Técnicas Computacionais em Probabilidade e Estatística I

Aula II

Chang Chiann
MAE 5704- IME/USP
1º Sem/2008

Alguns modelos de interesse prático:

a) Bernoulli; Binomial;

b) Poisson; Geométrica; Hipergeométrica;

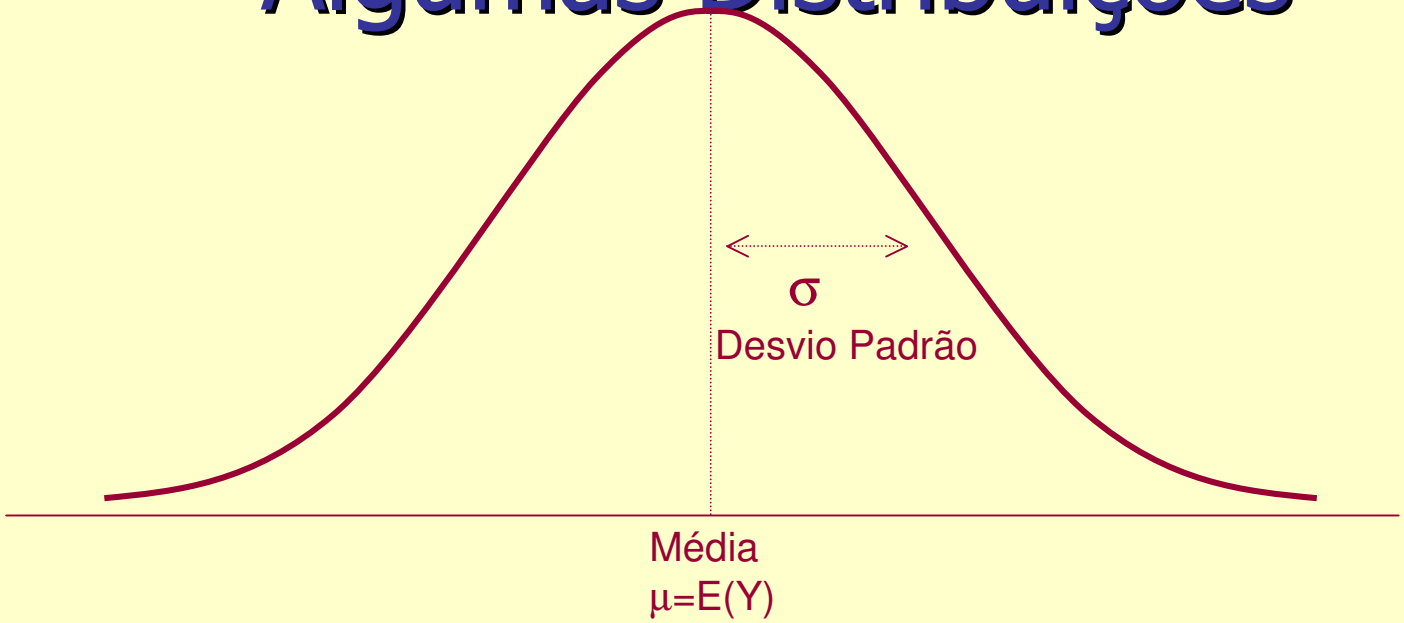
c) Uniforme; Exponencial;

d) Normal; Gama; quiquadrado; t

Caso discreto: a função de probabilidade $p(x)=P(X=x)$

Caso contínuo: f(x) f.d.p.

Algumas Distribuições

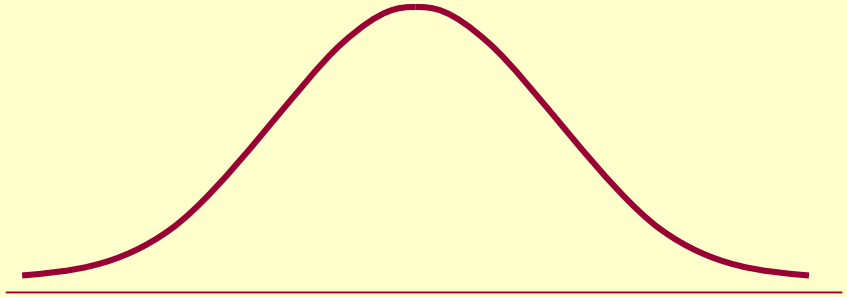


$$f_Y(y \mid \mu, \sigma) = \frac{1}{(2\pi)^{1/2}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{(y-\mu)}{\sigma}\right)^2\right\} \quad I_{[-\infty,+\infty]}(y), \sigma > 0, -\infty < \mu < +\infty$$

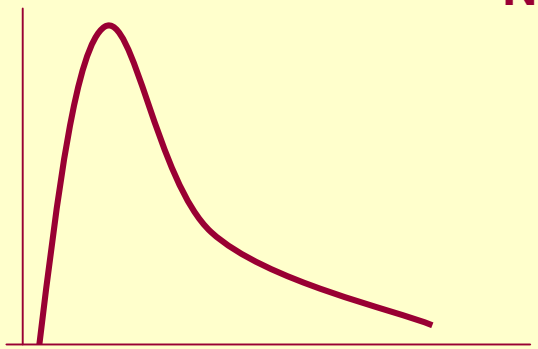
Densidade de uma Variável com Distribuição Normal

- Centralidade $\Rightarrow$ E(Y)
- Dispersão $\Rightarrow$ E(Y-E(Y))^2
- Simetria $\Rightarrow$ E(Y-E(Y))^3
- Curtose (caudas) $\Rightarrow$ E(Y-E(Y))^4

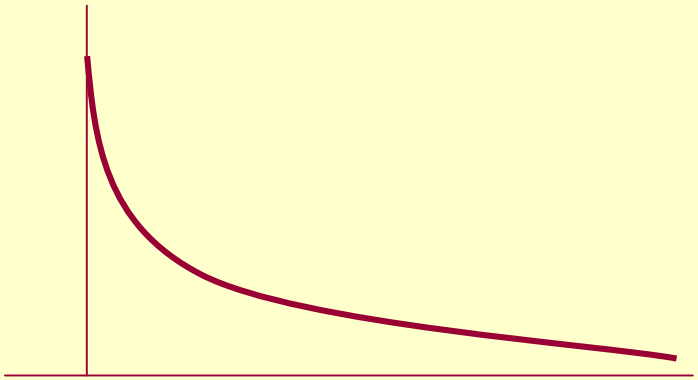
Forma das Distribuições



Distribuição Normal



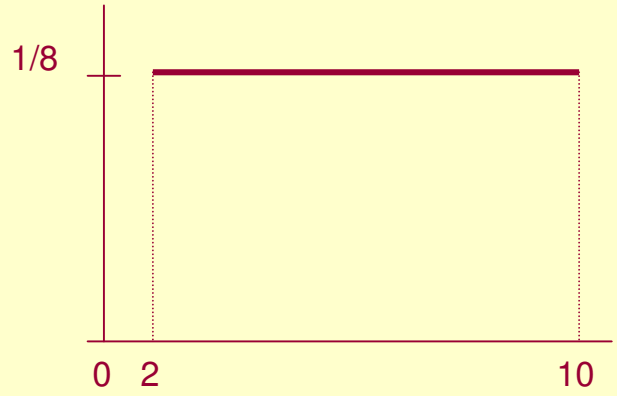
Distribuição Chi-Quadrado



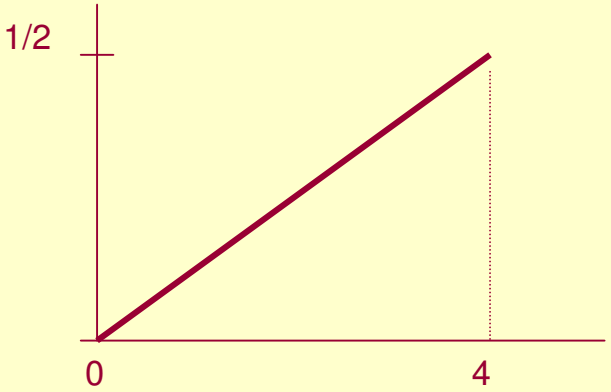
Distribuição Exponencial

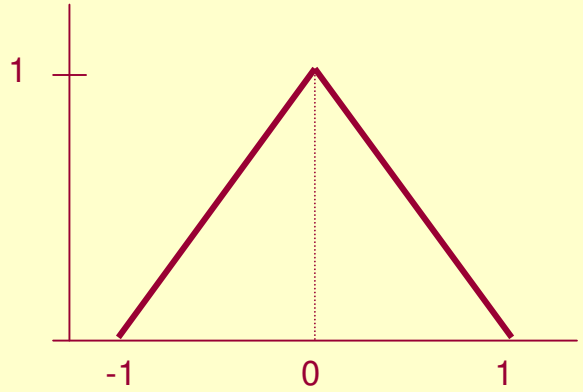
Quais “parâmetros” da distribuição serão adotados para investigar o problema real?

E(Y) ? Var(Y) ? …



Distribuição Uniforme (Retangular)





Distribuição Triangular

Distribuições Exatas

- $Y_1, Y_2, ..., Y_n$ é a.a. da Bernoulli(p)

$$\Rightarrow P\left(\overline{Y} = k/n\right) = \binom{n}{k} p^k (1-p)^{n-k} I_{[0,1,...,n]}(k) \quad \Rightarrow Bino(n; p)$$

- $Y_1, Y_2, ..., Y_n$ é a.a. da Poisson(λ)

$$\Rightarrow P\left(\overline{Y} = k/n\right) = \frac{e^{-n\lambda}(n\lambda)^k}{k!} I_{[0,1,2,...]}(k) \quad \Rightarrow P(n\lambda)$$

- $Y_1, Y_2, ..., Y_n$ é a.a. da Exponencial(θ)

$$\Rightarrow P\left(\overline{Y} \le y\right) = \int_0^y \frac{1}{\Gamma(n)} (nu)^{n-1} \theta^n e^{-n\theta u} n\, du \quad \Rightarrow \Gamma(n; n\theta)$$

Distribuição Qui-Quadrado

$$U \sim \chi^2_n \quad \text{se} \quad f(u) = \frac{1}{\Gamma(n/2)2^{n/2}} u^{(n-2)/2} e^{-u/2} \quad I_{(0,\infty)}(u)$$

- $E(U) = n \qquad E(U^2) = n(n+2) \qquad Var(U) = 2n$
- $Z_1, Z_2, ..., Z_n \quad iid \quad N(0;1) \Rightarrow \sum_{j=1}^{n} Z_j^2 \sim \chi^2_n$
- $Y_1, Y_2, ..., Y_n \quad iid \quad N(\mu;\sigma^2) \Rightarrow \sum_{j=1}^{n} \frac{(Y_j - \mu)^2}{\sigma^2} \sim \chi^2_n$

$$\Rightarrow \sum_{j=1}^{n} \frac{(Y_j - \bar{Y})^2}{\sigma^2} = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$$

$$\Rightarrow \bar{Y} \text{ e } s^2 \text{ são v.a. independentes}$$

Distribuição *t* de Student

$$T \sim t_n \quad \text{se} \quad f(t) = \frac{\Gamma\left((n+1)/2\right)}{\Gamma(n/2)\sqrt{n\pi}}\left(1+\frac{t^2}{n}\right)^{-(n+1)/2} I_{(-\infty,\infty)}(t)$$

- Para $n>1: E(T)=0$ Para $n>2: Var(T)=n/(n-2)$

- Seja $Z \sim N(0;1)$ e $U \sim \chi^2_n$; Z e U v.a. independentes. Então:

$$T = \frac{Z}{\sqrt{U/n}} \sim t_n$$

- $Y_1, Y_2, ..., Y_n$ *iid* $N(\mu;\sigma^2)$, com $\bar{Y}$ e s^2 a média e variância amostral

$$\Rightarrow \frac{\bar{Y}-\mu}{s/\sqrt{n}} \sim t_{n-1}$$

⇒ Justifique a estatística do teste "t"

Distribuição F de Snedecor

$$f(w)=\frac{\Gamma\left(\frac{n_1+n_2}{2}\right)\left(\frac{n_1}{n_2}\right)^{n_1/2}}{\Gamma(n_1/2)\Gamma(n_2/2)}w^{(\frac{n_1-2}{2})}\left(1+\frac{n_1}{n_2}w\right)^{-(\frac{n_1+n_2}{2})}I_{(0,\infty)}(w)$$

- $E(W)=n_2/(n_2-2)$ $\qquad Var(W)=\dfrac{2n_2^2(n_1+n_2-2)}{n_1(n_2-2)^2(n_2-4)^2}$

- Seja $U_1 \sim \chi^2_{n_1}$ e $U_2 \sim \chi^2_{n_2}$; U_1 e U_2 v.a. independentes. Então:

$$W=\frac{U_1/n_1}{U_2/n_2} \sim F_{n_1,n_2}$$

- Se $T \sim t_n$ então $T^2 \sim F_{1,n}$

Formulação do Problema (Clássico)

"Realização de um Experimento Aleatório"

População → **Dados**

Y: Variável Aleatória

Y = y P (y), f (y)

(Y1, Y2, ... , Yn) ***Amostra Aleatória***

PARÂMETROS

π

$\mu \quad \sigma^2$

ESTATÍSTICAS

$\hat{\pi} = p$

$\hat{\mu} = \bar{Y} \quad \hat{\sigma} = s$

$\Rightarrow$ Modelos estruturais: Y=*f* (X)

Estimadores e Estatísticas de Teste

- Função de Verossimilhança (Fisher, 1922)

$$Y_1, Y_2, ..., Y_n \quad iid \quad f(y;\theta)$$

$$L_n(\theta) = \prod_{j=1}^{n} f(Y_j;\theta)$$

- Método da Máxima verossimilhança

$$\frac{\partial}{\partial\theta}\log L_n(\theta) = 0$$

$$\Rightarrow \quad \hat{\theta}_n \,, \qquad 0 = \frac{\partial}{\partial\theta}\log L_n(\theta)\,|_{\theta=\hat{\theta}_n}$$

Soluções numéricas:

⇒Newton-Raphson

⇒Scoring

Estimadores e Estatísticas de Teste

Sob Condições de Regularidade

$$\hat{\theta}_n \sim N\left(\theta; n^{-1} I_\theta^{-1}\right); \qquad I_\theta = E_\theta\left[\frac{\partial}{\partial\theta}\log f(Y;\theta)\right]^2 \equiv -E_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(Y;\theta)\right]$$

- Estatística Razão de Verossimilhanças

$$H_0: \ \theta = \theta_0$$

$$\Lambda_n = 2\left[\ln L\left(\hat{\theta}\right) - \ln L\left(\theta_0\right)\right] \overset{sob H_0}{\sim} \chi^2_r$$

Propriedades de um Estimador

- Consistência: menor erro quadrático médio

$$E_\theta[\,T-\tau(\theta)\,]^2 = Var(T)+[\tau(\theta)-E_\theta(T)]^2$$

- Não Viciado: $E(T) = \tau(\theta)$
- Suficiente: $P(Y / T=t;\theta) = P(Y / T=t)$
- Variância Mínima (na classe dos não-viciados)
- Locação Invariante: $t(y+k)=t(y)+k$
- Escala invariante: $t(ky) = kt(y)$
- *Robustez, Resistência …*

Úteis em análises exploratórias!

Medidas de Posição

Primeiro quartil: valor que deixa 25% das observações abaixo dele

$$Q_1 = Y_{\left(\frac{1}{4}(n+1)\right)}$$

Segundo quartil: mediana

$$Q_2 = Y_{\left(\frac{2}{4}(n+1)\right)}$$

Terceiro quartil: valor que deixa 75% das observações abaixo dele

$$Q_3 = Y_{\left(\frac{3}{4}(n+1)\right)}$$

Resumo de 5 números: $Min \quad Q_1 \quad Q_2 \quad Q_3 \quad Max$

$$Trimédia = \frac{1}{4}Q_1 + \frac{1}{2}Q_2 + \frac{1}{4}Q_3$$

Outras Médias

Média Geométrica

$$\overline{Y}_G = \sqrt[n]{Y_1 Y_2 \ldots Y_n}$$

Média Harmônica

$$\frac{n}{\overline{Y}_H} = \frac{1}{Y_1} + \frac{1}{Y_2} + \ldots + \frac{1}{Y_n}$$

Média Geométrica

$$\overline{Y}_G = \sqrt[n]{Y_1 Y_2 \ldots Y_n}$$

⇒ Média Proporcional, Média das Taxas, das Razões

Ex. Taxa Média de Lucro

	2000	2001	2002
R$	500	650	900

$$\frac{650}{500} \qquad \frac{900}{650}$$

$$\overline{Y}_G = \sqrt{\frac{650}{500}\frac{900}{650}} = 1{,}34$$

Média Geométrica

Ex. Epidemia de Gripe

	1º Dia	2º Dia	3º Dia
# Casos	12	18	48

$$\frac{18}{12} \qquad \frac{48}{18}$$

1º Dia → 2º Dia : # de casos de gripe foi multiplicado por 18/12

2º Dia → 3º Dia : # de casos de gripe foi multiplicado por 48/18

Calcule a Média geométrica destas duas taxas de crescimento?

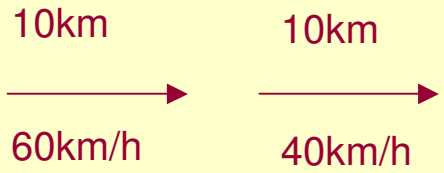
Estime o número de casos para o 4º e 5º Dia, assumindo que o padrão de contaminação se mantém constante.

Média Harmônica

$$\frac{n}{\overline{Y}_H} = \frac{1}{Y_1} + \frac{1}{Y_2} + \ldots + \frac{1}{Y_n}$$

Valoriza a regularidade (harmonia) $\Rightarrow$ é a média das ações de vários indivíduos, desenvolvidas quando ocorre a colaboração de uma ação com as outras.

Ex. Velocidade Média



$$\frac{1}{\overline{Y}_H} = \frac{1}{2}\left(\frac{1}{60} + \frac{1}{40}\right) = 48\,\text{km/h}$$

$\overline{Y}$ = 50 km/h ??

$\overline{Y}_H \Rightarrow$ 20km em 25min

Harmônico Global

$$\frac{1}{H} = \frac{1}{Y_1} + \frac{1}{Y_2} + \ldots + \frac{1}{Y_n}$$

Harmonia e Matemática

Ex. Uma pessoa demora 6h para construir um muro e outra leva 9h. Pondo-se as duas pessoas trabalhando juntas em quanto tempo o muro estará pronto?

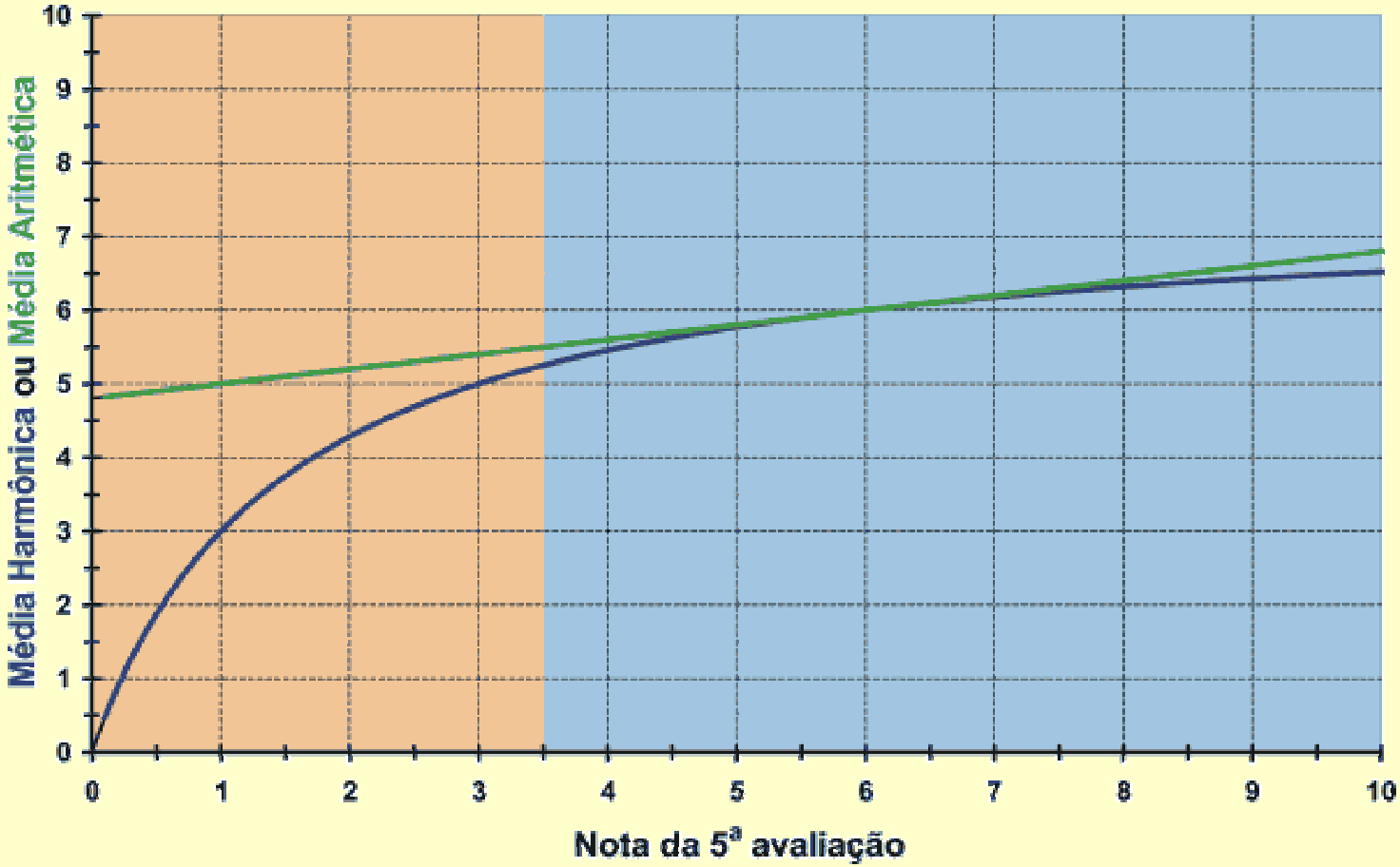
H = 3,6 h = 3 h 36 min

Comparação entre as Médias

$$\overline{Y}_H \leq \overline{Y}_G \leq \overline{Y}$$

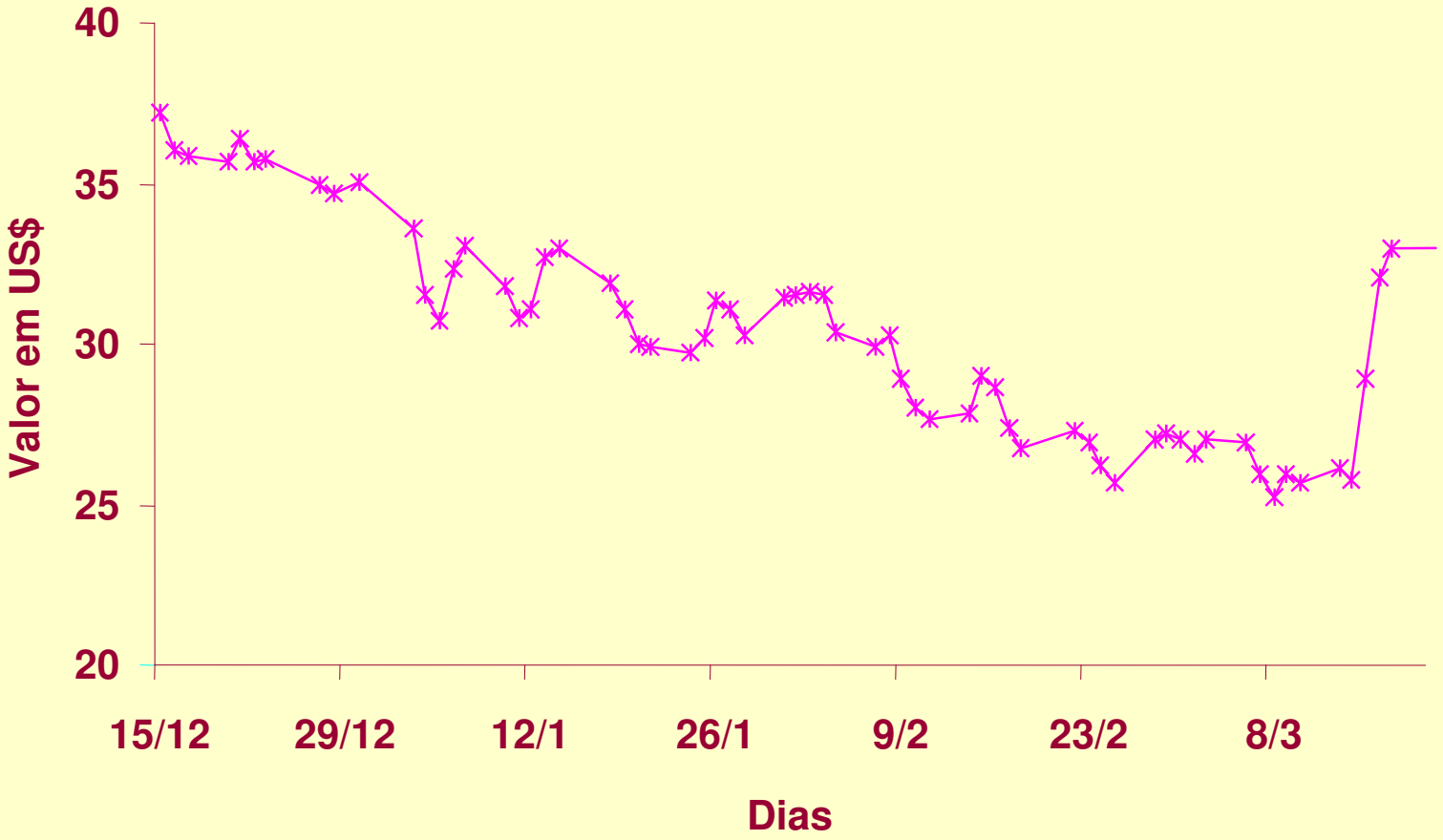
$$\overline{Y}_G \cong \frac{\overline{Y}_H + \overline{Y}}{2}$$ quando os valores da variável não diferirem muito

Aluno	P1	P2	P3	$\overline{Y}$	$\overline{Y}_G$	$\overline{Y}_H$
A1	7	7	7	7	7	7
A2	6	8	7	7	6,95	6,90
A3	4	10	7	7	6,54	6,08



Simulação: Dados = { 6 6 6 6 5a Nota variando entre 0-10 }

Gráficos de Linha – Séries Temporais



Médias Móveis Simples

Série temporal com *n* observações: Y_1 Y_2 ... Y_n

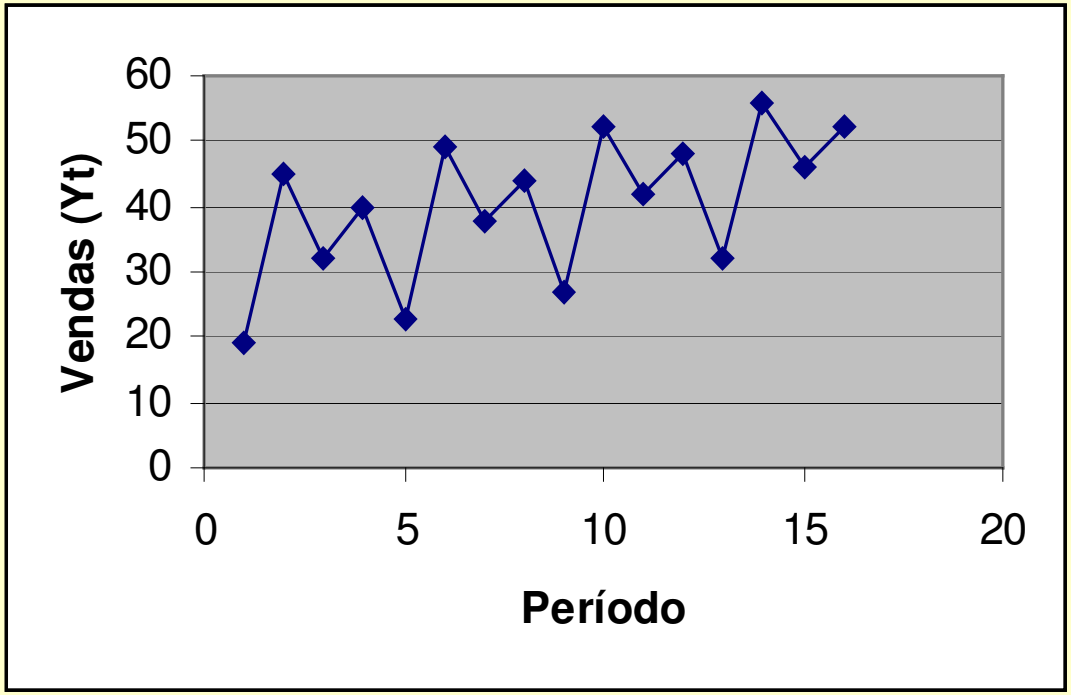
Média Móvel com amplitude k

$$MM_{T[h=k]} = MM_T = \frac{1}{k}\sum_{j=0}^{k-1} Y_{T-j}$$

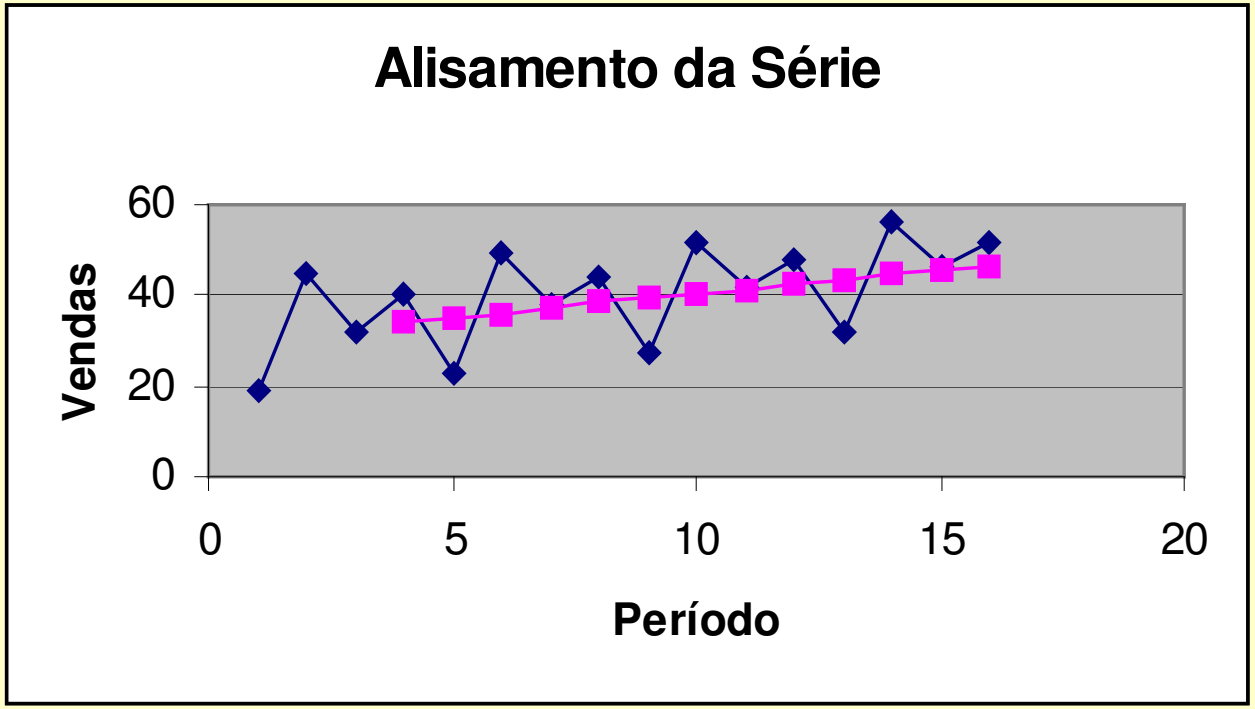
$$= \frac{Y_T + Y_{T-1} + \ldots + Y_{T-k+1}}{k}$$

$$= MM_{T-1} + \frac{Y_T - Y_{T-k}}{k}$$

Período	Vendas (Yt)
1	19
2	45
3	32
4	40
5	23
6	49
7	38
8	44
9	27
10	52
11	42
12	48
13	32
14	56
15	46
16	52



Período	Vendas (Yt)	MM(h=4)
1	19	-
2	45	-
3	32	-
4	40	34
5	23	35
6	49	36
7	38	37,5
8	44	38,5
9	27	39,5
10	52	40,25
11	42	41,25
12	48	42,25
13	32	43,5
14	56	44,5
15	46	45,5
16	52	46,5



Outras Médias

Média Quadrática

$$\bar{Y}_Q = \frac{1}{n}\sum_{j=1}^{n} D_j^2 \qquad D_j = \left(Y_j - \bar{Y}\right)$$

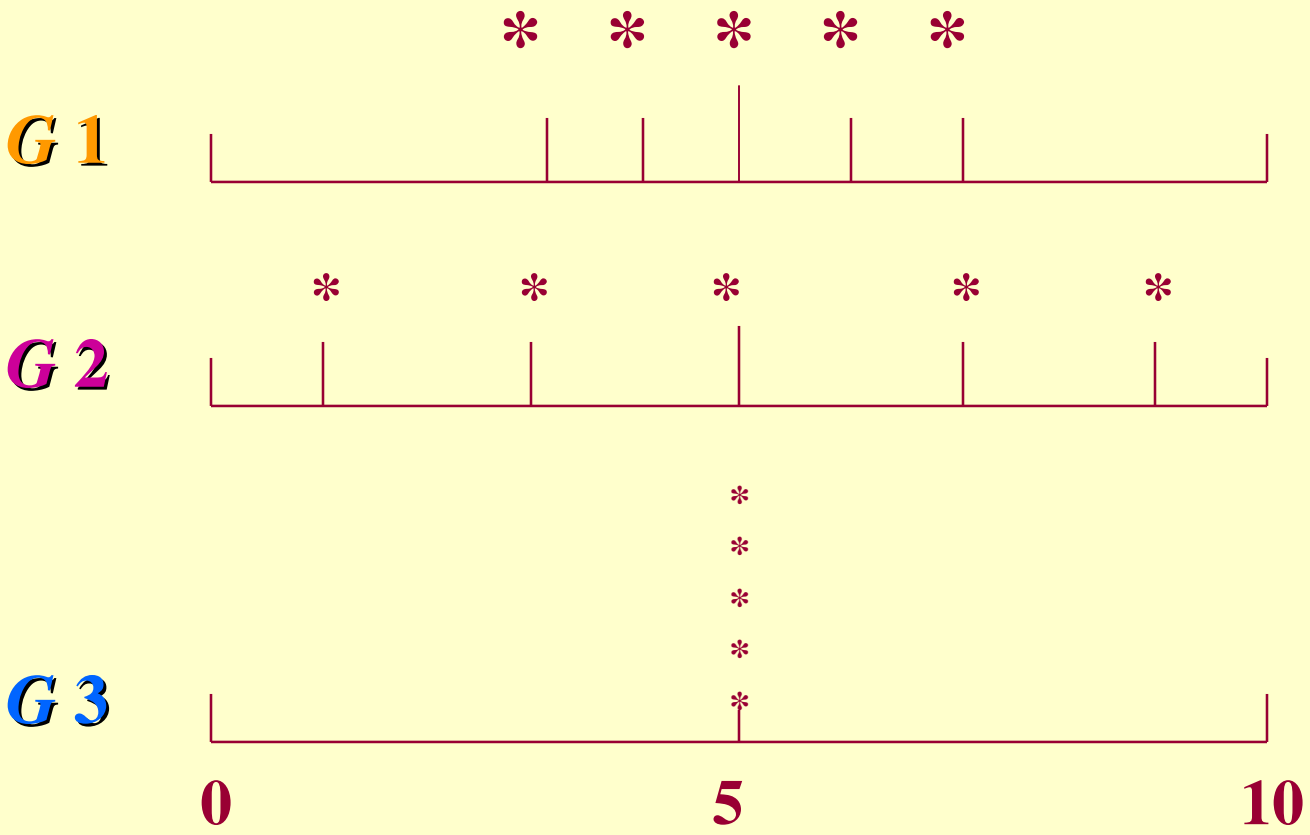
Média Absoluta

$$\bar{Y}_Q = \frac{1}{n}\sum_{j=1}^{n} \left|D_j\right| \qquad D_j = \left(Y_j - \bar{Y}\right)$$

e ainda existem outras propostas ...

Exemplo 2: Considere as notas de um teste de 3 grupos de alunos

Grupo 1: 3,4,5,6,7 **Grupo 2: 1, 3, 5, 7, 9** **Grupo 3: 5,5,5,5,5**



Temos: $\overline{x}_1 = \overline{x}_2 = \overline{x}_3 = 5$ e $md_1 = md_2 = md_3 = 5$

Medidas de Dispersão

Finalidade: encontrar um valor que resuma a variabilidade de um conjunto de dados

- **Amplitude (*A*):**

$$A = máx - min$$

Para os grupos anteriores, temos:

Grupo 1, $A = 4$

Grupo 2, $A = 8$

Grupo 3, $A = 0$

- **Variância**

$$Var(x) = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n} = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n}$$

- **Variância amostral:**

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1} = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$$

- **Desvio padrão:**

$$\text{Desvio Padrão} = s = \sqrt{\text{Variância}}$$

Cálculo para os grupos:

G1: $s^2 = \frac{(3-5)^2+(4-5)^2+(5-5)^2+(6-5)^2+(7-5)^2}{4}$

$\Rightarrow \quad s^2 = 10/4 = 2,5 \quad \Rightarrow s = 1,58$

G2: $s^2 = 10 \Rightarrow s = 3,16$

G3: $s^2 = 0 \Rightarrow s = 0$

Fórmula alternativa:

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

Em ***G***1: $\Sigma X_i^2 = 9 + 16 + 25 + 36 + 49 = 135$

$$\Rightarrow \quad S^2 = \frac{135 - 5\times(5)^2}{4} = 2{,}5$$

• Coeficiente de Variação (*CV*)

- é uma medida de dispersão relativa
- elimina o efeito da magnitude dos dados
- exprime a variabilidade em relação à média

$$CV = \frac{s}{\overline{x}} \times 100\ \%$$

Exemplo 3:

Altura e peso de alunos

	Média	Desvio Padrão	Coef. de Variação
Altura	**1,143m**	**0,063m**	**5,5%**
Peso	**50 kg**	**6kg**	**12%**

Conclusão: Os alunos são, aproximadamente, duas vezes mais dispersos quanto ao peso do que quanto à altura.

Exemplo 4:

Altura (em *cm*) de uma amostra de recém-nascidos e de uma amostra de adolescentes

	Média	Desvio padrão	Coef. de variação
Recém-nascidos	**50**	**6**	**12%**
Adolescentes	**160**	**16**	**10%**

Conclusão: Em relação às médias, as alturas dos adolescentes e dos recém-nascidos apresentam variabilidade quase iguais.

- **Desvio Médio**

$$dm(x) = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \bar{x}\right|$$

- Desvio mediano absoluto (uma medida robusta em relação à mediana)

$$dma(x) = md_{1 \le i \le n} \left| x_i - md(x) \right|$$

- Variância aparada ($S^2(\alpha)$)

•Intervalo-Interquartil:

É a diferença entre o terceiro quartil e o primeiro quartil, ou seja, $Q3 - Q1$.

Dados: 1,9 2,0 2,1 2,5 3,0 3,1 3,3 3,7 6,1 7,7

$Q1 = 2,05$ e $Q3 = 4,9$

$Q3 - Q1 = 4,9 - 2,05 = 2,85$

Para uma dist. Normal:

$$IQ = 1{,}349\sigma$$

Um estimador do desvio padrão populacional

$$S^* = IQ/1{,}349$$

Medidas de Variabilidade
Consistência

➡ $IQ = Q_3 - Q_1$

$$Y_j \sim N(\mu;1) \quad \Rightarrow \quad E(IQ) = 2 * qnorm\ (3/4) = 1{,}3490$$

$$\Rightarrow \frac{IQ}{1{,}3490}$$

Assegurar consistência

➡ $MAD = Mediana_{\ j} \left\{ \left| Y_j - Q_2 \right| \right\}$

$$Y_j \sim N(\mu;\sigma^2) \quad \Rightarrow \quad E(MAD) = \frac{\sigma}{qnorm\ (3/4)} = 1{,}4826\ \sigma$$

$$\Rightarrow \quad 1{,}4826\ MAD$$

Multiplicação por um fator de escala para assegurar consistência

Medidas de Variabilidade

$$AV = Y_{(n)} - Y_{(1)}$$ ⇒ Amplitude de Variação ("Range")

$$IQ = Q_3 - Q_1$$ ⇒ Intervalo Inter-Quartil

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$ ⇒ Variância (u.m.²) ⇒ s

$$dm = \frac{1}{n}\sum_{j=1}^{n}\left|Y_j - \bar{y}\right|$$ ⇒ Desvio Médio Absoluto

$$dma = Mediana_{\,j}\left\{\left|Y_j - Q_2\right|\right\}$$ ⇒ Desvio Mediano Absoluto

$$CV = \frac{s}{\bar{Y}} \times 100\ \%$$ ⇒ Coeficiente de Variação (adimensional)

Comparação entre Estimadores de Escala

Scores de desempenho escolar de estudantes da zona rural e urbana

Rural				Urbano			
Ind.	Y_j	$Y_{(j)}$	$\lvert Y_{(j)} - Q_2 \rvert$	Ind.	Y_j	$Y_{(j)}$	$\lvert Y_{(j)} - Q_2 \rvert$
1	800	500	312	1	900	675	225
2	974	700	112	2	803	751	149
3	500	725	87	3	1145	765	135
4	725	765	47	4	900	803	97
5	812	794	18	5	1225	825	75
6	794	800	12	6	751	850	50
7	765	812	0	7	825	900	0
8	900	826	14	8	1070	900	0
9	826	850	38	9	1128	1070	170
10	700	850	38	10	1080	1080	180
11	850	900	88	11	675	1128	228
12	945	945	133	12	850	1145	245
13	850	974	162	13	765	1225	325
Média=803.15		Mediana=812		Média=932.08		Mediana=900	
s=120.37		IQ=85		s=176.58		IQ=277	
AD=81.62		MAD=47		AD=144.54		MAD=149	

- os valores de MAD indicam uma grande variabilidade entre os desempenhos dos estudantes da zona urbana. Entre os estudantes da zona rural há mais homogeneidade

Estimadores de Escala
Medidas de Variabilidade

Scores de desempenho escolar de estudantes da zona rural e urbana

Medida	Rural	Urbano	U/R
s	120.37	176.58	1.47
AD	81.62	144.54	1.77
MAD	47	149	3.17
IQ	85	277	3.26
s'	82.20	176.58	2.15

s é influenciado pelo valor 500

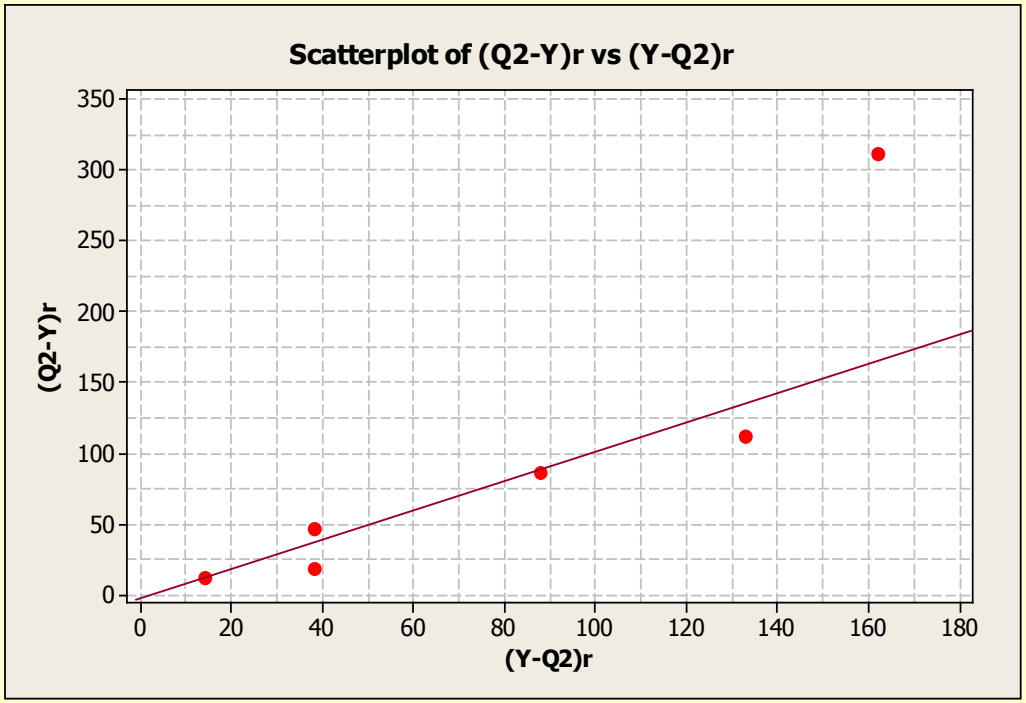
Super-estimam

⇒ IQ: aparar 50% das obs parece ser muito

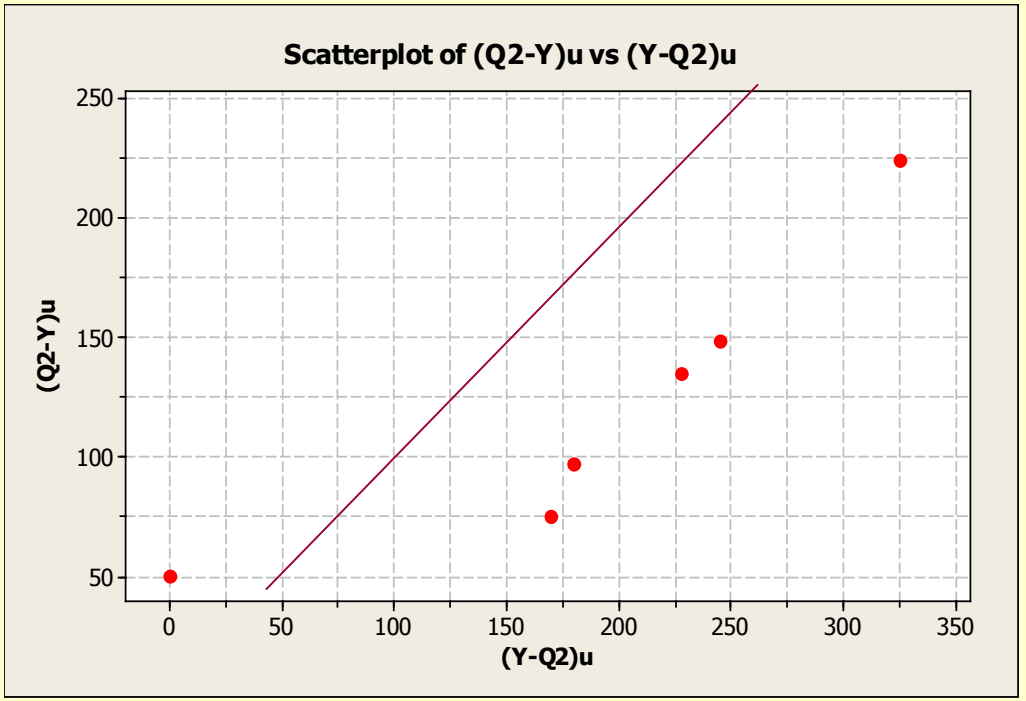
⇒ Escolha apropriada do estimador de escala não é simples

Gráficos de Simetria para os dados do Score de Estudantes

Rural



Urbana



⇒ Transformações para garantir simetria: transformações potência

$$Y \quad \Rightarrow \quad k\,Y^{p}$$

Estimadores de Escala
Medidas de Variabilidade

Calcule as medidas s e MAD para as amostras a seguir:

Amostra 1	Amostra 2
0	0
1	0
2	2
3	3
10	4

$\Rightarrow$ Note que na presença de obs atípicas (irregularidades locais nos dados) a escolha do estimador de escala não é simples.

Comparação de Estimadores de Escala

Estudos de Simulação (n=20)

Valor esperado do estimador (8000 "runs")

Grau crescente de contaminação para caudas pesadas →

Estimador	20 N(0;1)	19 N(0;1), 1 N(0;100)	20 N(0;1) / U(0;1)
s	0.98	2.23	23.26
MAD	0.64	0.68	1.51
IQ	1.35	1.41	3.25

- s é mais sensível a caudas pesadas
- MAD e IQ são mais resistentes
- Compararar estimadores ⇒ comparar suas variâncias (precisão)

⇒ A variância de um estimador depende de seu valor esperado ⇒ Como comparar estimadores se há variação dentro e entre distribuições?

⇒ Adotar a variância do ln do estimador, *V(ln T)*, que tem boas propriedades de invariância

$$Var[\ln(kT)] = Var[\ln(T) + \ln(k)] = Var[\ln(T)]$$

Estimadores de Escala

Estudos de Simulação (n=20)

Valor da *Var [ln (estimador)]* *(8000 "runs")*

Estimador	20 N(0;1)	19 N(0;1), 1 N(0;100)	20 N(0;1) / U(0;1)
s	0.026	0.271	1.1
AD	0.032	0.085	0.634
MAD	0.074	0.071	0.105
IQ	0.063	0.062	0.115
Mínima Var	0.026	0.029	0.099

Limite de Cramer-Rao

Simulação

Eficiência dos estimadores (razão entre variâncias)

Estimador	20 N(0;1)	19 N(0;1), 1 N(0;100)	20 N(0;1) / U(0;1)	Trieficiência
s	100	11	9	9
AD	81	34	16	16
MAD	35	41	94	35
IQ	41	47	86	41

IQ é um estimador robusto e teve a maior eficiência relativa (trieficiência)