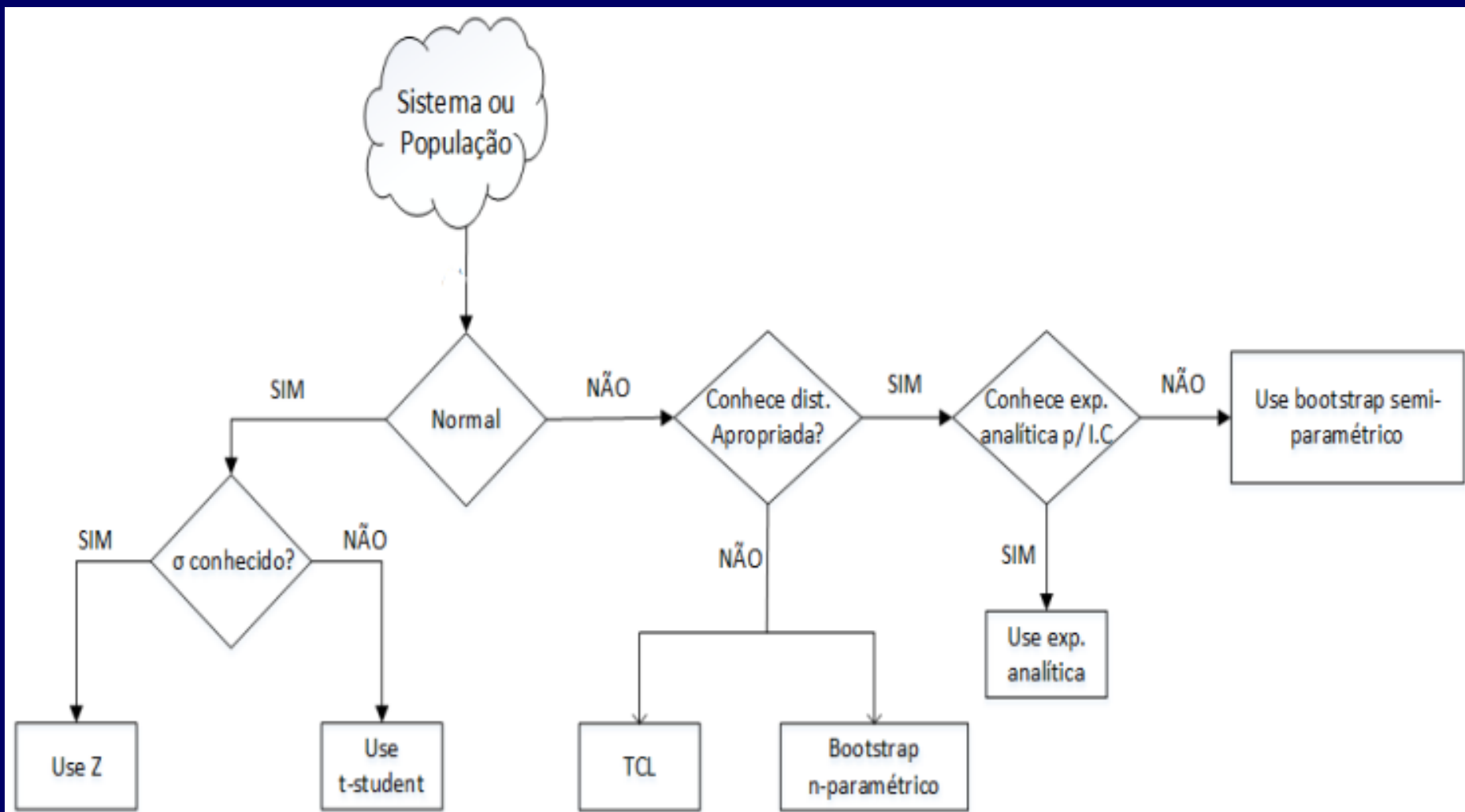
- Erro Tipo 1 (Erro do Consumidor)

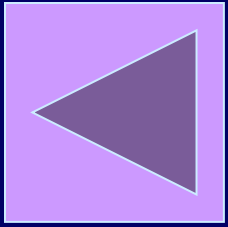
Rejeição de hipótese (H_0 – hipótese nula), quando esta é verdadeira (α - também chamado de nível de significância).

- Erro Tipo 2 (Erro do Produtor)

Falha em rejeitar hipótese (H_0 – hipótese nula), quando ela é falsa (β).

Orientação para Inferência – Média

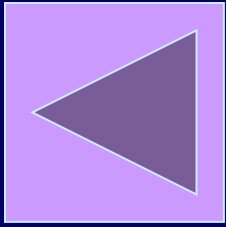




Inferência

□ Intervalo de Confiança

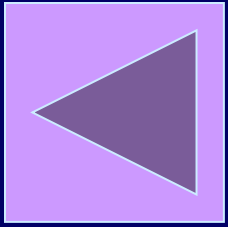
- $P(l \leq \mu \leq u) = 1 - \alpha$ (coeficiente de confiança).
- $l \leq \mu \leq u$ – Intervalo de confiança.



Inferência

□ Intervalo de Confiança

- Seja X uma variável aleatória com media μ e variância σ^2 .
- Suponha que seja extraída uma amostra aleatória de tamanho n , X_1, X_2, \dots, X_n .



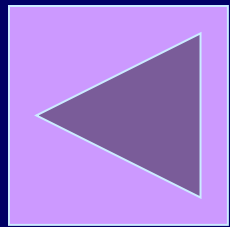
Inferência

□ Intervalo de Confiança

- Pode-se obter um intervalo de confiança de $100(1 - \alpha)\%$ de confiança para μ , considerando-se a distribuição amostral da média amostral \bar{X} .
- Observamos que a distribuição de \bar{X} é normal se X for normal ou aproximadamente normal se as condições do Teorema do Limite Central do forem verificadas.
- A média amostral de X tende para μ e a variância da média amostral é σ^2/\sqrt{n}
- Assim, a distribuição da estatística
$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

é tomada como a distribuição normal-padrão.

Inferência



Considerando a estatística:

$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

Como $E[Z] = z = 0$, e considerando:

Temos: $P(-z_{\alpha/2} \leq z \leq z_{\alpha/2}) = 1 - \alpha$

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

$$-z_{\alpha/2} \leq \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}$$

$$-z_{\alpha/2} \times \left(\frac{\sigma}{\sqrt{n}}\right) \leq \bar{x} - \mu \leq z_{\alpha/2} \times \left(\frac{\sigma}{\sqrt{n}}\right)$$

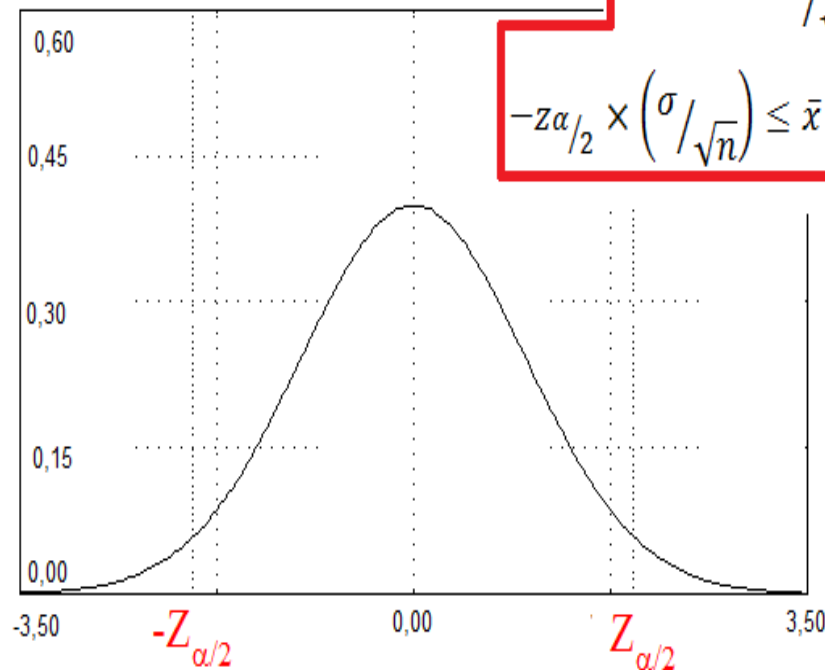
$$-z_{\alpha/2} \times \left(\frac{\sigma}{\sqrt{n}}\right) - \bar{x} \leq -\mu \leq z_{\alpha/2} \times \left(\frac{\sigma}{\sqrt{n}}\right) - \bar{x}$$

$$\times (-1) \quad z_{\alpha/2} \times \left(\frac{\sigma}{\sqrt{n}}\right) - \bar{x} \leq -\mu \leq z_{\alpha/2} \times \left(\frac{\sigma}{\sqrt{n}}\right) - \bar{x}$$

$$z_{\alpha/2} \times \left(\frac{\sigma}{\sqrt{n}}\right) + \bar{x} \geq \mu \geq -z_{\alpha/2} \times \left(\frac{\sigma}{\sqrt{n}}\right) + \bar{x}$$

$$\bar{x} + z_{\alpha/2} \times \left(\frac{\sigma}{\sqrt{n}}\right) \geq \mu \geq \bar{x} - z_{\alpha/2} \times \left(\frac{\sigma}{\sqrt{n}}\right)$$

$$\bar{x} - z_{\alpha/2} \times \left(\frac{\sigma}{\sqrt{n}}\right) \leq \mu \leq \bar{x} + z_{\alpha/2} \times \left(\frac{\sigma}{\sqrt{n}}\right)$$

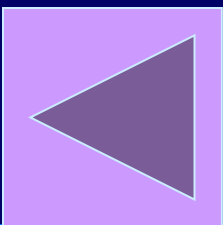


Como $z^* = -z_{\alpha/2}$, portanto:

$$\bar{x} - z^* \times \left(\frac{\sigma}{\sqrt{n}}\right) \leq \mu \leq \bar{x} + z^* \times \left(\frac{\sigma}{\sqrt{n}}\right)$$

Ou

$$\left(\bar{x} - z^* \times \left(\frac{\sigma}{\sqrt{n}}\right) , \bar{x} + z^* \times \left(\frac{\sigma}{\sqrt{n}}\right) \right)$$



Inferência

- Intervalo de Confiança para Média populacional com desvio-padrão conhecido

$$\bar{X} - Z^* \times \left(\sigma / \sqrt{n} \right) \leq \mu \leq \bar{X} + Z^* \times \left(\sigma / \sqrt{n} \right)$$

Ou

$$\left(\bar{X} - Z^* \times \left(\sigma / \sqrt{n} \right) , \bar{X} + Z^* \times \left(\sigma / \sqrt{n} \right) \right)$$

Valor crítico ($Z_{\alpha/2}$) da tabela da distribuição t quando o GL tende para infinito (última linha da tabela), considerando um dado nível de significância (α)

□ Intervalo de Confiança

- Exemplo: suponha que um conjunto de atividades, denominado aqui por A1, executadas por um departamento de uma organização seja normalmente distribuído com desvio padrão $\sigma=25\text{min}$. Uma amostra aleatória simples, com 100 medidas, relativa a mensuração do tempo associado a este conjunto de tarefas foi obtido. Estime o tempo médio associado a este conjunto de atividade com um nível de confiança de 95%.

Inferência

□ Intervalo de Confiança

- Exemplo: Suponha uma linha de produção que fabrica papel de comprimento 11 polegadas e o desvio padrão seja conhecido ($\sigma=0,02$ polegadas). Em intervalos periódicos, são selecionados amostras ($n=100$ folhas) para determinar se o comprimento do papel se manteve em 11 polegadas. Deseja-se uma estimativa com nível de confiança de 97,5%.

□ Intervalo de Confiança

- Exemplo: (cont.) Uma amostra aleatória foi obtida e o comprimento médio da amostra foi $\bar{X} = 10,998$.
- Solução: $\bar{X} - Z^*(\sigma / \sqrt{n}) \leq \mu \leq \bar{X} + Z^*(\sigma / \sqrt{n})$
 $10,998 - Z^*(0,02/10) \leq \mu \leq 10,998 + Z^*(0,02/10)$

Para nível de confiança de 97,5%, tem-se $Z^* = 1,96$, portanto:

$10,99408 \leq \mu \leq 11,00192$. Desta forma, conclui-se que o processo está operando de maneira apropriada.

Inferência

Statdisk
NormalPopSample

Minitab
NormalPopSample

□ Intervalo de Confiança

- Exemplo: desejamos estimar o tempo de serviço (normalmente distribuído) associado a uma determinada atividade de um departamento de prestação de serviço com um nível de confiança de 99%, considerando que se sabe o desvio padrão ($\sigma=10$ unidades de tempo) deste serviço (da população). Uma amostra aleatória simples, de tamanho igual a 100, foi adequadamente coletada. Forneça o intervalo de confiança para a média.

□ Intervalo de Confiança para Média populacional com desvio-padrão conhecido

– Escolha do Tamanho da Amostra

$$E \leq Z^* \left(\frac{\sigma}{\sqrt{n}} \right)$$
$$n \geq \left(\frac{Z^* \sigma}{E} \right)^2$$

$$\bar{X} \pm Z^* \left(\frac{\sigma}{\sqrt{n}} \right)$$

E

Margem de erro do intervalo de confiança.

Escolha do Tamanho da Amostra

$$E \leq Z^* \left(\frac{\sigma}{\sqrt{n}} \right)$$

$$n \geq \left(\frac{Z^* \sigma}{E} \right)^2$$

Se

$$Z^* \left(\frac{\sigma}{\sqrt{n}} \right) \leq E$$

Temos: $\left[Z^* \left(\frac{\sigma}{\sqrt{n}} \right) \right]^2 \leq E^2$

$$(Z^*)^2 \left(\frac{\sigma^2}{n} \right) \leq E^2$$

$$(Z^*)^2 \left(\frac{\sigma^2}{E^2} \right) \leq n$$

Portanto: $n \geq \left(\frac{Z^* \sigma}{E} \right)^2$

Inferência

Statdisk

Função:

SampleSizeDetermination

$1-\alpha=0.99$

$E=7$

$\sigma=10$

□ Intervalo de Confiança

- Exemplo: desejamos determinar o tamanho da amostra necessário se estimar o tempo de serviço (normalmente distribuído) de a uma determinada atividade de um departamento de prestação de serviço, com um nível de confiança de 99%, considerando que se sabe o desvio padrão ($\sigma=10$ unidades de tempo) deste serviço (da população) e considerando aceitável um erro de 7 unidades de tempo.
- Qual o tamanho necessário da amostra?



Inferência

- Intervalo de Confiança para Média populacional com desvio-padrão conhecido com população finita

$$- \bar{X} \pm Z^* (\sigma / \sqrt{n}) \left(\frac{(N-n)}{(N-1)} \right)^{1/2}$$

Fator de Correção
N – Tamanho da população
n – tamanho da amostra

Valor crítico ($Z_{\alpha/2}$) da tabela da distribuição t quando o GL tende para infinito (última linha da tabela), considerando um dado nível de significância (α)



Inferência

□ Teste de Hipótese

- Procedimento que permite decidir se se rejeita ou aceita uma hipótese baseada em informações contidas em uma amostra.
- A Hipótese Nula, H_0 , é hipótese que se tem interesse rejeitar. A hipótese contraditória, H_1 , é denominada Hipótese Alternativa.
- As n observações (amostra) são divididas em duas regiões, Região de Aceitação – $R(H_0)$ – e Região de Rejeição – $R(H_1)$.

Inferência

□ Teste de Hipótese



Inferência

□ Teste de Hipótese (bicaudal)

- O Teste de Hipótese Z para Média populacional com desvio-padrão conhecido

$$\square H_0 : \mu = e$$

$$\square H_1 : \mu \neq e$$

$$Z = (\bar{X} - \mu) / (\sigma / \sqrt{n})$$

□ Teste de Hipótese

- Passos para o Teste de Hipótese para Média populacional com **desvio-padrão conhecido**
 1. Declare a Hipótese Nula, H_0 .
 2. Declare a Hipótese alternativa, H_1 .
 3. Escolha o nível de significância, α .
 4. Encontre os valores críticos que sub-dividem as regiões de rejeição e não-rejeição.
 5. Escolha o tamanho da amostra.
 6. Colete os dados da amostra.
 7. Calcule a média da amostra.
 8. Calcule a estatística Z.
 9. Determine se a estatística Z se encontra na região de rejeição ou na região de não-rejeição.
 10. Tome a decisão e a apresente nos termos do problema.

Inferência

Teste de Hipótese

Hypothesis Testing: One Mean

1) Pop. Mean = Claimed Mean

Significance: 0,01

Claimed Mean: 98

Population St Dev. (if known): 10

Sample Size, n: 100

Sample Mean: 100,14

Sample St Dev, s:

Evaluate Print

Plot

Help ?

Claim: $\mu = \mu(\text{hyp})$

z Test

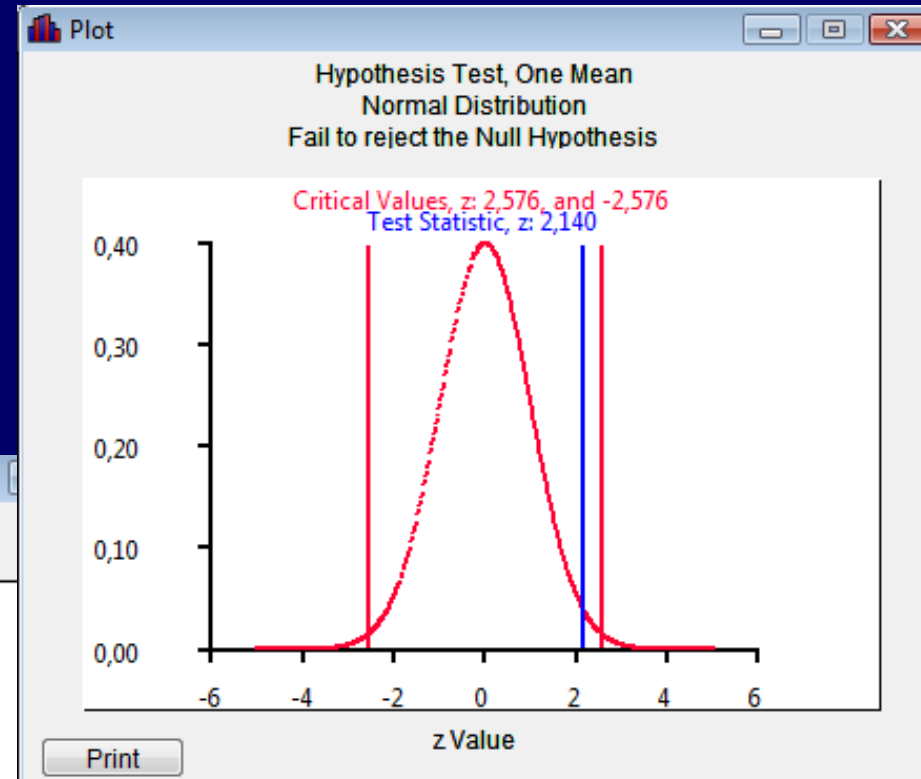
Test Statistic, z: 2,1400

Critical z: $\pm 2,5758$

P-Value: 0,0324

99% Confidence interval:
 $97.56417 < \mu < 102.7158$

Fail to Reject the Null Hypothesis
Sample does not provide enough evidence to reject the claim



Inferência

□ Teste de Hipótese

Observe que se eu não exijo um nível de confiança muito alto (99%), ou seja, se me contentar com um nível de confiança de 95%, temos evidências para rejeitar a hipótese (nula) da média ser igual a 98 ut.

The screenshot shows a software window titled "Hypothesis Testing: One Mean". It contains a dropdown menu set to "1) Pop. Mean = Claimed Mean". Below this, there are input fields for "Significance:" (0,05), "Claimed Mean:" (98), "Population St Dev: (if known)" (10), "Sample Size, n:" (100), "Sample Mean:" (100,14), and "Sample St Dev, s:" (empty). To the right of these fields is a text box containing the results of the test: "Claim: $\mu = \mu(\text{hyp})$ ", "z Test", "Test Statistic, z: 2,1400", "Critical z: $\pm 1,9600$ ", "P-Value: 0,0324", "95% Confidence interval: $98.18004 < \mu < 102.1$ ", and the conclusion "Reject the Null Hypothesis" and "Sample provides evidence to reject the claim". At the bottom of the window are buttons for "Evaluate", "Print", "Plot", and "Help ?".

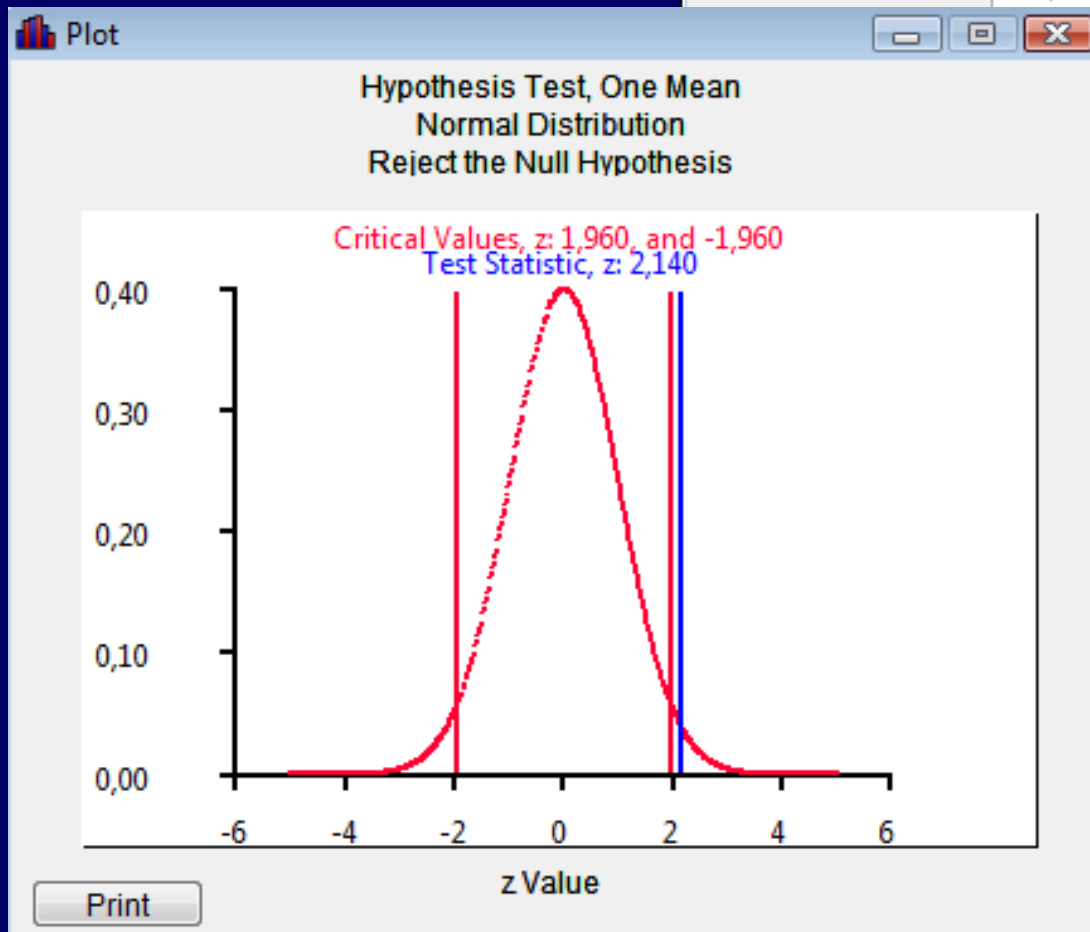
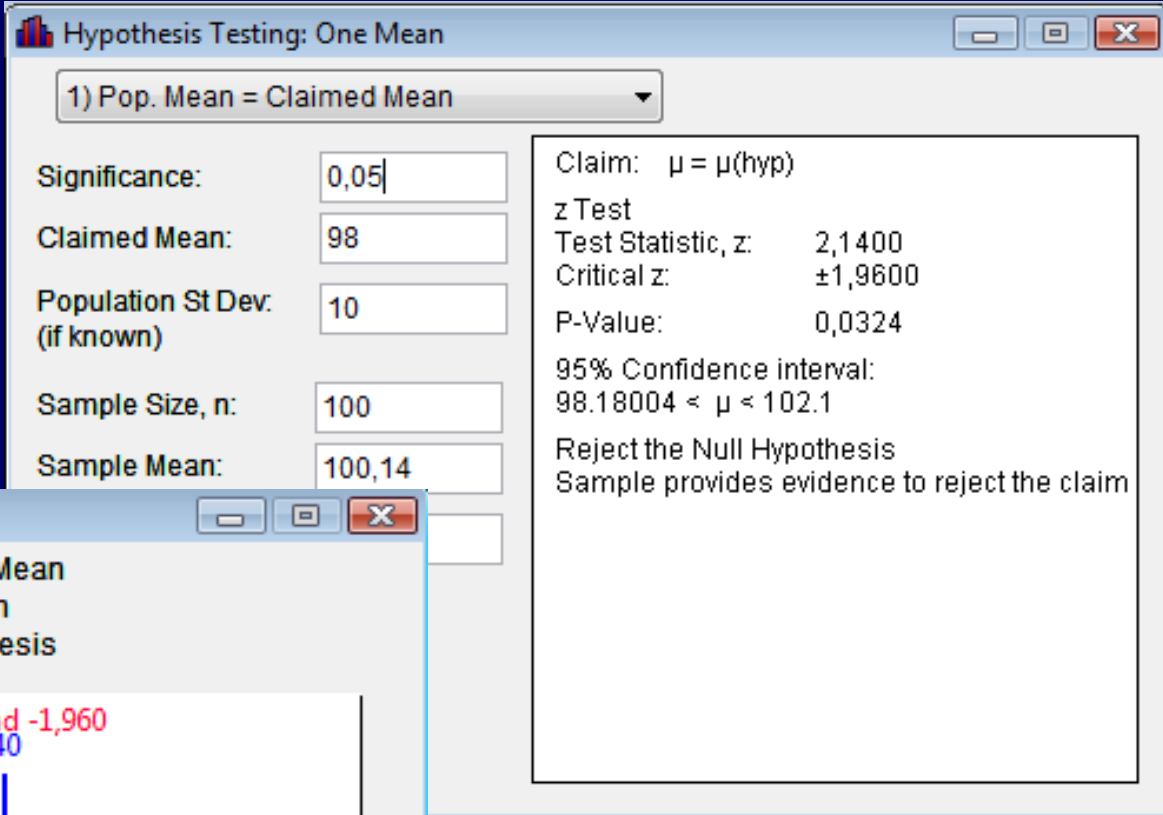
Parameter	Value
Significance	0,05
Claimed Mean	98
Population St Dev (if known)	10
Sample Size, n	100
Sample Mean	100,14
Sample St Dev, s	

Test Results:

- Claim: $\mu = \mu(\text{hyp})$
- z Test
- Test Statistic, z: 2,1400
- Critical z: $\pm 1,9600$
- P-Value: 0,0324
- 95% Confidence interval: $98.18004 < \mu < 102.1$
- Reject the Null Hypothesis
- Sample provides evidence to reject the claim

Inferência

Teste de Hipótese



□ Teste de Hipótese

Exemplo:

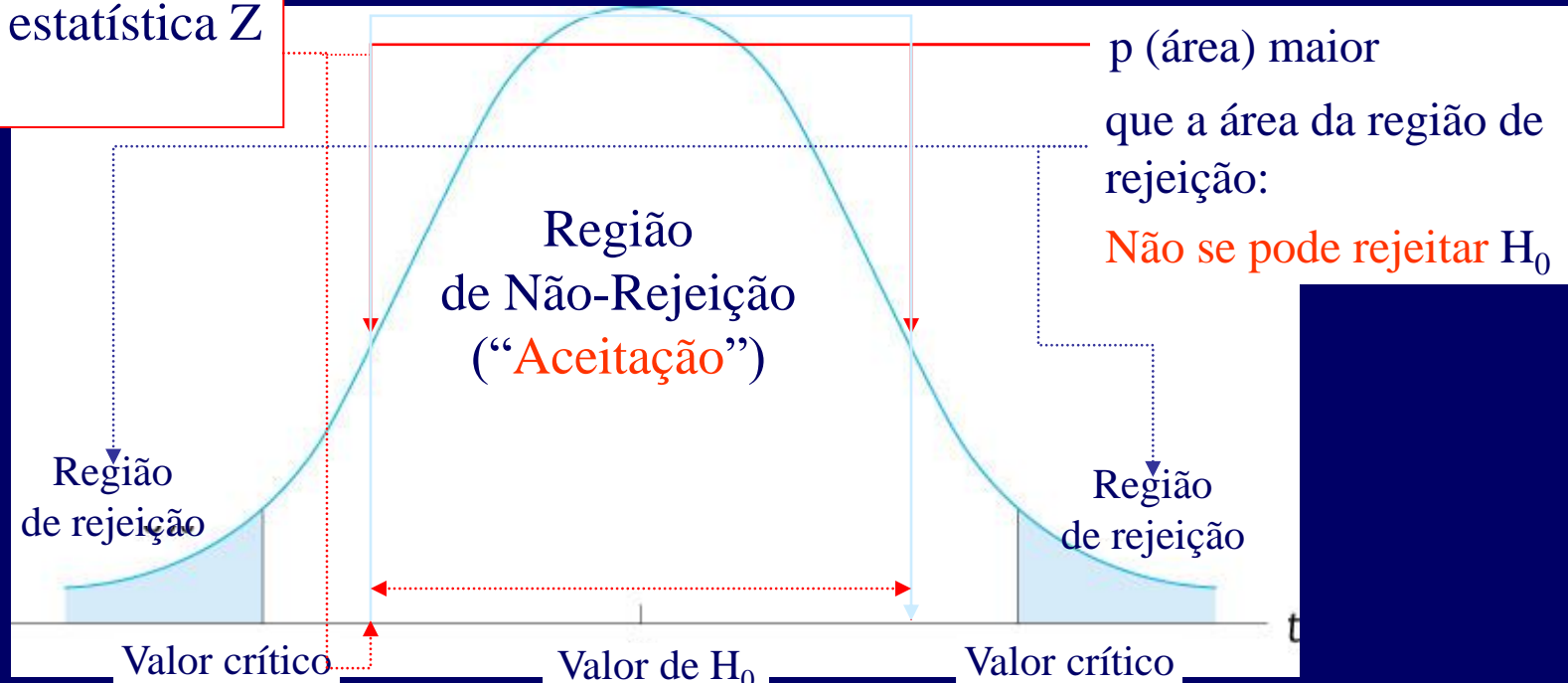
- Utilize os dados armazenados em NORMALPOPSAMPLE.MTW (NromalPopSample.sdd) e teste a hipótese de que a média seja 100 ut (unidades de tempo), sabendo-se que o desvio padrão populacional é 10, com 99% de confiança.

Inferência

□ Teste de Hipótese

- Nível de significância observado (O valor p):
é o menor nível de significância no qual H_0 pode ser rejeitado para a amostra.

Asociado a estatística Z
Calculada.



Inferência

□ Teste de Hipótese

- Passos para o Hipótese para Média populacional com **desvio-padrão conhecido** considerando o valor **p**
 1. Declare a Hipótese Nula, H_0 .
 2. Declare a Hipótese alternativa, H_1 .
 3. Escolha o nível de significância, α .
 4. Encontre os valores críticos que sub-dividem as regiões de rejeição e não-rejeição.
 5. Escolha o tamanho da amostra.
 6. Colete os dados da amostra.
 7. Calcule a média da amostra.
 8. Calcule a estatística Z.
 9. Calcule o valor de p com base na estatística Z.
 10. Compare o valor p com α .
 11. Tome a decisão e a apresente nos termos do problema.

Inferência

Minitab
NORMALPOPSAMPLE.MTW

□ Teste de Hipótese

- Exemplo: utilize os dados armazenados em NORMALPOPSAMPLE.MTW e teste a hipótese de que a média seja 100 ut (unidades de tempo), sabendo-se que o desvio padrão populacional é 10.

One-Sample Z: TempoServico

Test of $\mu = 100$ vs not = 100
The assumed standard deviation = 10

$p=0,584$ é muito maior que 1%,
portanto não se pode rejeitar a hipótese
nula ($\mu=100$).

Variable	N	Mean	StDev	SE Mean	99% CI	Z	P
TempoServico	100	99,4526	10,4051	1,0000	(96,8768; 102,0284)	-0,55	0,584

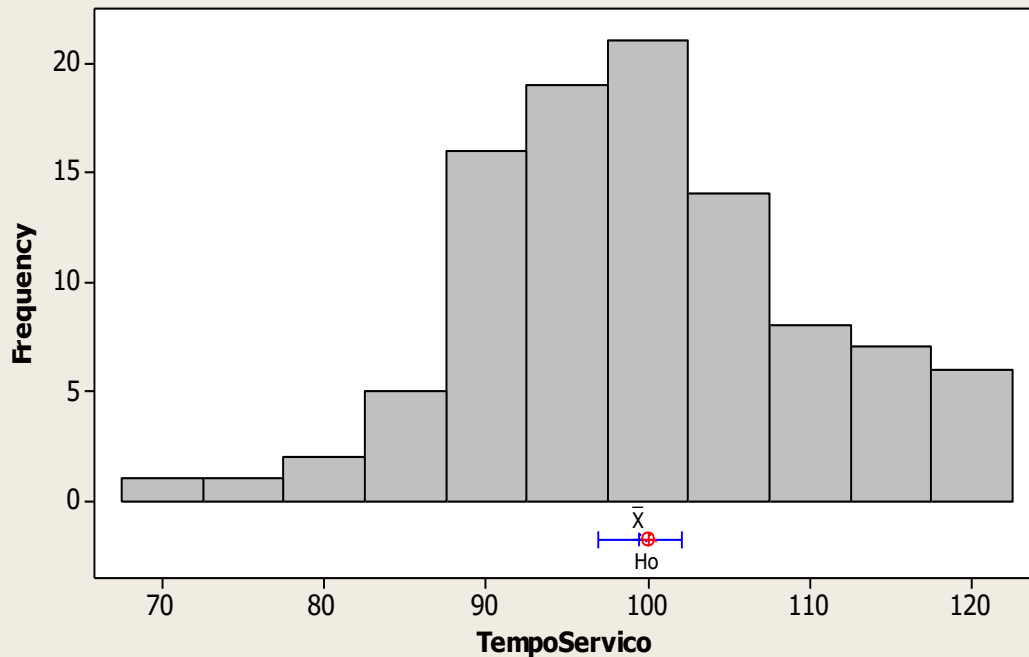
Inferência

Teste de Hipótese

- Exemplo:

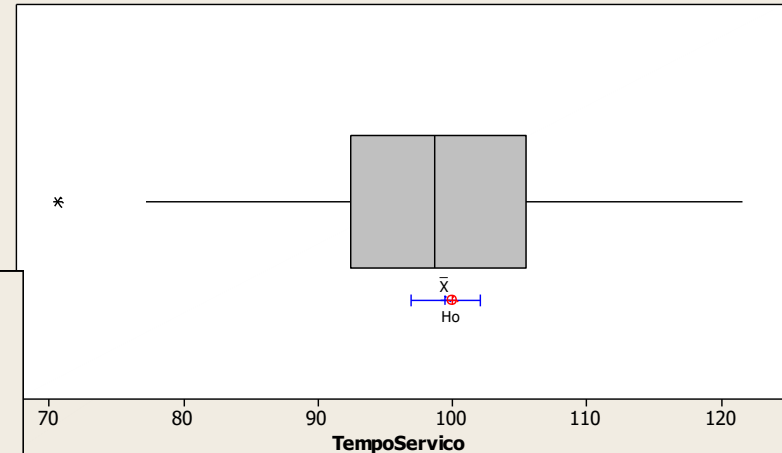
Histogram of TempoServico

(with H_0 and 99% Z-confidence interval for the Mean, and StDev = 10)



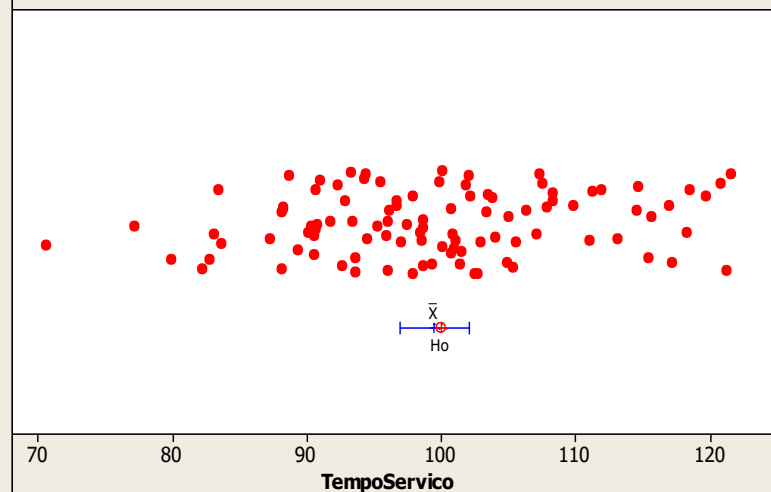
Boxplot of TempoServico

(with H_0 and 99% Z-confidence interval for the Mean, and StDev = 10)



Individual Value Plot of TempoServico

(with H_0 and 99% Z-confidence interval for the Mean, and StDev = 10)



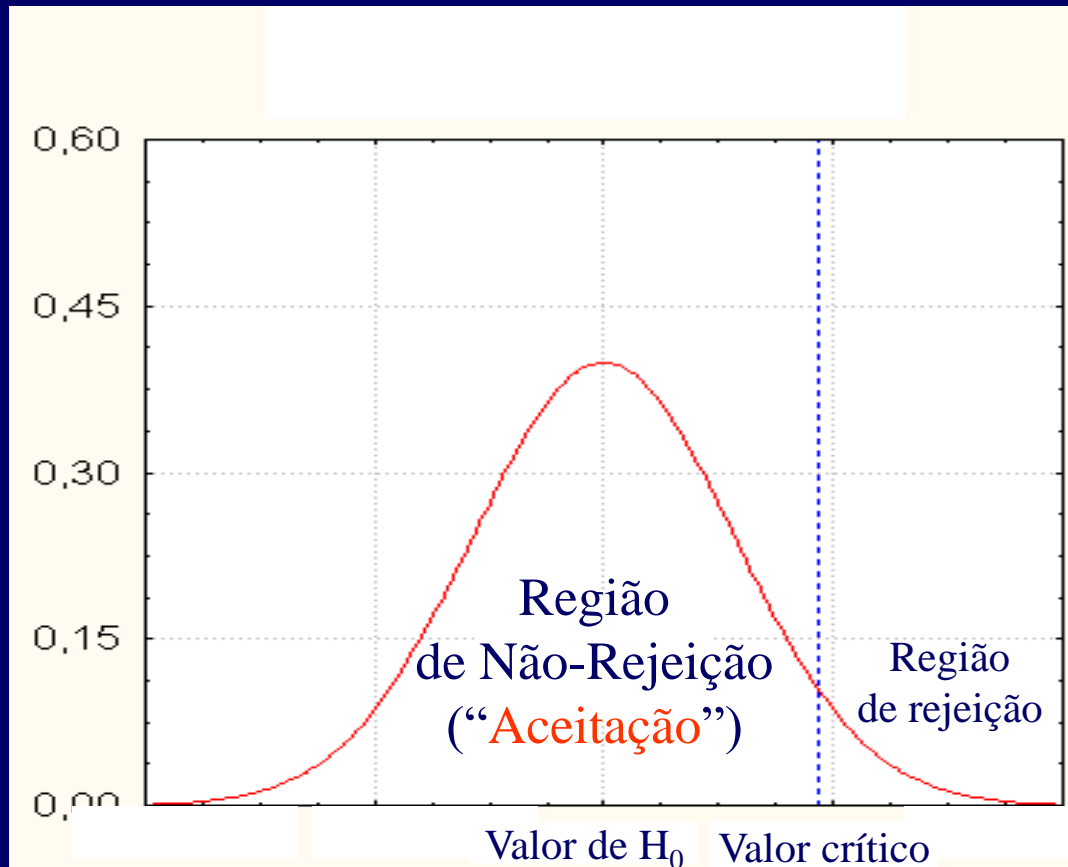
Inferência

□ Teste de Hipótese (Unicaudais)

O Teste de Hipótese Z para Média populacional com **desvio-padrão conhecido**

$$\square H_0 : \mu \leq n$$

$$\square H_1 : \mu > n$$



Inferência

□ Teste de Hipótese

Exemplo:

- Utilize os dados armazenados em NORMALPOPSAMPLE.MTW (NromalPopSample.sdd) e teste a hipótese de que a média seja menor ou igual 100 ut (unidades de tempo), sabendo-se que o desvio padrão populacional é 10, com 98% de confiança.

Inferência

Teste de Hipótese

NormalPopSampleHipTestMenorIgual

Hypothesis Test for Population Mean

Claim:
2) Pop. Mean < or = Claimed Mean

Significance, α : 0.01

Claimed Mean, μ_{hyp} : 100.0

Population St Dev, σ (if known): 10

Sample Size, n : 100

Sample mean, \bar{x} : 99.4526

Sample St Dev, s : 10.4551

Evaluate **Help** **Plot**

Claim	$\mu < \text{or} = \mu_{hyp}$
Pop St Dev Known	
Test Statistic, z	-0.5474
Critical z	2.3263
P-Value	0.7079
98% Confidence Interval:	
	$97.1263 < \mu < 101.7789$
Fail to Reject the Null Hypothesis	
Sample does not provide enough evidence to reject the claim	

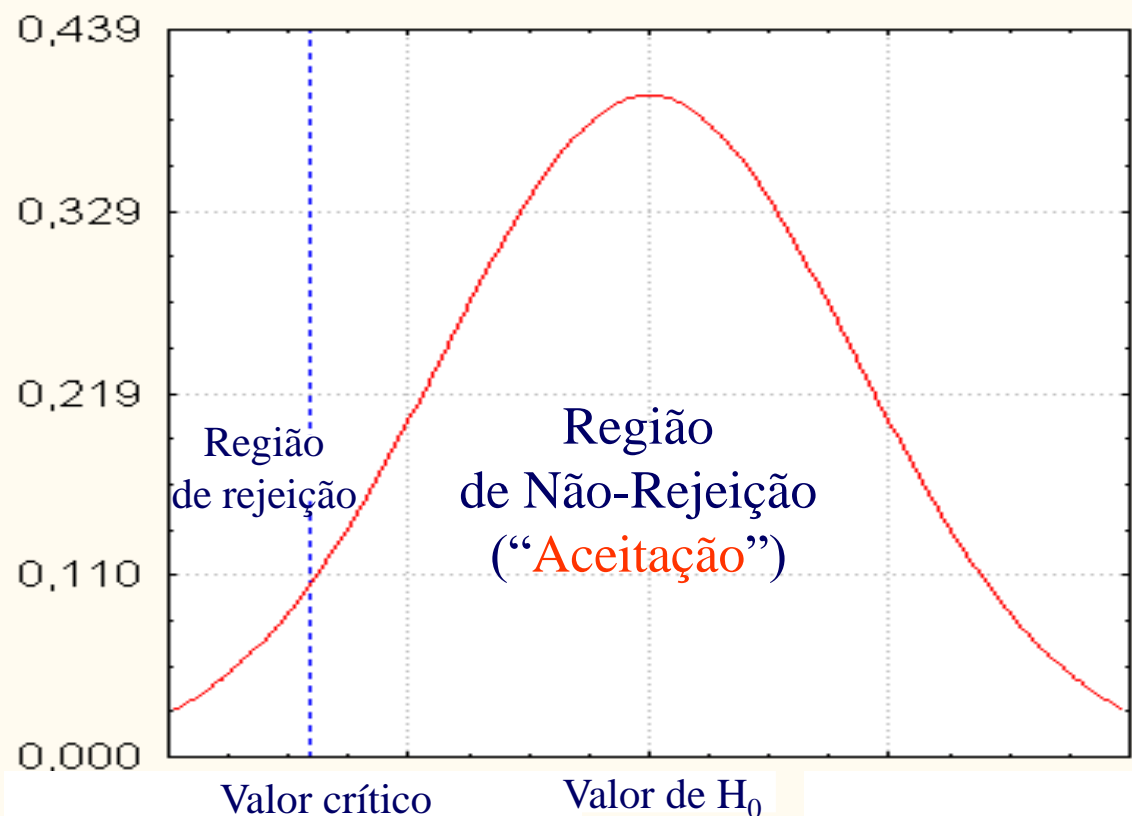
Inferência

□ Teste de Hipótese (Unicaudais)

O Teste de Hipótese Z para Média populacional com **desvio-padrão conhecido**

$$\square H_0 : \mu \geq n$$

$$\square H_1 : \mu < n$$



Inferência

□ Teste de Hipótese

Exemplo:

- Utilize os dados armazenados em NORMALPOPSAMPLE.MTW (NromalPopSample.sdd) e teste a hipótese de que a média seja maior ou igual 110 ut (unidades de tempo), sabendo-se que o desvio padrão populacional é 10, com 98% de confiança.

Inferência

Teste de Hipótese

NormalPopSampleHipTestMenorIgual

Hypothesis Test for Population Mean

Claim:
[3] Pop. Mean > or = Claimed Mean

Significance, α : [0.01]

Claimed Mean, μ_{hyp} : [110]

Population St Dev, σ (if known) [10]

Sample Size, n : [100]

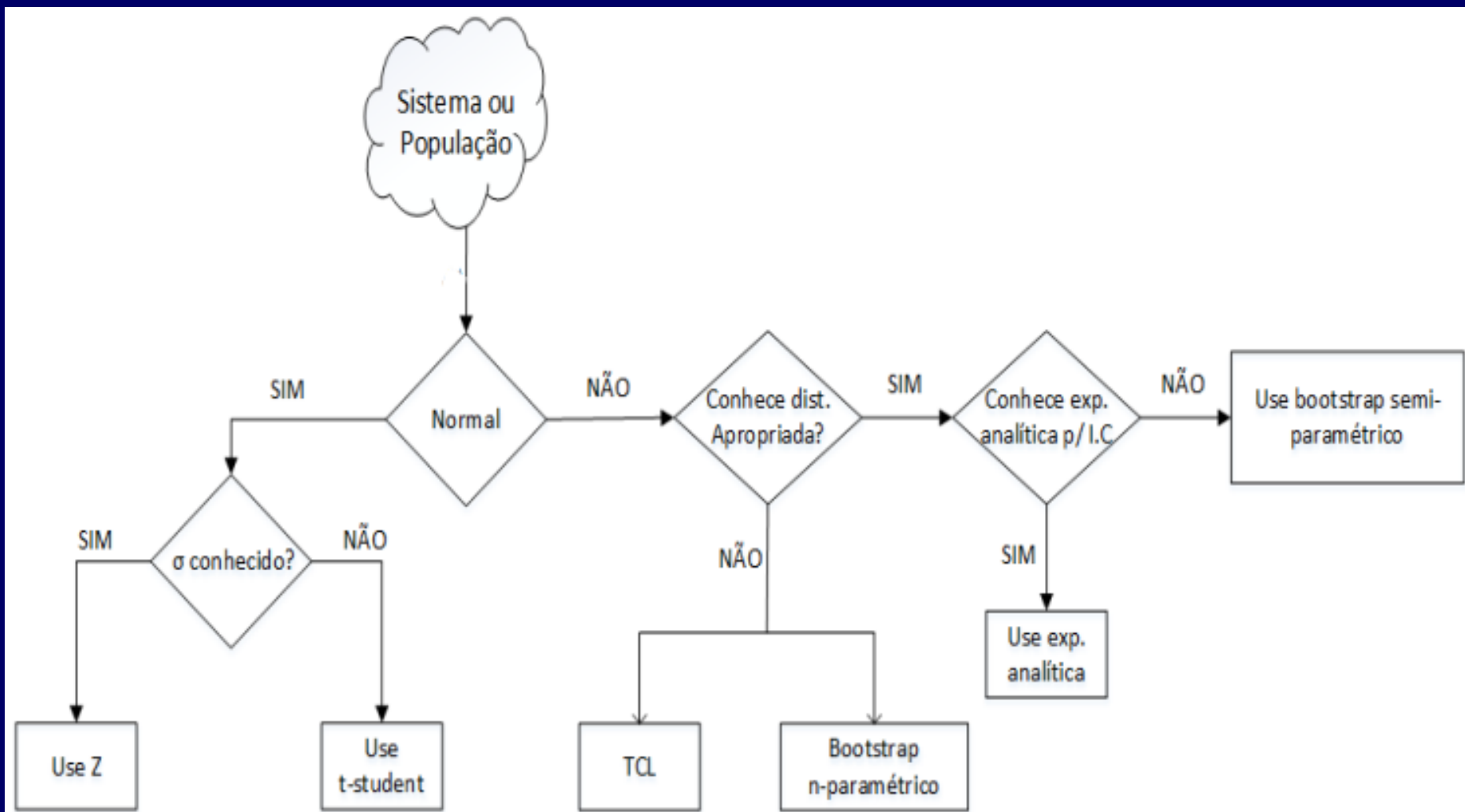
Sample mean, \bar{x} : [99.4526]

Sample St Dev, s : [10.4551]

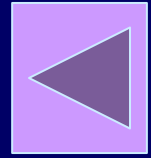
Evaluate **Help** **Plot**

Claim	$\mu > \text{or} = \mu_{hyp}$
Pop St Dev Known	
Test Statistic, z	-10.5474
Critical z	-2.3263
P-Value	0.0000
98% Confidence Interval:	
	$97.1263 < \mu < 101.7789$
Reject the Null Hypothesis	
Sample provides evidence to reject the claim	

Orientação para Inferência – Média



Inferência



□ Distribuições t

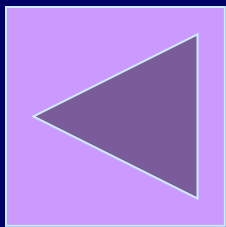
- Se consideramos uma variável aleatória X normalmente distribuída, a estatística

$$t = \frac{(\bar{X} - \mu)}{(s / \sqrt{n})}$$

(normalizado pelo erro padrão, só que utiliza-se s ao invés de σ - dado que não se conhece σ .)

tem distribuição **t** com n-1 graus de liberdade.

Illustrating Degrees of Freedom



Inferência

- Distribuições t
- Se consideramos uma população normalmente distribuída, a estatística t para uma amostra de tamanho n tem distribuição t com n-1 grau de liberdade (DF, GL).
 - É muito parecida com a distribuição Normal.
 - É simétrica, unimodal, tem forma de sino.
 - Tem área maior nas caudas e menor no centro do que a Normal, dado que se desconhece o σ , portanto os valores de t têm maior variabilidade.



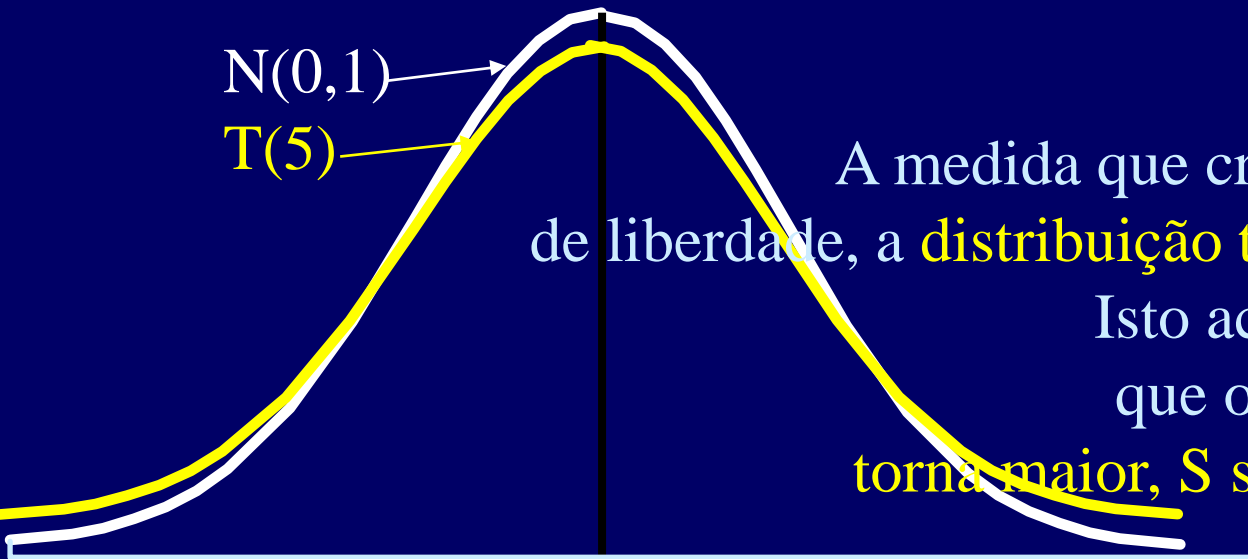
Inferência

Desenhar
a dist.

t.\..\Tools\Statistica\Sta_win.exe

□ Distribuições t

$$t = (\bar{X} - \mu) / (s / \sqrt{n})$$



A medida que cresce o número de graus de liberdade, a **distribuição t se aproxima da normal**.

Isto acontece porque a medida que o **tamanho da amostra se torna maior, S se torna mais semelhante a σ** .

Para uma **S maior** que 120 há **pouca diferença entre as distribuições Z e t**.

Inferência

Parte da tabela da distribuição t



Probabilidade de uma cauda						
	0.25	0.10	0.05	0.025	0.01	0.005
Probabilidade das duas caudas = α						
<i>df</i>	0.50	0.20	0.10	0.05	0.02	0.01
1	1.000	3.078	6.314	12.706	31.821	63.657
2	0.816	1.886	2.920	4.303	6.965	9.925
3	0.765	1.638	2.353	3.182	4.541	5.841
4	0.741	1.533	2.132	2.776	3.747	4.604
5	0.727	1.476	2.015	2.571	3.365	4.032
6	0.718	1.440	1.943	2.447	3.143	3.707

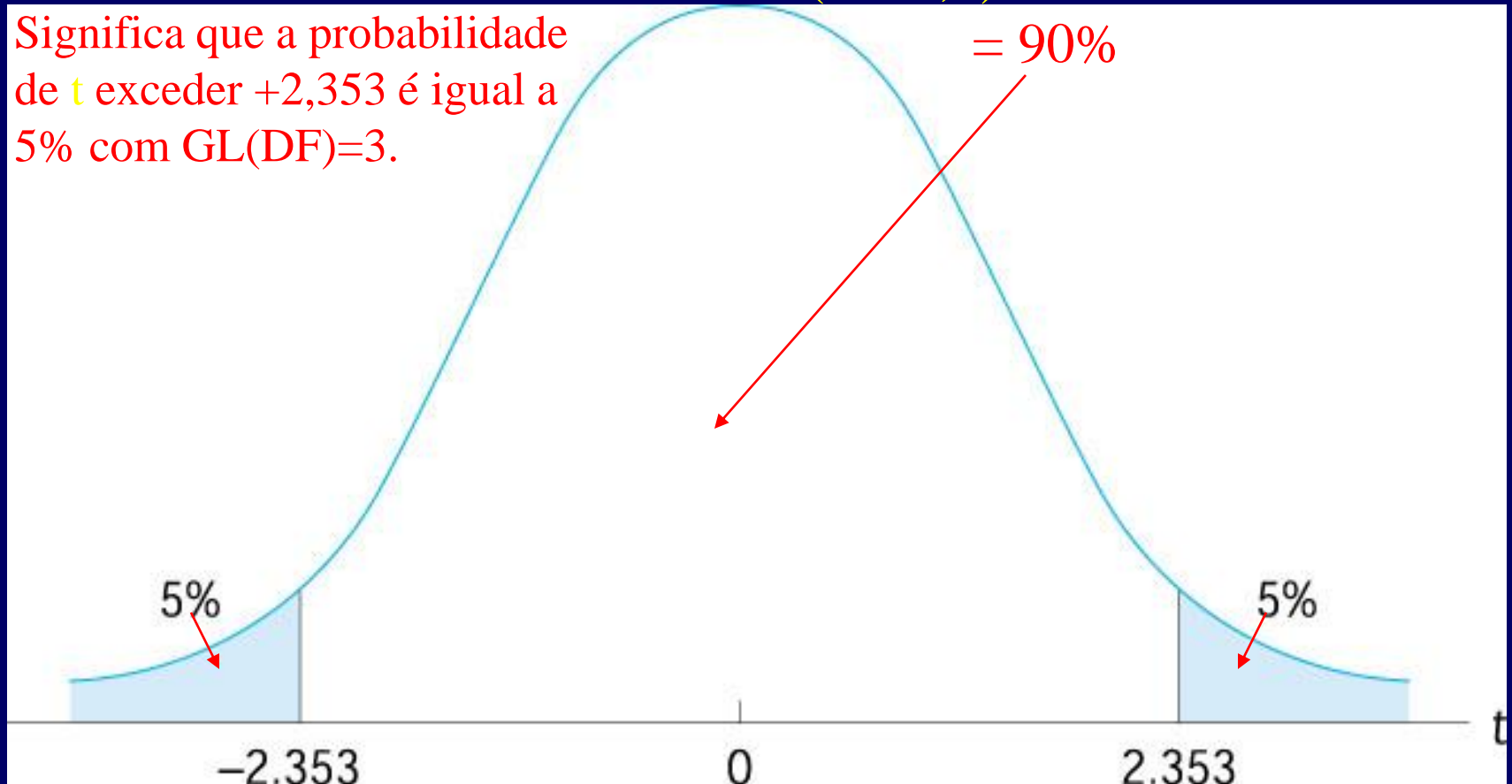
Significa que a probabilidade de t exceder $+2,353$ é igual a 5%
Com $GL(DF)=3$.

Inferência

Distribuição t com GL = 3

$$\begin{aligned}\text{Nível de Confiança} &= (1-\alpha) \times 100\% = \\ &= (1 - 0,1) \times 100\% = \\ &= 90\%\end{aligned}$$

Significa que a probabilidade de t exceder $+2,353$ é igual a 5% com GL(DF)=3.





Inferência

- Intervalo de Confiança para Média populacional com desvio-padrão desconhecido

$$\bar{X} - t_{\left(\frac{\alpha}{2}, n-1\right)} \times \left(S/\sqrt{n}\right) \leq \mu \leq \bar{X} + t_{\left(\frac{\alpha}{2}, n-1\right)} \times \left(S/\sqrt{n}\right)$$

Ou

$$\left(\bar{X} - t_{\left(\frac{\alpha}{2}, n-1\right)} \times \left(S/\sqrt{n}\right) \quad , \quad \bar{X} + t_{\left(\frac{\alpha}{2}, n-1\right)} \times \left(S/\sqrt{n}\right)\right)$$

Valor crítico da tabela da distribuição $t(n-1)$.

S é o desvio-padrão da amostra, considerando

Um nível de significância (α) .

□ Intervalo de Confiança

- Exemplo: suponha que um conjunto de atividades, denominado aqui por A1, seja executada por um departamento de uma organização. Uma amostra aleatória simples, com 100 medidas, relativa a mensuração do tempo associado deste conjunto de tarefas foi obtido. Estime o tempo médio associado a este conjunto de atividade com um nível de confiança de 95%.

□ Intervalo de Confiança

- Exemplo: suponha que um conjunto de atividades, denominado aqui por A1, executadas por um departamento de uma organização seja normalmente distribuído com desvio padrão desconhecido. Uma amostra aleatória simples, com 100 medidas, relativa a mensuração do tempo associado a este conjunto de tarefas foi obtido. Estime o tempo médio associado a este conjunto de atividade com um nível de confiança de 95%.

Inferência

- Intervalo de Confiança para Média populacional com desvio-padrão desconhecido com população finita

$$- \bar{X} \pm t^* (S/\sqrt{n}) ((N-n)/(N-1))^{1/2}$$

Fator de Correção
N – Tamanho da população
n – tamanho da amostra

Valor crítico da tabela da distribuição t(n-1).

S é o desvio-padrão da amostra, considerando

Um nível de significância (α). $t^* = t_{\alpha/2, n-1}$

Inferência

Sample Size

□ Intervalo de Confiança para Média populacional com desvio-padrão desconhecido

– Escolha do Tamanho da Amostra

$$\bar{X} \pm t_{\left(\frac{\alpha}{2}, n-1\right)} \times \left(S / \sqrt{n}\right) \Leftrightarrow \bar{X} \pm \epsilon$$

$$\epsilon = t_{\left(\frac{\alpha}{2}, n-1\right)} \times \left(S / \sqrt{n}\right) \Rightarrow \sqrt{n} = t_{\left(\frac{\alpha}{2}, n-1\right)} \times \left(S / \epsilon\right)$$

$$n' = \left(\frac{t_{\left(\frac{\alpha}{2}, n-1\right)} \times S}{\epsilon} \right)^2$$

— n — da amostra preliminar

— n' — da amostra

Margem de erro do intervalo de confiança.

Inferência

Excel

SampleSizeDetermination

$1-\alpha=0.95$

$E=7$

Statdisk

Função:

SampleSizeDetermination

$1-\alpha=0.95$

$E=7$

□ Intervalo de Confiança

- Exemplo: desejamos determinar o tamanho da amostra necessário para se estimar o tempo de serviço (normalmente distribuído) de a uma determinada atividade de um departamento de prestação de serviço, com um nível de confiança de 95%, considerando que o desvio padrão amostral é 10 ut e considerando aceitável um erro de 7 unidades de tempo.
- Qual o tamanho necessário da amostra?

Inferência

Statistica

Excel

StatDisk

Minitab

Mathematica

□ Intervalo de Confiança

- Exemplo: Suponha que se queira calcular o consumo médio anual de óleo (gl) usado para calefação em residências em uma determinada área. É selecionada uma amostra de 35 residências e o consumo médio destas residências está na tabela oiluse2T.txt.
- Abrir /Tools/SSP/Example/oiluse2T.txt

Inferência

□ Intervalo de Confiança

- Exemplo: Suponha que se queira calcular o consumo médio anual de óleo (gl) usado para calefação em residências em uma determinada área. É selecionada uma amostra de 35 residências e o consumo médio destas residências está na tabela oiluse2T.txt.
- Se se deseja ter 95% de confiança que o intervalo obtido contém a média da população (consumo médio da área), teremos: $\bar{X} = 1122,7$; $S = 295,72$; $t^* = t(n-1=34; p=0,025) = 2,0322$

Inferência

□ Intervalo de Confiança

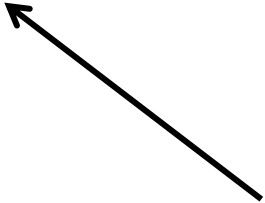
- Exemplo: Suponha que se queira calcular o consumo médio anual de óleo (gl) usado para calefação em residências em uma determinada área. É selecionada uma amostra de 35 residências e o consumo médio destas residências está na tabela oiluse.txt. Se se deseja ter 95% de confiança que o intervalo obtido contém a média da população (consumo médio da área), teremos:
 $X = 1122,73$; $S = 295,70$; $t^* = t(n-1=34; p=0,025) = 2,0322$
- $1122,73 \pm 101,58 \Leftrightarrow 1021,12 \leq \mu \leq 1224,28$



Inferência

- Teste de Hipótese para Média populacional com desvio-padrão desconhecido

$$\bar{X} \pm t_{\left(\frac{\alpha}{2}, n-1\right)} \times \left(S / \sqrt{n}\right) \Leftrightarrow \bar{X} - t_{\left(\frac{\alpha}{2}, n-1\right)} \times \left(S / \sqrt{n}\right) \leq \mu \leq \bar{X} + t_{\left(\frac{\alpha}{2}, n-1\right)} \times \left(S / \sqrt{n}\right)$$



Valor crítico da tabela da distribuição $t(n-1)$.
S é o desvio-padrão da amostra, considerando
Um nível de significância (α).

Inferência

□ Teste de Hipótese (bicaudal)

- O Teste de Hipótese t para Média populacional com desvio-padrão desconhecido.

$$\square H_0 : \mu = n$$

$$\square H_1 : \mu \neq n$$

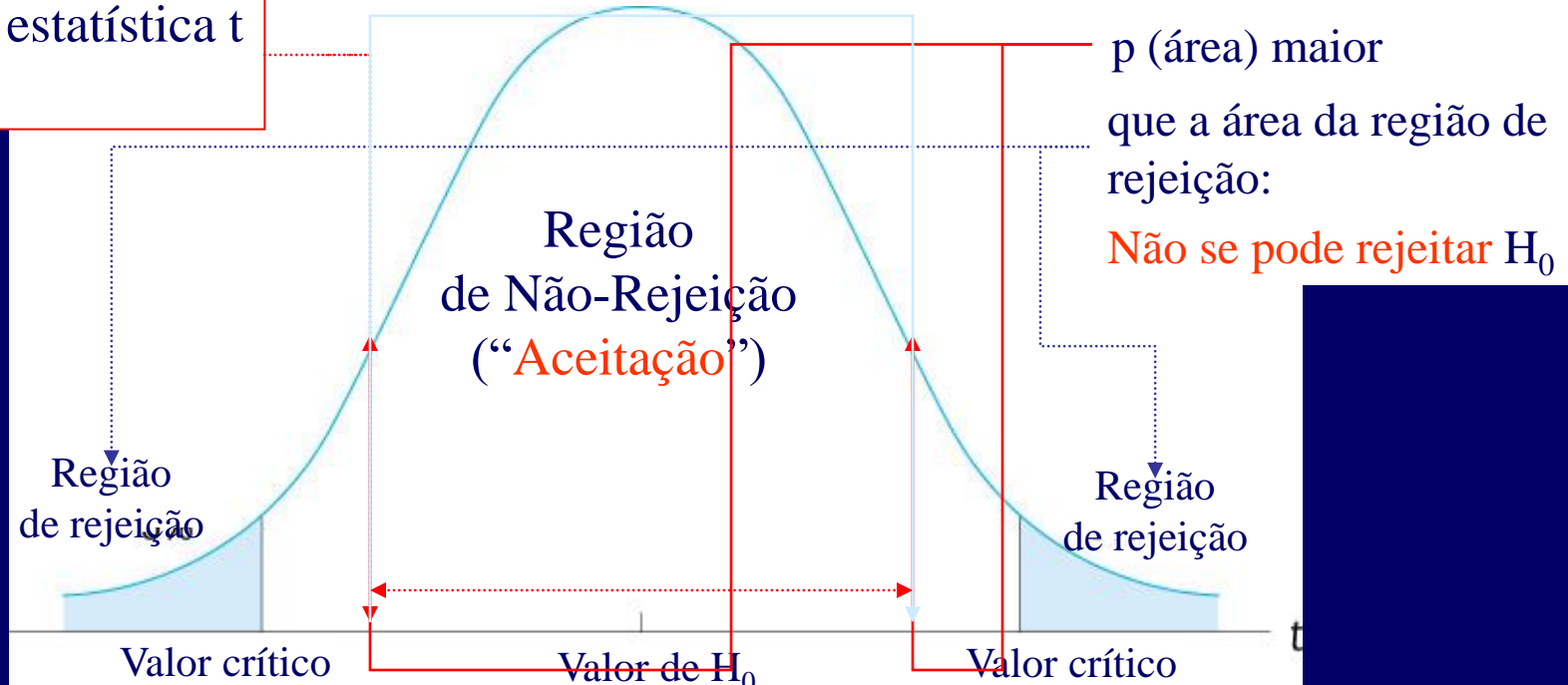
$$t = \frac{\bar{X} - \mu}{S / \sqrt{n}}$$

Inferência

□ Teste de Hipótese

- Nível de significância observado (O valor p):
é o menor nível de significância no qual H_0 pode ser rejeitado para a amostra.

Associado a estatística t
Calculada



Inferência

□ Teste de Hipótese

- Passos para o Hipótese para Média populacional com desvio-padrão desconhecido
 1. Declare a Hipótese Nula, H_0 .
 2. Declare a Hipótese alternativa, H_1 .
 3. Escolha o nível de significância, α .
 4. Escolha o tamanho da amostra.
 5. Encontre os valores críticos que sub-dividem as regiões de rejeição e não-rejeição.
 6. Colete os dados da amostra.
 7. Calcule a média da amostra.
 8. Calcule a estatística t .
 9. Determine se a estatística t se encontra na região de rejeição ou na região de não-rejeição.
 10. Tome a decisão e a apresente nos termos do problema.

Inferência

□ Teste de Hipótese

- Passos para o Hipótese para Média populacional com desvio-padrão desconhecido considerando o valor p
 1. Declare a Hipótese Nula, H_0 .
 2. Declare a Hipótese alternativa, H_1 .
 3. Escolha o nível de significância, α .
 4. Escolha o tamanho da amostra.
 5. Encontre os valores críticos que sub-dividem as regiões de rejeição e não-rejeição.
 6. Colete os dados da amostra.
 7. Calcule a média da amostra.
 8. Calcule a estatística t .
 9. Calcule o valor de p com base na estatística t .
 10. Compare o valor p com α .
 11. Tome a decisão e a apresente nos termos do problema.

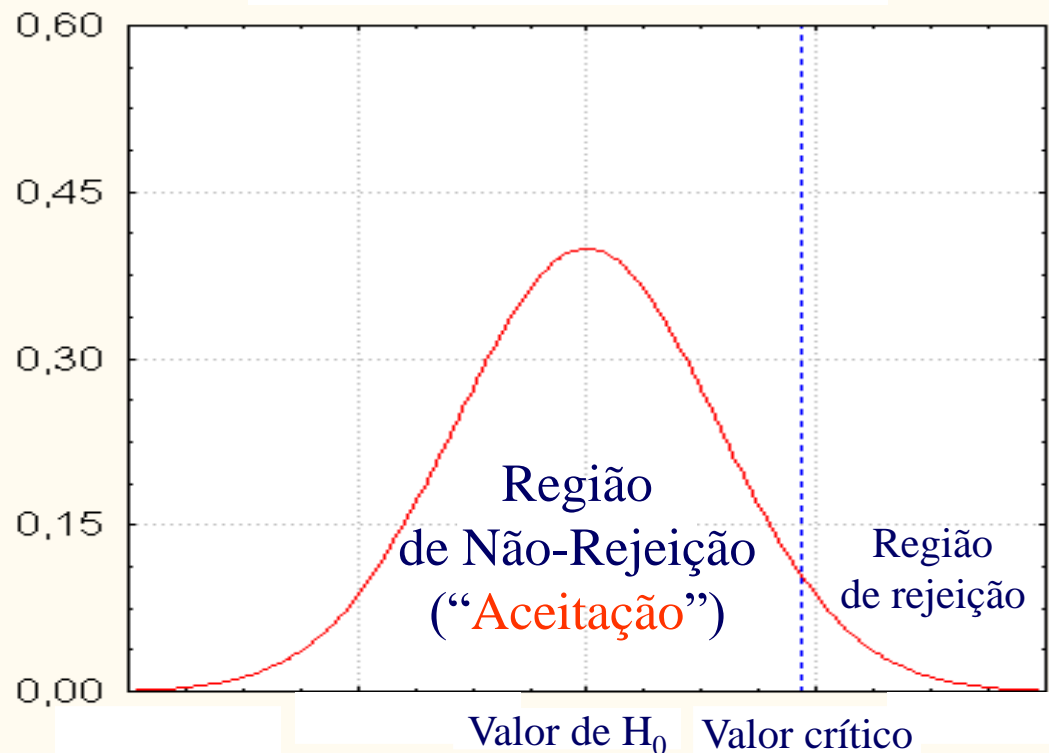
Inferência

□ Teste de Hipótese (Unicaudais)

O Teste de Hipótese t para Média populacional com desvio-padrão desconhecido

$$\square H_0 : \mu \leq n$$

$$\square H_1 : \mu > n$$



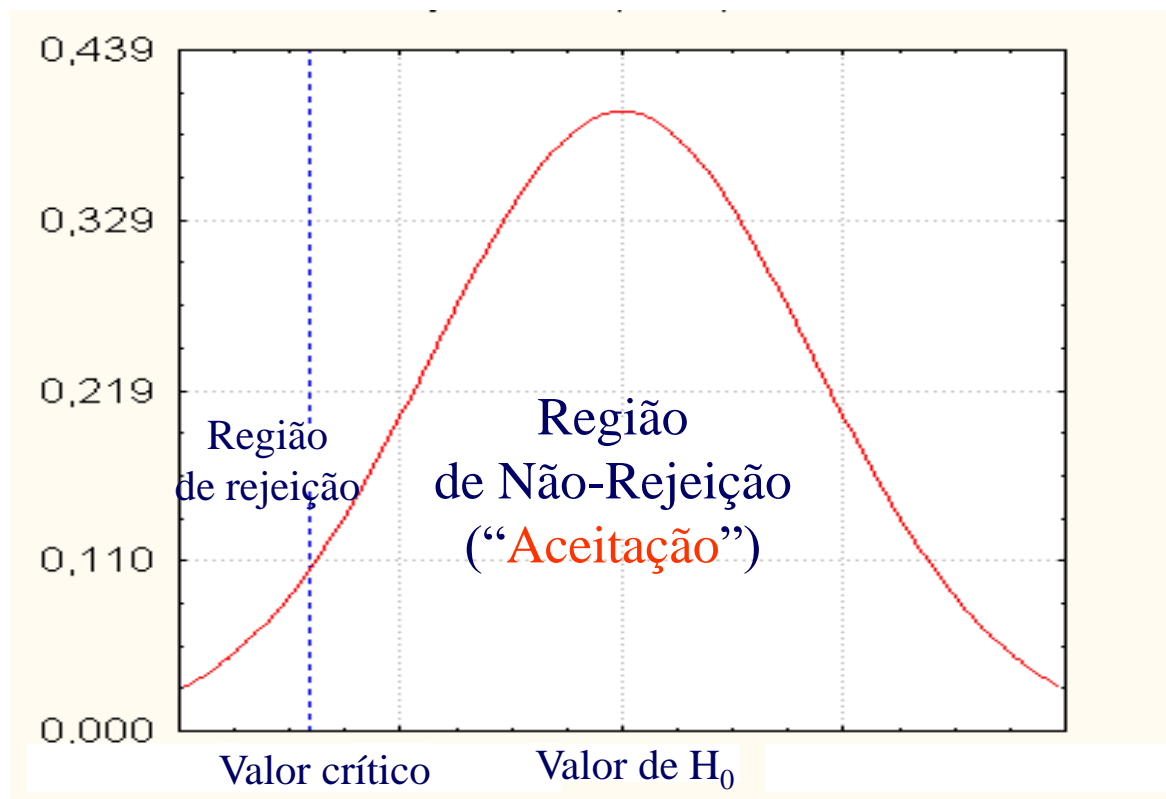
Inferência

□ Teste de Hipótese (Unicaudais)

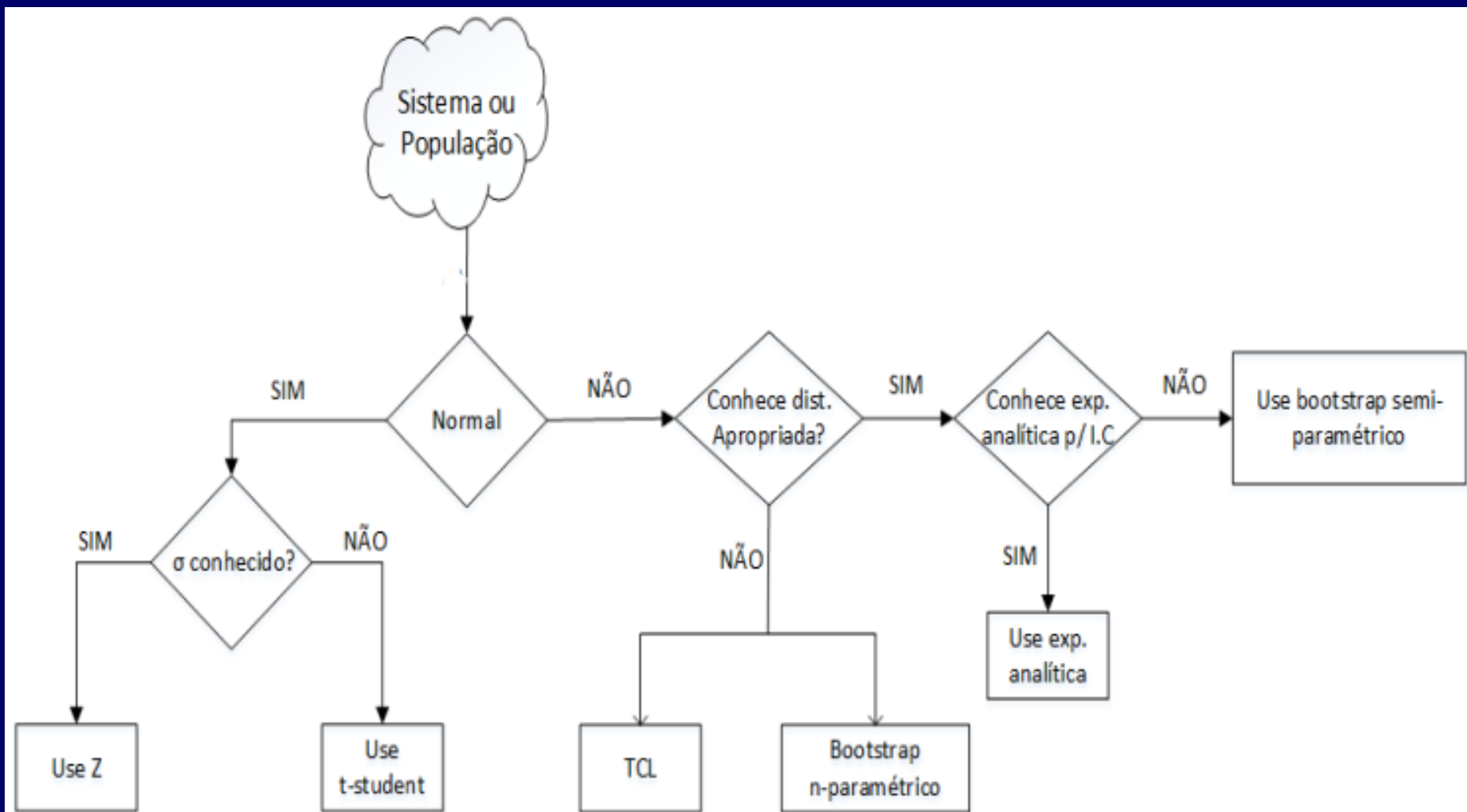
O Teste de Hipótese t para Média populacional com desvio padrão desconhecido

$$\square H_0 : \mu \geq n$$

$$\square H_1 : \mu < n$$



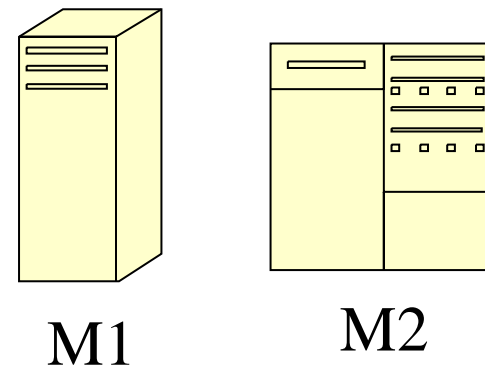
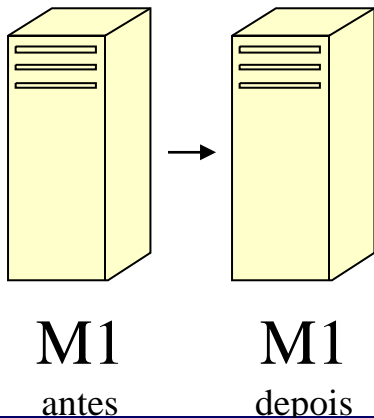
Orientação para Inferência – Média



Inferência

Comparação entre Alternativas

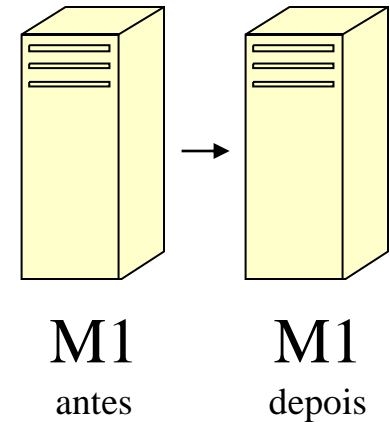
- Use de Intervalo de Confiança
 - Comparação Emparelhada (*Before-and-after comparisons*)
 - Medições não correspondentes



Inferência

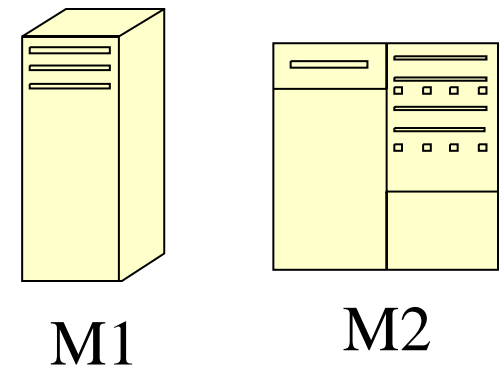
- Comparação Emparelhada
Before-and-after
(*t-paired test*)

A alteração feita provocou algum impacto estatisticamente significativo?



- **Medições não correspondentes**

Há diferença significativa entre os dois sistemas?



Inferência

□ *Before-and-After Comparison (t-paired test)*

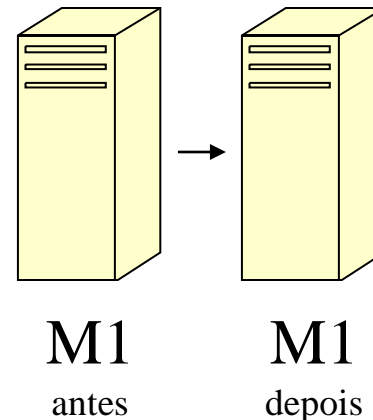
□ Premissas

- Medições *Before-and-after* não são independentes
- Variância entre os dois conjuntos de medições podem não ser iguais.

→ Medições são relacionadas

- Formam pares de medidas

□ Encontrar a diferença média



Inferência

- *Before-and-After Comparison*
(*t*-paired test)

b_i = medição antes

a_i = medição depois

$$d_i = a_i - b_i$$

\bar{d} = média de d_i

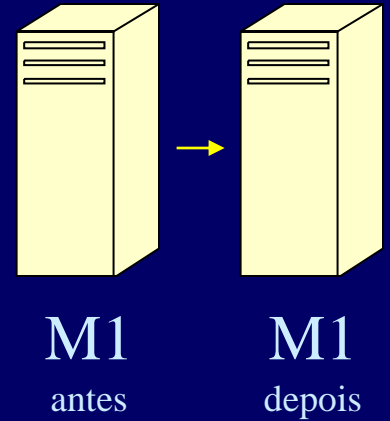
s_d = desvio padrão de d_i

$$(c_1, c_2) = \bar{d} \mp t_{1-\alpha/2; n-1} \frac{s_d}{\sqrt{n}}, \text{ se } n < 30$$

$$(c_1, c_2) = \bar{d} \mp z_{1-\alpha/2} \frac{s_d}{\sqrt{n}}, \text{ se } n \geq 30$$

Inferência

- *Before-and-After Comparison*



Medida (i)	Antes (b_i)	Depois (a_i)	Diferença ($d_i = b_i - a_i$)
1	85	86	-1
2	83	88	-5
3	94	90	4
4	90	95	-5
5	88	91	-3
6	87	83	4

Inferência

- ***Before-and-After Comparison
(t-paired test)***

Média das diferenças = $\bar{d} = -1$

Desvio padrão = $s_d = 4.15$

- Observando a média das diferenças, parece que o desempenho foi reduzido.
- No entanto, o desvio padrão é maior.

Inferência

•Intervalo de Confiança para diferença das média com 95% de confiança

$$t_{1-\alpha/2;n-1} = t_{0.975;5} = 2.571$$

	$\alpha/2$		
n	0.90	0.95	0.975
...
5	1.476	2.015	2.571
6	1.440	1.943	2.447
...
∞	1.282	1.645	1.960

Inferência

- **Intervalo de Confiança para diferença das média com 95% de confiança**

$$c_{1,2} = \bar{d} \mp t_{1-\alpha/2;n-1} \frac{s_d}{\sqrt{n}}$$

$$t_{1-\alpha/2;n-1} = t_{0.975;5} = 2.571$$

$$c_{1,2} = -1 \mp 2.571 \left(\frac{4.15}{\sqrt{6}} \right)$$

$$c_{1,2} = [-5.36, 3.36]$$

$$t_{1-\alpha/2;n-1} = t_{0.975;5} = 2.571$$

Inferência

• **Intervalo de Confiança para diferença das média com 95% de confiança**

□ $C_{1,2} = [-5.36, 3.36]$

□ O intervalo inclui 0

→ Com 95% de confiança, não existe diferença significativa entre os dois sistemas.

Inferência

Excel

t-paired_test

Suponha que tenhamos dois computadores denominados C1 e C2 e um *benchmark* com 15 aplicações. Gostaríamos de comparar o desempenho destes computadores com relação ao *benchmark*.

Os equipamentos foram isolados e os tempos de execução das aplicações foram medidos em cada computador. Os tempos médios de cada aplicação (cada aplicação foi medida diversas vezes) do *benchmark* foram calculados. Observa-se que as distribuições dos tempos de execução associada de cada aplicação não se afastam demasiadamente da distribuição Normal.

Os tempos médios de cada aplicação executada no computador C1 e C2 estão na tabela.

Podemos afirmar, com 95% de confiança, que um dos computadores tem melhor desempenho que o outro com relação à execução deste *benchmark*?

Inferência

Minitab

Excel

- **Outro exemplo:** suponha que tenhamos um determinado sistema computacional (S) que realize uma algumas atividades (A,B,C,D,E,F,G). Os tempos médios para execução destas atividades são distribuídos de forma aproximadamente Normal e tem os respectivos valores:

ATIVIDADES	Tempo Médio (ms)
A	107.27
B	108.04
C	111.05
D	114.35
E	113.24
F	112.41
G	108.49

Inferência

Minitab

Excel

- Outro exemplo (cont.): este sistema sofre ajuste que procuraram melhorar o seu desempenho. Após os ajustes os tempos médios das atividades A,B,C,D,E,F e G passaram ser:

ATIVIDADES	Tempo Médio (ms)
A	124.3
B	115.5
C	118.1
D	112.6
E	120.8
F	120.7
G	123.7

Inferência

- Outro exemplo (cont.). Podemos afirmar que os ajustes realizados resultaram em melhoria de desempenho do sistema?

Statdisk3

Hyp. Test for the Mean Difference: Matched Pairs

Claim:
1) Pop. Mean of Difference = 0

Significance, α : 0.01

Untitled			
1	107.27	1	124.3
2	108.04	2	115.5
3	111.05	3	118.1
4	114.35	4	112.6
5	113.24	5	120.8
6	112.41	6	120.7
7	108.49	7	123.7

Clear Copy Paste

Evaluate Help Plot

Claim $\mu_d = 0$

Sample Size, n 7

Diff. Mean, \bar{x}_d -8.6929

Diff. St Dev, s_d 6.1468

Test Statistic, t -3.7416

Critical t ± 3.7074

P-Value 0.0096

99% Confidence Interval:

$-17.3062 < \mu_d < -0.0795$

Reject the Null Hypothesis

Sample provides evidence to reject the claim

Inferência

- Outro exemplo (cont.). Podemos afirmar que os ajustes realizados resultaram em melhoria de desempenho do sistema?

Statdisk3

Hyp. Test for the Mean Difference: Matched Pairs

Claim:
1) Pop. Mean of Difference = 0

Significance, α : 0.005

Untitled			
1	107.27	1	124.3
2	108.04	2	115.5
3	111.05	3	118.1
4	114.35	4	112.6
5	113.24	5	120.8
6	112.41	6	120.7
7	108.49	7	123.7

Clear Copy Paste

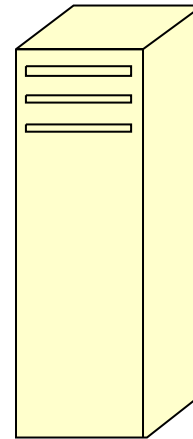
Evaluate Help Plot

Claim	$\mu_d = 0$
Sample Size, n	7
Diff. Mean, \bar{x}_d	-8.6929
Diff. St Dev, s_d	6.1468
Test Statistic, t	-3.7416
Critical t	± 4.3168
P-Value	0.0096
99.5% Confidence Interval:	
	$-18.7220 < \mu_d < 1.3363$
Fail to Reject the Null Hypothesis	
Sample does not provide enough evidence to reject the claim	

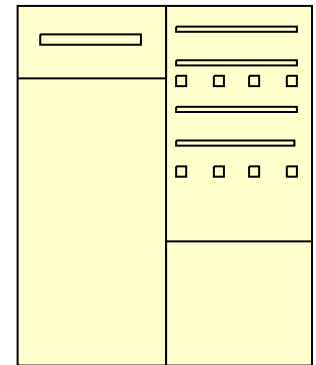
Inferência

□ Medições não correspondentes

- Quando não há correspondência entre pares de medidas
- Observações não-emparelhadas
- n_1 medições do sistema M1
- n_2 medições do sistema M2



M1



M2

Inferência

- Intervalo de Confiança para Diferença entre Médias
 1. Calcule as médias.
 2. Calcule a diferença das médias.
 3. Calcule o desvio padrão da diferença das médias.
 4. Calcule o intervalo de confiança para esta diferença.
 5. Se não houver diferença significativa entre os sistemas, o intervalo inclui o 0.

Inferência

Intervalo de Confiança para Diferença entre Médias

Diferença entre as médias :

$$\bar{x} = \bar{x}_1 - \bar{x}_2$$

Dado que para X_1 e X_2 mutuamente independentes

$$Var[X_1 - X_2] = Var[X_1] + Var[X_2]$$

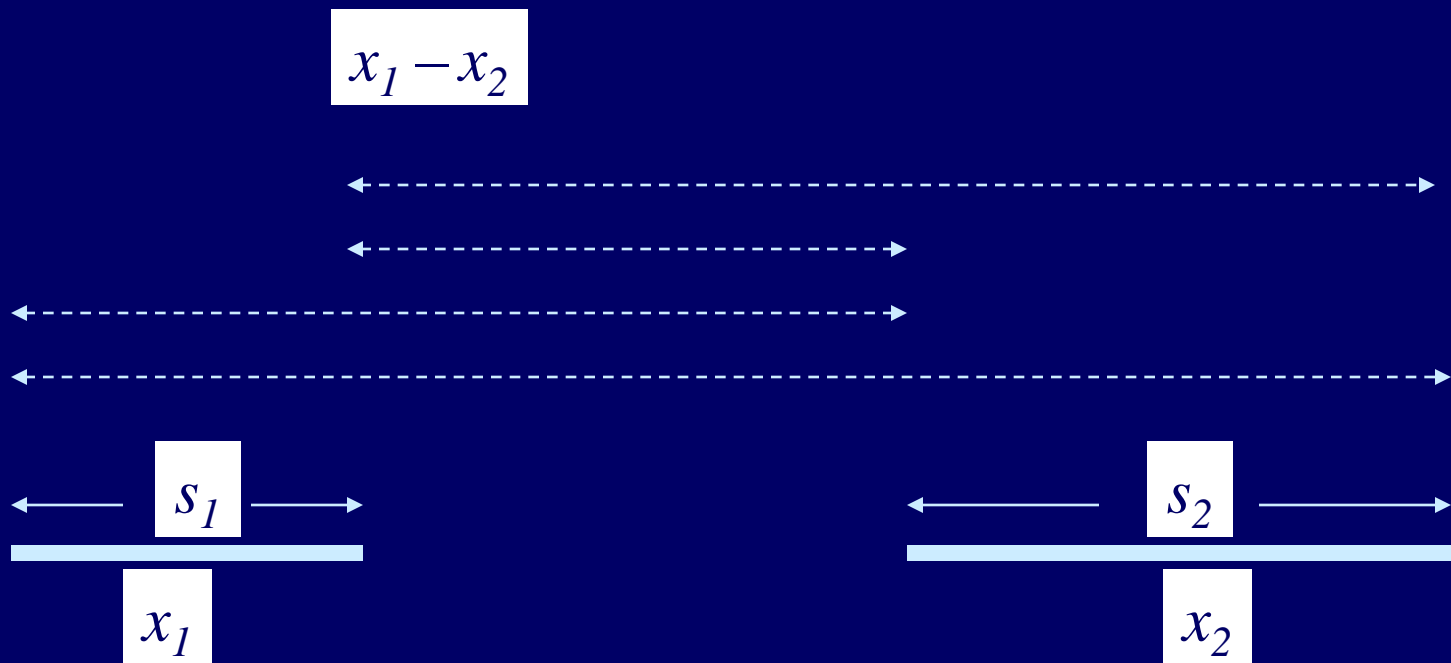
e que $s_x = \sqrt{Var[X_1] + Var[X_2]}$

Desvio padrão combinado :

$$s_x = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Inferência

- Por quê os desvios padrões são somados?



Inferência

- **Grau de Liberdade**

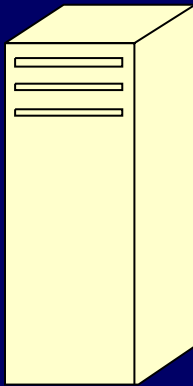
Não somente fazer $n_{df} = n_1 + n_2 - 2$

$$\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2$$

$$n_{df} = \frac{\left(s_1^2 / n_1 \right)^2}{n_1 - 1} + \frac{\left(s_2^2 / n_2 \right)^2}{n_2 - 1}$$

Inferência

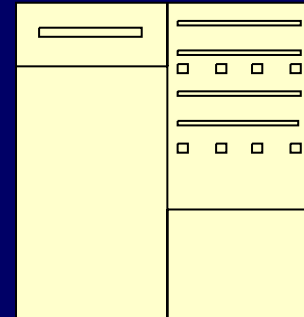
• Exemplo 1 – Medição não correspondentes



$$n_1 = 12 \text{ medidas}$$

$$\bar{x}_1 = 1243 \text{ s}$$

$$s_1 = 38.5$$



$$n_2 = 7 \text{ medidas}$$

$$\bar{x}_2 = 1085 \text{ s}$$

$$s_2 = 54.0$$

Inferência

•Exemplo 1– Medição não correspondentes (cont.)

$$\bar{x} = \bar{x}_1 - \bar{x}_2 = 1243 - 1085 = 158$$

$$s_x = \sqrt{\frac{38.5^2}{12} + \frac{54^2}{7}} = 23.24$$

$$\left(\frac{38.5^2}{12} + \frac{54^2}{7} \right)^2$$

$$n_{df} = \frac{\left(\frac{38.5^2}{12} \right)^2}{12-1} + \frac{\left(\frac{54^2}{7} \right)^2}{7-1} = 9.62 \rightarrow 10$$

Inferência

- **Exemplo 1 – Medição não correspondente.
Com 95% CI (cont.)**

$$c_{1,2} = \bar{x} \mp t_{1-\alpha/2; n_{df}} s_x$$

$$t_{1-\alpha/2; n_{df}} = t_{0.95; 10} = 1.813$$

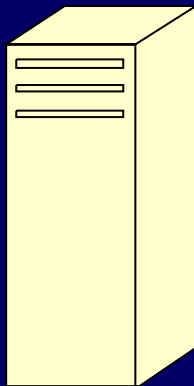
$$c_{1,2} = 158 \mp 1.813(23.24)$$

$$c_{1,2} = [116, 200]$$

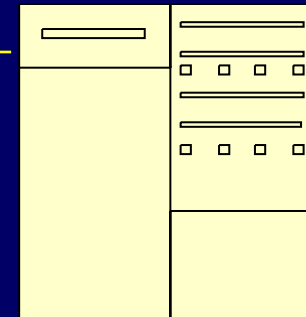
Inferência

Minitab
TwoIndMeansDifVar

• Exemplo – Medição não correspondentes (Minitab)



↓	C1	C2
1	1249,10	1231,14
2	1238,77	1228,64
3	1225,74	1187,84
4	1227,48	1206,00
5	1244,37	1160,85
6	1230,46	1218,09
7	1222,46	1209,79
8	1235,58	1198,74
9	1233,71	1176,97
10	1230,99	1248,89
11	1232,76	1185,71
12	1228,93	1224,04
13	1225,04	
14	1227,46	
15	1223,95	
16	1237,54	
17	1214,34	
18	1226,72	
19	1218,14	
20	1228,41	

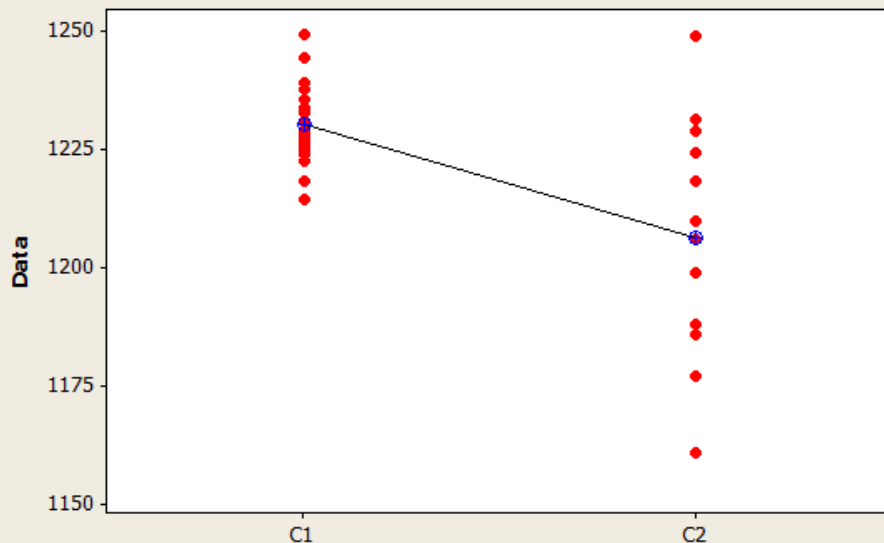


Inferência

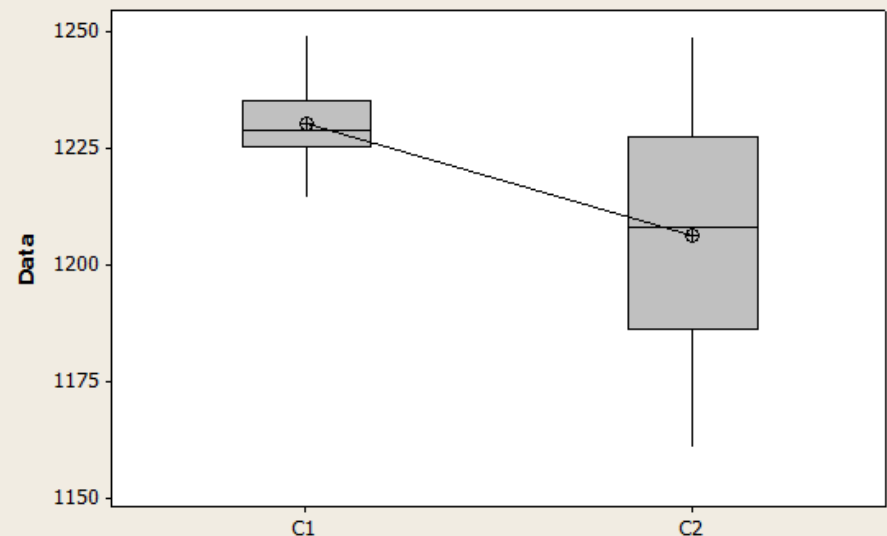
•Exemplo — Medição não correspondentes (Minitab)

```
Difference = mu (C1) - mu (C2)
Estimate for difference: 23,71
95% CI for difference: (7,18; 40,24)
T-Test of difference = 0 (vs not =): T-Value = 3,12  P-Value = 0,009  DF = 12
```

Individual Value Plot of C1; C2



Boxplot of C1; C2



Inferência

StatDisk
TwoIndMeasDiffVar

• Exemplo — Medição não correspondentes (StatDisk)

Hypothesis Test for the Mean of Two Independent Samples

Claim:
1) Pop. Mean 1 = Pop. Mean 2

Significance, α : 0.05

Sample 1

Sample Size, n_1 : 12

Sample mean 1: 1229.8

Sample St Dev, s_1 : 12.7

Pop. St Dev, σ_1 :
(if known)

Sample 2

Sample Size, n_2 : 7

Sample mean 2: 1085.17

Sample St Dev, s_2 : 9.68

Pop. St Dev, σ_2 :
(if known)

Claim $\mu_1 = \mu_2$

UNEQUAL Pop. Var's
Do not assume $\sigma_1^2 = \sigma_2^2$

Test Statistic, t 25.9401

Critical t ± 2.1098

P-Value 0.0000

95% Confidence Interval:

$132.8666 < \mu_1 - \mu_2 < 156.3934$

Reject the Null Hypothesis
Sample provides evidence to reject the claim

Not eq vars: NO POOL

Eq vars: POOL

Prelim F-test

Evaluate Help Plot

Inferência

•Caso Especial

- Se $n_1 < \approx 30$ or $n_2 < \approx 30$ e
 - erros são normalmente distribuídos,
 - e $s_1 = s_2$ (Os desvios padrão são iguais)

Ou

- Se $n_1 = n_2$ e
 - erros são normalmente distribuídos,
 - e mesmo que s_1 não seja igual a s_2
- Neste situação, tem-se:

Inferência

•Caso Especial

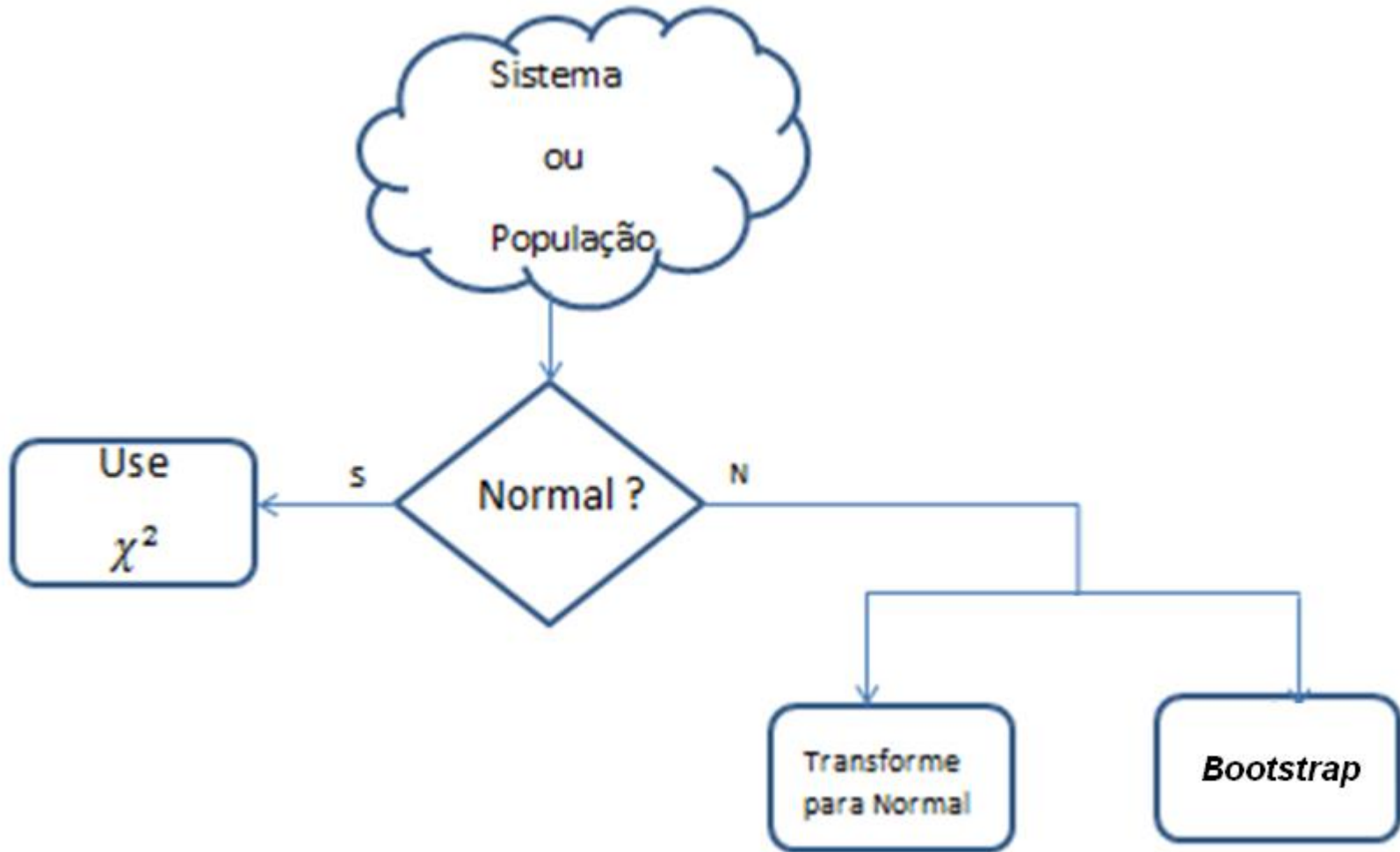
$$(c_1, c_2) = \bar{x} \mp t_{1-\alpha/2; n_{df}} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$n_{df} = n_1 + n_2 - 2$$

$$s_p = \sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}}$$

- Normalmente produz intervalos mais estreitos.
- Algumas vezes é também útil quando se realiza medições adicionais.

Orientação para Inferência - Variância



Inferência

Considere 30 AAS (de tamanho 20) obtidas de uma população normalmente distribuída com média 100 e desvio padrão 20. A amostra é apresentada na planilha DistVar_PopNormalDist.

Observe o histograma dos desvios padrão e da variância.

Amostra

Inferência

Em seguida, considere 100 AAS (de tamanho 20) obtidas da mesma população. A amostra é apresentada na planilha DistVar_PopNormalDist.

Observe o histograma dos desvios padrão e da variância.

Amostra

Inferência

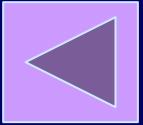
Agora considere 1000 AAS (de tamanho 20) obtidas da mesma população. A amostra é apresentada na planilha DistVar_PopNormalDist1000.

Observe o histograma dos desvios padrão e da variância.

Amostra - Minitab

Amostra - Excel

Inferência



□ Distribuição χ^2

- Por exemplo, considere uma população normalmente distribuída com variância conhecida 400. Seleccionamos aleatoriamente 1000 amostras independentes de tamanho 20 e calculamos as variâncias amostrais s^2_i . A estatística $\chi^2 = [(n-1) s^2_i] / \sigma^2$ tem distribuição qui-quadrado.

- n = tamanho da amostra,
- s^2_i = variância da amostra e
- σ^2 = variância populacional
- Média da distribuição χ^2 é $n-1$
- Variância da distribuição χ^2 é $2(n-1)$

χ^2

Observe
o Exemplo

Inferência

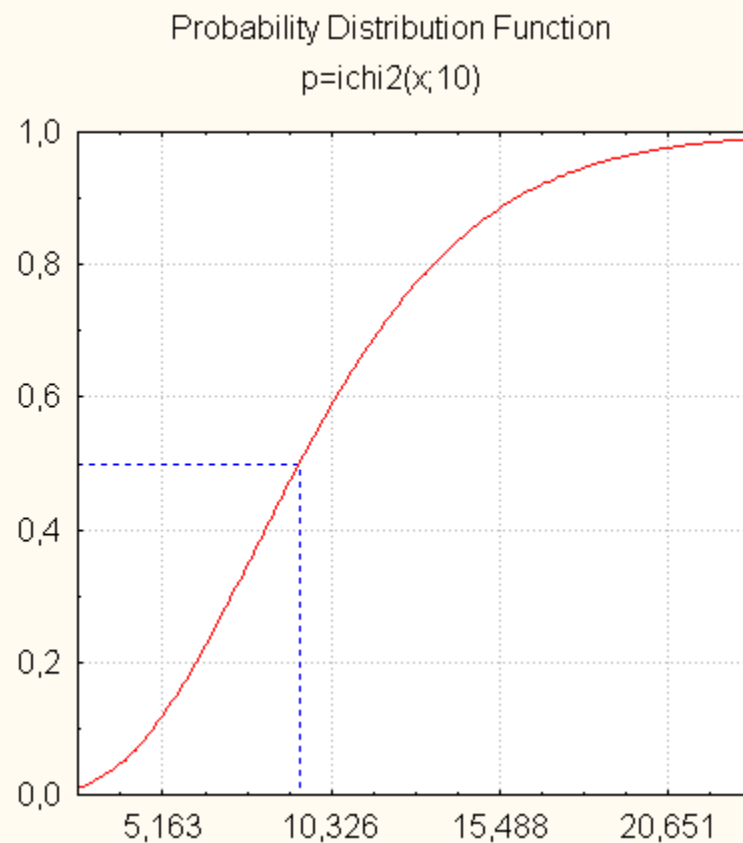
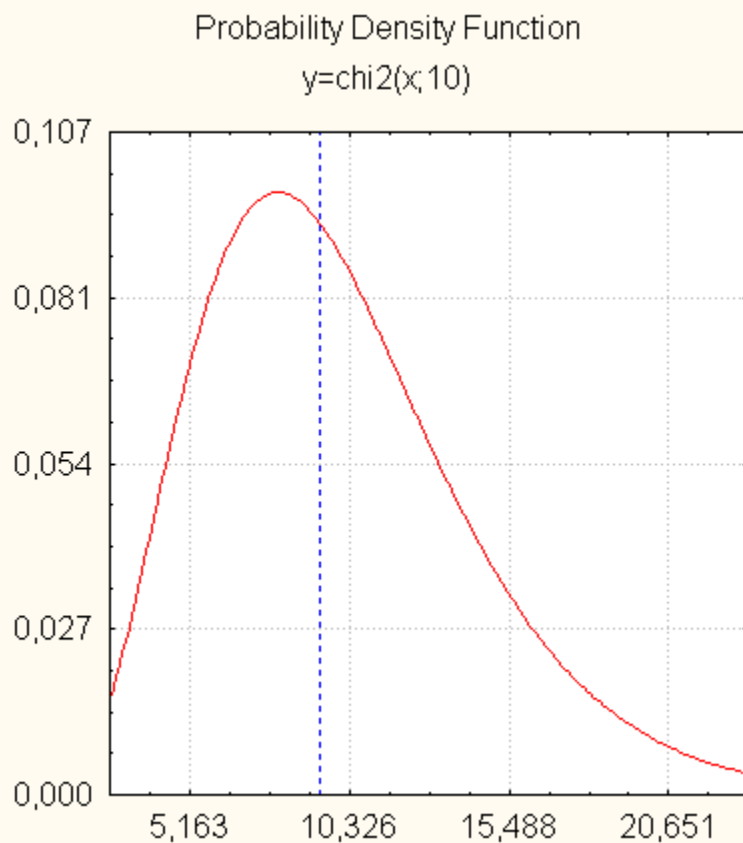
$$\chi^2 \rightarrow N$$

□ Distribuição χ^2

- Não é simétrica. Torna-se mais simétrica a medida que o número de graus de liberdade aumenta ($gl=n-1$),
- Os valores de X^2 podem ser 0 ou positivos (nunca negativos),
- A distribuição χ^2 é diferente para cada grau de liberdade ($gl=n-1$).
- A medida que o número de graus de liberdade aumenta a distribuição χ^2 se aproxima da Normal.

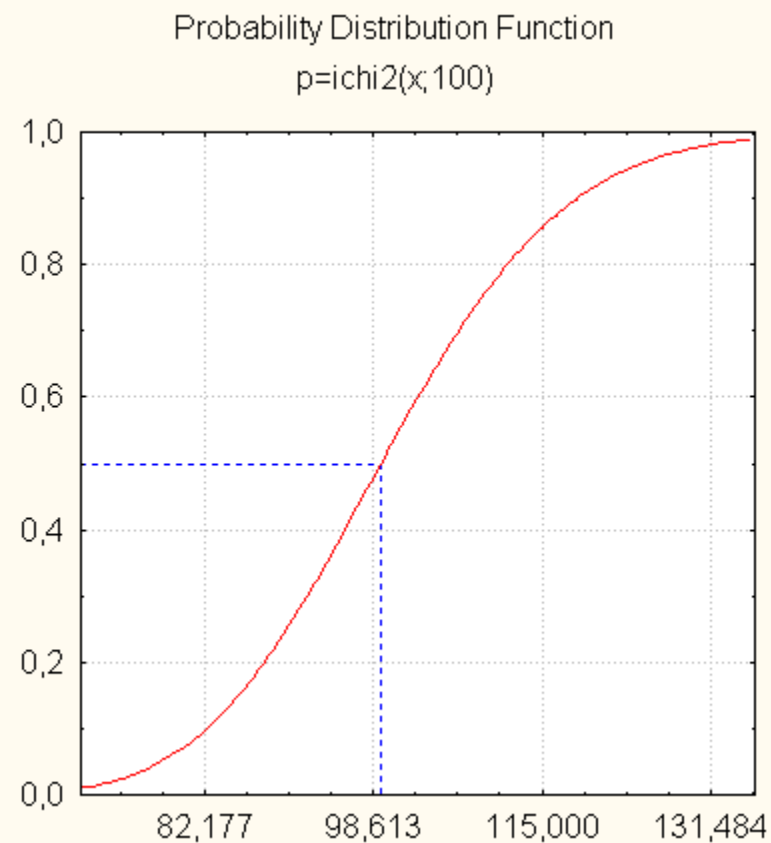
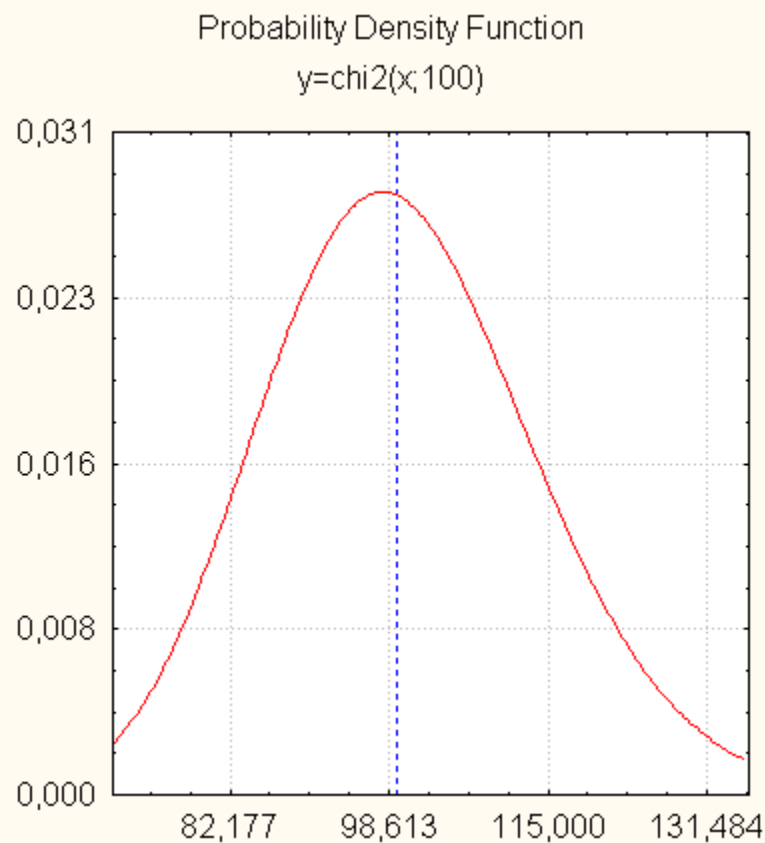
Inferência

□ Distribuição χ^2 (gl=10)



Inferência

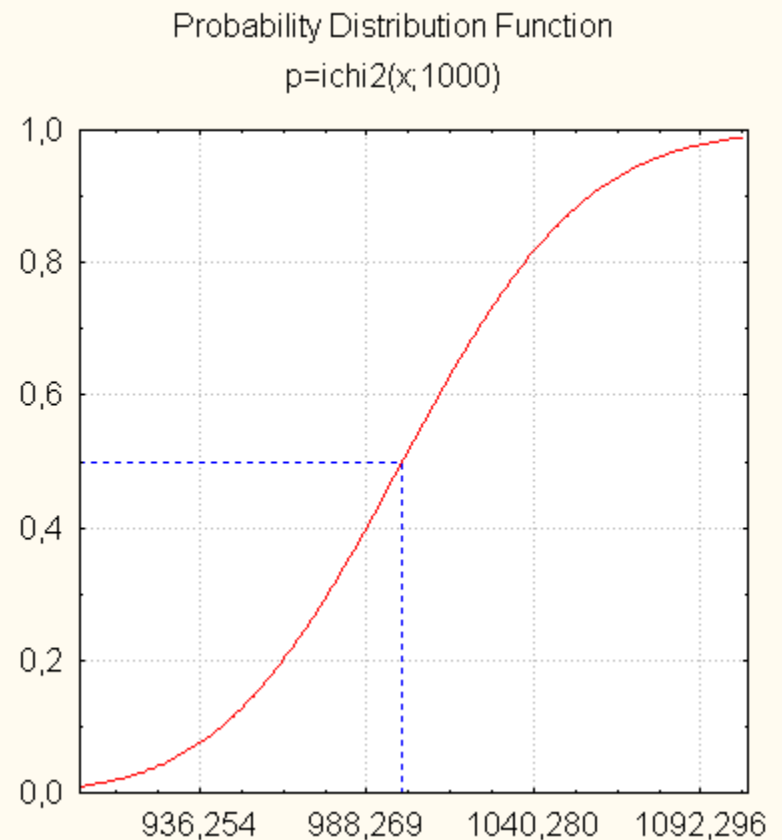
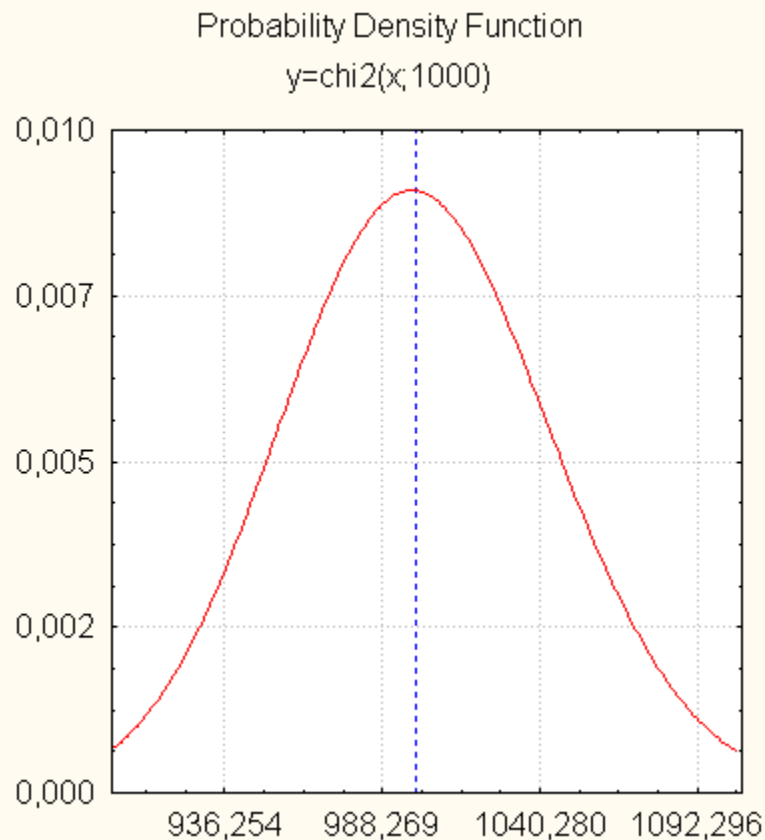
□ Distribuição χ^2 (gl=100)



Inferência

Geração de
Dist. χ^2

□ Distribuição χ^2 (gl=1000)

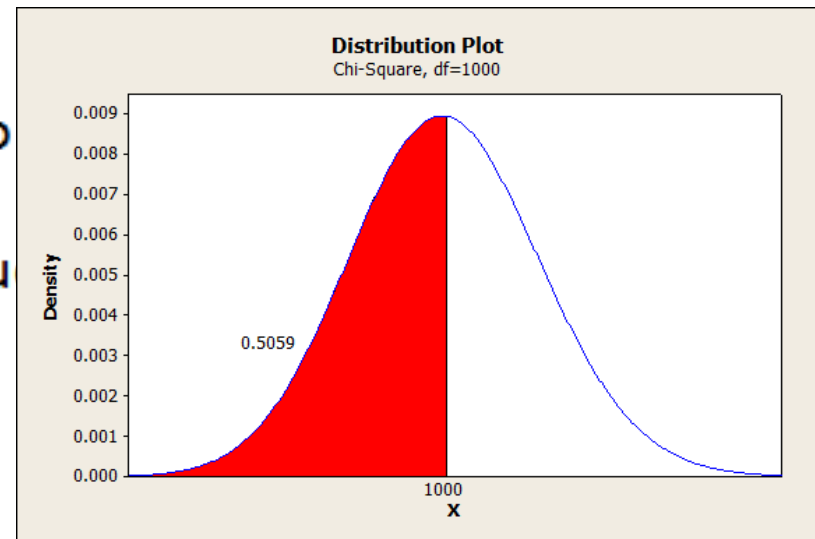
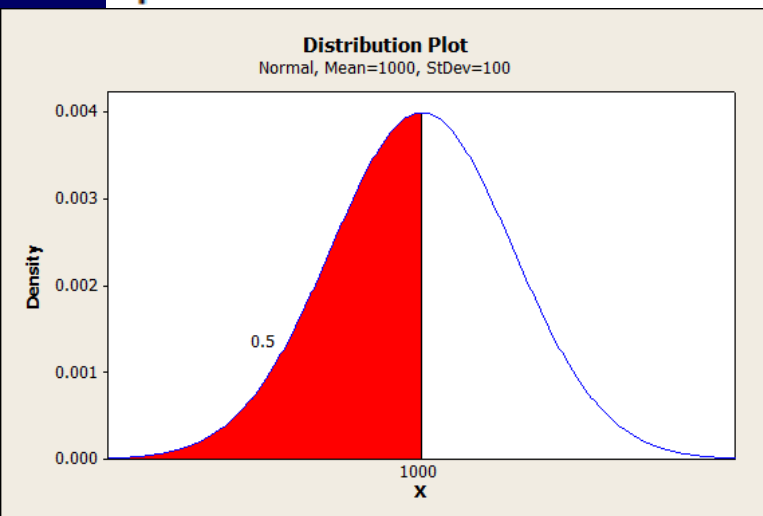


Inferência

Geração de
Dist. χ^2

□ Distribuição χ^2 (gl=1000)

Sabemos que $\mu = n - 1$ e $\sigma^2 = 2(n - 1)$, portanto para $n=1001$,
 $\mu = 1000$ e $\sigma = 100$.



Desta forma, esperamos que

$$P_{N(1000,100)}(X \leq 1000) \cong P_{\chi^2 (df=1000)}(X \leq 1000).$$

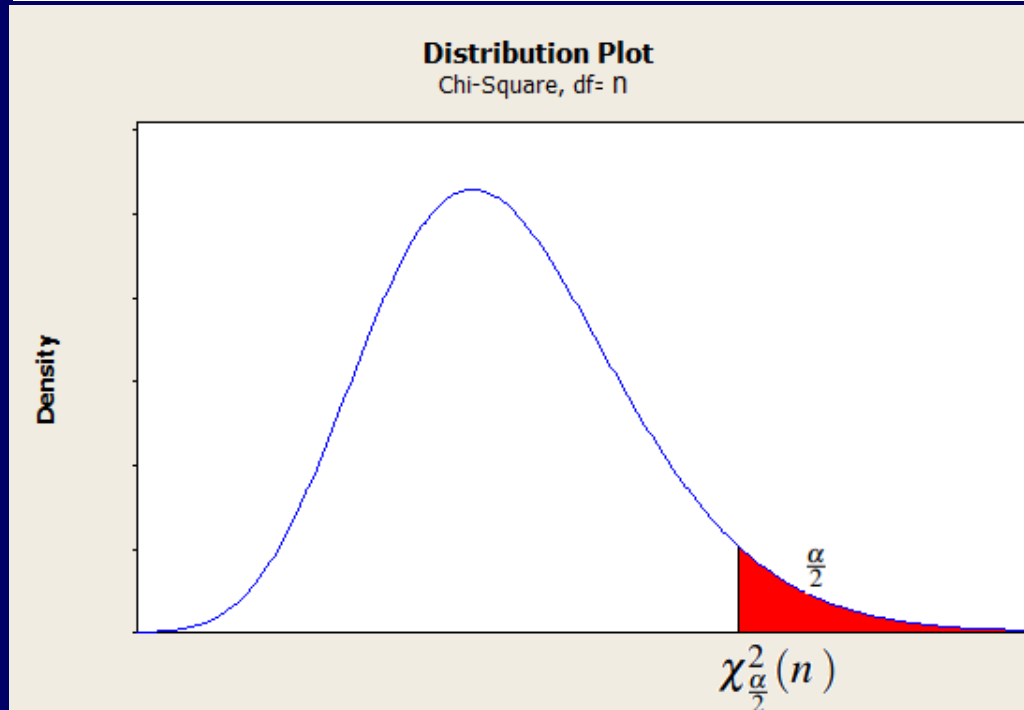
Inferência

Geração de
Dist. χ^2



Intervalo de confiança (IC) para σ^2

$\chi^2_{\frac{\alpha}{2}}(n)$ é um número real cuja área delimitada a sua direita e abaixo da curva de densidade (considerando o grau de liberdade n) é menor que α .



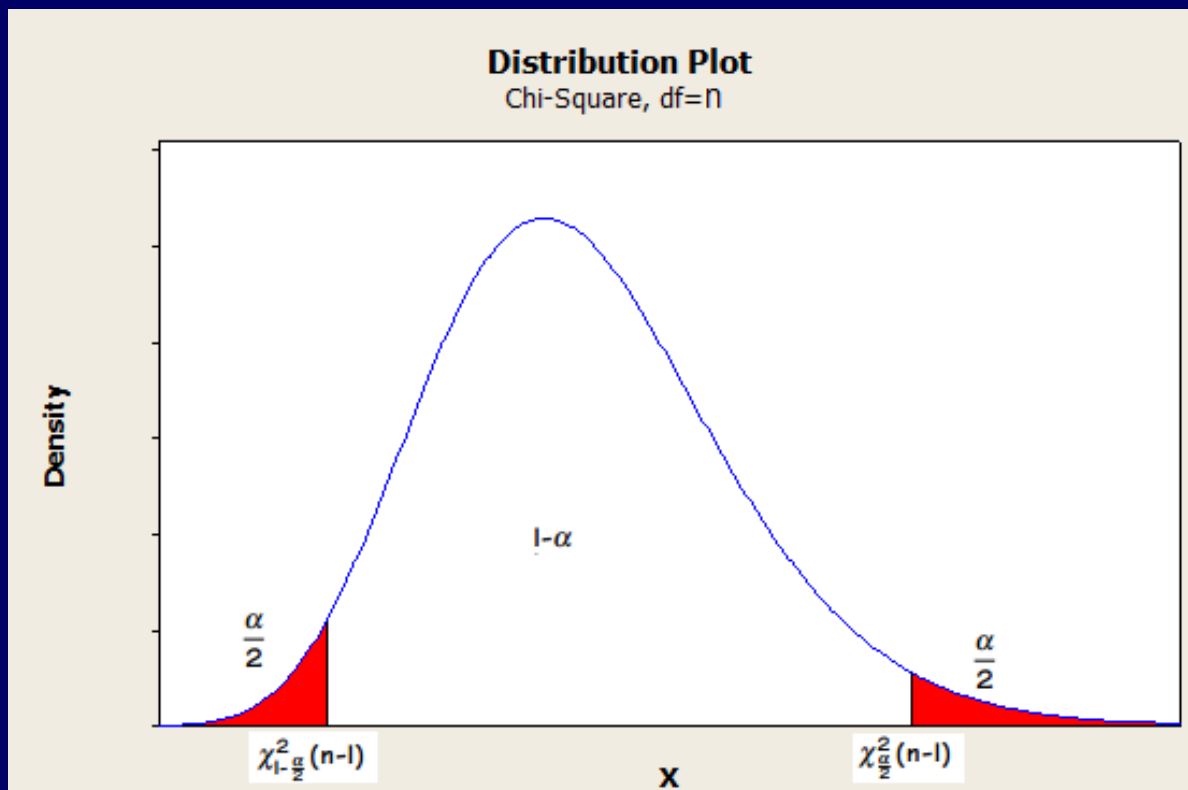
Inferência

Geração de
Dist. χ^2



Intervalo de confiança (IC) para σ^2

$\chi^2_{\frac{\alpha}{2}(n)}$ é um número real cuja área delimitada a sua direita e abaixo da curva de densidade (considerando o grau de liberdade n) é menor que α .



Inferência

Geração de
Dist. χ^2



Excel

Intervalo de confiança (IC) para σ^2

$$P\left(\chi^2_{\alpha/2, n-1} < \frac{(n-1) \times S^2}{\sigma^2} < \chi^2_{1-\alpha/2, n-1}\right) = 1 - \alpha$$

$$\frac{\chi^2_{\alpha/2, n-1}}{(n-1) \times S^2} < \frac{1}{\sigma^2} < \frac{\chi^2_{1-\alpha/2, n-1}}{(n-1) \times S^2}$$

$$\frac{(n-1) \times S^2}{\chi^2_{\alpha/2, n-1}} > \sigma^2 > \frac{(n-1) \times S^2}{\chi^2_{1-\alpha/2, n-1}}$$

Ou

$$\sigma^2 \in \left(\frac{(n-1) \times S^2}{\chi^2_{1-\alpha/2, n-1}}, \frac{(n-1) \times S^2}{\chi^2_{\alpha/2, n-1}} \right)$$

$$\frac{(n-1) \times S^2}{\chi^2_{1-\alpha/2, n-1}} < \sigma^2 < \frac{(n-1) \times S^2}{\chi^2_{\alpha/2, n-1}}$$

$$P\left(\frac{(n-1) \times S^2}{\chi^2_{1-\alpha/2, n-1}} < \sigma^2 < \frac{(n-1) \times S^2}{\chi^2_{\alpha/2, n-1}}\right) = 1 - \alpha$$

Inferência

Excel



Minitab

Intervalo de confiança (IC) para σ^2

$$\sqrt{\frac{(n-1) \times S^2}{\chi^2_{1-\alpha/2, n-1}}} < \sigma < \sqrt{\frac{(n-1) \times S^2}{\chi^2_{\alpha/2, n-1}}}$$

Ou

$$\sigma \in \left(\sqrt{\frac{(n-1) \times S^2}{\chi^2_{1-\alpha/2, n-1}}}, \sqrt{\frac{(n-1) \times S^2}{\chi^2_{\alpha/2, n-1}}} \right)$$

Inferência (resumo)



Excel

Minitab

□ Intervalo de confiança (IC) para σ^2

Resumindo:

$$\sigma^2 \in \left(\frac{(n-1) \times S^2}{\chi^2_{1-\alpha/2, n-1}}, \frac{(n-1) \times S^2}{\chi^2_{\alpha/2, n-1}} \right)$$

$$\sigma \in \left(\sqrt{\frac{(n-1) \times S^2}{\chi^2_{1-\alpha/2, n-1}}}, \sqrt{\frac{(n-1) \times S^2}{\chi^2_{\alpha/2, n-1}}} \right)$$

□ Intervalo de confiança (IC) para σ^2

1. Verifique se a amostra é uma AAS,
2. Verifique se os dados sugerem uma distribuição Normal,
3. Usando $n-1$ graus de liberdade e o nível de confiança desejado $(1-\alpha)$, encontre b e a .
4. Calcule os limites inferior e superior do intervalo

$$\sigma^2 \in \left(\frac{(n-1) \times S^2}{\chi^2_{1-\alpha/2, n-1}}, \frac{(n-1) \times S^2}{\chi^2_{\alpha/2, n-1}} \right)$$

- Se se deseja estimar o intervalo de confiança de σ , tome a raiz quadrada de σ^2 .

Inferência

- Intervalo de confiança (IC) para σ^2
- Exemplo: suponha que o tempo de execução da tarefa T1 foi medido 106 vezes (n), atendendo os requisitos de uma AAS. Os dados parecem provir de uma população normalmente distribuída e o valor médio do tempo da amostra é 98,2 s. O desvio padrão amostral $s = 0,62$ s. Não foram observados *outliers*. Calcule o intervalo de confiança para σ considerando o nível de confiança de 95% ($1 - \alpha$).
- Se $\alpha = 5\%$ e dividindo igualmente entre as duas caudas da distribuição χ^2 , devemos procurar por valores de χ^2 correspondentes as $a = \chi^2(105, (\alpha)/2)$ e
$$b = \chi^2[105, (1-\alpha)/2]$$

Portanto: $0,546 \text{ s} < \sigma < 0,717 \text{ s}$

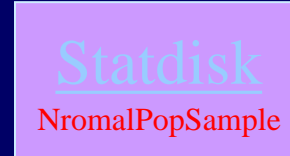
[Statdisk](#)

Statisc

□ Intervalo de Confiança

- Exemplo: suponha que um conjunto de atividades, denominado aqui por A1, executadas por um departamento de uma organização seja normalmente distribuído com desvio padrão desconhecido. Uma amostra aleatória simples, com 1000 medidas, relativa a mensuração do tempo associado a este conjunto de tarefas foi obtido. Estime o a variância e o desvio padrão associado ao deste conjunto de atividade com um nível de confiança de 95%.

Inferência

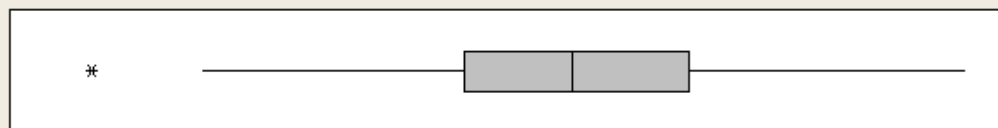
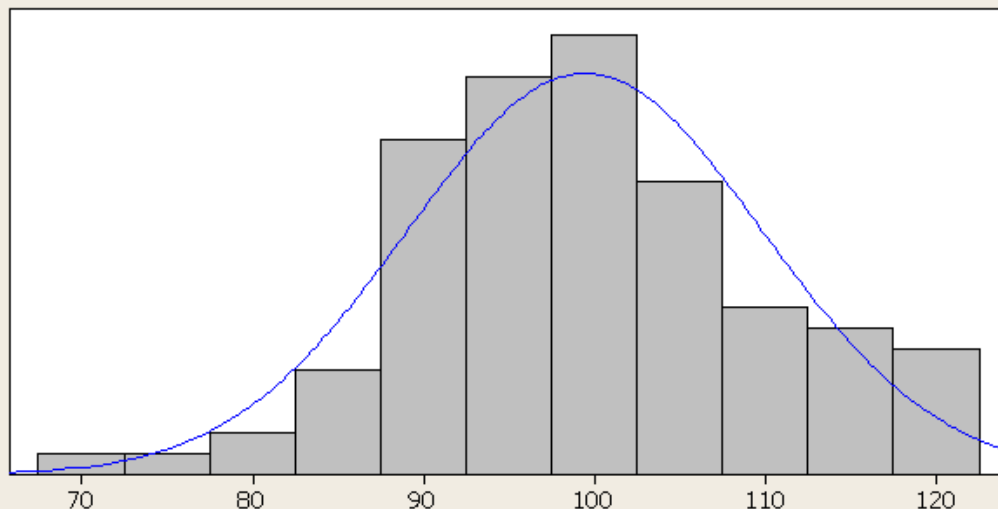


C:\Paulo Maciel\Tools\Statdisk104

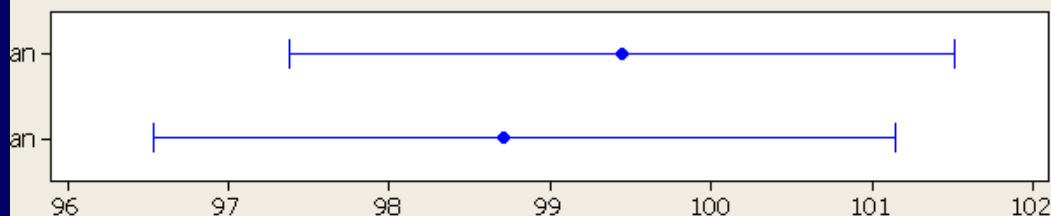
□ Intervalo de Confiança

- Exemplo: desejamos estimar a variância do tempo de serviço (normalmente distribuído) associado a uma determinada atividade de um departamento de prestação de serviço com um nível de confiança de 95%, considerando que o padrão amostral é $s=10$ de uma amostra aleatória simples, de tamanho igual a 100, foi adequadamente coletada.
- Forneça o intervalo de confiança para a variância.

Inferência



95% Confidence Intervals



Anderson-Darling Normality Test

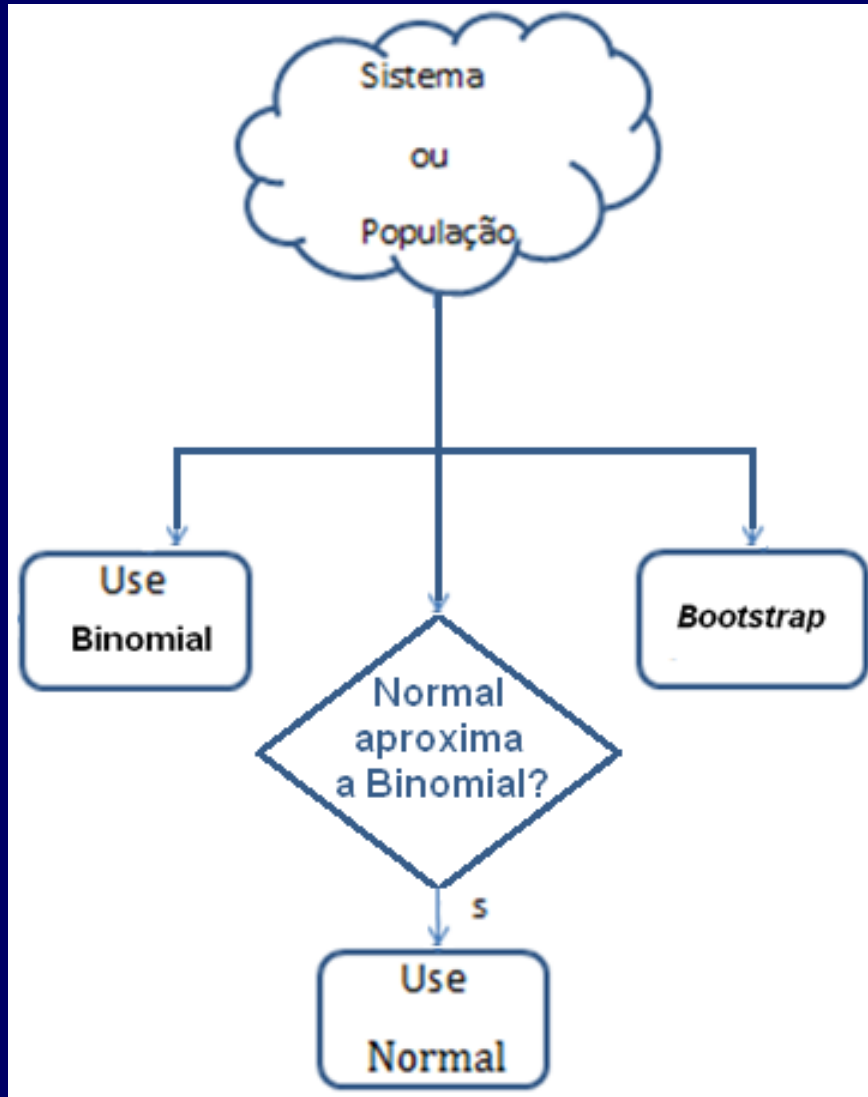
A-Squared	0,38
P-Value	0,403

Mean	99,453
StDev	10,405
Variance	108,267
Skewness	0,0917511
Kurtosis	-0,0554630
N	100

Minimum	70,580
1st Quartile	92,378
Median	98,710
3rd Quartile	105,495
Maximum	121,610

95% Confidence Interval for Mean	
97,388	101,517
95% Confidence Interval for Median	
96,543	101,148
95% Confidence Interval for StDev	
9,136	12,087

Orientação para Inferência - Proporção



Inferência



□ Discreta

– Bernoulli

□ Considere um experimento aleatório com dois resultados possíveis ($X=0, X=1$).

□ pmf(*probability mass function*) de X é dada por:
 $P(X=0) = 1-p$ e $P(X=1) = p$, $0 \leq p \leq 1$

Inferência

□ Discreta

– Bernoulli

□ Parâmetro: p ;

□ Valor Esperado = p ,

□ Variância = $p(1-p)$,

□ Coeficiente de variação = $(1-p)/p$

Inferência

Discreta

□ Binomial

- Considere um experimento aleatório independentes com dois resultados possíveis (0 e 1 por exemplo) realizados n vezes. A variável aleatória é o número de vezes que se tem resultado 1.

□ pmf de X é dada por: $P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$
 $k=0,1,\dots,n.$

Inferência

Binomial
Excel

Binomial

□ Discreta

– Binomial

- Parâmetros: n, p ;
- Valor Esperado= np ,
- Variância= $np(1-p)$,
- Coeficiente de variação= $(1-p)/np$

- A Normal como Aproximação da Binomial
Demonstração Intuitiva:

- *Vamos gerar números aleatórios segundo a distribuição Binomial de seguintes parâmetros:*

Binomial com $n = 20$ e $p = 0,1$

Binomial com $n = 80$ e $p = 0,1$

Binomial com $n = 200$ e $p = 0,1$

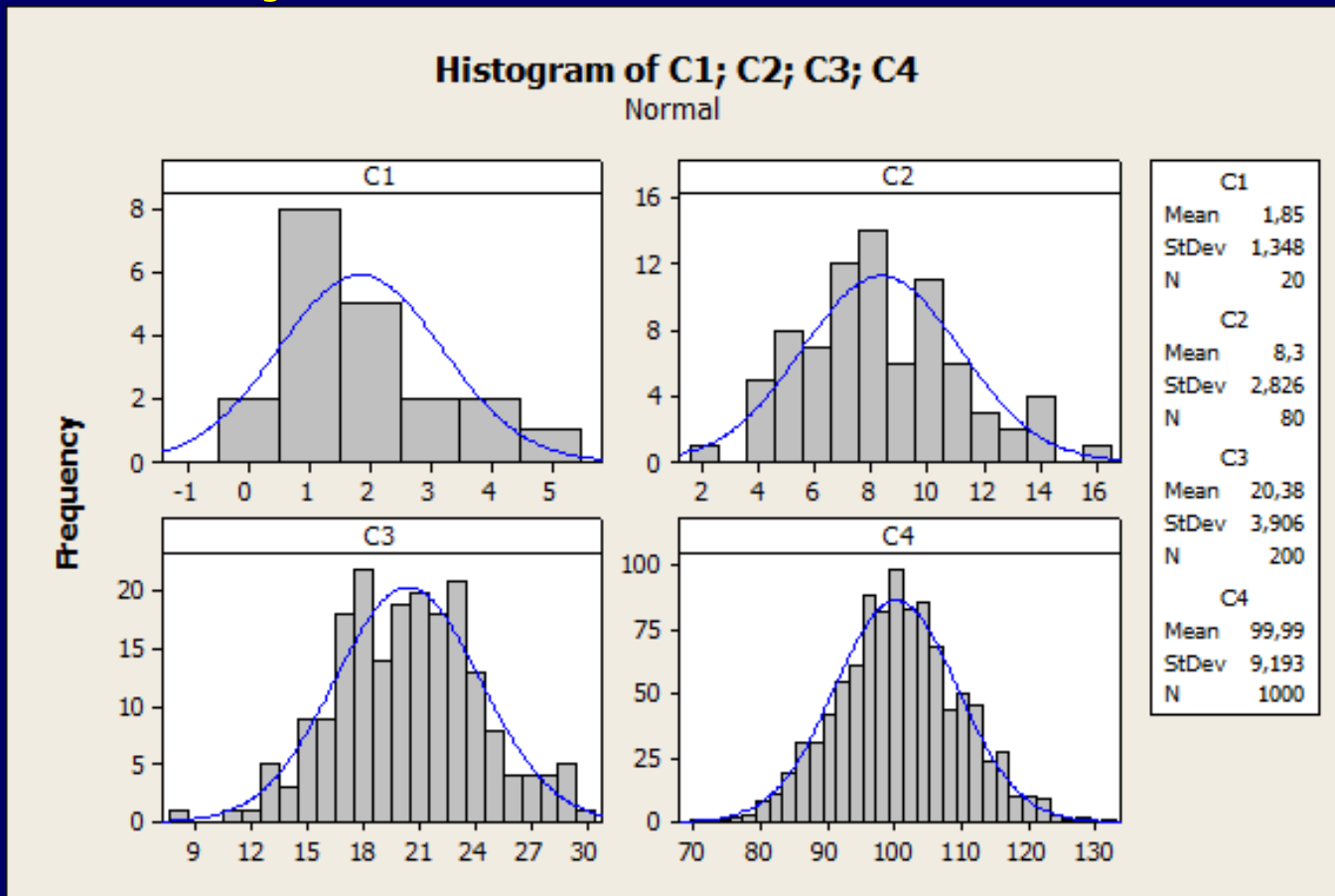
Binomial com $n = 1000$ e $p = 0,1$

Variáveis Aleatórias

Resumo

□ A Normal como Aproximação da Binomial

Demonstração Intuitiva:



□ A Normal como Aproximação da Binomial

Demonstração Intuitiva:

- *Se aumentarmos a probabilidade a distribuição se aproxima de uma Normal mesmo com tamanhos de amostras menores.*

□ *Vamos gerar números aleatórios segundo a distribuição Binomial de seguintes parâmetros:*

Binomial com $n = 20$ e $p = 0,4$

Binomial com $n = 40$ e $p = 0,4$

Binomial com $n = 80$ e $p = 0,4$

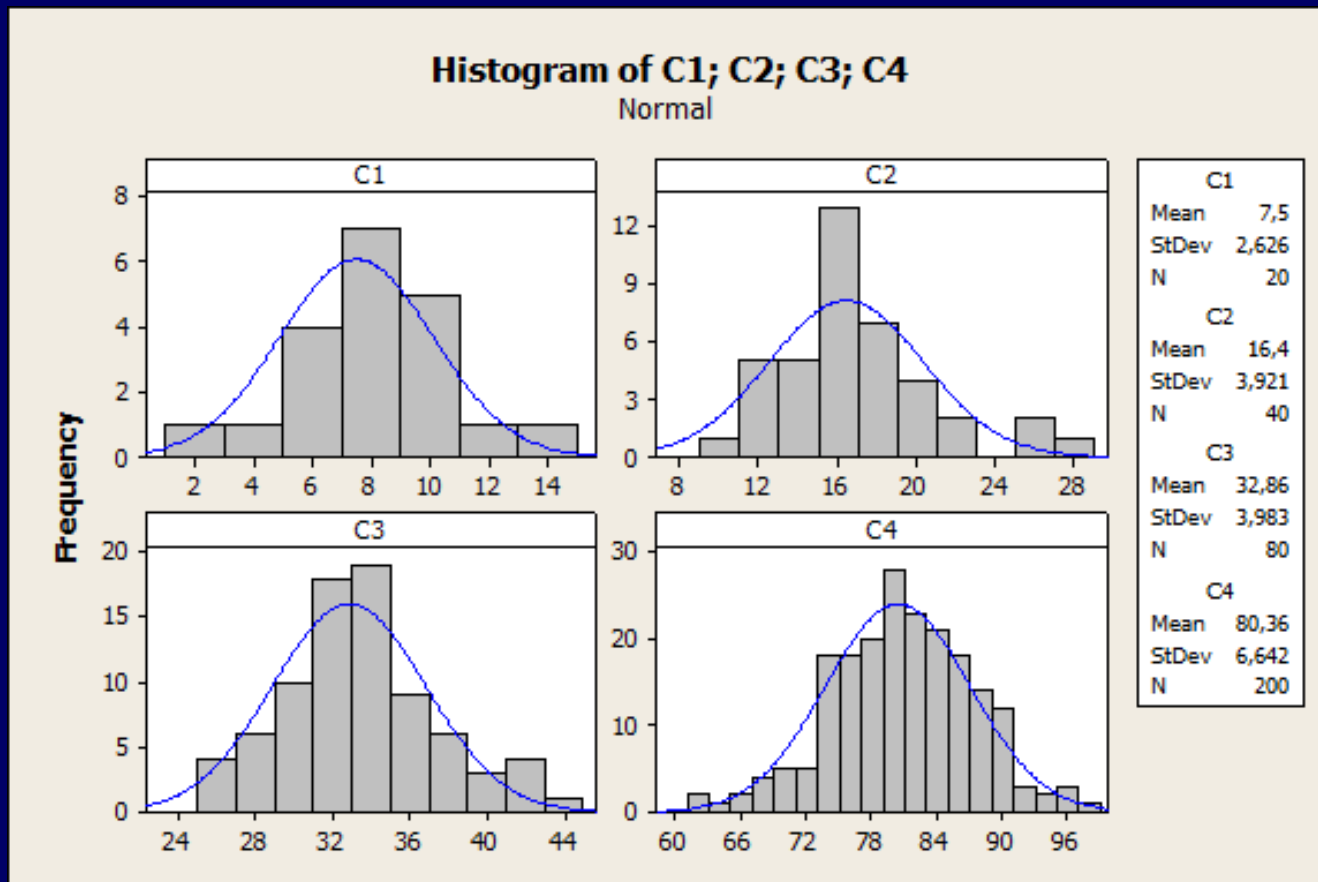
Binomial com $n = 200$ e $p = 0,4$

Variáveis Aleatórias

Resumo

A Normal como Aproximação da Binomial

Demonstração Intuitiva:



Variáveis Aleatórias

Resumo



□ A Normal como Aproximação da Binomial

Condições Necessárias da distribuição de probabilidade Binomial:

1. o procedimento deve ter um número fixo de repetições,
2. as repetições devem ser independentes,
3. cada repetição deve ter todos os resultados classificados em duas categorias,
4. as probabilidades devem permanecer constantes para cada repetição.
5. $np \geq 5$ ou $nq \geq 5$, onde $(p=1-q)$

Variáveis Aleatórias

Resumo





- A Normal como Aproximação da Binomial
 - $\mu = np$
 - $\sigma = \sqrt{npq}$
- A demonstração formal é conhecida como aproximação DeMoivre-Laplace



Inferência

□ Intervalo de Confiança para Proporção Populacional (p):

1. Verifique se a amostra é AAS,
2. Verifique as condições necessárias para Distribuição Binomial, 
3. Verifique se a Normal pode aproximar a Binomial, 
4. Encontre o valor crítico Z^* ($Z_{\alpha/2}$),
5. Calcule a margem de erro $E = Z^* \sqrt{(p'q')/n}$
(p' proporção amostral)
6. Encontre $p' - E < p < p' + E$ ou $p' \pm E$ ou $(p' - E, p' + E)$



- Intervalo de Confiança para Proporção Populacional (p):
- Exemplo: suponha que tenhamos uma amostra $n=829$, com 426 sucessos, $Z^*=1,96$ (95/% de confiança). Calcule o intervalo de confiança para a proporção.



Intervalo de Confiança para Proporção Populacional (p):

The point estimate of the proportion is obtained through $p' = \frac{k}{n}$.

Now, take into account that: $S_{\leq \frac{\alpha}{2}} = \left\{ k_l \mid k_l = \max_{k'} \{k'\}, P(X \leq k') \leq \frac{\alpha}{2}, k' \in \mathbb{N} \right\}$,

where $S_{\leq \frac{\alpha}{2}}$ is a unity set that has the maximal k' , that is k_l , satisfying $P(X \leq k') \leq \frac{\alpha}{2}$.

It is known that $P(X \leq k_l) \leq \frac{\alpha}{2} \iff P(X \leq k_l) = \sum_{k'=1}^{k_l} \binom{n}{k'} \times p'^{k'} \times (1 - p')^{n-k'} \leq \frac{\alpha}{2}$

Therefore, the lower limit is the confidence interval (two-tailed) with $(1 - \alpha)\%$ is

$$p_l = \frac{k_l}{n}.$$



Intervalo de Confiança para Proporção Populacional (p):

The point estimate of the proportion is obtained through $p' = \frac{k}{n}$.

Now, take into account that: $S_{\leq \frac{\alpha}{2}} = \left\{ k_l \mid k_l = \max_{k'} \{k'\}, P(X \leq k') \leq \frac{\alpha}{2}, k' \in \mathbb{N} \right\}$,

where $S_{\leq \frac{\alpha}{2}}$ is a unity set that has the maximal k' , that is k_l , satisfying $P(X \leq k') \leq \frac{\alpha}{2}$.

It is known that $P(X \leq k_l) \leq \frac{\alpha}{2} \iff P(X \leq k_l) = \sum_{k'=1}^{k_l} \binom{n}{k'} \times p'^{k'} \times (1 - p')^{n-k'} \leq \frac{\alpha}{2}$

Therefore, the lower limit is the confidence interval (two-tailed) with $(1 - \alpha)\%$ is

$$p_l = \frac{k_l}{n}.$$



Inferência

Intervalo de Confiança para Proporção Populacional (p):

Now consider

$$S_{\leq 1 - \frac{\alpha}{2}} = \left\{ k_u \mid k_u = \max_{k'} \{k'\}, P(X \geq k') \leq 1 - \frac{\alpha}{2}, k' \in \mathbb{N} \right\},$$

that is, $S_{\leq 1 - \frac{\alpha}{2}}$ is a unity set that has the maximal k' , that is k_u , which satisfies $P(X \leq k') \leq 1 - \frac{\alpha}{2}$.

Hence,
$$P(X \leq k_u) \leq 1 - \frac{\alpha}{2} \iff P(X \leq k_u) = \sum_{k'=1}^{k_u} \binom{n}{k'} \times p'^{k'} \times (1 - p')^{n-k'} \leq 1 - \frac{\alpha}{2}.$$

Therefore, the upper limit of the confidence interval (two-tailed) with the significance degree equal to (α) % is

$$p_u = \frac{k_u}{n}.$$

The confidence interval of the ratio $((1 - \alpha)\%$ confidence) is:

$$p \in (p_l, p_u) = \left(\frac{k_l}{n}, \frac{k_u}{n} \right).$$

Inferência

Intervalo de Confiança para Proporção Populacional (p):

Exemplo - Suponha que no último mês tenhamos testado os sensores de localização de 25843 telefones celulares. Observaram-se 342 equipamentos com defeito nos seus sistemas de localização. Desejamos calcular o intervalo de confiança para a proporção de defeitos com 95% de confiança.

Da descrição acima, temos $n = 25843$ celulares testados, dos quais $m = 342$ apresentaram defeitos. Portanto:

$$p' = \frac{m}{n} = \frac{342}{25843} = 0,013234.$$

Vamos considerar que todos os computadores são similares e que estavam em condições semelhantes durante todo o período de avaliação. Considerando um nível de confiança de $1 - \alpha = 0,95$, encontramos o k que¹ satisfaz:





Inferência

Intervalo de Confiança para Proporção Populacional (p):

$$p' = \frac{m}{n} = \frac{342}{25843} = 0.013234.$$

$$1 - \alpha = 95 \%$$

$$S_{\leq 0.025} = \{k | P(X \leq k) \leq 0.025\} = \{305\}.$$

$$S_{\leq 0.975} = \{k | P(X \geq k) \leq 0.975\} = \{378\}.$$

$$P(X \leq 305) = F_{b(25843)}(305) = 0.02197$$

$$P(X \leq 378) = F_{b(25843)}(378) = 0.97511$$

and

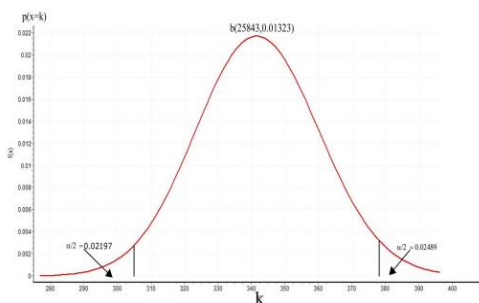
and

$$P(X \leq 306) = F_{b(25843)}(306) = 0.0251.$$

$$P(X \leq 377) = F_{b(25843)}(377) = 0.97188.$$

$$p_l = \frac{k}{n} = \frac{305}{25843} = 0.011802.$$

$$p_u = \frac{k}{n} = \frac{378}{25843} = 0.01463.$$



$$p \in (p_l, p_u) = \left(\frac{k_l}{n}, \frac{k_u}{n} \right)$$

$$p \in \left(\frac{305}{25843}, \frac{378}{25843} \right) = (0.011802, 0.01463)$$

Inferência



Session





Results for: PROPORINTCONF.MTW

Test and CI for One Proportion: C1

Event = 1

Variable	X	N	Sample p	95% CI
C1	426	829	0,513872	(0,479248; 0,548397)

Inferência

 Statdisk1   

Conf. Int. for Prop.

Confidence Level, (1- α):

Sample Size, n:

Num Successes, x:

Margin of error, E = 0.03403

95% confident that the prop.
is within the range:

$0.47622 < p < 0.54428$

Inferência

Considere uma situação em que precisamos estimar o tempo de execução que um processador passa executando uma determinada função A de uma aplicação computacional. No entanto os mecanismos de medição disponíveis não possuem resolução suficiente para mensurar diretamente a referida função (o tempo de execução da função é muito pequeno).

Dispomos de mecanismos que nos possibilitam verificar se o processador está executando a função ou não (observando o mapa de memória da aplicação e o valor do apontador de programa).

Utilizaremos um mecanismo de amostragem para periodicamente verificar se o processador está executando a função ou não. O tempo entre amostras é de 100ms.

Coletamos 1000 amostras. Portanto o tempo total de observação (TTO) foi de $1000 \times 100 \times 10^{-3} = 100s$.

Inferência

Indirect
Measurement

A amostra coletada está na tabela. 0 significa que o processador não estava executando a função A e 1 significa que o processador estava executando a função A.

p' – proporção amostral

$$q' = 1 - p'$$

$$E = \sqrt{\frac{p' \times q'}{n}}, \quad n \text{ é o tamanho da amostra} \leq Z_{\frac{\alpha}{2}} \sqrt{\frac{p' q'}{n}}$$

$$p' \pm E \Leftrightarrow p' - E \leq p \leq p' + E$$

Inferência

Indirect
Measurement

$TTO(p' - E; p' + E) = 100(0,026183; 0,050641) =$
 $(2,6183; 5,0641)s$ com 95% de confiança.

Se aumentarmos o número de amostras o intervalo diminui.

Observe a magnitude dos valores do intervalo e o TTO!

Se se sabe que o código foi executado 10 vezes, portanto o tempo médio de execução da função A está entre $(0,26183; 0,50641)s$ com 95% de confiança.

Inferência

Indirect
Measurement

Desta forma agora considere, que as amostras foram coletadas a cada 100ms (mesma frequência de amostragem) e 0 significa não estar executando a função A e 1 significa estar executando a função.

No entanto, o número de amostras foi 10000. Portanto o Tempo Total de Observação (TTO) foi de 1000s.

Desta forma

$TTO \times (0,045714; 0,054351) = 1000 (0,045714; 0,054351) = (45,714; 54,351)s$ com 95 % de confiança.

Se se sabe que o código foi executado 100 vezes, portanto o tempo médio

de execução da função A está entre (0,45714; 0,54351)s com 95% de confiança.

Inferência

Excel

Minitab

Considere uma máquina automática de venda que oferta 8 tipos de produtos, $TP = \{A, B, C, D, E, F, G, H\}$. Considere que se obteve uma amostra de tamanho 2000 coletada de forma apropriada e em período significativo.

Estime o percentual de venda do Produto C com 95 % de confiança.

A amostra está na planilha. Classificamos os produtos em duas classes: Produto C e Demais Produtos. Definimos uma variável aleatória C tal que:

$$C: TP \rightarrow \begin{cases} 1 & \text{se produto} = C \\ 0 & \text{se produto} \neq C \end{cases}$$

Inferência

PLA

Considere um setor de controle de qualidade de uma fábrica F. O setor de controle de qualidade coleta regularmente amostras dos produtos da linha de produção A para realização de testes. Diariamente uma amostra de 10 produtos é aleatoriamente coletada da linha de produção A para execução dos testes. A planilha (PLA) apresenta os resultados dos testes executados para cada um dos produtos das 30 amostras (a planilha apresenta informações relativas a amostras de 30 dias).

Test and CI for One Proportion: Teste

Event = 1

Variable	X	N	Sample p	95% CI
Teste	18	300	0,060000	(0,035944; 0,093170)

Inferência

ET>1
Minitab

ET>1
Excel

Considere a execução do programa A em um computador C. Estime o intervalo de confiança da proporção que contenha a probabilidade de que o tempo de execução do programa A seja maior que 215ms. Adote $\alpha = 5\%$.

Test and CI for One Proportion: ET>215

Event = 1

Variable	X	N	Sample p	95% CI
ET>215	11	100	0.110000	(0.048675, 0.171325)

Using the normal approximation.

Test and CI for One Proportion: ET>215

Event = 1

Variable	X	N	Sample p	95% CI
ET>215	11	100	0.110000	(0.056207, 0.188301)

Inferência

- Determinação do tamanho da amostra para estimar a Proporção Populacional (p):
 - $n = [(Z^*)^2 p'q']/E^2$ - quando se conhece a estimativa p' .
 - $n = [(Z^*)^2 0,25]/E^2$ - quando não se conhece a estimativa p' . (assume-se $p'=0,5$ e $q'=0,5$)

□ Determinação do tamanho da amostra para estimar a Proporção Populacional (p):

– Exemplo: $p'=0,2$; $E=0,04$ e $\alpha=5\%$, portanto o tamanho da amostra deve ser:

$$n = [(1,96)^2 \times 0,2 \times 0,8]/0,04 = 385$$

Inferência

Statdisk2

Sample Size Required to Estimate Proportion

Confidence Level, $(1-\alpha)$:

Margin of Error, E :

Estimate Proportion, p :
(if known)

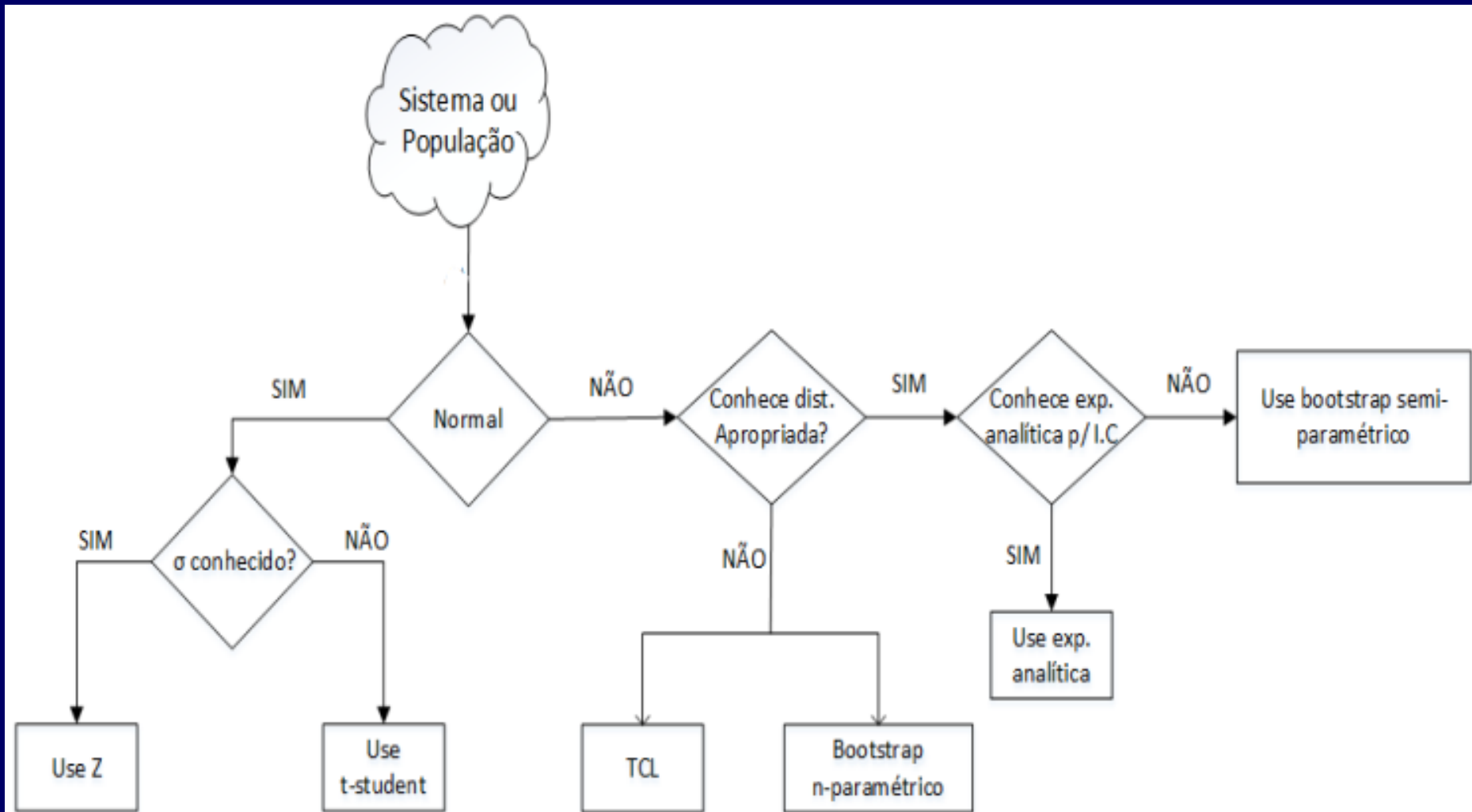
Population Size, N :
(if known)

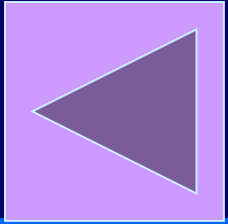
Required sample size is:

$n = 1537$

Assumed either infinite population or
the population was sampled with
replacement

Orientação para Inferência – Média

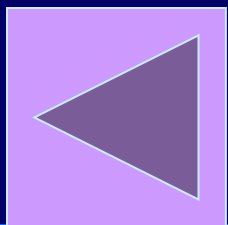




Inferência

□ *Bootstrap* (Re-amostragem)

- É um procedimento utilizado para obtermos aproximações de distribuições amostrais quando a teoria não pode dizer-nos qual a sua forma da distribuição da população. Pode ser aplicada também quando o tamanho da amostra é pequeno (dificuldade ou alto custo para obtenção de amostras maiores).



Inferência

Statdisk

Mathematica

Gerado
pelo
Statdisk
(Excel)

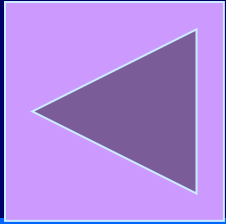
Excel_1

Semi-parametric
Bootstrap

C:\Paulo Maciel\Paulo Maciel\Tools\Minitab 14\Data\Data\Bootstrap.txt

□ *Bootstrap* (Re-amostragem)

1. Selecione uma amostra aleatória de tamanho n .
2. Selecione uma amostra da amostra (re-amostra) de tamanho n com reposição.
3. Calcule a estatística (a média, por exemplo) desta amostra.
4. Repita os passos de 2 a 3 m vezes. (m é grande)
5. Classifique as m estatísticas (médias, por exemplo) em ordem ascendente.
6. Em função do nível de confiança desejado $(1-\alpha)$, determine os valores que estejam a $(\alpha/2*100\%)$ acima do menor valor e $(\alpha/2*100\%)$ abaixo do maior valor.



Inferência

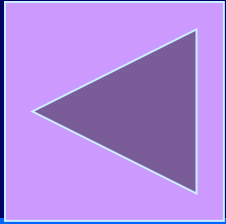
Mean

Skewness

□ *Bootstrap-t*

- Selecione uma amostra aleatória de tamanho n .
- Selecione uma amostra da amostra (re-amostra) de tamanho n com reposição.
- Calcule a estatística de interesse ($\hat{\theta}_j$) desta amostra (a média, por exemplo) .
- Repita os passos de 2 a 3 m vezes. (m é grande)
- Calcule a média da estatísticas de interesse $\bar{\theta} = \sum_{j=1}^m \frac{\hat{\theta}_j}{m}$
- Calcule o desvio padrão da estatística de interesse $SD_{\hat{\theta}_j} = \sqrt{\frac{\sum_{j=1}^m (\hat{\theta}_j - \bar{\theta})^2}{m-1}}$
- Obtenha $t_{(\frac{\alpha}{2}, m-1)}$
- Determine o intervalo de confiança desejado da estatística através de

$$\bar{\theta} - t_{(\frac{\alpha}{2}, m-1)} \times \frac{SD_{\hat{\theta}_j}}{\sqrt{m}} \leq \theta \leq \bar{\theta} + t_{(\frac{\alpha}{2}, m-1)} \times \frac{SD_{\hat{\theta}_j}}{\sqrt{m}}$$



Inferência

Semi-parametric
Bootstrap

C:\Paulo Maciel\Paulo Maciel\Tools\Minitab 14\Data\Data\Boo

□ *Bootstrap Semi-Paramétrico*

1. Selecione uma amostra aleatória de tamanho n .
2. Estime os parâmetros da distribuição, considerando a amostra.
3. Gere uma re-amostra de tamanho n .
4. Calcule a estatística (a média, por exemplo) desta amostra.
5. Repita os passos de 2 a 3 m vezes. (m é grande)
6. Classifique as m estatísticas (médias, por exemplo) em ordem ascendente.
7. Em função do nível de confiança desejado $(1-\alpha)$, determine os valores que estejam a $(\alpha/2*100\%)$ acima do menor valor e $(\alpha/2*100\%)$ abaixo do maior valor.

Prática

O *round trip time* (RTT) relativo a execução de uma transação foi registrado no arquivo da planilha1x. A planilha contém uma amostra de tamanho 500. Apresente o intervalo de confiança do RTT, considerando a amostra da planilha e $1 - \alpha = 95\%$. Adote inferência estatística paramétrica clássica se for possível e justique. Caso a amostra tenha que ser tratada e medidas adicionais sejam necessárias, considere que essas medidas adicionais estão disponível na planilha12x.

Prática

Considerando a amostra do exemplo anterior, apresente o intervalo de confiança do RTT, adotando o método bootstrap. Considere o nível de significância de 5%. Compare o intervalo obtido com o do exercício anterior.

Prática

Estime o intervalo de confiança do tempo de execução de um *benchmark*. Uma amostra de tamanho 400 com os tempos de execução medidos está disponível na planilha2. Adote a estatística paramétrica clássica se for possível e justique. Caso a amostra tenha que ser tratada e medidas adicionais sejam necessárias, considere que as medidas adicionais estão disponível na planilha12x.

Prática

O setor de qualidade de uma indústria alimentícia coleta periodicamente amostras para averiguar o peso líquido de um produto. O padrão de qualidade da empresa especifica que o desvio padrão dos pesos das embalagens não pode ser maior que 2% do peso líquido especificado. O peso líquido especificado para cada embalagem é 100g. Uma amostra foi obtida e os pesos líquidos de cada embalagem da amostra estão disponíveis na planilha peso.xlsx. Calcule o intervalo de confiança do desvio padrão, considerando a amostra e informe se o sistema de produção está sob controle ou se está produzindo itens fora da especificação. Se for possível, use inferência estatística paramétrica clássica, caso contrário adote bootstrap.

Prática

Considere que um sistema de medição baseado em *pc-sampling* foi adotado para estimar o tempo de execução da função A de uma aplicação de software. A aplicação foi monitorada até se obter 1000 amostras. A frequência através da qual as amostras foram coletadas foi 0,1 KHz. Sabe-se que a aplicação foi executada 5 vezes durante o período de monitoração.

A planilha3 apresenta os dados obtidos durante a monitoração. 1 significa que o valor do *program counter* (PC) continha um endereço que correspondia a um endereço da função A. 0 denota que o PC continha um valor que não correspondia a um endereço da função A.

Apresente o intervalo de confiança do tempo de execução da função A, considerando $\alpha = 5\%$.

Projeto

<http://www.mrtc.mdh.se/projects/wcet/benchmarks.html>

<https://github.com/screwtop/diskbench>

Elabore uma metodologia para avaliar e comparar o desempenho de dois computadores da família X86. Apresente o fluxo de atividades e o documento que descreve detalhadamente os pré-requisitos, os insumos, ações, produtos e condições que sinalizam a finalização de cada uma das atividades do fluxo de atividades. O documento deve descrever a carga adotada, as métricas analisadas, as ferramentas a serem utilizadas, as técnicas a serem adotadas, as fórmulas e os procedimentos de execução do processo de medição, análise e diagnóstico.

Seguindo a metodologia estabelecida, avaliem, apresentem resultados e o diagnóstico do desempenho de dois computadores de fabricantes distintos com mesmo sistema operacional, cujos processadores devem ser da família X86. A avaliação deve considerar pelo menos dez (10) programas do WCET Project Benchmark (<http://www.mrtc.mdh.se/projects/wcet/benchmarks.html>) e do DiskBench (<https://github.com/screwtop/diskbench>) – para operações em disco. Selecione programas que “exercitem” operações inteira, de ponto flutuante, loops, recursão, operações com matrizes, dados estruturados e não-estruturados e operações em disco.

Inferência

□ Teste de Postos com Sinais de Wilcoxon

1. Deseja-se testa a hipótese $H_0: \tilde{u} = \tilde{u}_0$ contra alternativas ($H_1: \tilde{u} \neq \tilde{u}_0$)
2. Suponha que X_1, X_2, \dots, X_n seja uma amostra aleatória com média (mediana) igual a \tilde{u} .
3. Calcule as diferenças $X_i - \tilde{u}$, $i=1, 2, \dots, n$.
4. Ordene os valores absolutos da diferença,
5. $|X_i - \tilde{u}|$, $i=1, 2, \dots, n$, em ordem crescente.
6. Faça R^+ a soma dos postos positivos e R^- a soma dos postos negativos.
7. $R = \min(R^+, R^-)$.
8. Considerando-se, o tamanho da amostra (n) e o nível de significância (α), encontra-se o valor crítico de R^*_α .
9. Se $R > R^*_\alpha$ não se pode rejeitar a hipótese nula ($\tilde{u} = \tilde{u}_0$).

Inferência

Sinais com
Postos de
Wilcoxon

□ Teste de Postos com Sinais de Wilcoxon

Obs.	Xi	Xi-m
1	2158.70	158.70
2	1678.15	-321.85
3	2316.00	316.00
4	2061.30	61.30
5	2207.50	207.50
6	1708.30	-291.70
7	1784.70	-215.30
8	2575.00	575.00
9	2357.90	357.90
10	2256.70	256.70
11	2165.20	165.20
12	2399.55	399.55
13	1779.80	-220.20
14	2336.75	336.75
15	1765.30	-234.70
16	2053.50	53.50
17	2414.40	414.40
18	2200.50	200.50
19	2654.20	654.20
20	1753.70	-246.30

Testar a hipótese:

$H_0: \tilde{\mu} = 2000$

$H_1: \tilde{\mu} \neq 2000$

Com 95% de confiança

Inferência

Sinais com
Postos de
Wilcoxon

□ Teste de Postos com Sinais de Wilcoxon

Obs	Xi-m (Classificada)	Posto com Sinal
16	53.50	1
4	61.30	2
1	158.70	3
11	165.20	4
18	200.50	5
5	207.50	6
7	-215.30	-7
13	-220.20	-8
15	-234.70	-9
20	-246.30	-10
10	256.70	11
6	-291.70	-12
3	316.00	13
2	-321.85	-14
14	336.75	15
9	357.90	16
12	399.55	17
17	414.40	18
8	575.00	19
19	654.20	20

150	-60
R+	R-
R	150
R*	52

H0 não pode ser rejeitada

Inferência

Goodness of fitting

Excel

Excel

Mathematica

Graphical Methods

1. Collect the sample $s = \{t\}$, $|s| = n$.
2. Sort the sample in ascending order. Therefore, we have $s = \{t_1, t_2, \dots, t_i, \dots, t_n\}$, where t^1 is the smallest measure, t_2 is the second smallest measure, and so forth; and t_n is the largest measure of the sample.
3. Compute the empirical distribution - $F_E(t_i) = i/(n+1)$ or i/n or $(i-0.3)/(n+0.4)$.
4. Choose the theoretical distribution, F_T .
5. Estimate the parameters of F_T from the sample s .
6. Plot F_T against F_E .
7. Check how close the dots are from a 45° straight line. In case they deviate, the distributions differ. Therefore, check the angular coefficient (a) of the linear regression, and the intercept (b). The coefficient of determination, r^2 , shows how close the dots are from a straight line.

Inferência

Goodness of fitting

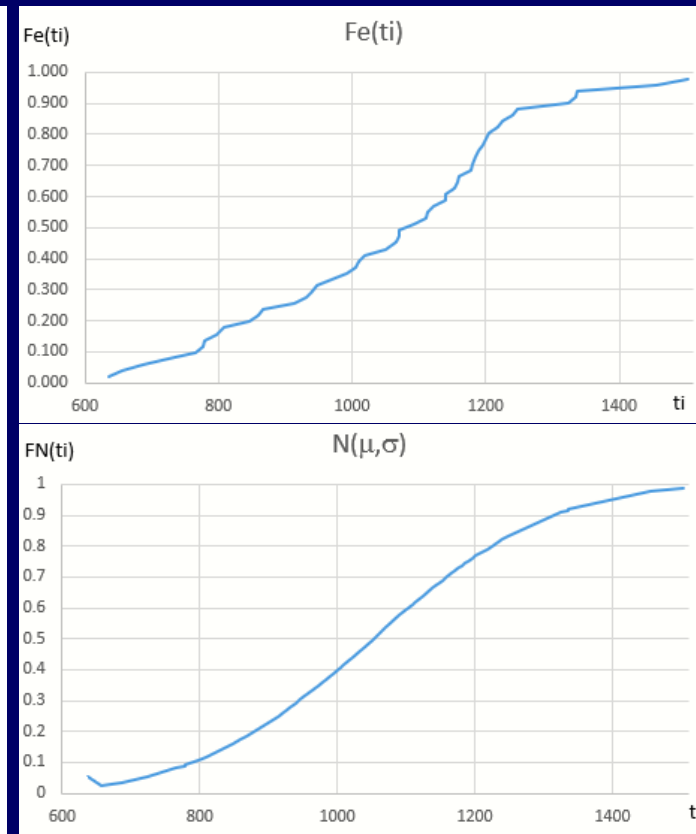
Excel

Mathematica

Graphical Methods

i	t_i	$F_E(t_i)$	$F_N(t_i, \mu, \sigma)$	$F_{Exp}(t_i, \lambda)$
1	636	0.02	0.0301	0.4540
2	656	0.039	0.0365	0.4638
3	686	0.059	0.0490	0.4791
4	725	0.078	0.0696	0.4979
5	766	0.098	0.0987	0.5174
6	778	0.118	0.1080	0.5226
7	778	0.137	0.1081	0.5227
8	798	0.157	0.1256	0.5315
9	809	0.176	0.1368	0.5366
10	847	0.196	0.1783	0.5531
11	859	0.216	0.1932	0.5583
12	867	0.235	0.2033	0.5616
13	913	0.255	0.2669	0.5802
14	932	0.275	0.2961	0.5878
15	940	0.294	0.3083	0.5908
16	947	0.314	0.3196	0.5935
17	969	0.333	0.3564	0.6019
18	991	0.353	0.3941	0.6102
19	1005	0.373	0.4190	0.6154
20	1009	0.392	0.4269	0.6170
21	1020	0.412	0.4454	0.6207
22	1050	0.431	0.4993	0.6313
23	1066	0.451	0.5288	0.6370
24	1070	0.471	0.5367	0.6385
25	1071	0.49	0.5372	0.6386
26	1093	0.51	0.5769	0.6461
27	1110	0.529	0.6074	0.6519

i	t_i	$F_E(t_i)$	$F_N(t_i, \mu, \sigma)$	$F_{Exp}(t_i, \lambda)$
28	1113	0.549	0.6128	0.6529
29	1121	0.569	0.6270	0.6556
30	1140	0.588	0.6589	0.6617
31	1140	0.608	0.6591	0.6617
32	1152	0.627	0.6789	0.6656
33	1158	0.647	0.6884	0.6674
34	1160	0.667	0.6913	0.6680
35	1177	0.686	0.7181	0.6734
36	1181	0.706	0.7240	0.6746
37	1186	0.725	0.7310	0.6760
38	1189	0.745	0.7363	0.6771
39	1197	0.765	0.7479	0.6795
40	1201	0.784	0.7544	0.6809
41	1205	0.804	0.7594	0.6819
42	1218	0.824	0.7770	0.6857
43	1225	0.843	0.7863	0.6878
44	1241	0.863	0.8069	0.6925
45	1248	0.882	0.8157	0.6946
46	1324	0.902	0.8939	0.7161
47	1335	0.922	0.9023	0.7189
48	1336	0.941	0.9032	0.7192
49	1456	0.961	0.9674	0.7493
50	1504	0.98	0.9804	0.7605

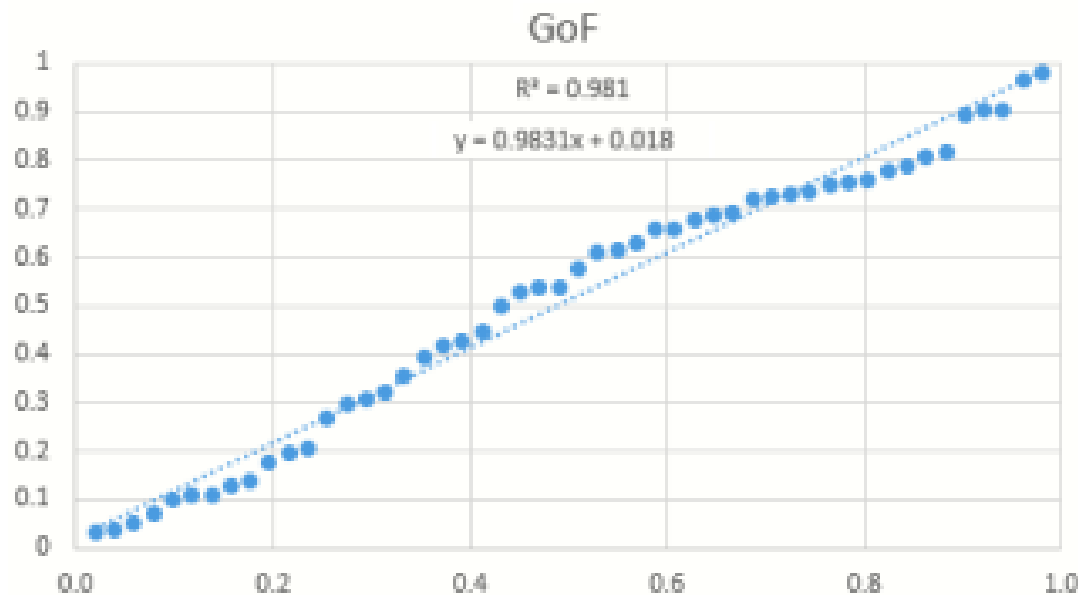


Inferência

Goodness of fitting

Mathematica

Graphical Methods



Probability-Probability Plot - $F_E(t_i) \times F_N(t_i, 1050ms, 220ms)$

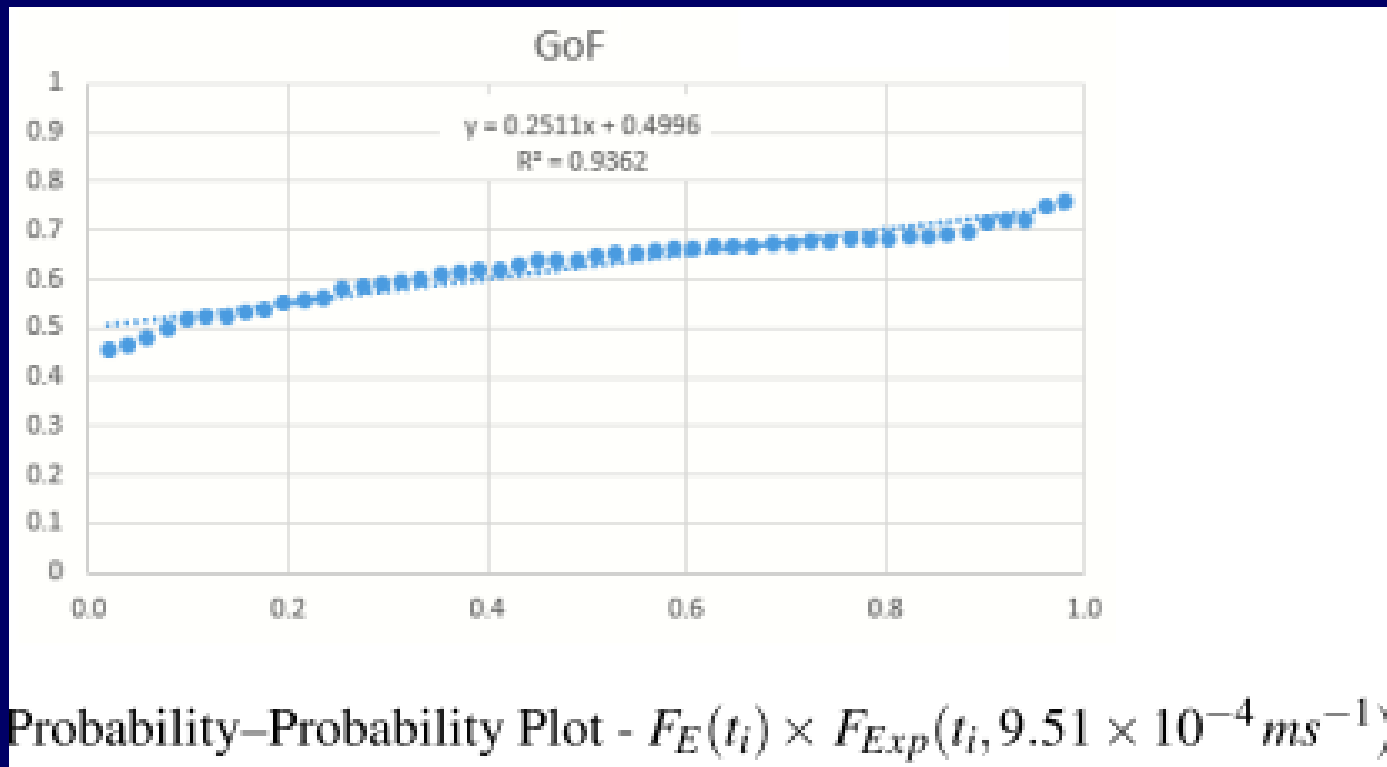
The angular coefficient calculated is $a = 0.9831$ and the linear coefficient is $b = 0.018$. It is worth noting the dots are close to a 45° degree ($a = 0.9831$) straight line with the intercept at $b = 0.018$. Besides, the coefficient of determination, r^2 , is 0.981.

Inferência

Goodness of fitting

Mathematica

Graphical Methods



The angular coefficient calculated is $a = 0.2511$ and the linear coefficient is $b = 0.4996$. Hence, the dots are far from a 45° degree straight line, even having a coefficient of determination, r^2 , for a linear model as 0.9362.

Goodness of fitting - Pearson Chi-Squared Method

The χ^2 method should be adopted for checking if a sample comes from a population with a specific probability distribution function. The method considers that data is binned, hence the test statistics depends on how the data is grouped.

The χ^2 test is based on Q^2 statistics. This statistic is a measure of the distance between the observed values and those that we would expect to obtain if the sample came from a particular probability distribution. In other words, the fitting test compares the frequency of sample and respective data that comes from a theoretical probability distribution. This method is well suited for categorical data, but may also be applied to numerical data. This test presents better results for discrete distributions, since for applying the method for continuous data, the sample should be first grouped into bins, and the test is very sensitive to the choice of bins. Moreover, the test is dependent on:

- number of bins,
- sample size, and
- bin size.

$$Q^2 = \sum_{z=1}^k \frac{(o_z - e_z)^2}{e_z}$$

Goodness of fitting - Pearson Chi-Squared Method

There is no optimal method for defining the bin width and the number of bins, since the optimal bin width depends on the distribution. Even though, most reasonable alternatives should produce similar outcome. This method should not be applied if samples are small or if some bins count are too small (less than five). In such a case, you may combine some bins in the tails.

The χ^2 test aim is:

- compare observed frequencies with expected frequencies, and
- to decide whether the observed frequencies seem to agree or disagree with the expected frequencies assigned to a probability distribution of interest.

Observed and Expected Data

bin	o_i	e_i	$\frac{(o_i - e_i)^2}{e_i}$
(0 – 14)	7	8.56	0.28
(15 – 17)	20	19.72	0.00
(18 – 20)	30	29.06	0.03
(21 – 23)	26	24.71	0.07
(24 – 26)	11	12.77	0.24
(27 – 100)	6	5.18	0.13

Goodness of fitting - Pearson Chi-Squared Method

- 1: Select the random sample of size n , $S = \{x_1, x_2, \dots, x_n\}$. This sample is obtained from a population with distribution function $F(\Theta, t)$ of unknown parameters. $\theta_i \in \Theta$.
- 2: Divide S into set of disjoint intervals or categories, that is, $IS = \{I_i \mid I_i = (lb_i, ub_i), \quad lb_i, ub_i \in \mathbb{R}, \quad lb_i < ub_i, \quad I_i \cap I_j = \emptyset, \quad \forall I_i, I_j \in IS, \quad I_i \neq I_j\}$. And, let o_z be the number of values that belongs to the interval I_z ($z = 1, 2, \dots, k$).
- 3: Find $P(I_z) = F(\Theta, ub_z) - F(\Theta, lb_z)$ and let $e_z = n \times P(I_z)$ ¹¹ be the expected frequency of the interval I_z .
- 4: Calculate the test statistic $Q^2 = \sum_{z=1}^k \frac{(o_z - e_z)^2}{e_z}$.
- 5: The statistic Q^2 is distributed approximately according to χ_{k-1}^2 , where k is the number of classes ($k - 1$ degree of freedom.).
- 6: Considering a confidence degree equal to $1 - \alpha$, reject the null hypothesis (H_0) if $Q^2 \geq \chi_{1-\alpha, k-1}^2$.

Inferência

Goodness of fitting - Pearson Chi-Squared Method

Example 6.6.2. Consider that the software company tested every day one hundred ($n = 100$) appliances. Considering of period of one hundred (100) days, the total number of appliances tested was $100 \times 100 = 10000$. The daily mean number of defects (MND) observed in 100 days was $MND = \sum_{i=1}^{100} d_i / 100 = 19.86$, where d_i is the number of observed defects in day i . Since we tested 100 products every day, the point estimate of the defect probability is $\hat{p} = 19.86/100 = 0.1986$.

Table 6.10 presents the number of days in which we observed the respective number of defects that belongs to a bin¹². We adopted a bin size equal to (see footnote). Therefore (0 – 14), (15 – 17), (18 – 20), (21 – 23), (24 – 26) and (27 – 100) defects per day. Hence, we have six (6) number of bins (classes), that is $i = \{1, 2, 3, 4, 6\}$. It is worth mentioning that the first and the last bins (classes) are wider since the number of defects is much rarer. The number of defects observed in each bin is depicted by o_i .

Our aim is to check if we have evidences to refute the binomial distribution probability with parameters $Bin(n, p)$, where $n = 100$ and $p = 0.1986$, as a suitable distribution for representing the observed data. The expected number of defects for each bin i is therefore $e_i = n \times P_B(bin_i, \hat{p}, n)$ where n is the daily sample size, bin_i is a bin or class, \hat{p} is the point estimate of the defect probability and $P_B(bin_i, \hat{p}, n)$ is the probability of the number of defects of a bin_i .

Observed and Expected Data

bin	o_i	e_i	$\frac{(o_i - e_i)^2}{e_i}$
(0 – 14)	7	8.56	0.28
(15 – 17)	20	19.72	0.00
(18 – 20)	30	29.06	0.03
(21 – 23)	26	24.71	0.07
(24 – 26)	11	12.77	0.24
(27 – 100)	6	5.18	0.13

The test statistic Q^2 is calculated by summing up the values of column $(o_i - e_i)^2 / o_i$, that is, $Q^2 = \sum_{i=1}^k (o_i - e_i)^2 / e_i$, where $k = 6$. In this example, $Q^2 = 0.76$. If we consider a confidence degree of $1 - \alpha = 95\%$, then $\chi^2_{1-\alpha, k-1} = \chi^2_{95\%, 5} = 1.1455$, where $k = 6$ is the number of bins. As $Q^2 < \chi^2_{\alpha, k-1}$, we have no evidence to reject the binomial distribution as suitable for representing the observed data taking into account a 95% of confidence degree. It is worth mentioning, however, that the bin range was carefully chosen. If we change the ranges, we may obtain a result in which we have enough evidence to reject the $Bin(n, p)$, $n = 100$ and $p = 0.1986$, as suitable distribution to represent the data.

Inferência

Goodness of fitting

Pearson

Chi-Squared Method

Goodness of Fitting - Binomial distribution - $B(n = 100, p = 0.2)$

Class: (i, xl_i, xh_i)			o_i	$B((xl_i \leq X \leq xh_i, n, p))$	e_i	$(o_i - e_i)^2 / e_i$	Q^2	$\chi^2_{1-\alpha, k-1}$
1	0	14	7	0.1137	11.37	1.680	13.74	11.07
2	15	17	15	0.2273	22.73	2.628		
3	18	20	18	0.2970	29.70	4.607		
4	21	23	14	0.2238	22.38	3.137		
5	24	100	9	0.1382	13.82	1.683		

A set of servers were observed for one hundred (100) weeks. The observation period was divided into five ($k = 5$) classes. The first class range is fourteen (14) weeks, the three (3) subsequent classes have a range of three (3) weeks. The last class has a range of seventy seven (77) weeks. The number of failure observed in each class is shown in Column o_i of Table 6.11, $i \in \{1, 2, 3, 4, 5\}$. The aim is to check if there is enough evidence to refute the binomial distribution $B(p, n)$, where $p = 0.2$ and $n = 100$ as a suitable distribution to represent the observed server failures. Column $B((xl_i \leq X \leq xh_i, n, p))$ depicts the probability of each class i , that is, $B((X \leq xh_i, n, p) - B((X \leq xl_i, n, p))$, where xl_i and xh_i are respectively the lower and upper bound of each class i . Column e_i is the expected value for the class i , considering $B(p, n)$, that is $B((xl_i \leq X \leq xh_i, n, p)) \times n$. Column $(o_i - e_i)^2 / e_i$ presents the normalized squared difference between the observed data and the expected value of the class i .

Column $Q^2 = \sum_{i=1}^k \frac{(o_i - e_i)^2}{e_i}$ is the calculated statistic from the data and $\chi^2_{1-\alpha, k-1}$ is obtained from χ^2_{k-1} distribution, considering $1 - \alpha = 95\%$ confidence degree and $k = 5$ as the number of classes. As $Q^2 = 13.74 > \chi^2_{1-\alpha, k-1} = 9.49$, there is evidence to reject $B(p, n)$, $p = 0.2$ and $n = 100$, as a suitable distributions to represent the data set. We even have enough evidence to reject $B(p, n)$, $p = 0.2$ and $n = 100$, as fitting to the data since $\chi^2_{1-\alpha, k-1} = 13.28$ for $1 - \alpha = 99\%$. Nevertheless, if we want only reject $B(p, n)$, $p = 0.2$ and $n = 100$, only if we are $1 - \alpha = 99.9\%$ sure of its inadequacy, thus we have not enough evidence to achieve such a high degree of confidence since $\chi^2_{1-\alpha, k-1} = 18.47$, and in such a case $Q^2 < \chi^2_{1-\alpha, k-1}$ ($13.74 < 18.47$).

Inferência

Goodness of fitting - Pearson Chi-Squared Method

GoF - Server Time to Failure in Accelerate Reliability Test - $N(196 \text{ days}, 30 \text{ days})$							
Class: i, xl_i, xh_i	o_i	$N((xl_i \leq X \leq xh_i, \mu, \sigma)$	e_i	$(o_i - e_i)^2 / e_i$	Q^2	$\chi^2_{1-\alpha, nc-1}$	
1 0 151	9	0.0668	6.67	0.8138	6.6770	16.9190	
2 151 163	8	0.0689	6.88	0.1832			
3 163 175	8	0.1063	10.62	0.6462			
4 175 187	12	0.1401	14.00	0.2867			
5 187 199	12	0.1577	15.77	0.9013			
6 199 211	13	0.1516	15.17	0.3094			
7 211 223	8	0.1245	12.46	1.5939			
8 223 235	7	0.0873	8.74	0.3450			
9 235 271	8	0.0906	9.08	0.1279			

A set of one hundred servers (100) were observed during an accelerate reliability test. The observation period was divided into nine ($k = 9$) classes. The first class range is one hundred and fifty 151 days, whereas subsequent seven (7) classes have a range of twelve (12) days. The last class has a range of 36 days. The number of failure observed in each class is shown in Column o_i of Table 6.12, $i \in \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$. The aim is to check if there is enough evidence to refute the normal distribution $N(\mu, \sigma)$, where $\mu = 196$ days and $\sigma = 30$ days as a suitable distribution to represent the observed lifetime data. Column $N((xl_i \leq X \leq xh_i, \mu, \sigma)$ depicts the probability of each class i , that is, $N((X \leq xh_i, \mu, \sigma) - N((X \leq xl_i, \mu, \sigma)$, where xl_i and xh_i are respectively the lower and upper bound of each class i . Column e_i is the expected value for the class i , considering $N(\mu, \sigma)$, that is $N((X \leq xh_i, \mu, \sigma) - N((X \leq xl_i, \mu, \sigma) \times n$. Column $(o_i - e_i)^2 / e_i$ depicts the normalized squared difference between the observed data and the expected value of the class i .

Column $Q^2 = \sum_{i=1}^{nc} \frac{(o_i - e_i)^2}{e_i}$ is the calculated statistic from the data and $\chi^2_{1-\alpha, k-1}$ is obtained from χ^2_{k-1} distribution, considering $1 - \alpha = 95\%$ confidence degree and $k = 9$ as the number of classes. As $Q^2 = 5.2075$ and $\chi^2_{1-\alpha, k-1} = 15.5073$ ($Q^2 < \chi^2_{1-\alpha, k-1}$), there is no evidence to reject $N(\mu, \sigma)$, $\mu = 36$ days and $\sigma = 30$ days, as a suitable distributions to represent the the servers' time to failure under the accelerate reliability test with such a degree of confidence.

Inferência

Example

Goodness of fitting - Chi-Squared Method

Goodness of Fitting - Normal distribution - $N(196 \text{ days}, 30 \text{ days})$

Class: i, xl_i, xh_i			o_i	$N((xl_i \leq X \leq xh_i), \mu, \sigma)$	e_i	$(o_i - e_i)^2 / e_i$	χ^2	$\chi_{1-\alpha, nc-1}$
1	0	151	9	0.0668	6.6807	0.8052	6.6770	16.9190
2	151	163	7	0.0689	6.8859	0.0019		
3	163	175	8	0.1063	10.6298	0.6506		
4	175	187	17	0.1401	14.0125	0.6369		
5	187	199	12	0.1577	15.7739	0.9029		
6	199	211	18	0.1516	15.1635	0.5306		
7	211	223	8	0.1245	12.4477	1.5892		
8	223	235	12	0.0873	8.7260	1.2284		
9	235	271	8	0.0906	9.0591	0.1238		

Testes de Aderência

Excel

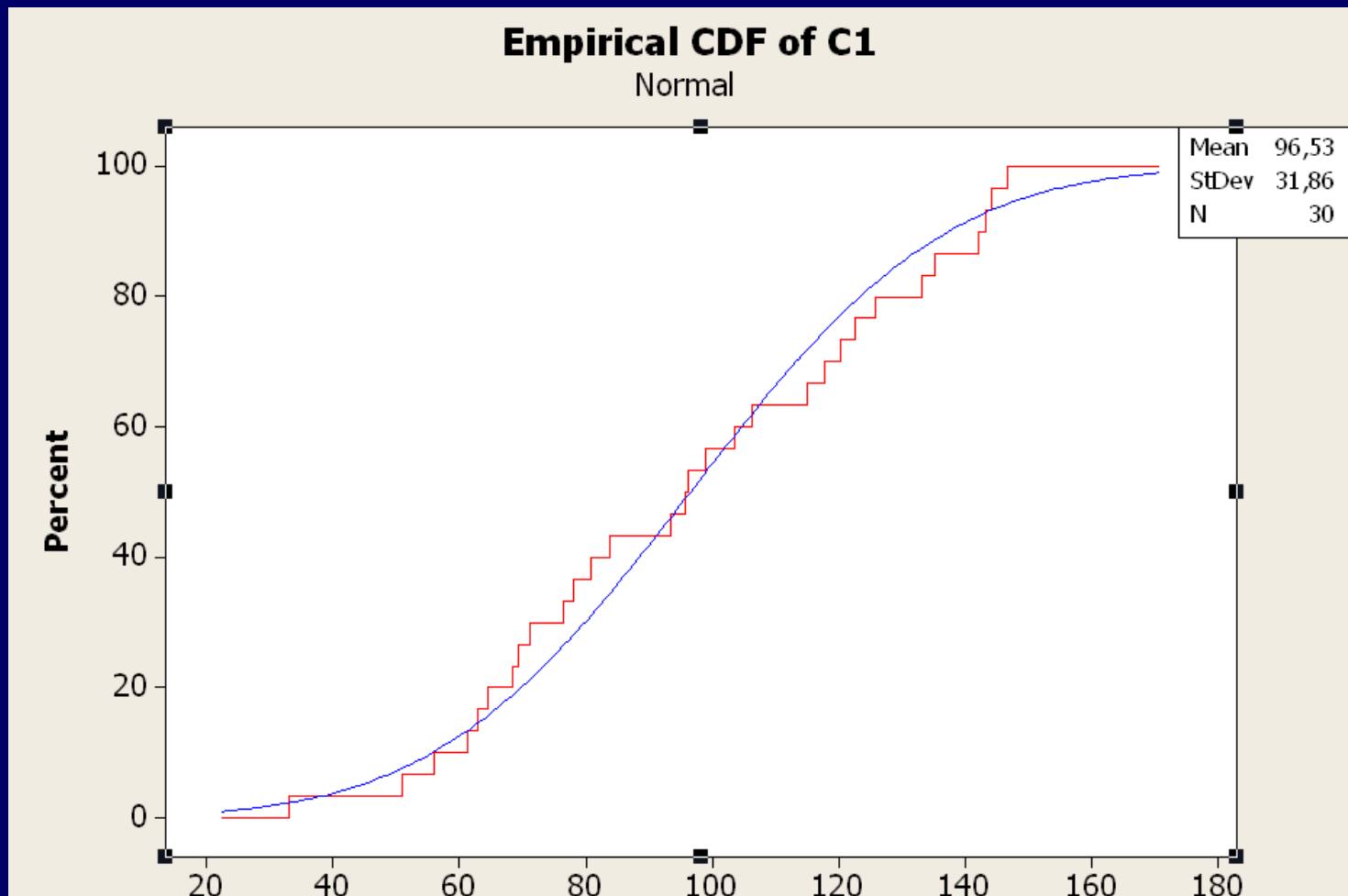
Minitab

□ Teste KS

- Estabelecer duas hipóteses:
 - H_0 : As distribuições das amostras se aproximam
 - H_a : As distribuições das amostras não se aproximam
- A partir de cada amostra $Y=(y_1, y_2, \dots, y_n)$, calcula-se a função de distribuição acumulada empírica (ecdf).

Testes de Aderência

Teste KS



Testes de Aderência

□ Teste KS

- Estabelecer duas hipóteses:
 - H0: As distribuições das amostras se aproximam
 - Ha: As distribuições das amostras não se aproximam
- A partir de cada amostra $Y=(y_1, y_2, \dots, y_n)$, calcula-se a função de distribuição acumulada empírica (ecdf).
- Calcula-se:

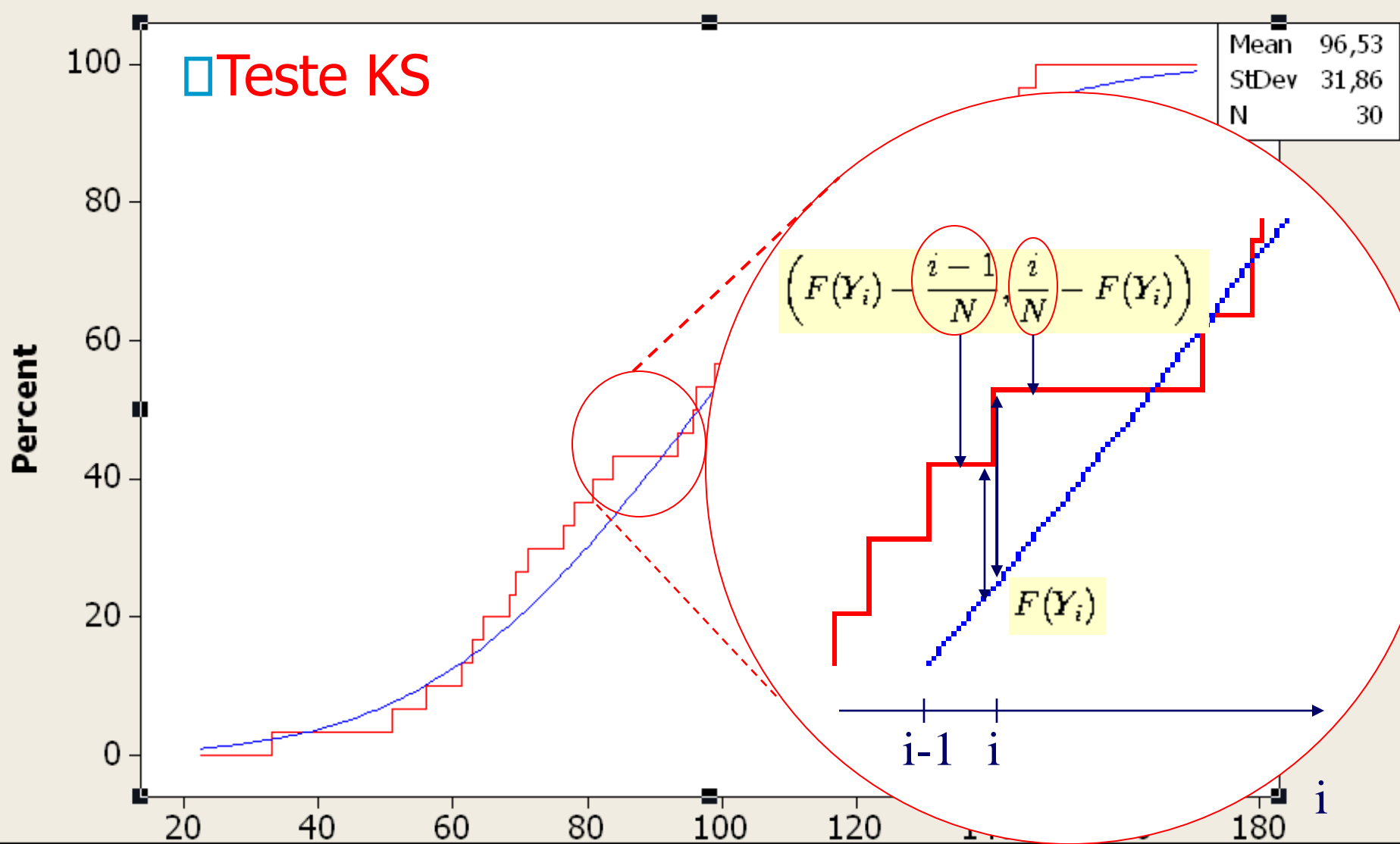
$$D = \max_{1 \leq i \leq N} \left(F(Y_i) - \frac{i-1}{N}, \frac{i}{N} - F(Y_i) \right)$$
- Compara-se $D_{\text{observado}}$ com o um $D_{\text{crítico}}$
 - Caso $D_{\text{observado}}$ observado seja menor que o $D_{\text{crítico}}$ a hipótese nula não será rejeitada.

O $D_{\text{crítico}}$ pode ser calculado por:

$$D_{\text{crítico}, \alpha} = \frac{c(\alpha)}{\sqrt{n}}$$

Testes de Aderência

Teste KS



Testes de Aderência

Teste KS

$1 - \alpha$ n	0.9	0.95	0.99
1	0.950	0.975	0.995
2	0.776	0.842	0.929
3	0.636	0.708	0.829
4	0.565	0.624	0.734
5	0.510	0.563	0.669
6	0.468	0.520	0.617
7	0.436	0.483	0.576
8	0.410	0.454	0.542
9	0.387	0.430	0.513
10	0.369	0.409	0.489
11	0.352	0.391	0.468
12	0.338	0.375	0.450
13	0.325	0.361	0.432
14	0.314	0.349	0.418
15	0.304	0.338	0.404
16	0.295	0.327	0.392
17	0.286	0.318	0.381
18	0.279	0.309	0.371
19	0.271	0.301	0.361
20	0.265	0.294	0.352

$1 - \alpha$ n	0.9	0.95	0.99
21	0.259	0.287	0.344
22	0.253	0.281	0.337
23	0.247	0.275	0.330
24	0.242	0.269	0.323
25	0.238	0.264	0.317
26	0.233	0.259	0.311
27	0.229	0.254	0.305
28	0.225	0.250	0.300
29	0.221	0.246	0.295
30	0.218	0.242	0.290
31	0.214	0.238	0.285
32	0.211	0.234	0.281
33	0.208	0.231	0.277
34	0.205	0.227	0.273
35	0.202	0.224	0.269
> 35	$\frac{1.224}{\sqrt{n}}$	$\frac{1.358}{\sqrt{n}}$	$\frac{1.628}{\sqrt{n}}$
	$D_{crítico,\alpha}$		

Testes de Aderência

□ Teste KS

$$D = \max_{1 \leq i \leq N} \left(F(Y_i) - \frac{i-1}{N}, \frac{i}{N} - F(Y_i) \right)$$

- Compara-se $D_{observado}$ com o um $D_{crítico}$
 - Caso $D_{observado}$ observado seja maior que o $D_{crítico}$ a hipótese nula será rejeitada

$$D_{crítico,\alpha} = \frac{c(\alpha)}{\sqrt{n}}$$

$c(\alpha)$ é provido pela tabela:

α	0.10	0.05	0.025	0.01	0.005	0.001
$c(\alpha)$	1.22	1.36	1.48	1.63	1.73	1.95

Testes de Aderência

□ Teste KS

i	t ~ FE	FE ⁻ (i)	FE ⁺ (i)	Ft(t)	Ft(t)-FE ⁻ (i)	FE ⁺ (i)-Ft(t)	Max {Ft(t)-FE ⁻ (i), FE ⁺ (i)-Ft(t)}
1	142.773	0.000	0.010	0.001823	0.002	0.008	0.008
2	146.217	0.010	0.020	0.01002	0.000	0.010	0.010
3	147.676	0.020	0.030	0.018805	-0.001	0.011	0.011
4	147.740	0.030	0.040	0.019305	-0.011	0.021	0.021
5	149.016	0.040	0.050	0.031954	-0.008	0.018	0.018
6	149.105	0.050	0.060	0.033053	-0.017	0.027	0.027
7	150.476	0.060	0.070	0.054119	-0.006	0.016	0.016
8	151.284	0.070	0.080	0.070823	0.001	0.009	0.009
9	151.461	0.080	0.090	0.074959	-0.005	0.015	0.015

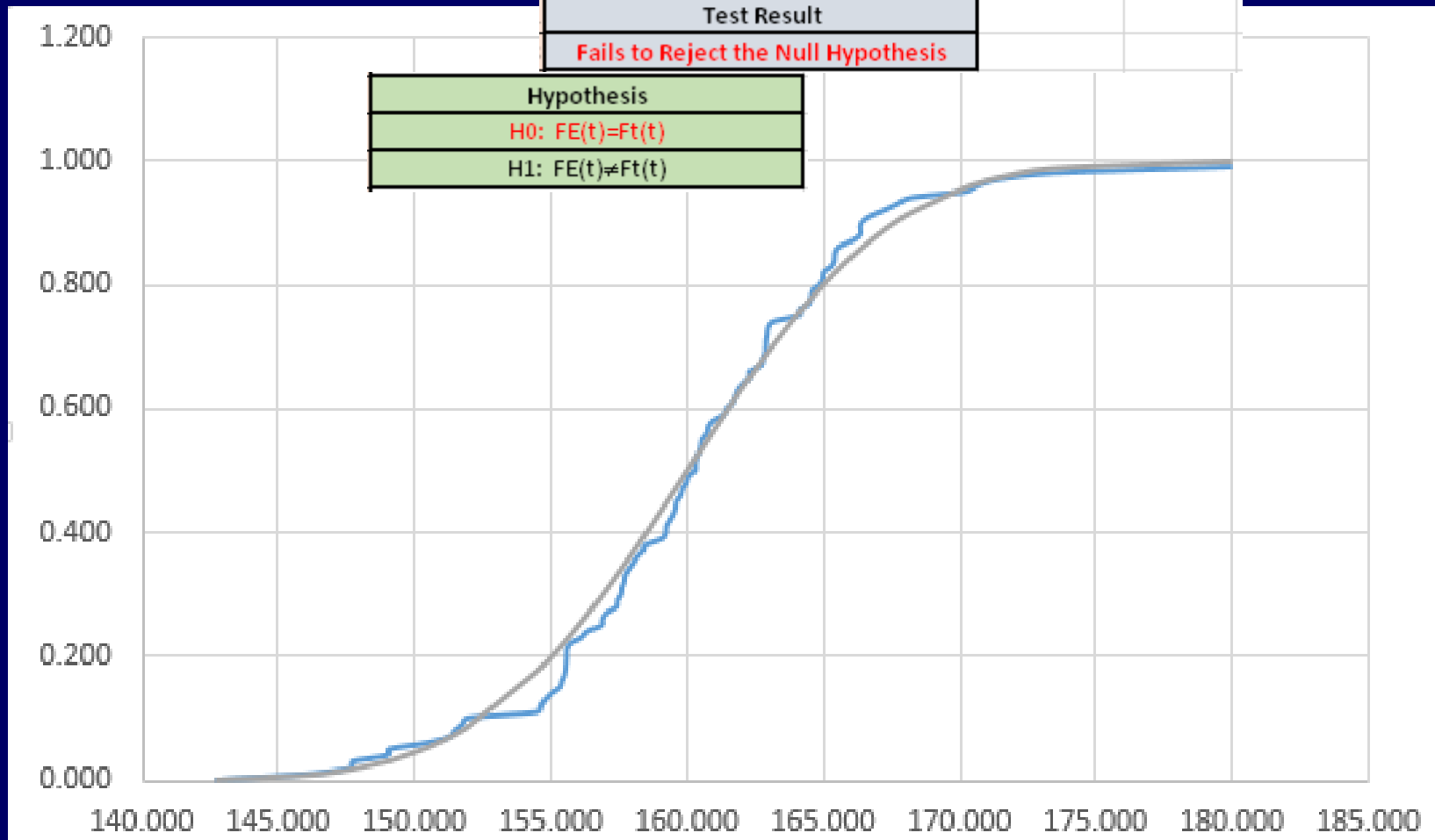
Dcal=MAX{Max {Ft(t)-FE ⁻ (i), FE ⁺ (i)-Ft(t)}}	Statistics		Hypothesis
0.068	Mean (B)=	159.9844	H0: FE(t)=Ft(t)
Dcritical	SD (B)=	5.920048	H1: FE(t)≠Ft(t)
0.1358	N	100	
Test Result			
Fails to Reject the Null Hypothesis			

Testes de Aderência

Teste KS

Dcal=MAX{Max {Ft(t)-FE-(i),FE+(i)-Ft(t))}		Statistics	
0.068		Mean (B)=	159.9844
Dcritical		SD (B)=	5.920048
0.1358		N	100
Test Result			
Fails to Reject the Null Hypothesis			

Hypothesis
H0: $FE(t)=Ft(t)$
H1: $FE(t)\neq Ft(t)$



Data Fitting

□ Linear Regression

A linear model that relates a response variable y (dependent variable) with a single regressor x (independent variable or predictor).

Before entering into details, it is worth mentioning, that the observations on the response variable x are assumed to be random observations from the population, and the regressor variables (y) are assumed to be constants for a given population of interest.

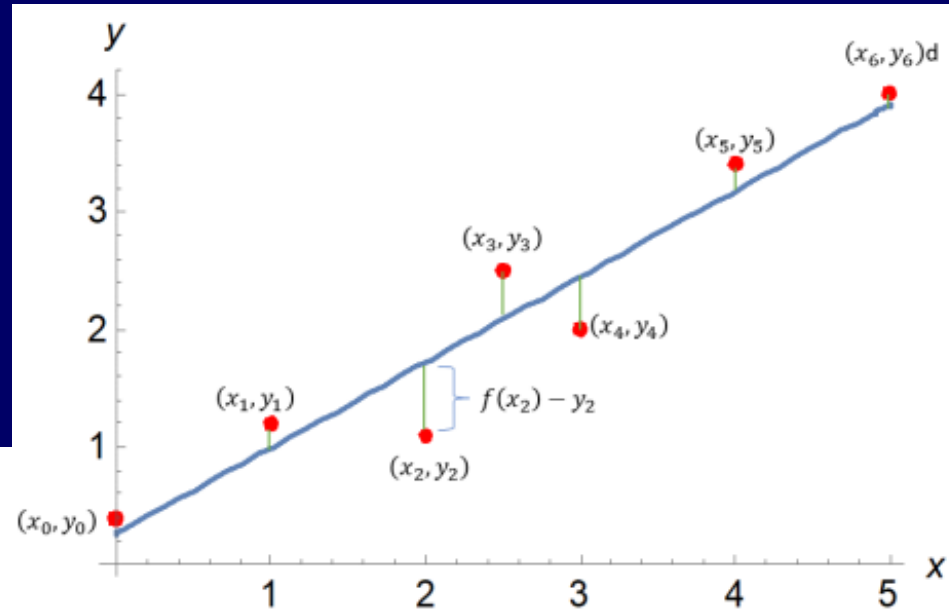
Assume a straight-line function represented by

$$f(x) = Ax + B$$

where A and B are the unknown angular coefficient and the intercept.

Data Fitting

Linear Regression



Now, consider a data set of size n represented by pairs of points, such as $(x_1, y_1), (x_1, y_1), \dots, (x_k, y_k), \dots, (x_n, y_n)$ that has a linear trend, where $X = \{x_1, x_2, \dots, x_n\}$ and $Y = \{y_1, y_2, \dots, y_n\}$ are the set of ordered values of x_k and y_k , respectively.

The general aim here is figure out what is the best values of A and B that minimize the error, E , between the real points and respective points obtained through the linear function

$$f(x) = Ax + B$$

Data Fitting

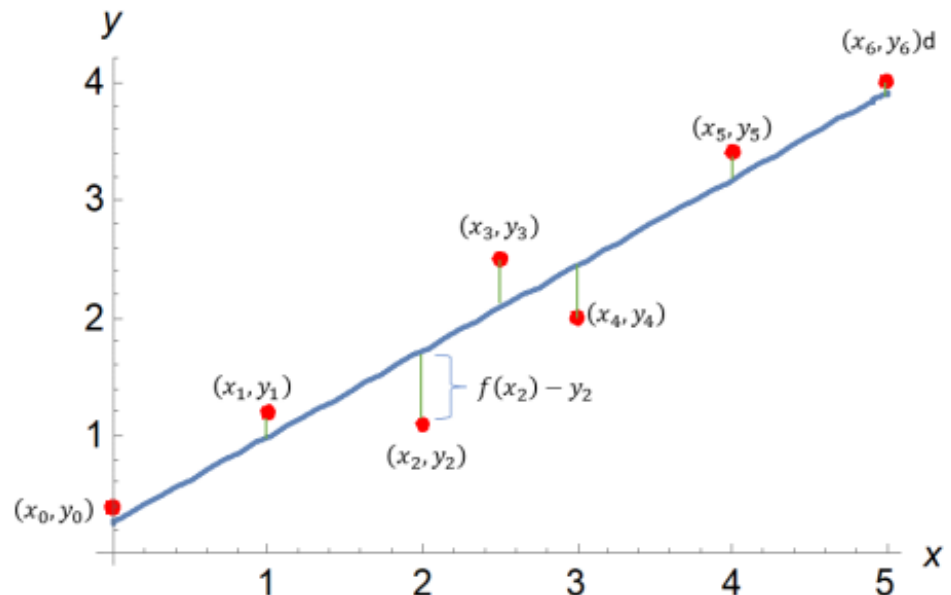
Linear Regression

The error may be quantified by many distinct metrics. One of the most adopted is the Least Square Mean (LSM) error. The LSM is defined by the square root of the average of squares of distances between data points and the respective points provided by the straight line.

More formally,

$$E = \sqrt{\frac{1}{n} \sum_{k=1}^n (f(x_k) - y_k)^2}$$

$$f(x) = Ax + B$$



Data Fitting

□ Linear Regression

Our aim is to find estimates for A and B that minimize the error E . Hence,
 $\min E \Rightarrow \min \varepsilon = \min \sum_{k=1}^n (f(x_k) - y_k)^2$, where

$$\varepsilon = \sum_{k=1}^n ((Ax_k + B) - y_k)^2$$

where ε is usually named *residual* or *sum of square errors*.

Therefore, deriving ε for A and B and equally each expression to zero, allows us to obtain the minimal values of A and B that minimize the error. Thus, deriving ε for A and B , we get

$$\frac{\partial \varepsilon}{\partial A} = 0 \quad \text{and} \quad \frac{\partial \varepsilon}{\partial B} = 0$$

Data Fitting

Linear Regression

equalling them to zero, and then solving both equations for the two unknowns (A and B), we have

$$\frac{\partial}{\partial A} \left(\sum_{k=1}^n ((Ax_k + B) - y_k)^2 \right), \text{ which leads to } \sum_{k=1}^n (Ax_k^2 + Bx_k - y_k x_k) = 0. \text{ Hence}$$

$$A \sum_{k=1}^n x_k^2 + B \sum_{k=1}^n x_k = \sum_{k=1}^n y_k x_k.$$

Likewise,

$$\frac{\partial}{\partial B} \left(\sum_{k=1}^n ((Ax_k + B) - y_k)^2 \right) \text{ take us to } \sum_{k=1}^n 2(Ax_k + B - y_k) = 0$$

Thus, $A \sum_{k=1}^n x_k + nB = \sum_{k=1}^n y_k$, which is equivalent to this linear matrix system

$$\begin{pmatrix} \sum_{k=1}^n x_k^2 & \sum_{k=1}^n x_k \\ \sum_{k=1}^n x_k & n \end{pmatrix} \begin{pmatrix} A \\ B \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^n y_k x_k \\ \sum_{k=1}^n y_k \end{pmatrix}$$

Then, we estimate A and B by

$$\hat{A} = \frac{n \sum_{k=1}^n y_k x_k - \sum_{k=1}^n y_k \sum_{k=1}^n x_k}{n \sum_{k=1}^n x_k^2 - (\sum_{k=1}^n x_k)^2} \text{ and } \hat{B} = \frac{\sum_{k=1}^n y_k \sum_{k=1}^n x_k^2 - \sum_{k=1}^n x_k \sum_{k=1}^n y_k x_k}{n \sum_{k=1}^n x_k^2 - (\sum_{k=1}^n x_k)^2}$$

Data Fitting

□ Linear Regression

$$Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$$

$$Cov(X, Y) = E((X - E(X))(Y - E(Y)))$$

If X and Y are independent, then

$$Cov(X, Y) = 0; \text{ hence}$$

$$Var(X + Y) = Var(X) + Var(Y).$$

The correlation coefficient is dimensionless measure that quantifies the linear relationship between two random variables.

The correlation coefficient between the data sets x_k and y_k , $k \in \{1, 2, \dots, n\}$, is a values in the interval $(-1, 1)$ that indicates a linear relationship between the data sets.

$$\rho_{XY} = \frac{Cov(X, Y)}{\sqrt{Var(X) Var(Y)}}.$$

Likewise, the coefficient of determination also provide information about the linear relationship between the data sets, where values of ρ_{XY}^2 close to one indicate to a strong linear relation between X and Y , uncorrelated X and Y have ρ_{XY}^2 close to zero.

$$\rho_{XY}^2 = \left(\frac{Cov(X, Y)}{\sqrt{Var(X) Var(Y)}} \right)^2$$

Data Fitting

Linear Regression

Example 6.7.1. Assume a data set composed of seven pairs depicted in Columns x_k and y_k of Table 6.16. The respective points are also shown in Figure 6.21. Table 6.17 shows the data summary calculated from data in Table 6.16.

Table 6.16: Data Set (x_k, y_k)

k	x_k	y_k	x_k^2	$x_k y_k$	$f(x_k)$	$(f(x_k) - y_k)^2$
1	0.00	0.40	0.00	0.00	0.2643	0.0184
2	1.00	1.20	1.00	1.20	0.9929	0.0429
3	2.00	1.10	4.00	2.20	1.7214	0.3862
4	2.50	2.50	6.25	6.25	2.0857	0.1716
5	3.00	2.00	9.00	6.00	2.4500	0.2025
6	4.00	3.40	16.00	13.60	3.1786	0.0490
7	5.00	4.00	25.00	20.00	3.9071	0.0086

Data Fitting

Linear Regression



Data Summary

$\sum_{k=1}^n x_k$	$\sum_{k=1}^n y_k$	$\sum_{k=1}^n x_k^2$	$\sum_{k=1}^n x_k y_k$	$\sum_{k=1}^n x_k$ $\sum_{k=1}^n y_k$	n	A	B
17.50	14.60	61.25	49.25	255.50	7	0.7286	0.2643

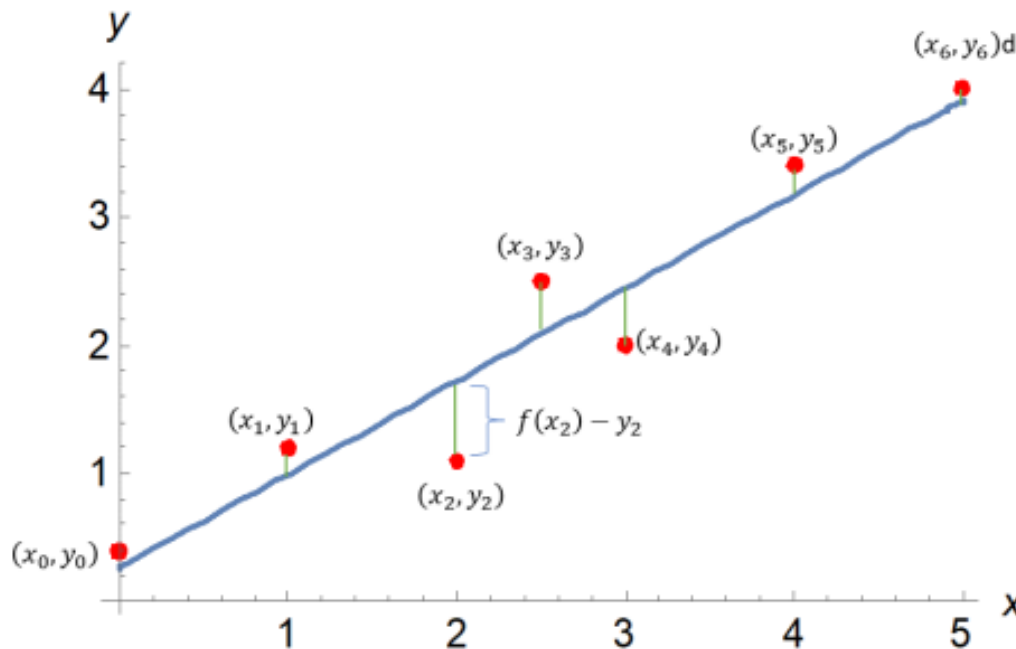
$$A = 0.7286 \text{ and } B = 0.2643$$

$$\hat{y} = \hat{A}x + \hat{B},$$

$$\rho = 0.9558$$

$$\hat{y} = 0.7286x + 0.2643$$

$$\rho^2 = 0.9355$$



Linear Regression Data Fitting

Example 6.7.2. Let us consider a data set composed of twenty pairs, (x_k, y_k) , depicted in Columns x_k and y_k of Table 6.18. The respective points are also shown in Figure 6.22.

Table 6.18: Data Set Composed of Twenty Pairs

k	x_k	y_k	k	x_k	y_k
1	0.3855	78.2658	11	9.4302	35.7762
2	0.4484	73.6320	12	10.3725	35.6596
3	1.3584	71.9883	13	10.3732	34.5777
4	1.4620	70.7748	14	11.7744	30.7041
5	5.2846	63.8734	15	16.7460	30.6425
6	5.3833	58.0750	16	16.9578	26.7791
7	6.1245	57.3510	17	17.7499	22.2907
8	6.8050	50.2785	18	18.1667	22.0439
9	8.7149	44.4071	19	19.1345	10.4765
10	9.4110	39.1079	20	19.1875	4.4540

Table 6.19: Data Summary - Data Set Composed of Twenty Pairs

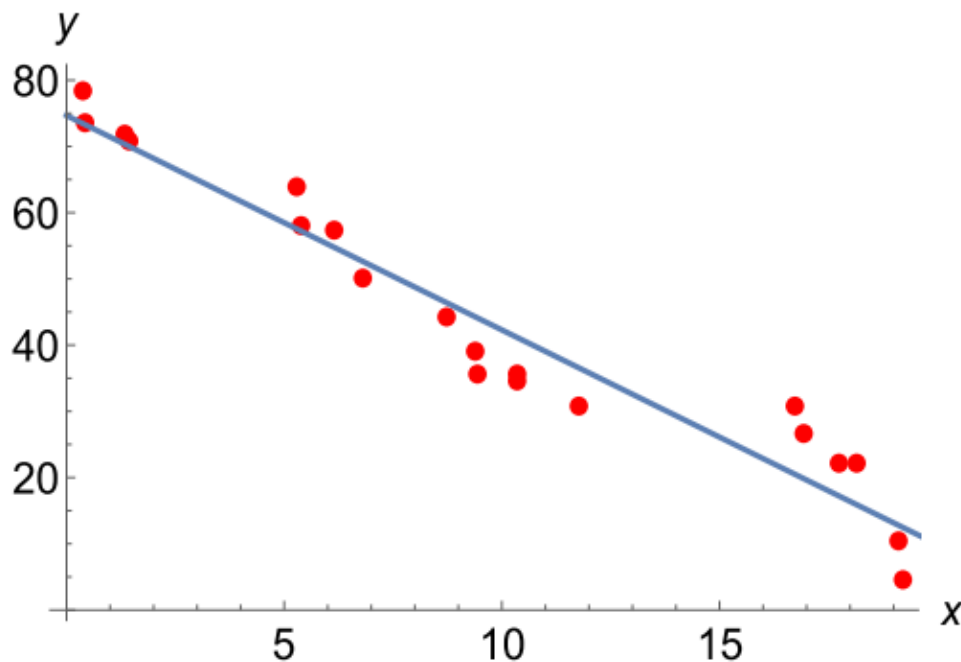
$\sum x_k$	$\sum y_k$	$\sum x_k^2$	$\sum x_k y_k$	$\sum x_k \sum y_k$	n	A	B
195.2701	861.1584	2699.701195	5839.8223	168158.5116	20	-3.2377	74.6695
E	ϵ	ρ	ρ^2				
5.3549	573.4950	-0.9672	0.9355				

Data Fitting

Linear Regression

Table 6.19: Data Summary - Data Set Composed of Twenty Pairs

$\sum x_k$	$\sum y_k$	$\sum x_k^2$	$\sum x_k y_k$	$\sum x_k \sum y_k$	n	A	B
195.2701	861.1584	2699.701195	5839.8223	168158.5116	20	-3.2377	74.6695
E	ϵ	ρ	ρ^2				
5.3549	573.4950	-0.9672	0.9355				



$$\hat{y} = -3.2377x + 74.6695$$

$$\rho = -0.9672$$

$$\rho^2 = 0.9355$$

Figure 6.22: Linear Regression

Multiple Linear Regression

y with more than one regressors, $\mathbf{x} = (x_1, x_2, \dots, x_m)$.

$$f(\mathbf{x}) = f(\mathbf{x}, C_0, C_1, \dots, C_m) = C_1 x_1 + C_2 x_2 + C_3 x_3 + \dots C_m x_m + C_0.$$

minimize the error, \mathcal{E} .

$$\frac{\partial \mathcal{E}(C_0, C_1, \dots, C_m)}{\partial C_0} = 0, \quad \frac{\partial \mathcal{E}(C_0, C_1, \dots, C_m)}{\partial C_2} = 0,$$

$$\frac{\partial \mathcal{E}(C_0, C_1, \dots, C_m)}{\partial C_1} = 0, \quad \begin{array}{c} \vdots \\ \frac{\partial \mathcal{E}(C_0, C_1, \dots, C_m)}{\partial C_m} = 0, \end{array}$$

Multiple Linear Regression

which is the system of $m + 1$ linear equations with $m + 1$ unknowns (C_0, C_1, \dots, C_m) .

Hence, solving

this system of equations, we find C_0, C_1, \dots, C_m that minimize $\varepsilon(C_0, C_1, \dots, C_m)$.

$$\frac{\partial \varepsilon(C_0, C_1, \dots, C_m)}{\partial C_0} = 0, \quad \frac{\partial \varepsilon(C_0, C_1, \dots, C_m)}{\partial C_2} = 0,$$

$$\frac{\partial \varepsilon(C_0, C_1, \dots, C_m)}{\partial C_1} = 0, \quad \vdots \quad \frac{\partial \varepsilon(C_0, C_1, \dots, C_m)}{\partial C_m} = 0,$$

$$\min \varepsilon(C_0, C_1, \dots, C_m) = \min \sum_{k=1}^n (f(\mathbf{x}, C_0, C_1, \dots, C_m) - y_k)^2$$

Data Fitting

Mathematica

Excel

Multiple Linear Regression

Example 6.7.3. Assume a data set composed of fifteen tuples depicted in Columns x_k , y_k and z_k of Table 6.20. The respective points are also shown in Figure 6.23. The format of linear model is

$$f(x, y) = Ax + By + C.$$

Our aim is finding estimates of A , B , and C that minimize ε . Thus,

$$\min \varepsilon(A, B, C) = \min \sum_{k=1}^{15} (f(x_k, y_k) - z_k)^2.$$

Table 6.20: Data Set $\{(x_k, y_k, z_k)\} \ k \in \mathbb{N}, k \in [0, 15]$

Multiple Linear Regression

k	x_k	y_k	z_k	k	x_k	y_k	z_k
1	1.00	4.00	8.00	9	9.00	9.80	23.00
2	2.00	1.80	9.00	10	10.00	9.20	19.00
3	3.00	2.70	10.00	11	11.00	11.80	25.00
4	4.00	4.30	8.00	12	12.00	12.90	22.00
5	5.00	5.20	13.00	13	13.00	13.30	26.00
6	6.00	5.70	16.00	14	14.00	14.80	24.00
7	7.00	7.60	14.00	15	15.00	13.00	27.00
8	8.00	7.80	18.00				

$$\frac{\partial \varepsilon(A, B, C)}{\partial A} = 0,$$

$$\frac{\partial \varepsilon(A, B, C)}{\partial B} = 0,$$

$$\frac{\partial \varepsilon(A, B, C)}{\partial C} = 0,$$

Data Fitting

Mathematica

Excel

Multiple Linear Regression

Example 6.7.3. Assume a data set composed of fifteen tuples depicted in Columns x_k , y_k and z_k of Table 6.20. The respective points are also shown in Figure 6.23. The format of linear model is

$$f(x,y) = Ax + By + C.$$

Our aim is finding estimates of A , B , and C that minimize ε . Thus,

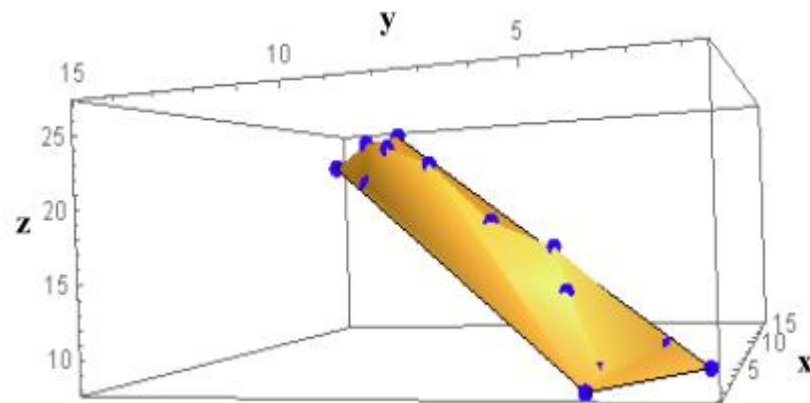
$$\min \varepsilon(A,B,C) = \min \sum_{k=1}^{15} (f(x_k, y_k) - z_k)^2.$$

$$f(x,y) = Ax + By + C.$$

$$\hat{A} = -1.0110, \hat{B} = 2.3714, \text{ and } \hat{C} = 17.4395$$

$$f(x,y) = -1.0110x + 2.3714y + 17.4395$$

$$E = 2.9132 \text{ and } \varepsilon = 127.301.$$



Data Fitting

Mathematica

Excel

Polynomial Regression

The strategy adopted for linear regression may also be broadly adopted as a general setting to estimate parameters of more complex curve fitting. The whole idea behind curve fitting is taking

your data, (x_k, y_k) , $k \in \{1, 2, \dots, n\}$, and to assume you have a function, $f(x) = f(x, C_0, C_1, \dots, C_m)$.

our aim is to minimize $\varepsilon(C_0, C_1, \dots, C_m)$

$$\varepsilon(C_0, C_1, \dots, C_m) = \sum_{k=1}^n (f(x_k, C_0, C_1, \dots, C_m) - y_k)^2$$

Hence,

$$\frac{\partial \varepsilon(C_0, C_1, \dots, C_m)}{\partial C_j} = 0, \quad j = 0, 1, \dots, m,$$

Data Fitting

Mathematica

Excel

Polynomial Regression

Example 6.7.4. Consider the data set composed of fifteen pairs depicted in Columns x_k and y_k of Table 6.21. The respective points are also shown in Figure 6.24. Let us assume the following function is a good model for representing such data set

$$f(x) = Ax^2 + Bx + C.$$

Table 6.21: Data Set $(x_k, y_k), k = \{1, 2, \dots, 15\}$

k	x_k	$f(x)$	y_k	k	x_k	$f(x)$	y_k
0	0	8	12	8	8	176	163
1	1	15	10	9	9	215	190
2	2	26	23	10	10	258	305
3	3	41	35	11	11	305	250
4	4	60	80	12	12	356	350
5	5	83	60	13	13	411	420
6	6	110	102	14	14	470	540
7	7	141	95	15	15	533	550

Data Fitting

Mathematica

Excel

Polynomial Regression

Our aim is finding estimates of A , B and C that minimize $\varepsilon(A, B, C)$. Thus,

$$\min \varepsilon(A, B, C) = \min \sum_{k=1}^{15} (f(x_k, y_k) - z_k)^2.$$

Hence, deriving $\varepsilon(A, B, C)$ for A , B and C , and equally them to zero,

$$\frac{\partial \varepsilon(A, B, C)}{\partial A} = 0, \quad \frac{\partial \varepsilon(A, B, C)}{\partial B} = 0, \quad \frac{\partial \varepsilon(A, B, C)}{\partial C} = 0,$$

we get

$$\frac{\partial \varepsilon(A, B, C)}{\partial A} = 16(-55982 + 22289A + 1800B + 155C) = 0,$$

$$\frac{\partial \varepsilon(A, B, C)}{\partial B} = 16(-4529 + 1800A + 155B + 15C) = 0,$$

$$\frac{\partial \varepsilon(A, B, C)}{\partial C} = -6350 + 2480A + 240B + 32C = 0,$$

Data Fitting

Polynomial Regression

which is equivalent to

$$\begin{pmatrix} 22289 & 1800 & 155 \\ 1800 & 155 & 15 \\ 2480 & 240 & 32 \end{pmatrix} \begin{pmatrix} A \\ B \\ C \end{pmatrix} = \begin{pmatrix} 55982 \\ 4529 \\ 6350 \end{pmatrix}$$

Solving this system of linear equation for A , B , and C , we obtain

$\hat{A} = 2.71376$, $\hat{B} = -4.17847$, and $\hat{C} = 19.4596$.

Therefore, $f(x) = 2.71376x^2 + -4.17847x + 19.4596$.

and $E = 27.3372$ and $\varepsilon = 11957.2$.

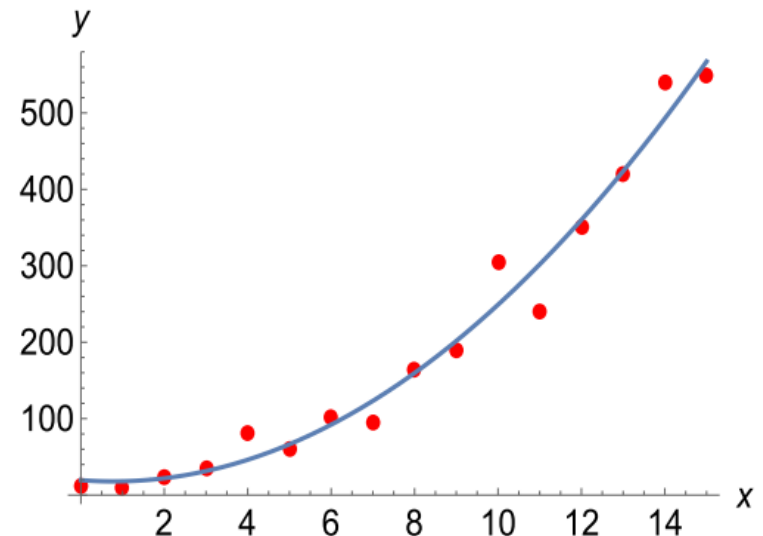


Figure 6.24: Data Set (x, y) - Polynomial Regression

Data Fitting

Mathematica

Excel

Exponential Regression

Now let us apply the general data fitting approach to find parameters of a model that fit a data set that may be represented by an exponential behavior. The function is generically represented by

$$f(x) = y = C e^{Ax}.$$

The function that specifies \mathcal{E} is defined as

$$\mathcal{E}(A, C) = \sum_{k=1}^n (C e^{Ax_k} - y_k)^2$$

Data Fitting

Mathematica

Excel

Exponential Regression

The problem with the model above is that when $\varepsilon(A, C)$ is derived for A and C , we no longer obtain a system of linear equations as in previous cases. For instance, let us take the derivative of $\varepsilon(A, C)$ for A ; thus

$$\frac{\partial \varepsilon(A, C)}{\partial A} = \sum_{k=0}^n 2C e^{Ax_k} x_k (C e^{Ax_k} - y_k),$$

and then equaling the derivative to zero, we get

$$\sum_{k=0}^n 2C e^{Ax_k} x_k (C e^{Ax_k} - y_k) = 0.$$

Likewise, if we derive $\varepsilon(A, C)$ for C and then equaling the expression to zero, we obtain

$$\sum_{k=0}^n 2e^{Ax_k} (C e^{Ax_k} - y_k) = 0.$$

This two equations is a non-linear system of equation, which may be too complex to find a global minimum.

Data Fitting

Mathematica

Excel

Exponential Regression

However, for this particular case, the exponential models may be transformed into a linear model by applying a logarithm.

Thus, let us define a function $Y = g(y) = \ln(y)$.

$$f(x) = y = Ce^{Ax}.$$

then $\ln(y) = \ln(Ce^{Ax}).$

$$\ln(y) = \ln(C) + \ln(e^{Ax}).$$

$$\ln(y) = \ln(C) + Ax \ln(e).$$

$$\ln(y) = \ln(C) + Ax.$$

Therefore, the original data set is transformed into

$$(x_k, \ln(y_k)), \quad k \in \{1, 2, \dots, n\}.$$

Now, we may use the linear regression

to estimate A and B , from C since $B = \ln C$; thus

$$C = e^B.$$

Now, let us define $B = \ln(C)$; hence

$$\ln(y) = Ax + B,$$

which is $Y = Ax + B.$

Data Fitting

Mathematica

Excel

Exponential Regression

Example 6.7.5. Let us consider a data set composed of fifteen pairs, (x_k, y_k) , depicted in Columns x_k and y_k of Table 6.22. The respective points are also shown in Figure 6.25. For this data set, we adopted an exponential regression is considered. Hence, we need to estimate the parameters C and A of the the model (see Function 6.7.32)

$$y = Ce^{Ax}.$$

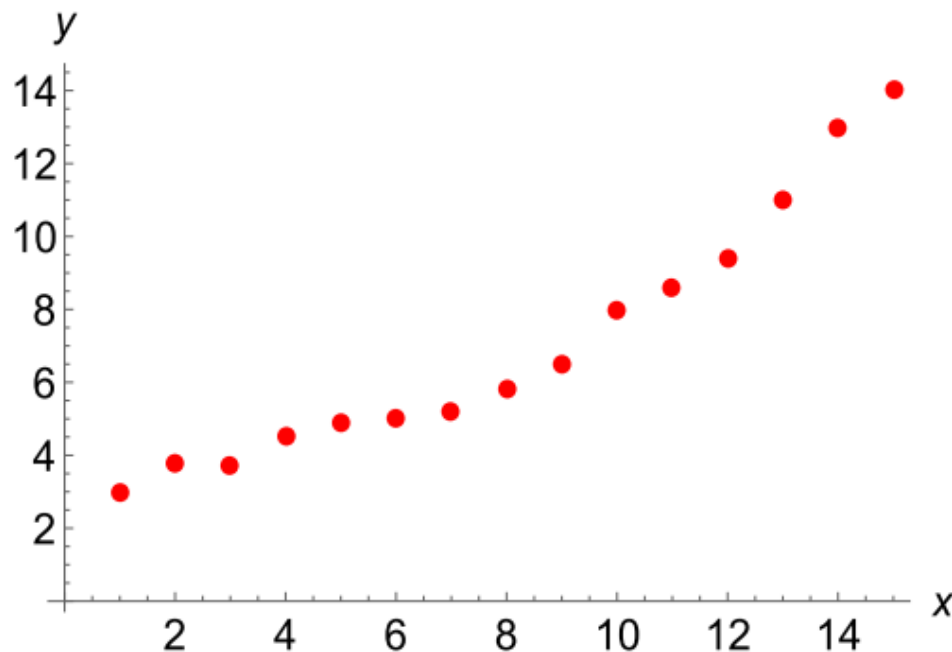


Figure 6.25: Data Set Composed of Fifteen Pairs

Data Set Composed of Fifteen Pairs
Exponential Regression

k	x_k	y_k	$\ln(y_k)$	x_k^2	$x_k y_k$
1	1	3	1.0986	1	1.0986
2	2	3.8	1.3350	4	2.6700
3	3	3.7	1.3083	9	3.9250
4	4	4.5	1.5041	16	6.0163
5	5	4.9	1.5892	25	7.9462
6	6	5	1.6094	36	9.6566
7	7	5.2	1.6487	49	11.5406
8	8	5.8	1.7579	64	14.0629
9	9	6.5	1.8718	81	16.8462
10	10	8	2.0794	100	20.7942
11	11	8.6	2.1518	121	23.6694
12	12	9.4	2.2407	144	26.8885
13	13	11	2.3979	169	31.1726
14	14	13	2.5649	196	35.9093
15	15	14	2.6391	225	39.5859

Data Fitting

Exponential Regression

Data Summary and \hat{A} , \hat{B} , and \hat{C} .

$\sum x_k$	$\sum y_k$	$\sum x_k^2$	$\sum x_k y_k$	$\sum x_k \sum y_k$	n	A	B	C
120.00	27.80	1240.00	251.78	3335.62	15	0.1050	1.0129	2.7536

Hence, $\hat{C} = e^{1.0129} = 2.7536$. Therefore, we get

$$Y = 2.7536e^{0.1050x},$$

where $E = 0.0631$ and $\varepsilon = 0.0598$.

This curve is presented in Figure 6.27.

$$\rho = 0.9905$$

$$\rho^2 = 0.9810$$

Data Set Composed of Fifteen Pairs
Exponential Regression

k	x_k	y_k	$\ln(y_k)$	x_k^2	$x_k y_k$
1	1	3	1.0986	1	1.0986
2	2	3.8	1.3350	4	2.6700
3	3	3.7	1.3083	9	3.9250
4	4	4.5	1.5041	16	6.0163
5	5	4.9	1.5892	25	7.9462
6	6	5	1.6094	36	9.6566
7	7	5.2	1.6487	49	11.5406
8	8	5.8	1.7579	64	14.0629
9	9	6.5	1.8718	81	16.8462
10	10	8	2.0794	100	20.7942
11	11	8.6	2.1518	121	23.6694
12	12	9.4	2.2407	144	26.8885
13	13	11	2.3979	169	31.1726
14	14	13	2.5649	196	35.9093
15	15	14	2.6391	225	39.5859

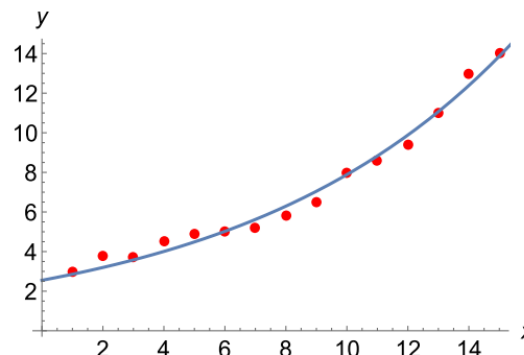
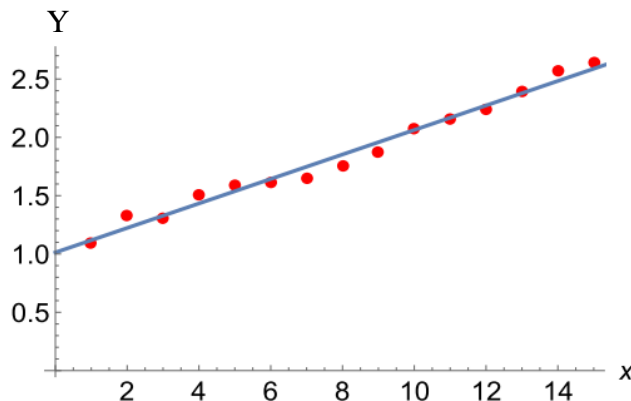


Figure 6.27: Exponential Curve - $y = 2.7536e^{0.1050x}$

Data Fitting

Interpolating Polynomial

The approach discussed here is a polynomial fitting approach that is not the best fit, in the sense that it reduces the error; instead, the polynomial goes through all data points of the data set. Hence, the error obtaining when using such an approach is zero.

A polynomial of order n has $n - 1$ turning points. For instance, a polynomial of degree equal to six has five turning points.

Therefore, if we have a data set with one hundred points, the polynomial may have ninety-nine turns to fit the data. The phenomenon is usually called polynomial *wiggle*.

Data Fitting

Interpolating Polynomial

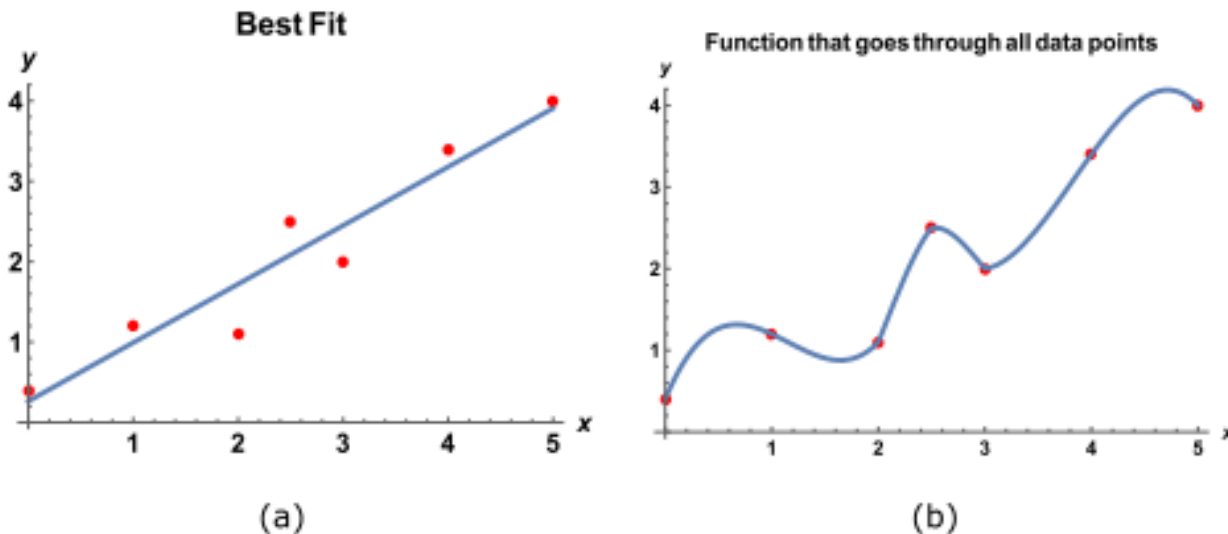


Figure 6.28: Error Reduction (a) Function that goes through all data points (b)

Hence, such an approach is very sensitive to the presence of outliers and data fluctuation. Besides, this approach may provide reasonable estimates for data interpolation, that is for estimating missing points within the collected data set, but we may see huge fluctuation in the data set edges

Data Fitting

Interpolating Polynomial

Consider a data set $\{(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n)\}$ and assume a polynomial $P_n(x) = a_0 + a_1x + a_2x^2 + \dots, a_nx^n$. Then, we have a polynomial system with $n+1$ unknowns (a_0, a_1, \dots, a_n) and $n+1$ constraints $(\{(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n)\})$. Thus, considering all data point $(x_i, y_i), i \in \{0, 1, \dots, n\}$, we get

$$P_n(x_0) = y_0 = a_0 + a_1x_0 + a_2x_0^2 + \dots, +a_nx_0^n,$$

$$P_n(x_1) = y_1 = a_0 + a_1x_1 + a_2x_1^2 + \dots, +a_nx_1^n,$$

$$\vdots$$

$$P_n(x_n) = y_n = a_0 + a_1x_n + a_2x_n^2 + \dots, +a_nx_n^n,$$

Data Fitting

Interpolating Polynomial

which is a system of $n + 1$ linear equations with $n + 1$ unknowns. This system, represented in matrix form, is

$$\begin{pmatrix} 1 & x_0 & x_0^2 & \dots & x_0^n \\ 1 & x_1 & x_1^2 & \dots & x_1^n \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^n \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{pmatrix},$$

which is equal to

$$XA = Y$$

As we have the data set $\{(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n)\}$, we obtain a_0, a_1, \dots, a_n by solving the system of linear equations.

One drawback of such an approach is that if we add one more data point, we have to reconstruct the whole system of linear equations.

Data Fitting

Interpolating Polynomial

An alternative to the process presented above adopts the *Lagrange's Polynomial*. Lagrange's polynomial is defined as

$$\begin{aligned}
 P_n(x) = & \frac{(x - x_1)(x - x_2) \dots (x - x_{n-1})}{(x_0 - x_1)(x_0 - x_2) \dots (x_0 - x_{n-1})} y_0 + \\
 & \frac{(x - x_0)(x - x_2) \dots (x - x_{n-1})}{(x_1 - x_0)(x_1 - x_2) \dots (x_1 - x_{n-1})} y_1 + \\
 & \vdots \\
 & + \frac{(x - x_0)(x - x_2) \dots (x - x_{n-2})}{(x_{n-1} - x_0)(x_{n-1} - x_2) \dots (x_{n-1} - x_{n-2})} \cdot y_{n-1},
 \end{aligned}$$

which may be encapsulated by

$$P(x) = \sum_{j=0}^{n-1} \left(y_j \prod_{k=0, k \neq j}^{n-1} \frac{x - x_k}{x_j - x_k} \right).$$

Data Fitting

Interpolating Polynomial

Example 6.7.6. Assume a data set composed of seven pairs depicted in Columns x_k and y_k of Table 6.16. The respective points are also shown in Figure 6.28. Using this data set on Function 6.7.44, we obtain the polynomial

$$\begin{aligned}
 P(x) = & 0.0013(x-5)(x-4)(x-3)(x-2.5)(x-2)(x-1) + \\
 & 0.1833(x-5)(x-4)(x-3)(x-2.5)x(x-1) - \\
 & 0.7111(x-5)(x-4)(x-3)(x-2)x(x-1) + \\
 & 0.3333(x-5)(x-4)(x-2.5)(x-2)x(x-1) - \\
 & 0.0944(x-5)(x-3)(x-2.5)(x-2)x(x-1) - \\
 & 0.0333(x-5)(x-4)(x-3)(x-2.5)(x-2)x,
 \end{aligned}$$

k	x_k	y_k
1	0.00	0.40
2	1.00	1.20
3	2.00	1.10
4	2.50	2.50
5	3.00	2.00
6	4.00	3.40
7	5.00	4.00

which may be simplified to

$$P(x) = 0.4 + 39.71x - 87.4369x^2 + 70.5833x^3 - 26.3756x^4 + 4.62667x^5 - 0.307556x^6.$$

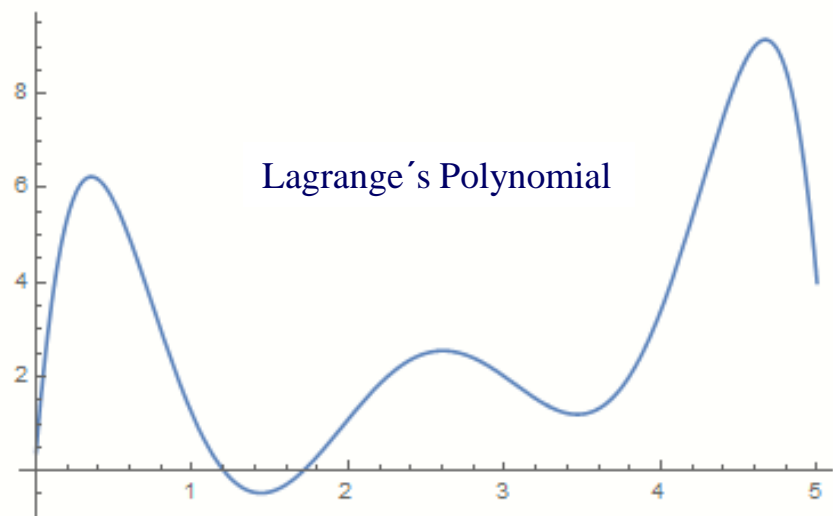
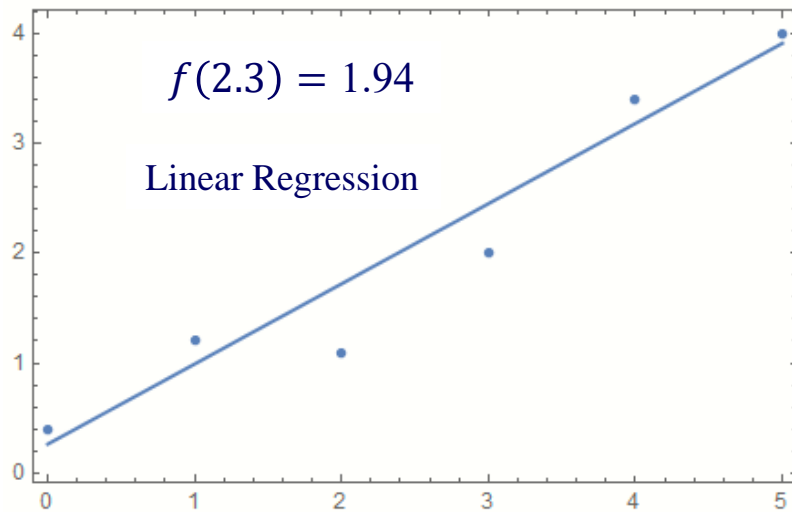
The interpolation at $x = 2.3$ is $P(2.3) = 2.1448$.

Data Fitting

Lagrange's Polynomial

$$P(x) = 0.4 + 39.71x - 87.4369x^2 + 70.5833x^3 - 26.3756x^4 + 4.62667x^5 - 0.307556x^6.$$

The interpolation at $x = 2.3$ is $P(2.3) = 2.1448$.



Data Fitting

Lagrange's Polynomial

Table 6.24: Data Set - (x_k, y_k) , $k = \{1, 2, \dots, 15\}$

k	x_k	y_k	k	x_k	y_k
1	1.0	7.50	9	800.0	16.35
2	100.0	12.50	10	900.0	16.41
3	200.0	14.50	11	1000.0	16.31
4	300.0	15.00	12	1100.0	16.21
5	400.0	15.50	13	1200.0	16.32
6	500.0	15.70	14	1300.0	16.48
7	600.0	16.20	15	1400.0	16.46
8	700.0	16.32	16		

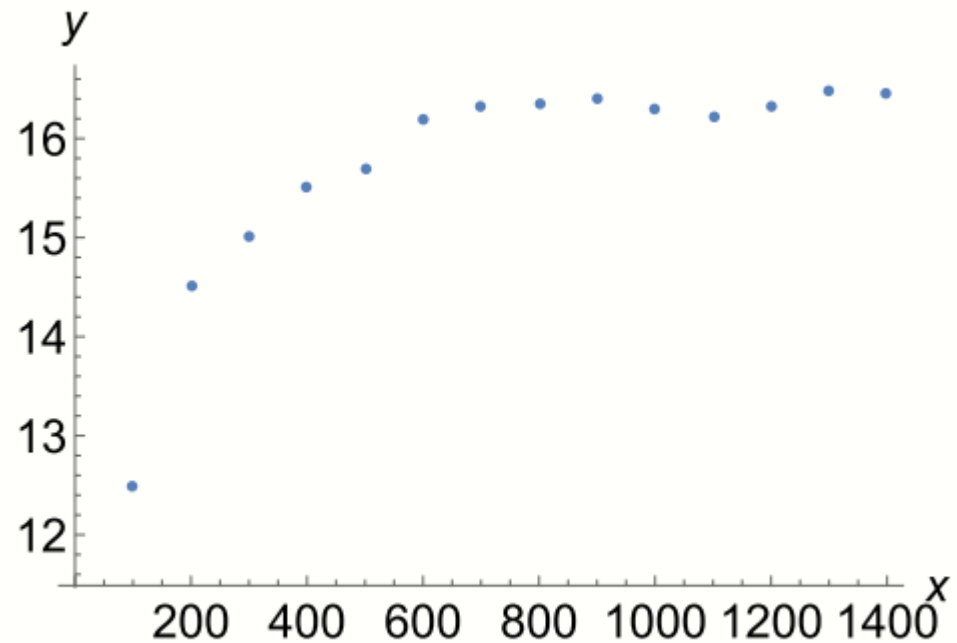


Figure 6.29: Data Plot - (x_k, y_k) , $k = \{1, 2, \dots, 15\}$

Data Fitting

Lagrange's Polynomial

$$\begin{aligned}
 P(x) = & 8.2884 - 0.8142x + 0.0260x^2 - 3.3 \times 10^{-4}x^3 \\
 & + 2.3 \times 10^{-6}x^4 - 1.0 \times 10^{-8}x^5 + 3.2 \times 10^{-11}x^6 \\
 & - 6.8 \times 10^{-14}x^7 + 1.1 \times 10^{-16}x^8 - 1.2 \times 10^{-19}x^9 \\
 & + 9.3 \times 10^{-23}x^{10} - 5.2 \times 10^{-26}x^{11} + 1.9 \times 10^{-29}x^{12} \\
 & - 4.2 \times 10^{-33}x^{13} + 4.2 \times 10^{-37}x^{14}.
 \end{aligned}$$

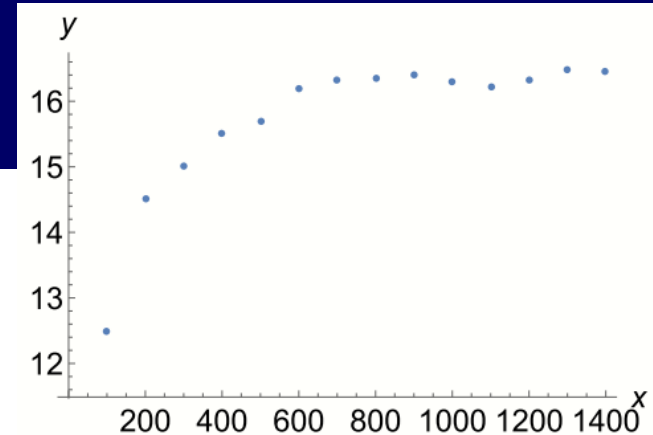


Figure 6.29: Data Plot - $(x_k, y_k), k = \{1, 2, \dots, 15\}$

Assume $x = 650, P(650) = 16.3098$.

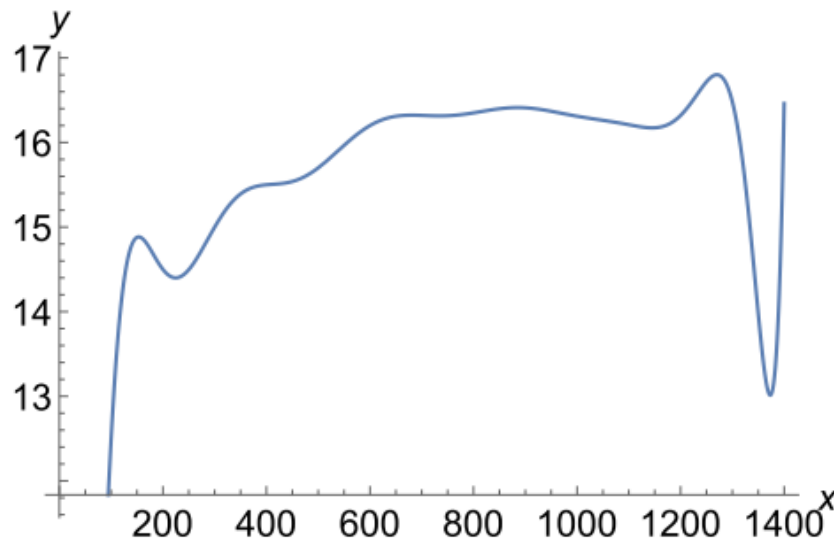


Figure 6.30: Interpolation Graph - $x \in (0, 1400)$

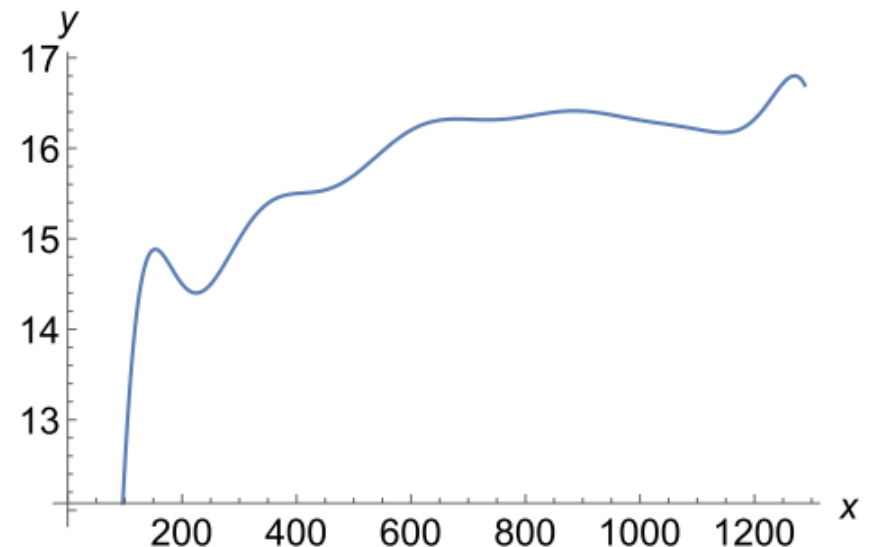


Figure 6.31: Interpolation Graph - $x \in (0, 1288)$