Distribuições Contínuas Univariadas

jlborges@fe.up.pt

Uma variável aleatória é utilizada para estudar experiências aleatórias

Se o número de resultados possíveis for finito temos uma variável aleatórias discreta

- Lançamento de um dado
- 1,2,3,4,5,6

Se o número de resultados possíveis for infinito temos uma **variável aleatória contínua**

- o peso exato de um aluno da FEUP selecionado ao acaso
- entre 40.0(0) e 110.0(0)

Vamos estudar um conjunto de **distribuições contínuas univariadas** pela frequência com que ocorrem fenómenos aleatórios

- Uniforme
- Exponencial negativa
- Normal

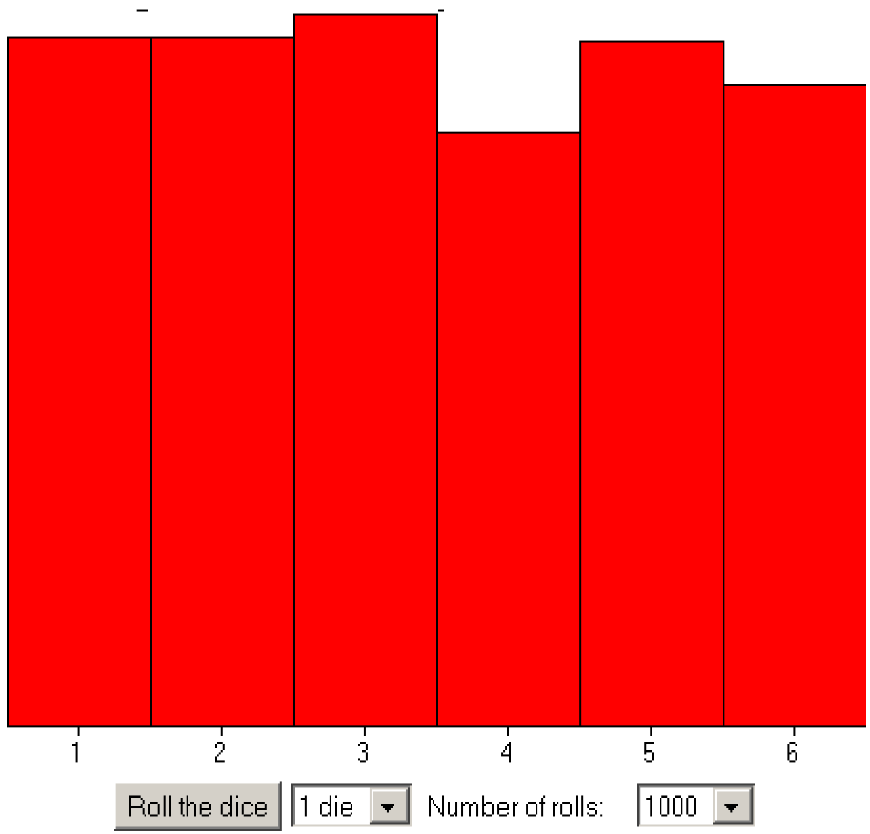
Mais tarde vamos utilizar três distribuições que são necessárias para a resolução de muitos problemas de inferência estatística

- Qui-quadrado
- T de student
- F

jlborges@fe.up.pt

Distribuição Uniforme

A distribuição Uniforme representa variáveis com probabilidade constante



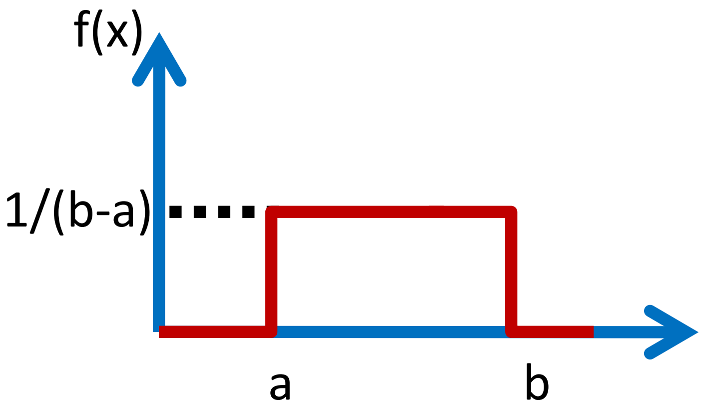
http://www.amstat.org/publications/jse/v6n3/applets/CLT.html

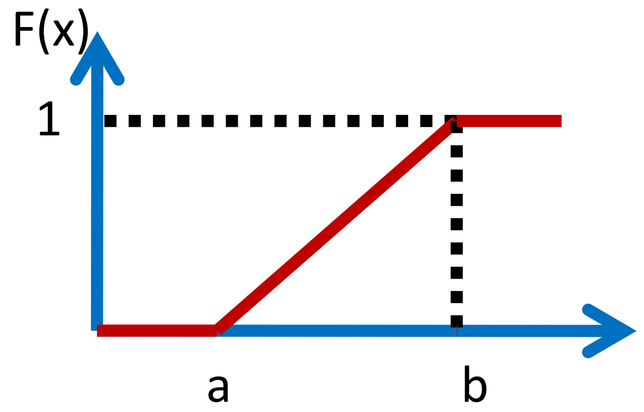
Distribuição Uniforme

$$X \rightarrow U(a,b)$$

$$f(x)=\begin{cases} 1/(b-a), & se \quad a \le x \le b \\ 0, & se \quad x<a \quad ou \quad x>b \end{cases}$$

$$F(x)=\begin{cases} 0, & se \quad x<a \\ (x-a)/(b-a), & se \quad a \le x \le b \\ 1, & se \quad x>b \end{cases}$$





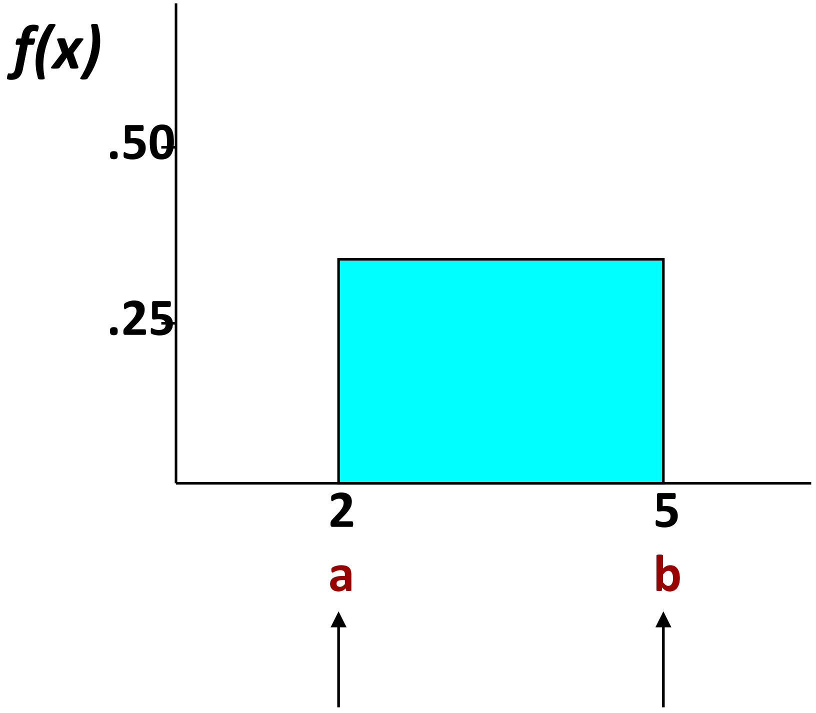
Média e Variância

$$\mu=\frac{a+b}{2} \qquad \sigma^2=\frac{1}{12}\cdot(b-a)^2$$

Distribuição Uniforme

$$f(x) = \frac{1}{5-2} = \frac{1}{3} = 0.33$$

for $2 \leq x \leq 5$

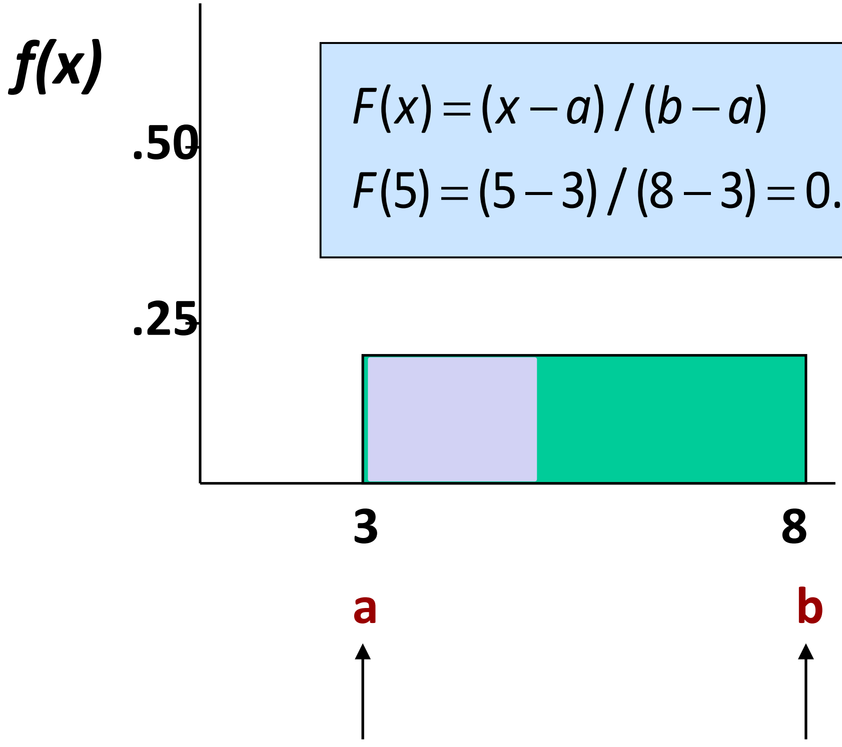


$$f(x) = \frac{1}{8-3} = \frac{1}{5} = 0.20$$

for $3 \leq x \leq 8$

$$F(x) = (x-a)/(b-a)$$

$$F(5) = (5-3)/(8-3) = 0.4$$



Exemplo

O atraso de um comboio, expresso em minutos, segue uma distribuição U(0,12)

Qual a probabilidade de ocorrer um atraso compreendido entre 5 e 10 minutos?

$$X \rightarrow U\left(0,12\right)$$

$$P(5 < X < 10) = F(10) - F(5) = \frac{(10-0)-(5-0)}{12-0} = 0.412$$

Qual a média e a variância do atraso?

$$\mu = \frac{0+12}{2} = 6 \text{ minutos} \qquad \sigma^2 = \frac{1}{12} \cdot (12-0)^2 = 12 \text{ minutos}^2$$

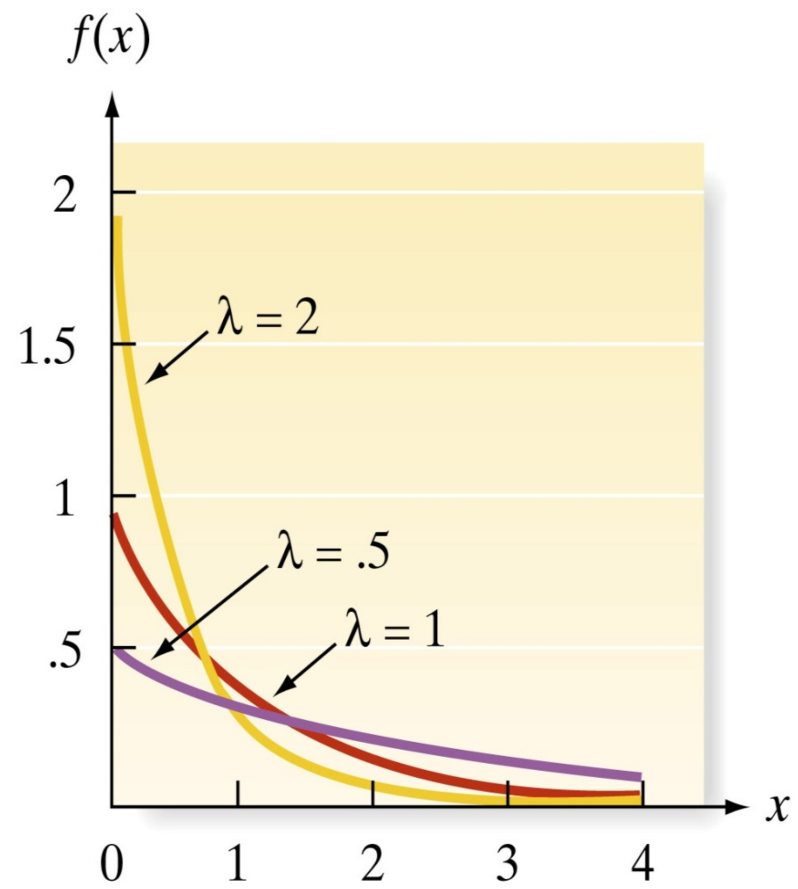
Importância da Distribuição Uniforme

- Representa experiências em que todos os resultados são igualmente prováveis
- É fácil gerar números aleatórios de acordo com a distribuição uniforme
  - gerador de números aleatórios disponíveis nas máquinas calcular
  - tômbola com bolas numeradas
- É possível gerar números aleatórios de acordo com outras distribuições a partir de números uniformemente distribuídos

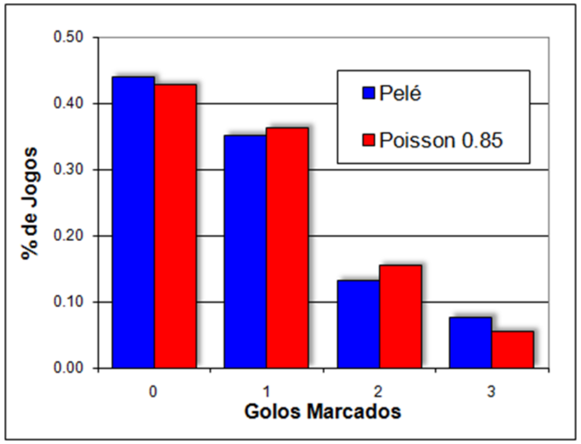
Distribuição Exponencial Negativa

- Utilizada para descrever o **tempo entre ocorrências** de acontecimentos sucessivos de uma distribuição de **Poisson**

- A forma da distribuição é determinada pelo valor de λ

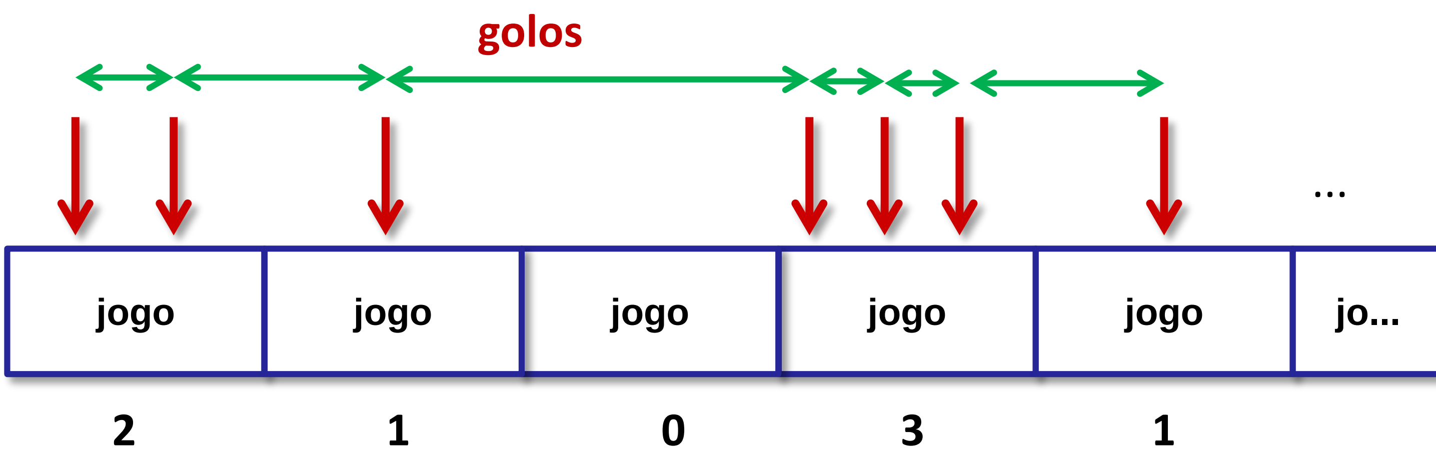


$$f\left(x\right) = \lambda e^{-\lambda x}, (x > 0)$$



Poisson: número de golos por jogo

Exponencial Negativa: tempo (minutos) entre golos



Distribuição Exponencial Negativa

Função densidade de probabilidade: $$\boxed{f(x) = \lambda e^{-\lambda x}, (x > 0)}$$

Função distribuição: $$\boxed{F(x) = 1 - e^{-\lambda x}, (x > 0)}$$

Média: $$\boxed{\mu = \frac{1}{\lambda}}$$

Desvio Padrão: $$\boxed{\sigma = \frac{1}{\lambda}}$$

Exemplo

O **nº de avarias diárias** de um dado equipamento segue uma **distribuição Poisson(λ=0.5)**

Qual a probabilidade de o **tempo entre duas avarias** sucessivas ser superior a 1 dia?

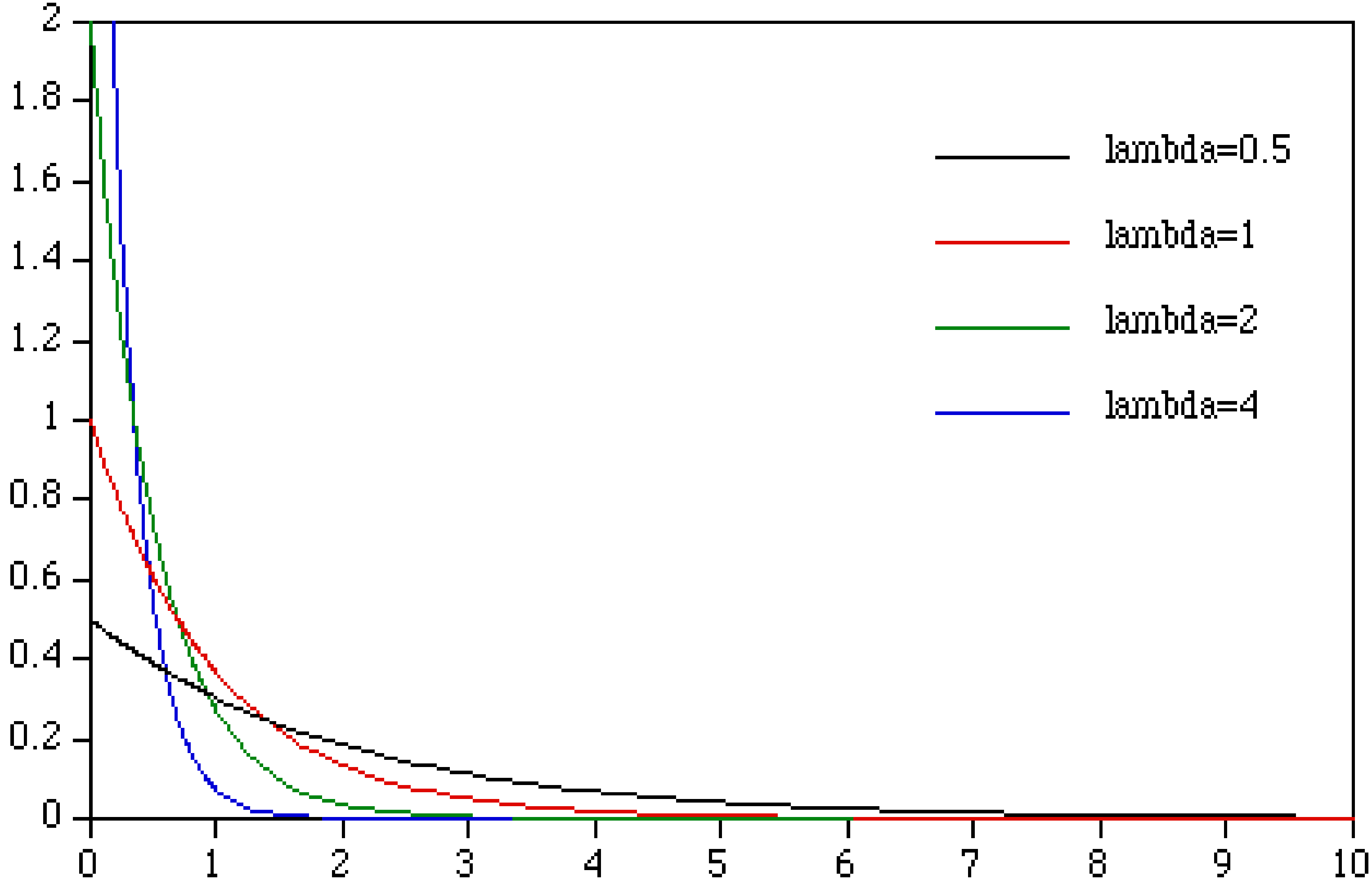
$$X \quad \rightarrow EN(0.5)$$

$$P(X > 1) = 1 - F(1) = 1 - (1 - e^{-0.5 \cdot 1}) = 0.607$$

Qual a média e a variância do tempo entre avarias sucessivas?

$$\mu = 1/0.5 = 2 \text{ dias} \qquad \sigma^2 = (1/0.5)^2 = 4 \text{ dias}^2$$

Função densidade de probabilidade para diferentes valores do parâmetro da distribuição Exponencial Negativa



Retirado de um artigo cientifico

Jean Tourrilhes. *Packet Frame Grouping : Improving IP multimedia performance over CSMA/CA*. Proc. of ICUPC `98 (IEEE International Conference on Universal Personal Communications)

...

Traffic models

There are usually two ways to model the traffic over a network: by using a statistical model or by using network traces (recorded on a real network). As there are no traces over 802.11 or at 2 Mb/s, statistical models have been used for this set of simulations.

5.3.1 Random traffic

The first traffic model used is the well known *random* traffic generator: all packet sizes are **uniformly distributed in *]0 ; max packet size]*** and packets arrive at the MAC following a **Poisson** process (random interarrival time with **negative exponential distribution**).

The main advantage of this traffic model is that it is already widely used and it allows to explore the full range of network loads and packet sizes

...

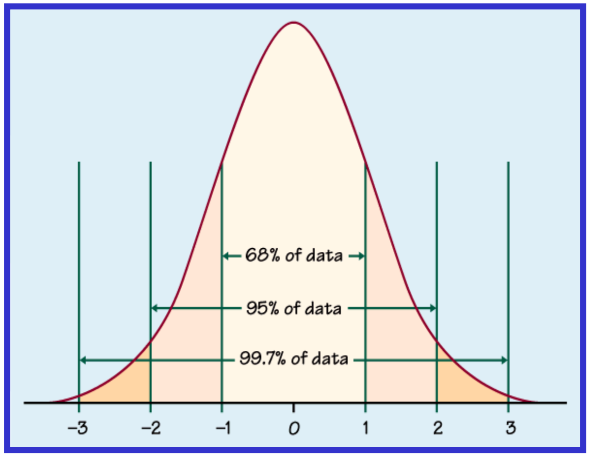
Importância da distribuição Exponencial Negativa

- Fenómenos de Poisson são muito frequentes
- EN permite caracterizar tempo entre ocorrências de Poisson
- Importante para jogos
  - Poisson permite caracterizar o número de golos por jogo, terramotos por século, ...
  - A distribuição Exponencial Negativa permite caracterizar o tempo entre essas ocorrências, o que é essencial para situar as ocorrências no tempo

Distribuição Normal (ou de Gauss)

Perfeitamente Simétrica

Média = Mediana = Moda



Uma variável que é afetada por muitos efeitos pequenos e independentes pode ser aproximada por uma distribuição normal

Isto explica porque a distribuição Normal é tão frequente
Cotações de mercado, altura dos alunos, médias de temperaturas anuais...

A altura de um aluno é resultado da genética, nutrição, doenças,...

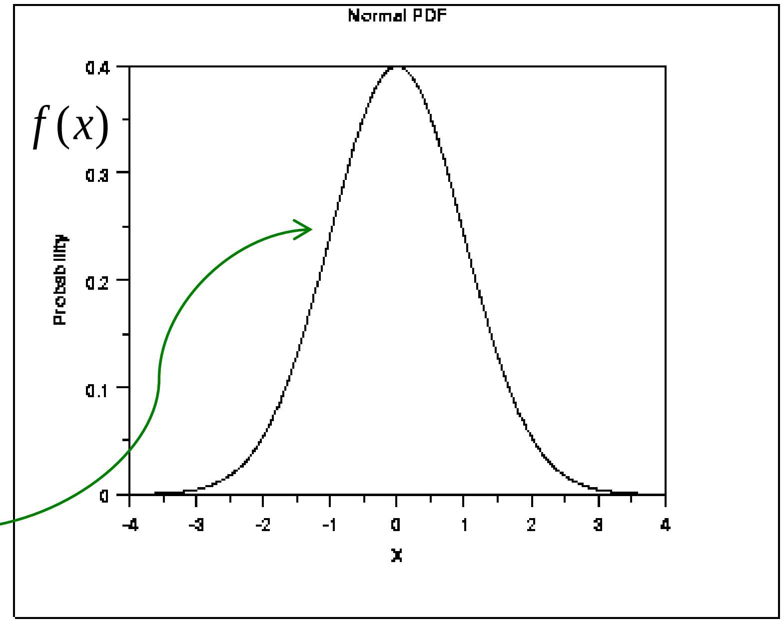
http://webphysics.davidson.edu/Applets/Galton/BallDrop.html

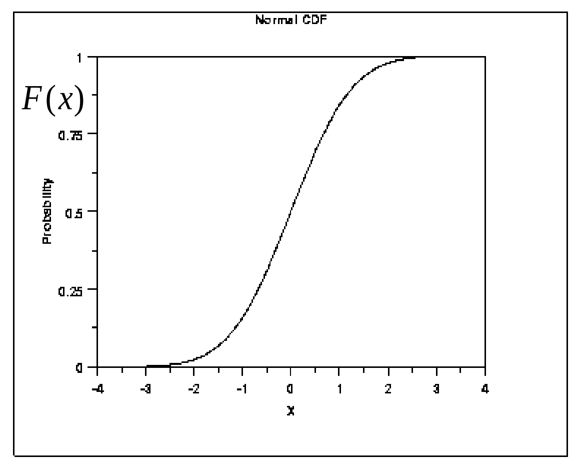
jlborges@fe.up.pt

Distribuição Normal

$$X \rightarrow N(\mu, \sigma^2)$$

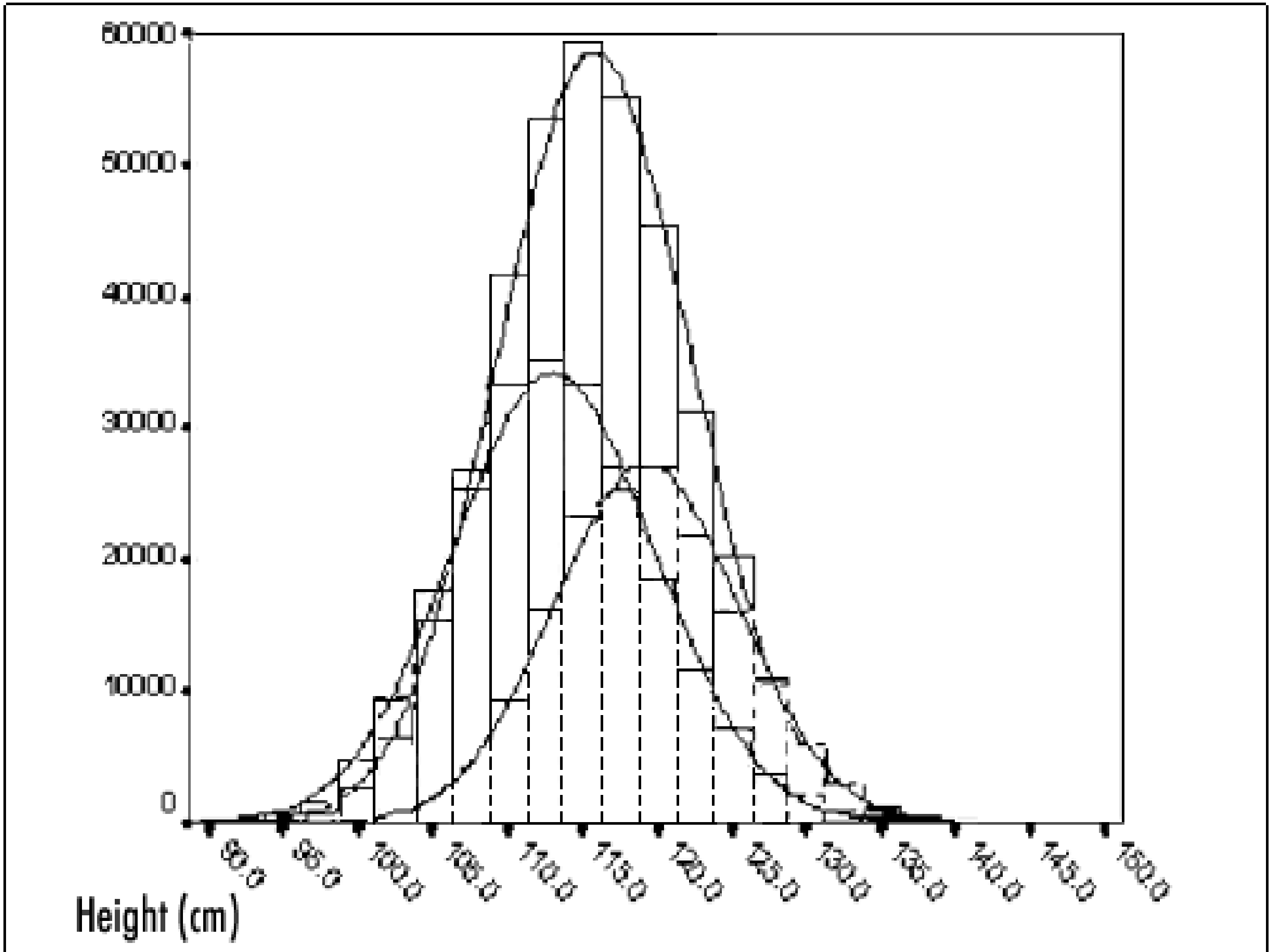
$$f(x) = \frac{1}{\sqrt{2 \cdot \pi \cdot \sigma^2}} \cdot e^{-\frac{1}{2} \cdot \left(\frac{x-\mu}{\sigma}\right)^2}$$





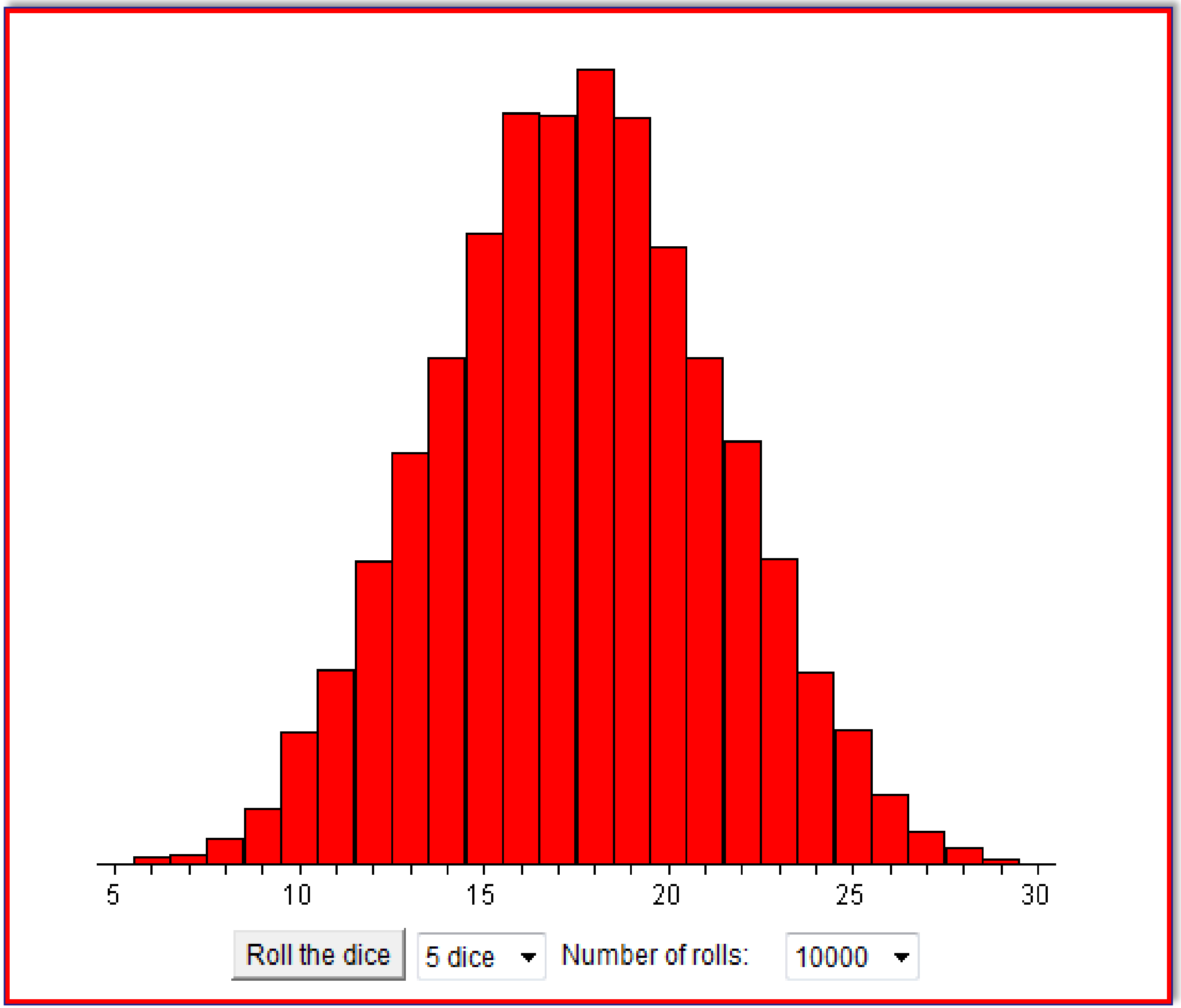
$F(x)$ – só pode ser obtida por via numérica

jlborges@fe.up.pt



Distribution of height in first-grade children in El Salvador, Guatemala, and Honduras from National Census around 2000.

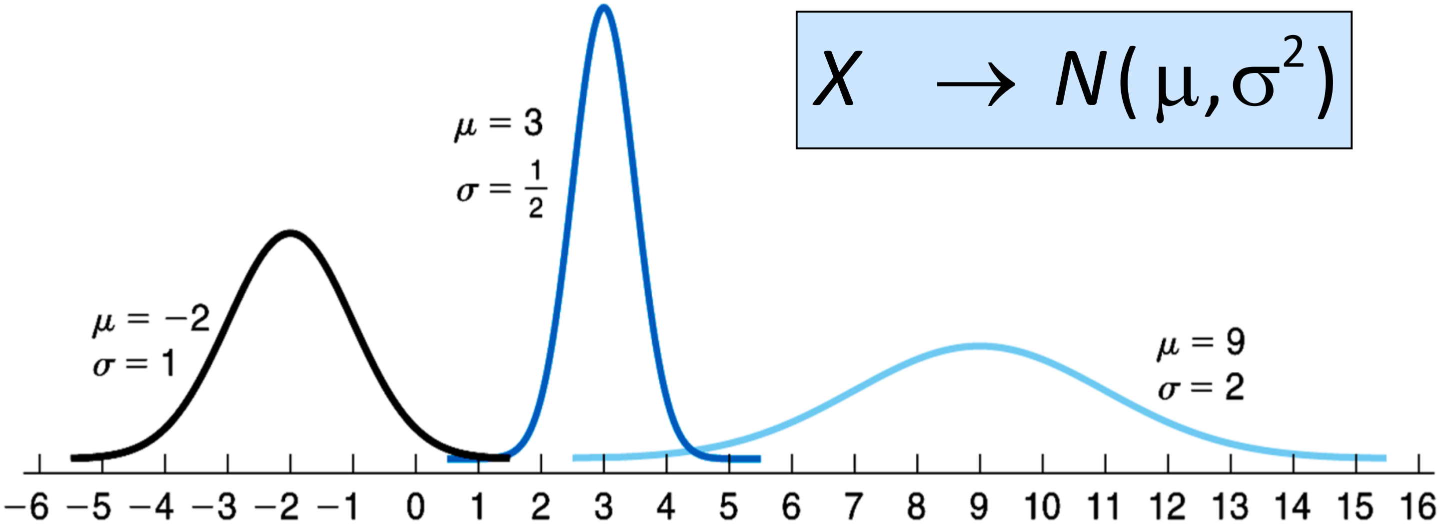
jlborges@fe.up.pt

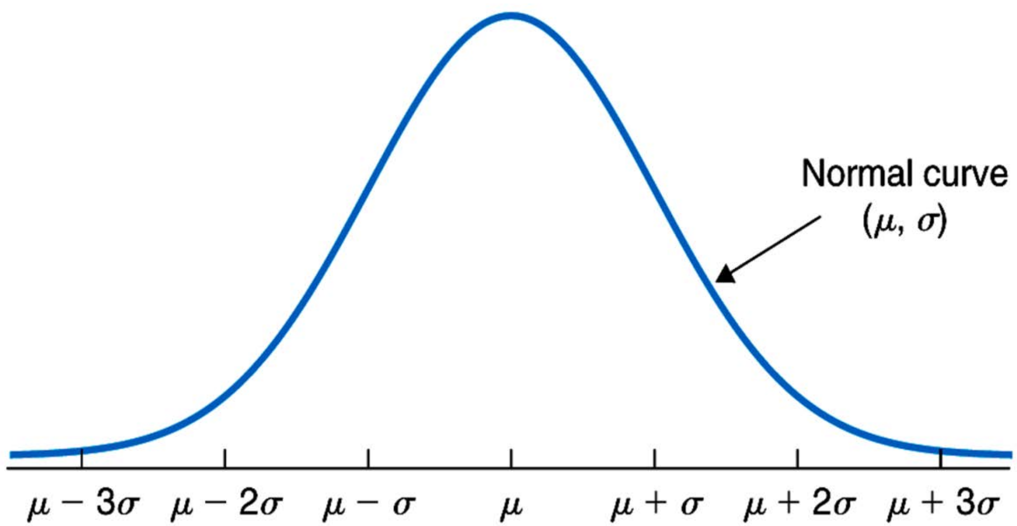


http://www.amstat.org/publications/jse/v6n3/applets/CLT.html

Efeito da variação dos parâmetros da distribuição Normal

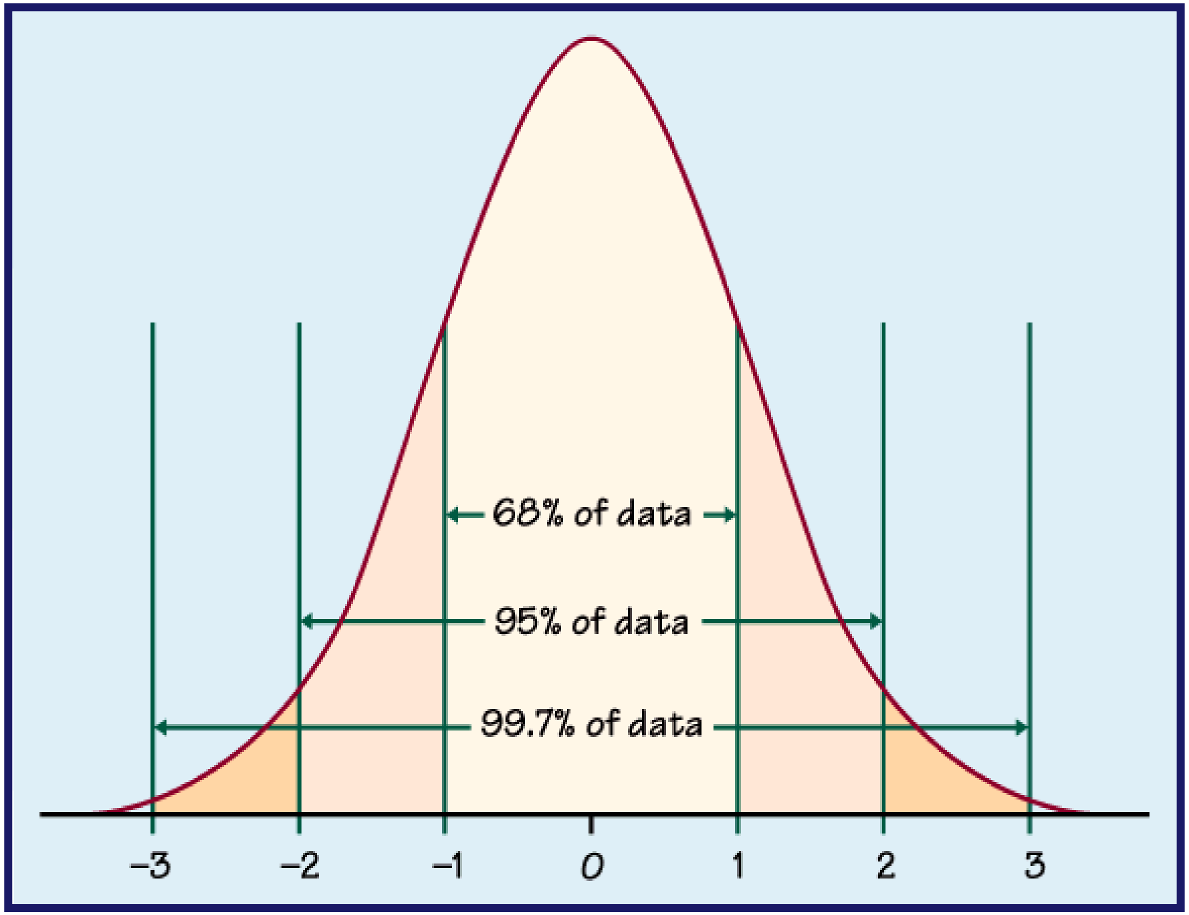
$$X \rightarrow N(\mu,\sigma^2)$$





jlborges@fe.up.pt

The 68-95-99.7 rule for Normal distributions



If we know how many standard deviations away from the mean a score is, assuming a normal distribution, we know what percentage of scores falls above or below that score.

68% of the observations fall within σ of the mean μ

95% of the observations fall within 2σ of the mean μ

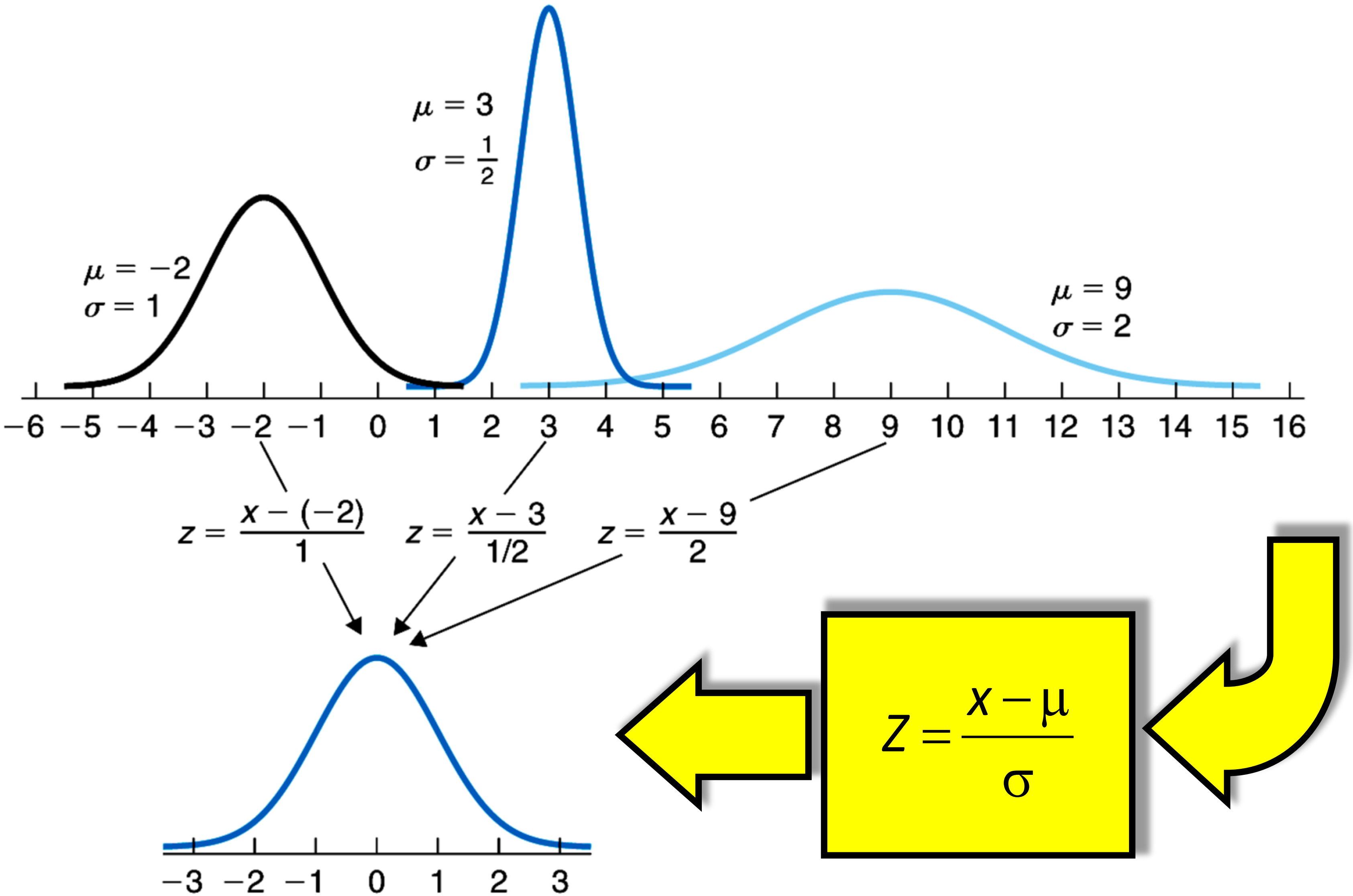
99.7% of the observations fall within 3σ of the mean μ

f(x) não é integrável analiticamente e não é possível construir uma tabela geral para a distribuição Normal

Para tal, seria necessário um número infinito de tabelas

A solução será transformar a distribuição Normal em análise numa determinada distribuição Normal cujas características sejam perfeitamente conhecidas, isto é, a distribuição **Normal Padronizada**

Transformação de uma variável Normal na variável Normal Padronizada



$X \quad \rightarrow N(\mu, \sigma^2)$ variável Normal

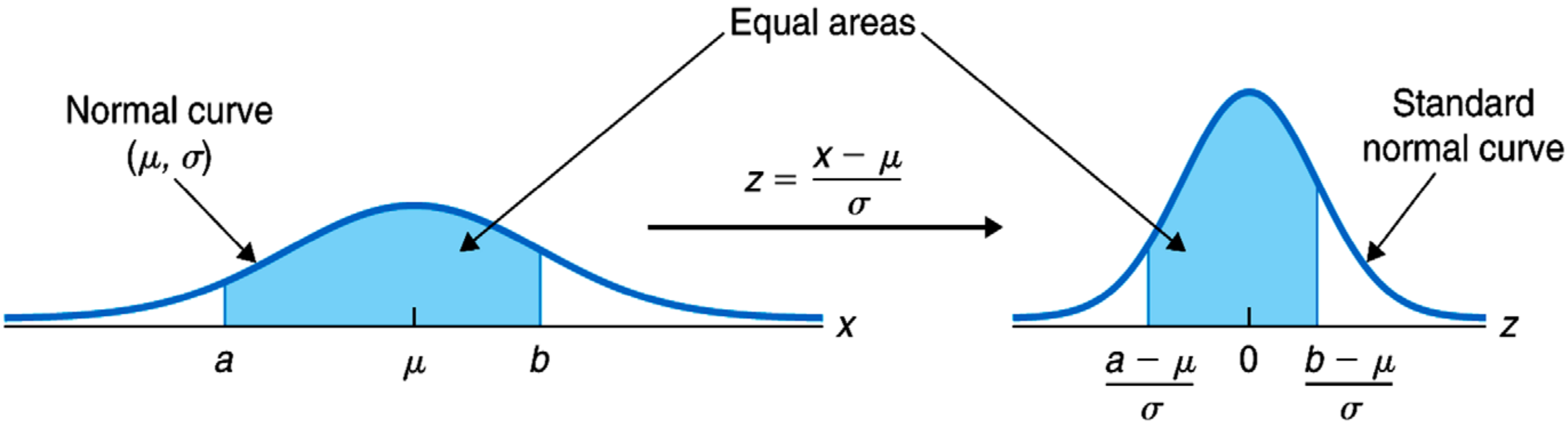
$$\boxed{\begin{array}{l} Z = \dfrac{X - \mu}{\sigma} = -\dfrac{\mu}{\sigma} + \dfrac{1}{\sigma} \cdot X \\ a = -\dfrac{\mu}{\sigma} \quad e \quad b = \dfrac{1}{\sigma} \end{array}}$$

transformação linear

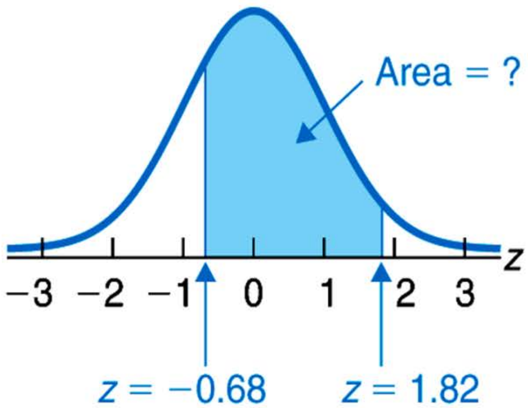
$$\left.\boxed{\begin{array}{l} \mu_Z = E\left(\dfrac{x-\mu}{\sigma}\right) = \dfrac{E(x) - \mu}{\sigma} = 0 \\ \sigma_Z^2 = Var\left(\dfrac{x-\mu}{\sigma}\right) = \dfrac{1}{\sigma^2} Var(x) = 1 \end{array}}\right\}$$

parâmetros da variável transformada

podendo demonstrar-se que **Z** segue uma distribuição **Normal N(0,1)**, que se designa por **Normal Padronizada**


Equal areas
Normal curve
(μ, σ)
$z = \frac{x - \mu}{\sigma}$
Standard
normal curve
x
a
μ
b
z
$\frac{a - \mu}{\sigma}$
0
$\frac{b - \mu}{\sigma}$

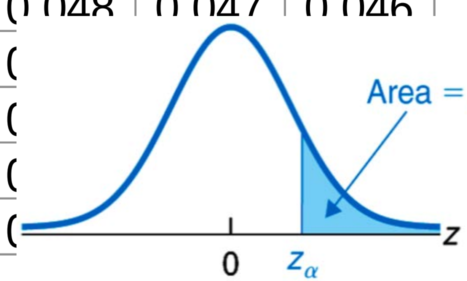
Cálculo de probabilidades para uma distribuição **Normal padronizada**



(a)

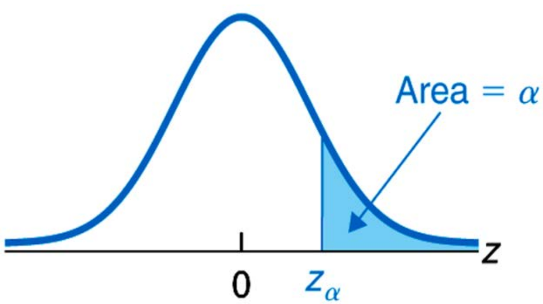
$$P(-0.68 < z < 1.82) = (1 - 0.248) - 0.034$$

z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
...	...	...	...	...	...	...	...	...	...	...
0.5	0.309	0.305	0.302	0.298	0.295	0.291	0.288	0.284	0.281	0.278
0.6	0.274	0.271	0.268	0.264	0.261	0.258	0.255	0.251	**0.248**	0.245
0.7	0.242	0.239	0.236	0.233	0.23	0.227	0.224	0.221	0.218	0.215
...	...	...	...	...	...	...	...	...	...	...
1.6	0.055	0.054	0.053	0.052	0.051	0.05	0.049	0.048	0.047	0.046
1.7	0.045	0.044	0.043	0.042	0.041	0.04	0.039	[illegible]	[illegible]	[illegible]
1.8	0.036	0.035	**0.034**	0.034	0.033	0.032	0.031	[illegible]	[illegible]	[illegible]
1.9	0.029	0.028	0.027	0.027	0.026	0.026	0.025	[illegible]	[illegible]	[illegible]
2	0.023	0.022	0.022	0.021	0.021	0.02	0.02	[illegible]	[illegible]	[illegible]
...	...	...	...	...	...	...	...	...	...	...



jlborges@fe.up.pt

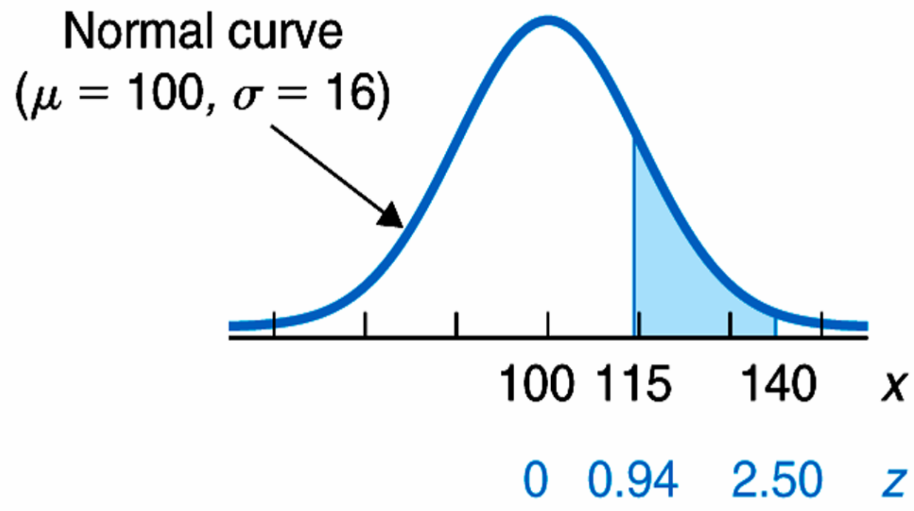
TABELA 3 do livro:
Probabilidades associadas à cauda direita da distribuição normal padronizada



Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.50000	0.49601	0.49202	0.48803	0.48405	0.48006	0.47608	0.47210	0.46812	0.46414
0.1	0.46017	0.45620	0.45224	0.44828	0.44433	0.44038	0.43644	0.43251	0.42858	0.42465
0.2	0.42074	0.41683	0.41294	0.40905	0.40517	0.40129	0.39743	0.39358	0.38974	0.38591
0.3	0.38209	0.37828	0.37448	0.37070	0.36693	0.36317	0.35942	0.35569	0.35197	0.34827
0.4	0.34458	0.34090	0.33724	0.33360	0.32997	0.32636	0.32276	0.31918	0.31561	0.31207
0.5	0.30854	0.30503	0.30153	0.29806	0.29460	0.29116	0.28774	0.28434	0.28096	0.27760
0.6	0.27425	0.27093	0.26763	0.26435	0.26109	0.25785	0.25463	0.25143	0.24825	0.24510
0.7	0.24196	0.23885	0.23576	0.23270	0.22965	0.22663	0.22363	0.22065	0.21770	0.21476
0.8	0.21186	0.20897	0.20611	0.20327	0.20045	0.19766	0.19489	0.19215	0.18943	0.18673
0.9	0.18406	0.18141	0.17879	0.17619	0.17361	0.17106	0.16853	0.16602	0.16354	0.16109
1.0	0.15866	0.15625	0.15386	0.15151	0.14917	0.14686	0.14457	0.14231	0.14007	0.13786
1.1	0.13567	0.13350	0.13136	0.12924	0.12714	0.12507	0.12302	0.12100	0.11900	0.11702
1.2	0.11507	0.11314	0.11123	0.10935	0.10749	0.10565	0.10383	0.10204	0.10027	0.09853
1.3	0.09680	0.09510	0.09342	0.09176	0.09012	0.08851	0.08691	0.08534	0.08379	0.08226
1.4	0.08076	0.07927	0.07780	0.07636	0.07493	0.07353	0.07215	0.07078	0.06944	0.06811
1.5	0.06681	0.06552	0.06426	0.06301	0.06178	0.06057	0.05938	0.05821	0.05705	0.05592
1.6	0.05480	0.05370	0.05262	0.05155	0.05050	0.04947	0.04846	0.04746	0.04648	0.04551
1.7	0.04457	0.04363	0.04272	0.04182	0.04093	0.04006	0.03920	0.03836	0.03754	0.03673
1.8	0.03593	0.03515	0.03438	0.03362	0.03288	0.03216	0.03144	0.03074	0.03005	0.02938
1.9	0.02872	0.02807	0.02743	0.02680	0.02619	0.02559	0.02500	0.02442	0.02385	0.02330
2.0	0.02275	0.02222	0.02169	0.02118	0.02068	0.02018	0.01970	0.01923	0.01876	0.01831
2.1	0.01786	0.01743	0.01700	0.01659	0.01618	0.01578	0.01539	0.01500	0.01463	0.01426
2.2	0.01390	0.01355	0.01321	0.01287	0.01255	0.01222	0.01191	0.01160	0.01130	0.01101
2.3	0.01072	0.01044	0.01017	0.00990	0.00964	0.00939	0.00914	0.00889	0.00866	0.00842
2.4	0.00820	0.00798	0.00776	0.00755	0.00734	0.00714	0.00695	0.00676	0.00657	0.00639
2.5	0.00621	0.00604	0.00587	0.00570	0.00554	0.00539	0.00523	0.00508	0.00494	0.00480
2.6	0.00466	0.00453	0.00440	0.00427	0.00415	0.00402	0.00391	0.00379	0.00368	0.00357
2.7	0.00347	0.00336	0.00326	0.00317	0.00307	0.00298	0.00289	0.00280	0.00272	0.00264
2.8	0.00256	0.00248	0.00240	0.00233	0.00226	0.00219	0.00212	0.00205	0.00199	0.00193
2.9	0.00187	0.00181	0.00175	0.00169	0.00164	0.00159	0.00154	0.00149	0.00144	0.00139
3.0	0.00135	0.00131	0.00126	0.00122	0.00118	0.00114	0.00111	0.00107	0.00104	0.00100

jlborges@fe.up.pt

Cálculo de probabilidades para uma distribuição Normal qualquer



z-score computations: | Area to the left of *z*:

$$x = 115 \longrightarrow z = \frac{115 - 100}{16} = 0.94 \qquad 0.8264$$

$$x = 140 \longrightarrow z = \frac{140 - 100}{16} = 2.50 \qquad 0.9938$$

Shaded area = 0.9938 − 0.8264 = 0.1674

Verifica-se que uma **combinação linear de variáveis Normais independentes** segue uma distribuição **Normal**

Exemplo

Uma empresa desenvolve um conjunto restrito de actividades A_n (n = 1, ... , n). Admite-se que os lucros L_n associados às diferentes actividades A_n são variáveis independentes que seguem distribuições

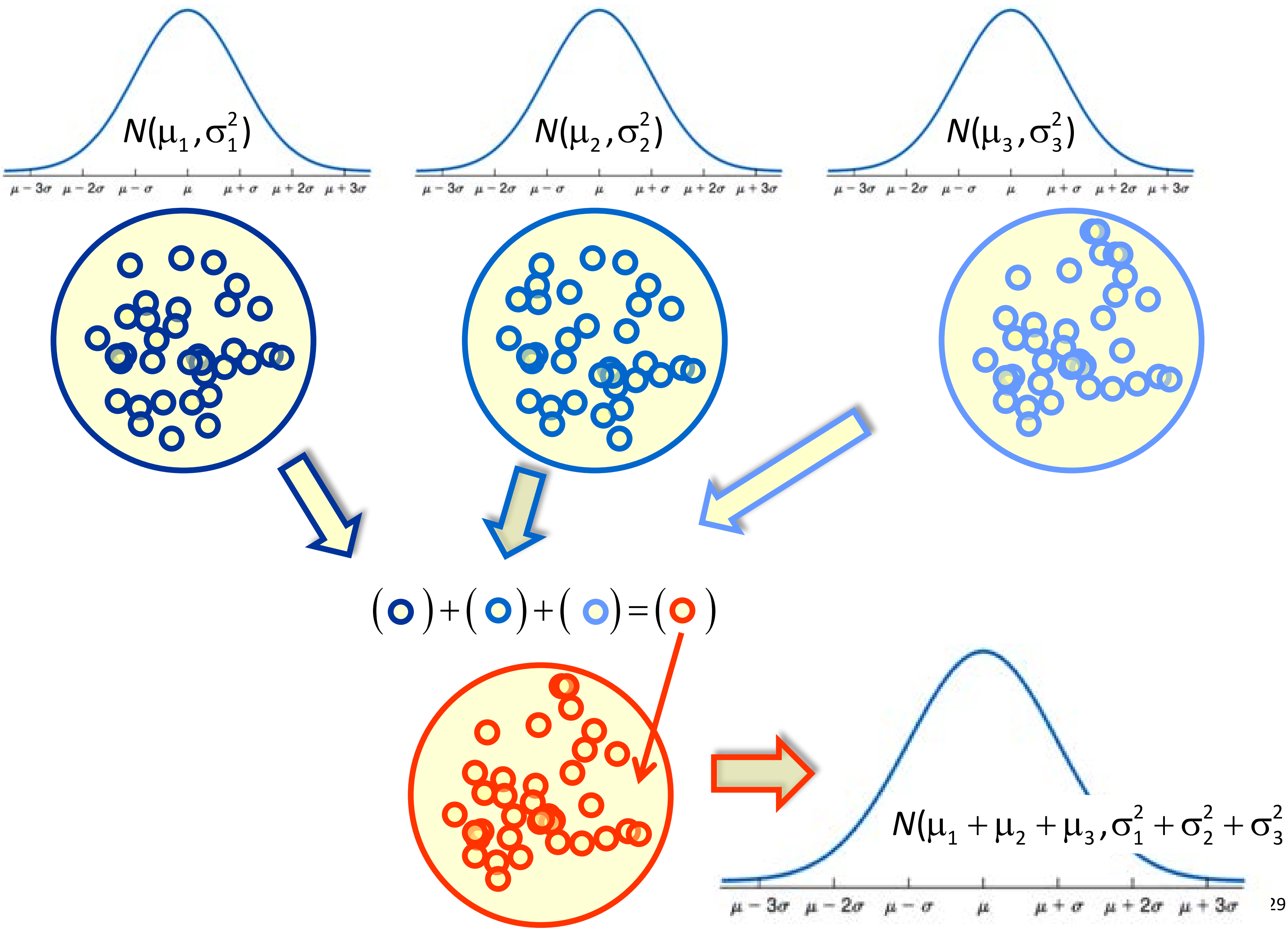
$$N(\mu_n, \sigma_n^2)$$

Então, lucro global da empresa, $L = \sum_{n=1}^{N} L_n$

segue uma distribuição normal, cujos parâmetros são dados por

$$\mu = \sum_{n=1}^{N} \mu_n \qquad \sigma^2 = \sum_{n=1}^{N} \sigma_n^2$$

jlborges@fe.up.pt



Uma distribuição

B(n,p)

pode ser aproximada razoavelmente pela distribuição

N(N·p,N·p·(1-p))

quando

N≥20 , **N·p>7** e **N·(1-p) >7**

A distribuição **Normal** poderá ainda ser utilizada para aproximar as distribuições **Hipergeométrica** e de **Poisson** sempre que estas, por sua vez, sejam **aproximáveis** por distribuições **Binomiais** (aproximadamente simétricas).

Importância da Distribuição Normal

- Distribuição Normal **muito frequente**
- Representa variáveis que são afetadas por muitos efeitos pequenos e independentes
  - ***soma de muitos efeitos em que nenhum é predominante***
- Muitas distribuições tendem para a distribuição Normal dentro de certas condições

Normal Probability Plots

One useful tool for assessing normality is the **normal quantile plot** (or normal probability plots)

How are they built?

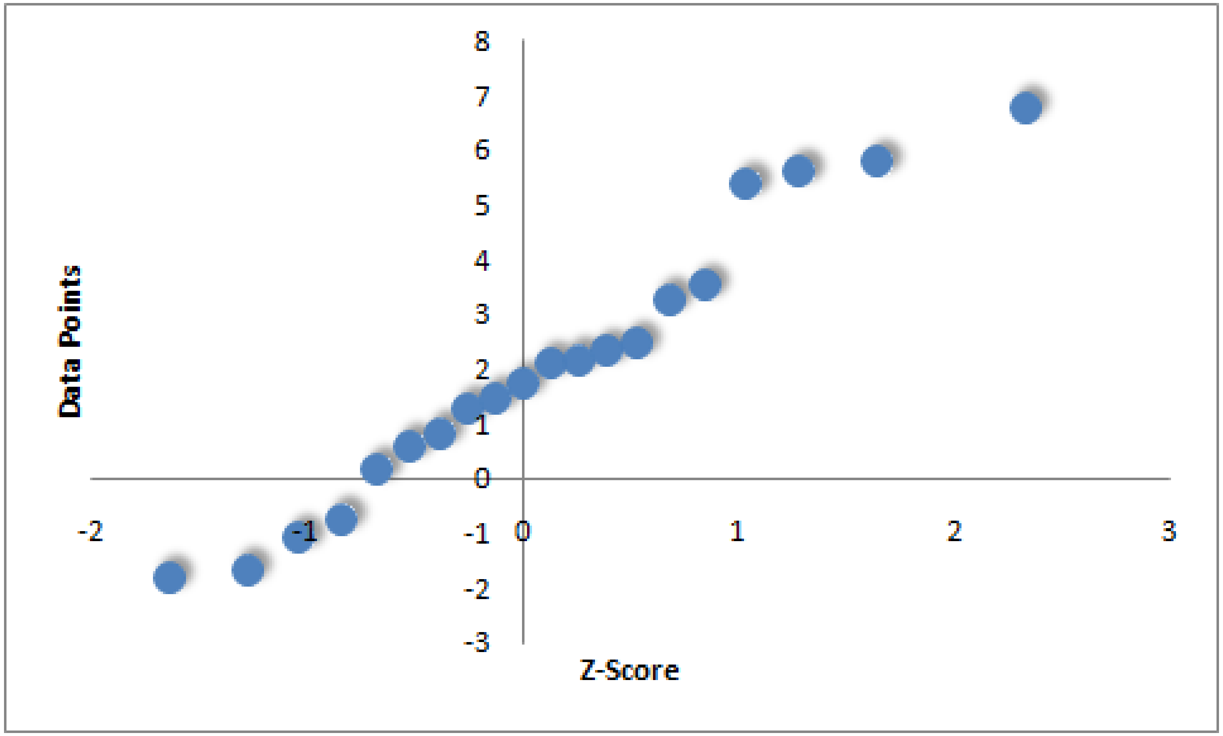
1. Arrange the data values from the smallest to the largest and record what percentile of the data each value occupies
2. Find the Z-score at the same percentiles
3. Plot each original data value against the corresponding Z-score.

A normal distribution will produce a straight line. Note that standardizing is a linear transformation that can change the slope and intercept of the line in the plot but cannot turn a line into a curved pattern

jlborges@fe.up.pt

n	N(3,2.2)	Percentil	Z-score
1	-1.804	0.05	-1.645
2	-1.659	0.1	-1.282
3	-1.063	0.15	-1.036
4	-0.719	0.2	-0.842
5	0.189	0.25	-0.674
6	0.609	0.3	-0.524
7	0.849	0.35	-0.385
8	1.298	0.4	-0.253
9	1.482	0.45	-0.126
10	1.751	0.5	0.000
11	2.111	0.55	0.126
12	2.196	0.6	0.253
13	2.339	0.65	0.385
14	2.485	0.7	0.524
15	3.297	0.75	0.674
16	3.537	0.8	0.842
17	5.409	0.85	1.036
18	5.636	0.9	1.282
19	5.808	0.95	1.645
20	6.813	1	2.326

amostra recolhida de uma população Normal(3, 2.2)

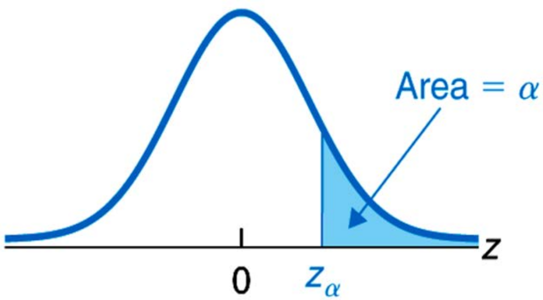


Método visual para verificar se uma variável segue uma distribuição aproximadamente Normal

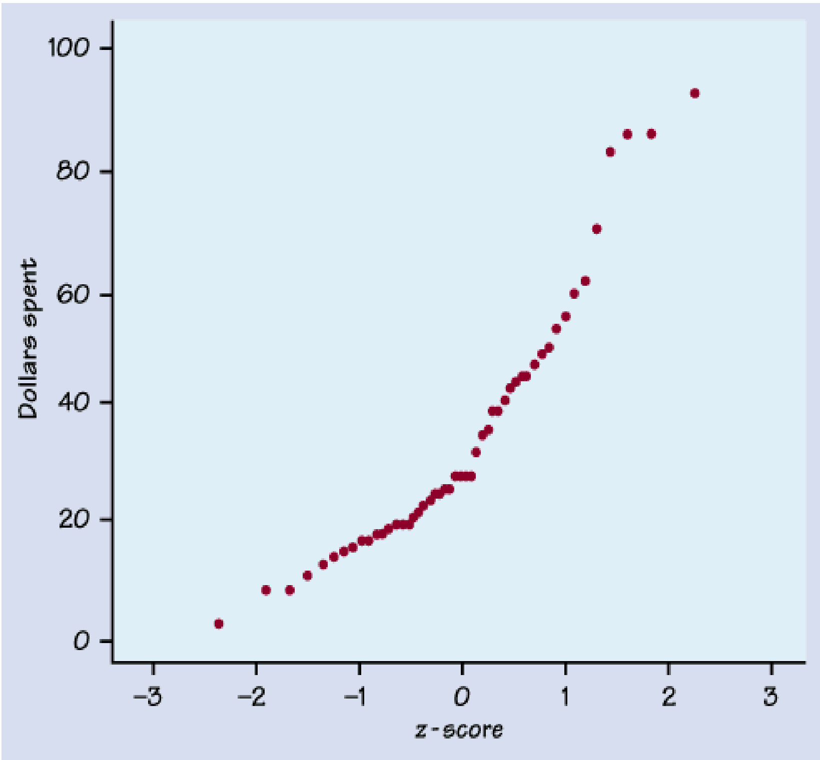
Se for Normal os pontos ficam alinhados numa reta

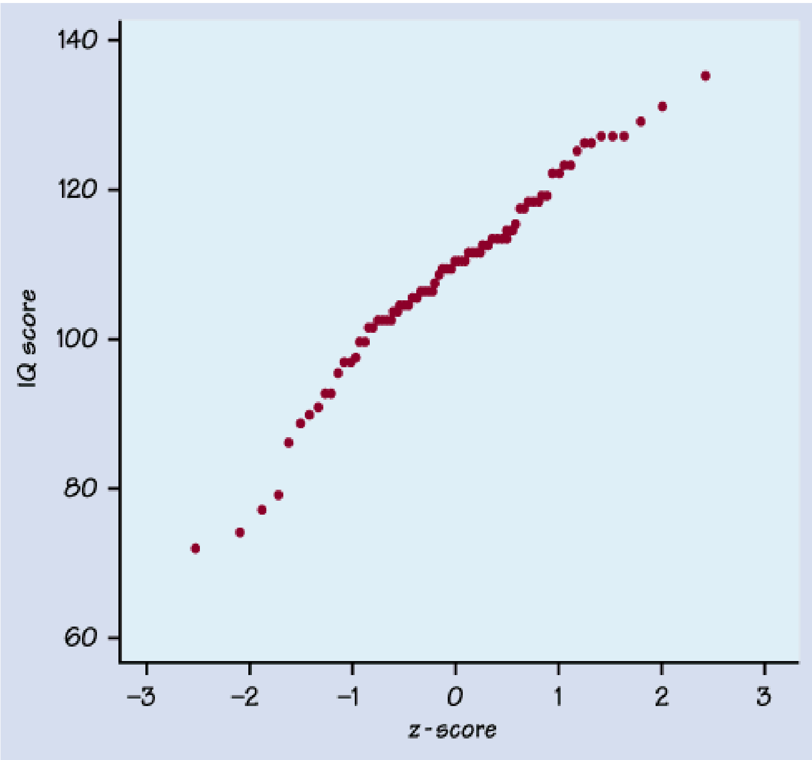
jlborges@fe.up.pt

TABELA 3 do livro:
Probabilidades associadas à cauda direita da distribuição normal padronizada



Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.50000	0.49601	0.49202	0.48803	0.48405	0.48006	0.47608	0.47210	0.46812	0.46414
0.1	0.46017	0.45620	0.45224	0.44828	0.44433	0.44038	0.43644	0.43251	0.42858	0.42465
0.2	0.42074	0.41683	0.41294	0.40905	0.40517	0.40129	0.39743	0.39358	0.38974	0.38591
0.3	0.38209	0.37828	0.37448	0.37070	0.36693	0.36317	0.35942	0.35569	0.35197	0.34827
0.4	0.34458	0.34090	0.33724	0.33360	0.32997	0.32636	0.32276	0.31918	0.31561	0.31207
0.5	0.30854	0.30503	0.30153	0.29806	0.29460	0.29116	0.28774	0.28434	0.28096	0.27760
0.6	0.27425	0.27093	0.26763	0.26435	0.26109	0.25785	0.25463	0.25143	0.24825	0.24510
0.7	0.24196	0.23885	0.23576	0.23270	0.22965	0.22663	0.22363	0.22065	0.21770	0.21476
0.8	0.21186	0.20897	0.20611	0.20327	0.20045	0.19766	0.19489	0.19215	0.18943	0.18673
0.9	0.18406	0.18141	0.17879	0.17619	0.17361	0.17106	0.16853	0.16602	0.16354	0.16109
1.0	0.15866	0.15625	0.15386	0.15151	0.14917	0.14686	0.14457	0.14231	0.14007	0.13786
1.1	0.13567	0.13350	0.13136	0.12924	0.12714	0.12507	0.12302	0.12100	0.11900	0.11702
1.2	0.11507	0.11314	0.11123	0.10935	0.10749	0.10565	0.10383	0.10204	0.10027	0.09853
1.3	0.09680	0.09510	0.09342	0.09176	0.09012	0.08851	0.08691	0.08534	0.08379	0.08226
1.4	0.08076	0.07927	0.07780	0.07636	0.07493	0.07353	0.07215	0.07078	0.06944	0.06811
1.5	0.06681	0.06552	0.06426	0.06301	0.06178	0.06057	0.05938	0.05821	0.05705	0.05592
1.6	0.05480	0.05370	0.05262	0.05155	0.05050	0.04947	0.04846	0.04746	0.04648	0.04551
1.7	0.04457	0.04363	0.04272	0.04182	0.04093	0.04006	0.03920	0.03836	0.03754	0.03673
1.8	0.03593	0.03515	0.03438	0.03362	0.03288	0.03216	0.03144	0.03074	0.03005	0.02938
1.9	0.02872	0.02807	0.02743	0.02680	0.02619	0.02559	0.02500	0.02442	0.02385	0.02330
2.0	0.02275	0.02222	0.02169	0.02118	0.02068	0.02018	0.01970	0.01923	0.01876	0.01831
2.1	0.01786	0.01743	0.01700	0.01659	0.01618	0.01578	0.01539	0.01500	0.01463	0.01426
2.2	0.01390	0.01355	0.01321	0.01287	0.01255	0.01222	0.01191	0.01160	0.01130	0.01101
2.3	0.01072	0.01044	0.01017	0.00990	0.00964	0.00939	0.00914	0.00889	0.00866	0.00842
2.4	0.00820	0.00798	0.00776	0.00755	0.00734	0.00714	0.00695	0.00676	0.00657	0.00639
2.5	0.00621	0.00604	0.00587	0.00570	0.00554	0.00539	0.00523	0.00508	0.00494	0.00480
2.6	0.00466	0.00453	0.00440	0.00427	0.00415	0.00402	0.00391	0.00379	0.00368	0.00357
2.7	0.00347	0.00336	0.00326	0.00317	0.00307	0.00298	0.00289	0.00280	0.00272	0.00264
2.8	0.00256	0.00248	0.00240	0.00233	0.00226	0.00219	0.00212	0.00205	0.00199	0.00193
2.9	0.00187	0.00181	0.00175	0.00169	0.00164	0.00159	0.00154	0.00149	0.00144	0.00139
3.0	0.00135	0.00131	0.00126	0.00122	0.00118	0.00114	0.00111	0.00107	0.00104	0.00100





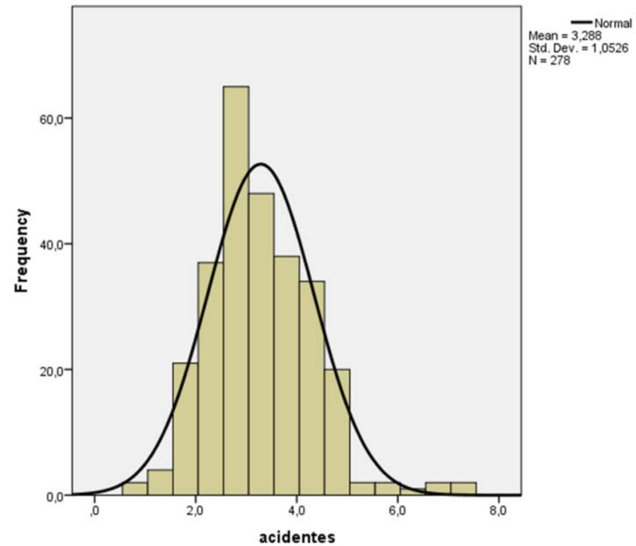
Amount spent in a supermarket by a group of people. The patterns bends up at the right, showing right skewness

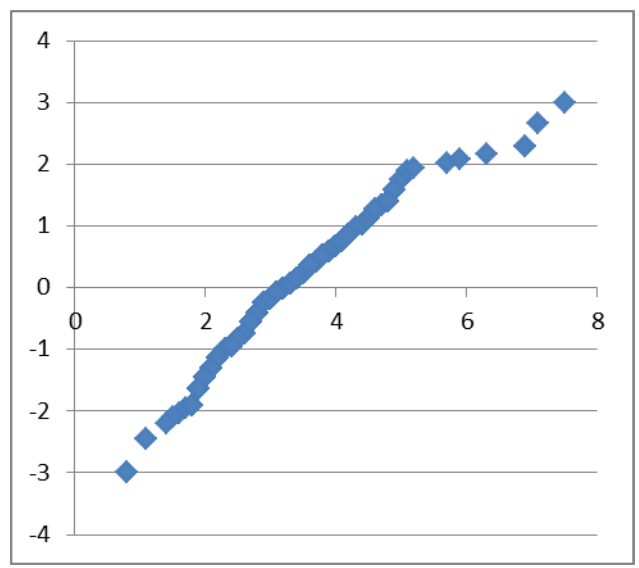
IQ scores of 78 seventh-grade students. The straight line shows that this distribution is close to normal

Acidentes de viação por município com vítimas por mil habitantes (2011)

www.pordata.pt

	A	B	C	D
1				
2	Acidentes com vítimas por mil hab.		2011	
3	município	acidentes	percentil	zscore
4	Arcos de Valdevez	2.6	0.231	-0.73556
5	Caminha	1.9	0.05	-1.64485
6	Melgaço	2.2	0.126	-1.14551
7	Monção	3.4	0.555	0.138304
8	Paredes de Coura	2.8	0.342	-0.40701
9	Ponte da Barca	2	0.072	-1.46106
10	Ponte de Lima	3.1	0.465	-0.08784
11	Valença	4.5	0.87	1.126391
12	Viana do Castelo	3.1	0.465	-0.08784
13	Vila Nova de Cerveira	4.5	0.87	1.126391
14	Amares	4.2	0.801	0.845199
15	Barcelos	2.9	0.404	-0.24301
16	Braga	2.8	0.342	-0.40701
17	Esposende	4.6	0.898	1.270238
18	Terras de Bouro	1.8	0.028	-1.91104
19	Vila Verde	3.5	0.592	0.232693
20	Fafe	3.4	0.555	0.138304
21	Guimarães	3.2	0.49	-0.02507
22	Póvoa de Lanhoso	4.9	0.942	1.571787
23	Santo Tirso	3.2	0.49	-0.02507
24	Trofa	3.3	0.53	0.07527
25	Vieira do Minho	2.5	0.202	-0.8345
26	Vila Nova de Famalicão	3.7	0.664	0.423405
278	Silves	4.2	0.801	0.845199
279	Tavira	3.6	0.638	0.353118
280	Vila do Bispo	5.9	0.981	2.074855
281	Vila R. Santo António	3.1	0.465	-0.08784



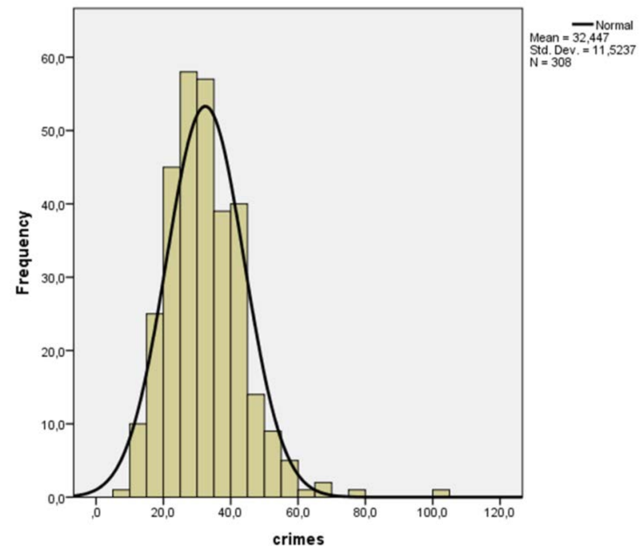


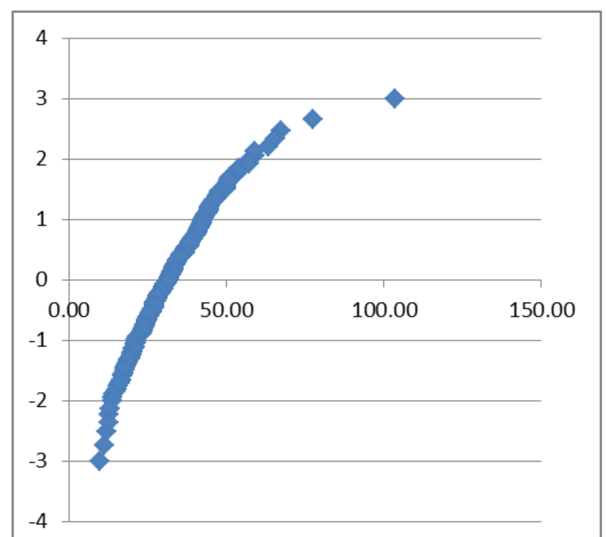
jlborges@fe.up.pt

Crimes registados por município por mil habitantes (2011)

www.pordata.pt

	A	B	C	D
1	Crimes registados por mil habitantes		**2011**	
2	**município**	**crimes**	**percentil**	**zscore**
3	Arcos de Valdevez	34.60	0.628	0.326561
4	Caminha	47.50	0.921	1.41183
5	Melgaço	24.70	0.25	-0.67449
6	Monção	29.30	0.42	-0.201893
7	Paredes de Coura	36.70	0.674	0.450985
8	Ponte da Barca	45.30	0.895	1.253565
9	Ponte de Lima	33.40	0.583	0.209574
10	Valença	58.80	0.98	2.053749
11	Viana do Castelo	31.80	0.511	0.027576
12	Vila Nova de Cerveira	33.60	0.592	0.232693
13	Amares	46.00	0.902	1.293032
14	Barcelos	23.00	0.185	-0.896473
15	Braga	37.20	0.7	0.524401
16	Esposende	34.00	0.599	0.25076
17	Terras de Bouro	47.10	0.918	1.391744
18	Vila Verde	34.10	0.605	0.266311
19	Fafe	27.10	0.342	-0.407011
301	Câmara de Lobos	26.30	0.299	-0.527279
302	Funchal	31.30	0.491	-0.022562
303	Machico	35.80	0.664	0.423405
304	Ponta do Sol	17.80	0.071	-1.468384
305	Porto Moniz	27.10	0.342	-0.407011
306	Ribeira Brava	35.40	0.651	0.388022
307	Santa Cruz	19.40	0.1	-1.281552
308	Santana	20.80	0.13	-1.126391
309	São Vicente	25.80	0.289	-0.556308
310	Porto Santo	24.00	0.224	-0.758754
311				



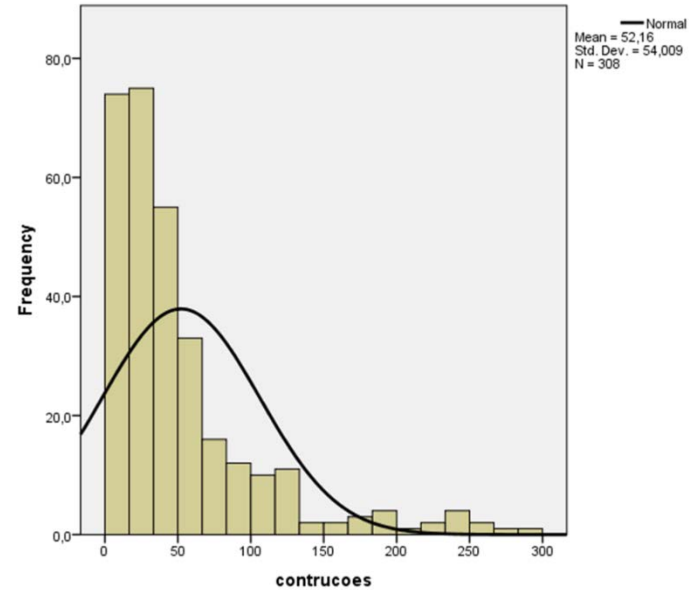


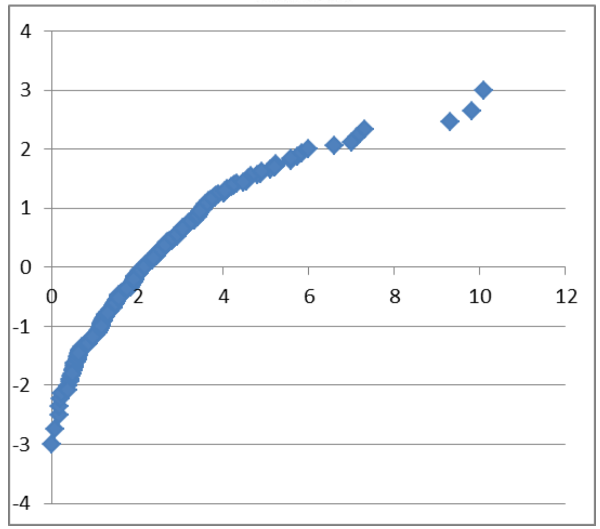
jlborges@fe.up.pt

Edifícios licenciados por município - Construções novas (2011)

www.pordata.pt

	A	B	C	D	E	F
1						
2	Edifícios licenciados - Construções novas		2011			
3	**município**	**construções**	**Pop. residente**	**rácio**	**percentil**	**zscore**
4	**Arcos de Valdevez**	59	22741	2.594432962	0.622	0.310738
5	**Caminha**	59	16643	3.545033948	0.84	0.994458
6	**Melgaço**	11	9172	1.199302224	0.175	-0.93459
7	**Monção**	64	19158	3.340640985	0.788	0.799501
8	**Paredes de Coura**	29	9175	3.160762943	0.758	0.699884
9	**Ponte da Barca**	25	12031	2.077965256	0.475	-0.06271
10	**Ponte de Lima**	156	43423	3.592566152	0.856	1.062519
11	**Valença**	43	14080	3.053977273	0.736	0.631062
12	**Viana do Castelo**	144	88642	1.624512082	0.328	-0.44544
13	**Vila Nova de Cerveira**	32	9234	3.465453758	0.82	0.915365
14	**Amares**	54	18851	2.864569519	0.69	0.49585
15	**Barcelos**	276	120339	2.293520804	0.54	0.100434
16	**Braga**	264	181600	1.453744493	0.257	-0.65262
17	**Esposende**	122	34268	3.560172756	0.846	1.019428
18	**Terras de Bouro**	23	7218	3.186478249	0.762	0.712751
19	**Vila Verde**	118	47839	2.466606743	0.579	0.199336
20	**Fafe**	170	50569	3.361743361	0.798	0.834499
21	**Guimarães**	242	158080	1.530870445	0.289	-0.55631
306	**Porto Moniz**	1	2698	0.370644922	0.019	-2.07485
307	**Ribeira Brava**	21	13329	1.575512041	0.306	-0.50722
308	**Santa Cruz**	50	43005	1.162655505	0.159	-0.99858
309	**Santana**	12	7683	1.561889887	0.299	-0.52728
310	**São Vicente**	11	5693	1.932197435	0.413	-0.21983
311	**Porto Santo**	19	5468	3.474762253	0.824	0.930717



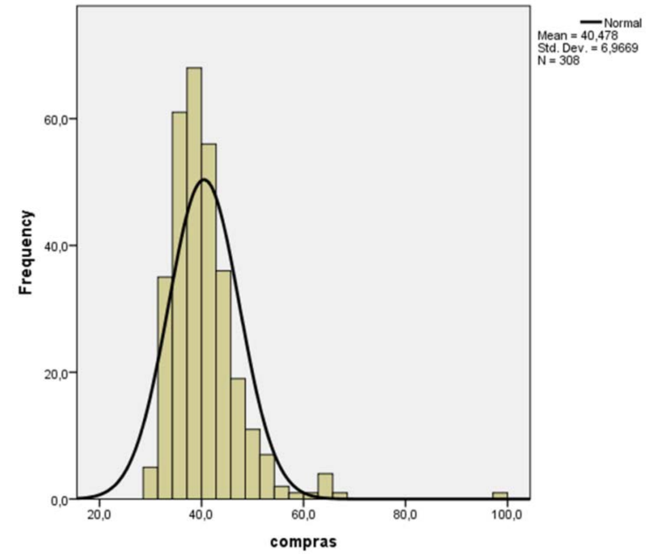


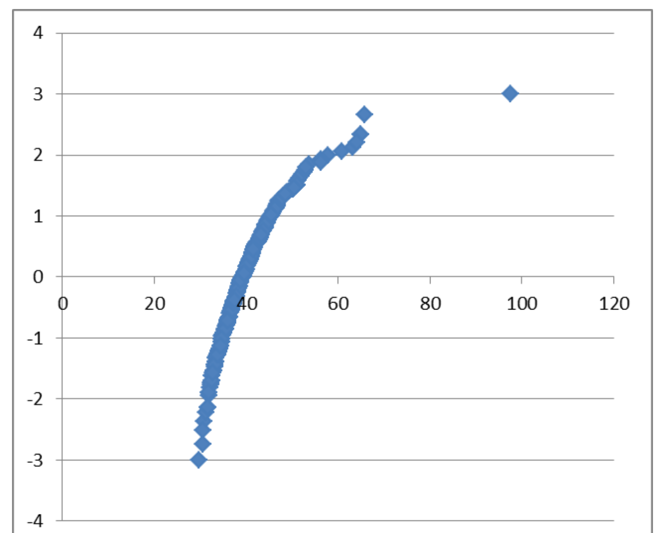
jlborges@fe.up.pt

Valor médio por município das compras efetuadas em terminais de pagamento automático (2011)

www.pordata.pt

	A	B	C	D
1				
2	Valor médio das compras efectuadas em terminais de pagamento automático		2011	
3	**município**	**compras**	**percentil**	**zscore**
4	**Arcos de Valdevez**	42.8	0.729	0.60979
5	**Caminha**	40.6	0.592	0.23269
6	**Melgaço**	53.7	0.967	1.83842
7	**Monção**	46	0.859	1.07584
8	**Paredes de Coura**	39.2	0.495	-0.01253
9	**Ponte da Barca**	41.2	0.631	0.3345
10	**Ponte de Lima**	44.9	0.827	0.94238
11	**Valença**	49.3	0.921	1.41183
12	**Viana do Castelo**	44.1	0.788	0.7995
13	**Vila Nova de Cerveira**	40.1	0.55	0.12566
14	**Amares**	41.5	0.657	0.40429
15	**Barcelos**	43.3	0.755	0.69031
16	**Braga**	43.1	0.742	0.64952
17	**Esposende**	38.6	0.456	-0.11052
18	**Terras de Bouro**	51.6	0.947	1.61644
19	**Vila Verde**	42.6	0.719	0.57987
20	**Fafe**	39.1	0.488	-0.03008
21	**Guimarães**	41.1	0.628	0.32656
22	**Póvoa de Lanhoso**	44.9	0.827	0.94238
23	**Santo Tirso**	35.8	0.221	-0.76882
24	**Trofa**	46.8	0.882	1.18504
304	**Machico**	38.9	0.475	-0.06271
305	**Ponta do Sol**	50.5	0.928	1.46106
306	**Porto Moniz**	47.8	0.902	1.29303
307	**Ribeira Brava**	39.6	0.524	0.0602
308	**Santa Cruz**	37.5	0.348	-0.39073
309	**Santana**	45.6	0.846	1.01943
310	**São Vicente**	48.6	0.912	1.35317
311	**Porto Santo**	43.3	0.755	0.69031
312				



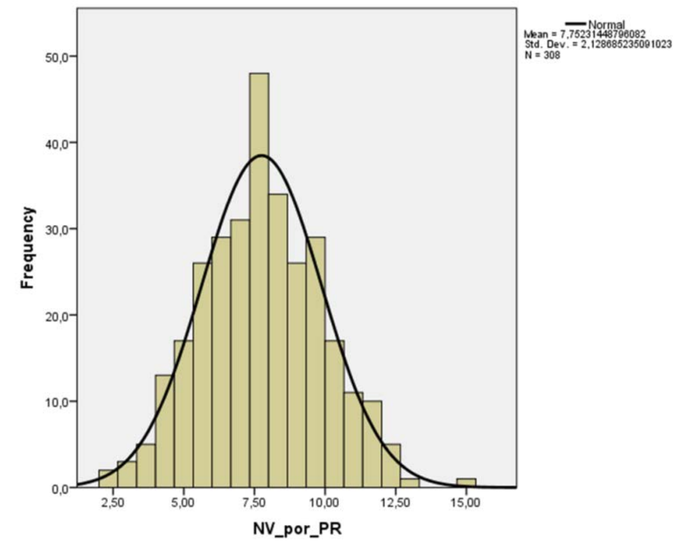


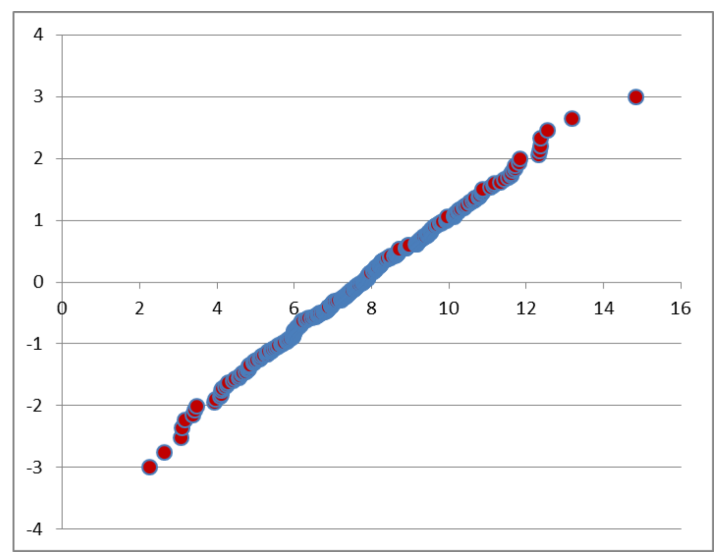
jlborges@fe.up.pt

Número de nados vivos por mil habitantes nos municípios de portugal (2011)

www.pordata.pt

	A	B	C	D	E	F
1	Número de nados vivos por residente nos municípios de portugal.			2011		
2	**Município**	**Nados_vivos**	**Pop_resid**	**NV_por_PR**	**percentile**	**z-score**
3	Oleiros	15	5692	2.635277583	0.003	-2.747781385
4	Góis	13	4233	3.071107961	0.006	-2.512144328
5	Arcos de Valdevez	139	22741	6.112308166	0.237	-0.71598599
6	Figueiró dos Vinhos	19	6155	3.086921202	0.009	-2.365618127
7	Penamacor	18	5659	3.180773988	0.013	-2.226211769
8	Pampilhosa da Serra	15	4462	3.361721201	0.016	-2.144410621
9	Vinhais	31	9025	3.434903047	0.019	-2.074854734
10	Boticas	20	5727	[illegible]	0.022	2.014090812
303	Odivelas	1785	144705	12.33544107	0.98	2.053748911
304	Alcochete	217	17573	12.34848916	0.983	2.12007169
305	Mafra	949	76748	12.36514307	0.986	2.197286377
306	São Roque do Pico	42	3395	12.37113402	0.99	2.326347874
307	Vila Franca do Campo	141	11235	12.55006676	0.993	2.45726339
308	Montijo	676	51251	13.18998654	0.996	2.652069808
309	Arronches	7	3096	2.260981912	0	-3
310	Ribeira Grande	477	32151	14.83624149	1	3



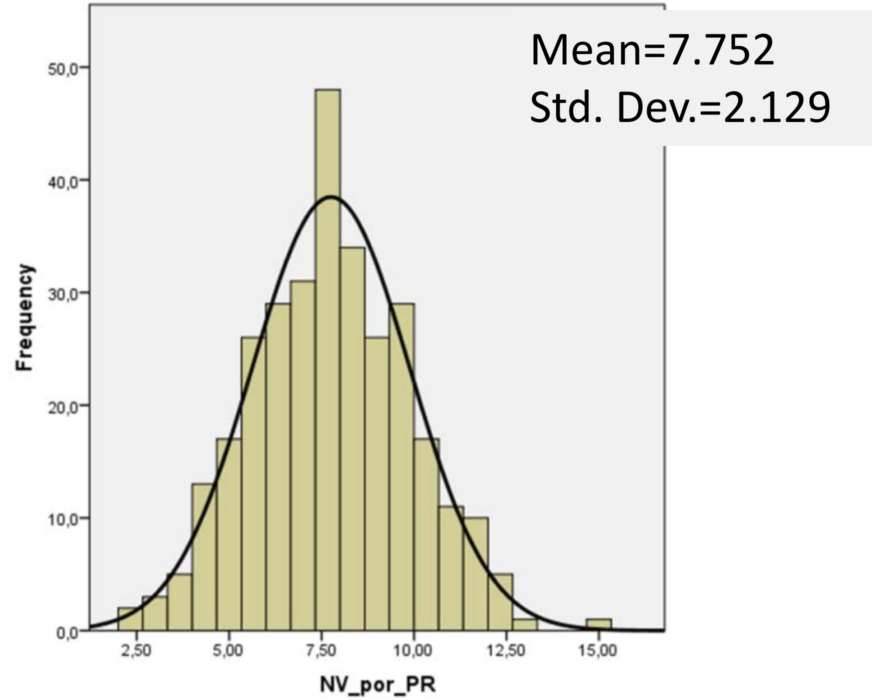


jlborges@fe.up.pt

Número de nados vivos por mil habitantes nos municípios de portugal (2011)

www.pordata.pt

	A	B	C	D	E	F
1	Número de nados vivos por residente nos municípios de portugal.			2011		
2	**Município**	**Nados_vivos**	**Pop_resid**	**NV_por_PR**	**percentile**	**z-score**
3	Oleiros	15	5692	2.635277583	0.003	-2.747781385
4	Góis	13	4233	3.071107961	0.006	-2.512144328
5	Arcos de Valdevez	139	22741	6.112308166	0.237	-0.71598599
6	Figueiró dos Vinhos	19	6155	3.086921202	0.009	-2.365618127
7	Penamacor	18	5659	3.180773988	0.013	-2.226211769
8	Pampilhosa da Serra	15	4462	3.361721201	0.016	-2.144410621
9	Vinhais	31	9025	3.434903047	0.019	-2.074854734
10	Boticas	20	[illegible]	[illegible]	[illegible]	[illegible]
303	Odivelas	1785	144705	12.33544107	0.98	2.053748911
304	Alcochete	217	17573	12.34848916	0.983	2.12007169
305	Mafra	949	76748	12.36514307	0.986	2.197286377
306	São Roque do Pico	42	3395	12.37113402	0.99	2.326347874
307	Vila Franca do Campo	141	11235	12.55006676	0.993	2.45726339
308	Montijo	676	51251	13.18998654	0.996	2.652069808
309	Arronches	7	3096	2.260981912	0	-3
310	Ribeira Grande	477	32151	14.83624149	1	3



Podemos admitir que a variável segue uma distribuição Normal.

$$X \rightarrow N(\mu = 7.752, \sigma^2 = 2.129\ ^2)$$

Qual a probabilidade de um município selecionado ao acaso ter mais de 9 nascimentos por mil habitante?

Número de nados vivos por mil habitantes nos municípios de portugal (2011)
www.pordata.pt

Qual a probabilidade de um município selecionado ao acaso ter mais de 9 nascimentos por mil hab?

Assumindo $X \rightarrow N(\mu = 7.752, \sigma^2 = 2.129\,^2)$

$$P(X > 9) = P\left(Z > \frac{9 - 7.75}{2.129} = 0.587\right) = 27.9\%$$

Usando a definição frequencista de probabilidade

$$P(X > 9) = \frac{num.\,casos > 9}{num.\,casos} = \frac{84}{308} = 27.3\%$$

Número de nados vivos por mil habitantes nos municípios de portugal (2011)
www.pordata.pt

Qual a probabilidade de um município selecionado ao acaso ter entre 5 e 10 nascimentos...?

Assumindo $X \to N(\mu = 7.752, \sigma^2 = 2.129\ ^2)$

$$P(5 < X < 10) = P(X > 5) - P(X > 10) =$$

$$P\left(Z > \frac{5 - 7.752}{2.129} = -1.29\right) - P\left(Z > \frac{10 - 7.752}{2.129} = 1.06\right) = 0.902 - 0.145 = 0.757$$

Usando a definição frequencista de probabilidade

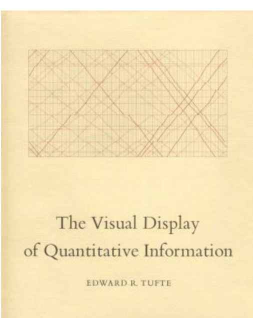
$$P(5 < X < 10) = P(X > 5) - P(X > 10) =$$

$$= \frac{num.casos\ \ > 5}{num.casos} - \frac{num.casos\ \ > 10}{num.casos} = \frac{274}{308} - \frac{45}{308} = 75.0\%$$

Visualization is an essential step in data analysis

Anscombe’s quartet

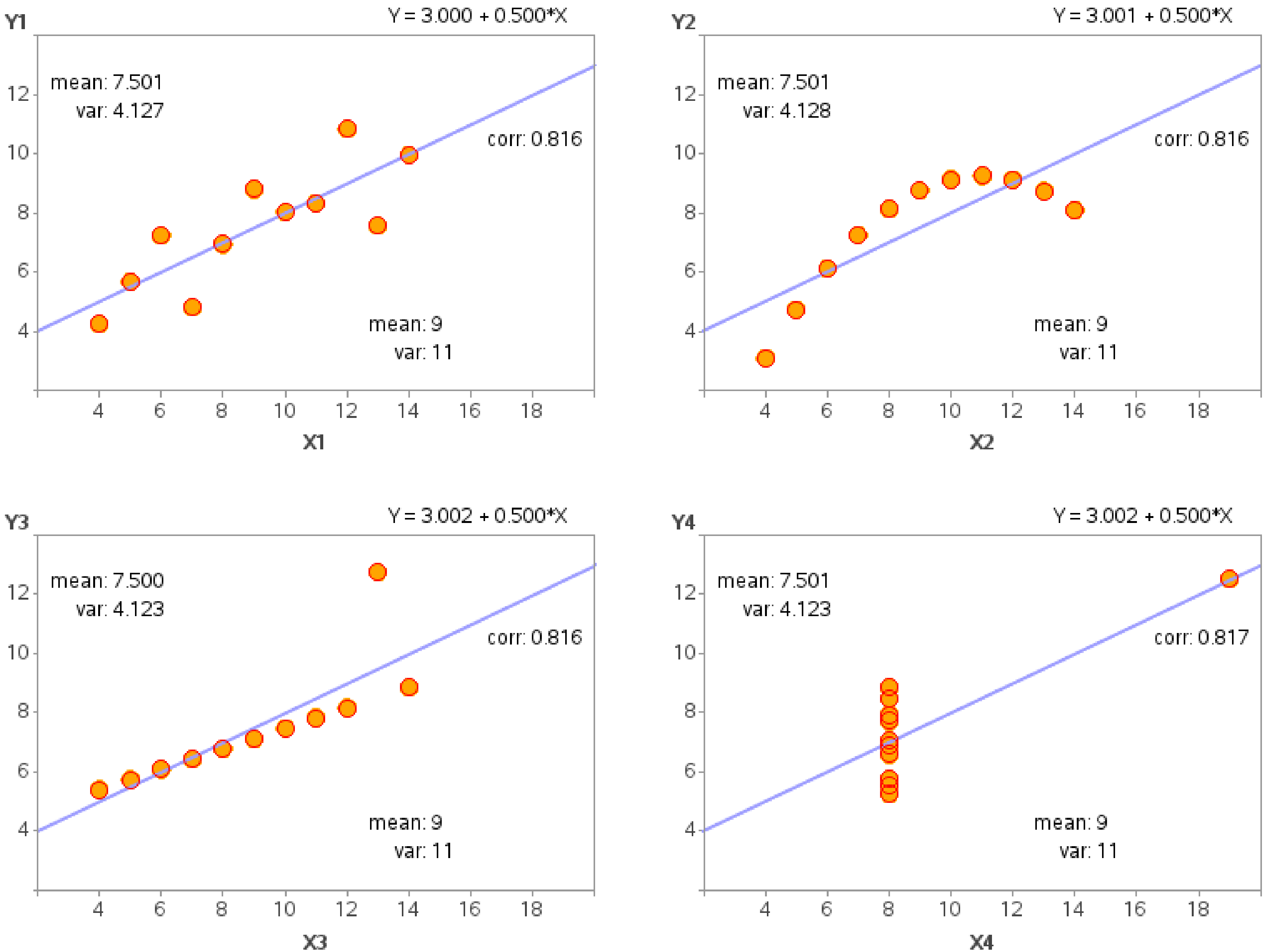
1		2		3		4	
X	**Y**	**X**	**Y**	**X**	**Y**	**X**	**Y**
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89



N	11
Mean of X	9.0
Mean of Y	7.5
Regression	y = 3 + 0.5x
Correlation coefficient (r)	0.82
Level of Explanation (r2)	0.67

The four datasets have the same statistics

F.J. Anscombe, “Graphs in Statistical Analysis” American Statistician, 27, pp 17-21, February 1973


Y1
Y = 3.000 + 0.500*X
mean: 7.501
var: 4.127
corr: 0.816
mean: 9
var: 11
X1
Y2
Y = 3.001 + 0.500*X
mean: 7.501
var: 4.128
corr: 0.816
mean: 9
var: 11
X2
Y3
Y = 3.002 + 0.500*X
mean: 7.500
var: 4.123
corr: 0.816
mean: 9
var: 11
X3
Y4
Y = 3.002 + 0.500*X
mean: 7.501
var: 4.123
corr: 0.817
mean: 9
var: 11
X4

Edward J. Tufte, has coined the memorable term ***chartjunk*** to refer to all of the extraneous elements that convey no information and yet litter many contemporary charts and graphs.

This one reports the results of an online poll of "self-selected respondents" who were apparently asked, "How much will you or do you owe in student loans?"

The graph in Figure 6 contains bars, represented by stylized greenbacks; the bills seem to be sticking out from a graduation cap. But there is no way of telling where the bars in the graph begin.

We can tell that 18 percent is greater than 13 percent, which is in turn greater than 11 percent but then we already knew that.

What we don't get is any clear, visual sense of the relative proportions of these quantities, because some unknown part of each bar is hidden from us, and the uneven, peaked contours of the cap suggest that the obscured proportion probably differs from bar to bar.

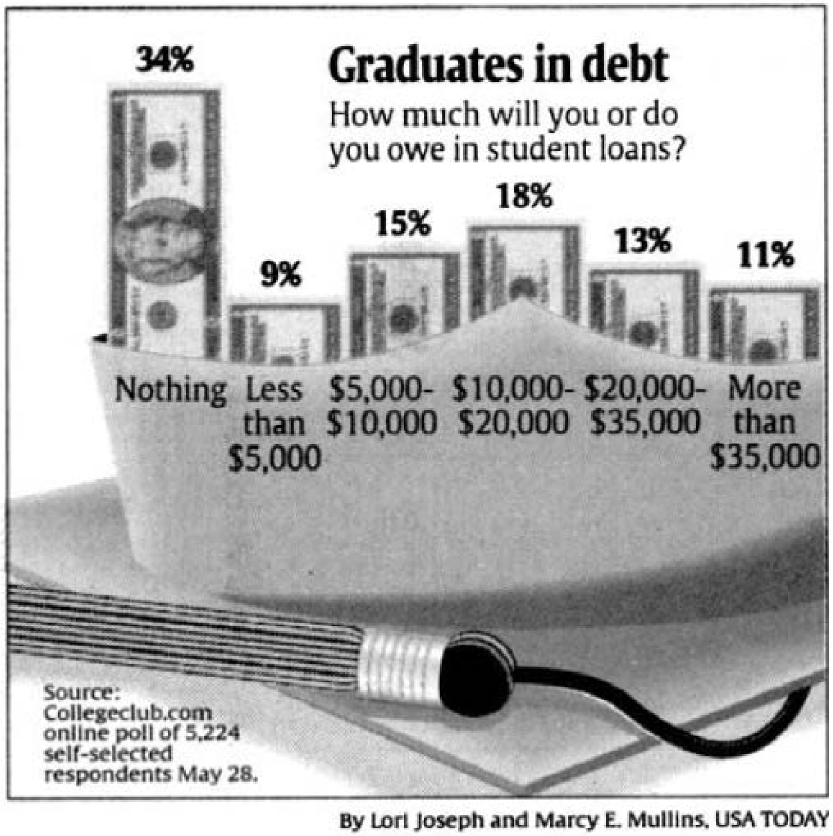
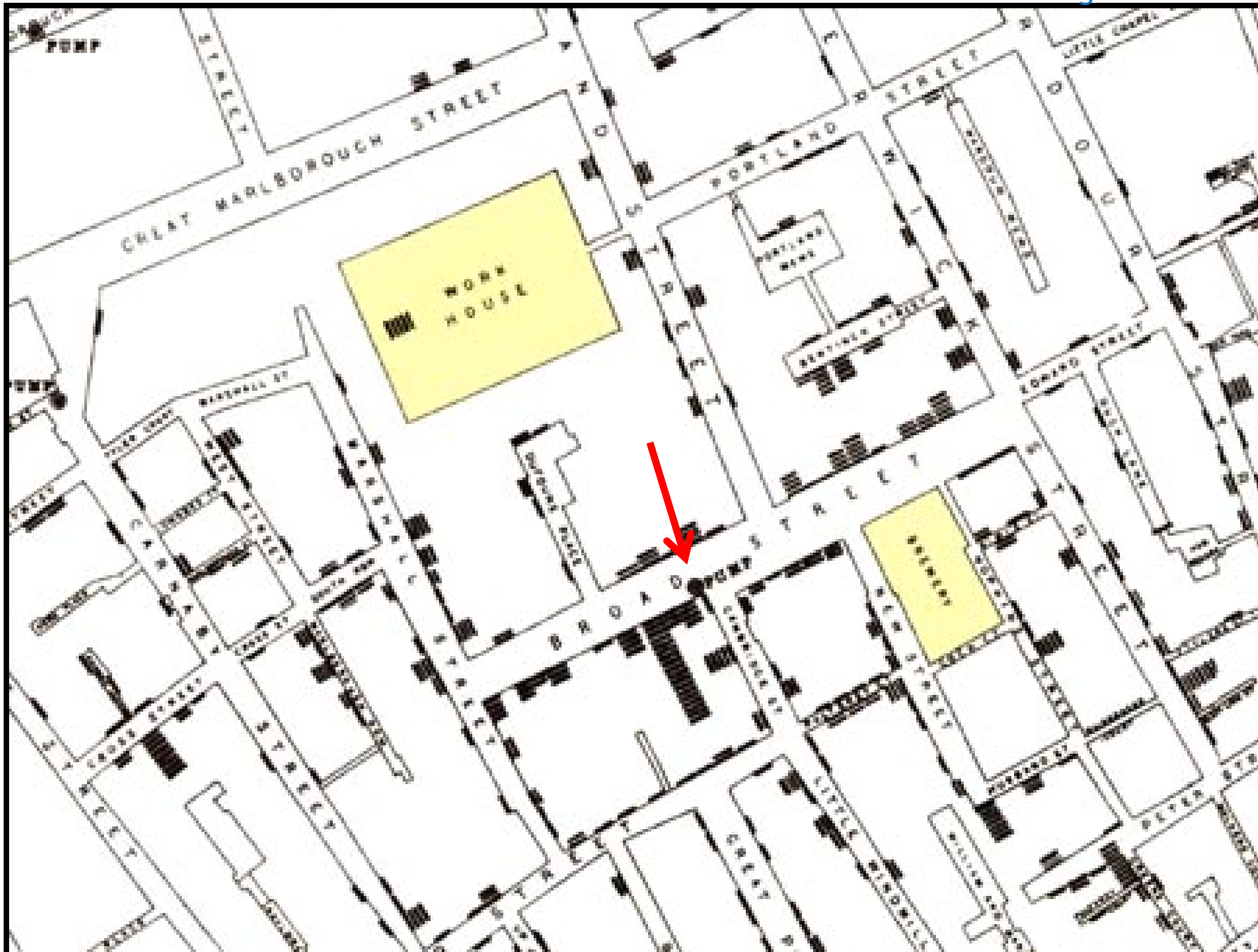


FIGURE 6. The meaning of this graphic is obscured by chartjunk. (*Source: USA Today,* August 6, 2002, p. 1A; © USA Today, reprinted by permission.)

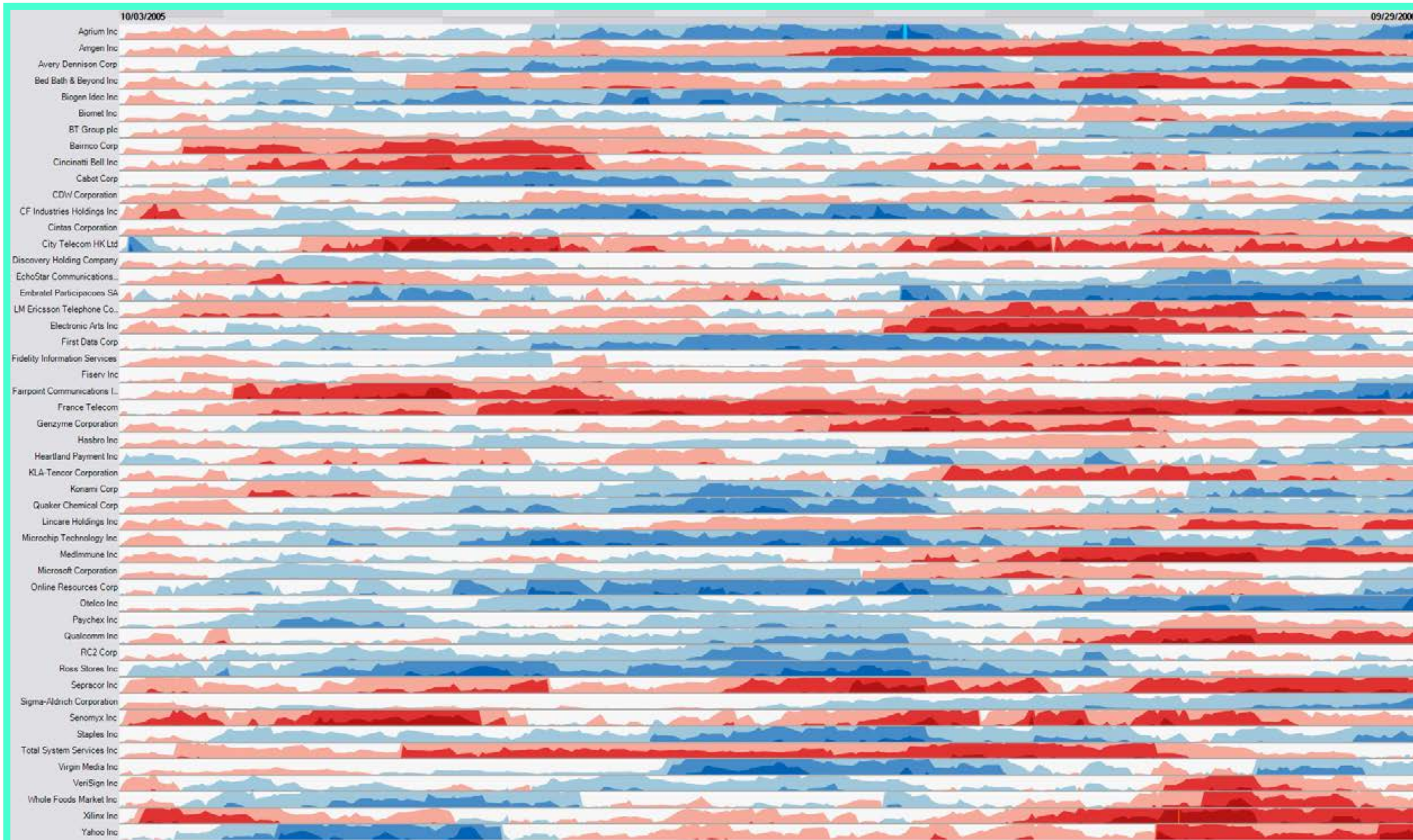

Russia
Sakhalin
La Perouse Strait
China
Vladivostok
Hokkaido
Japan Basin
Tsugaru Strait
North Korea
P'yongyang
Sea of Japan
(East Sea)
Yamato Basin
Seoul
South Korea
Tsushima Basin
(Ulleung Basin)
Japan
Honshu
Tokyo
Pacific Ocean
Shikkoku
Kyushu



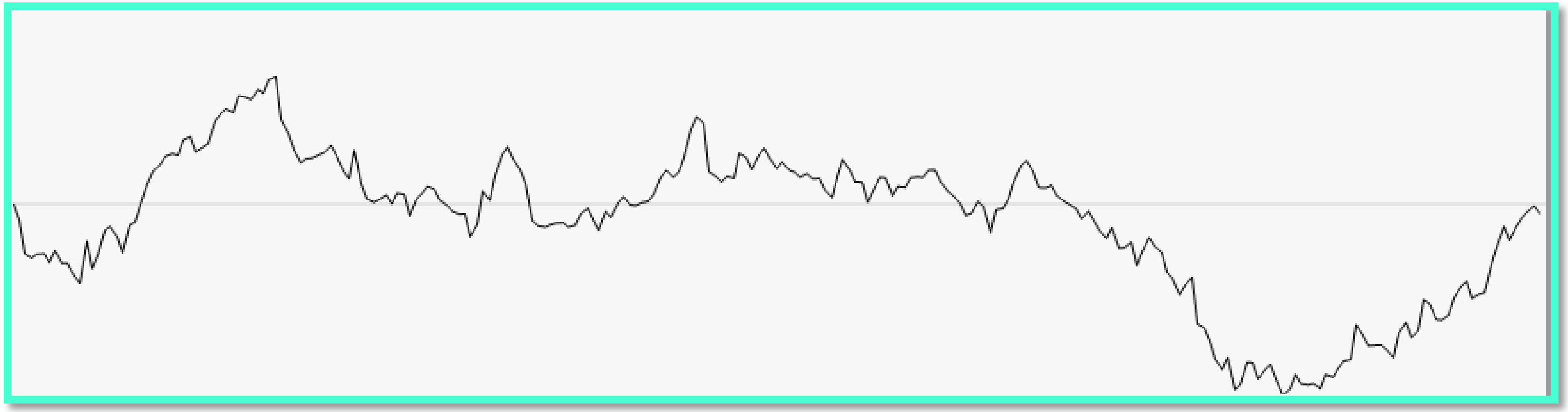
John Snow’s map in which each death from cholera is indicated by a black bar.
Cholera is spread by infected water. There were outbreaks of cholera in 1832, 1848 and 1854 which killed thousands of people across Britain. People at the time believed that cholera was caused by miasma (bad smells) in the air which caused disease. In 1843, John Snow made a breakthrough. He proved that there was a link between cholera and the water supply. Snow studied a particluar water pump in London on Broad Street. He drew up a map showing the number of deaths surrounding the water pump. He then removed the handle from the water pump so that people coud not drink from it. Unsuprisingly, the number of deaths in the aread dropped dramatically.

Horizon plot - (how to visualize lots of time series?)

a year’s worth of stock prices for 50 separate stocks (one per row)



http://www.perceptualedge.com/articles/visual_business_intelligence/time_on_the_horizon.pdf a line encodes the changing price of a single stock over the course of roughly one year



jlborges@fe.up.pt

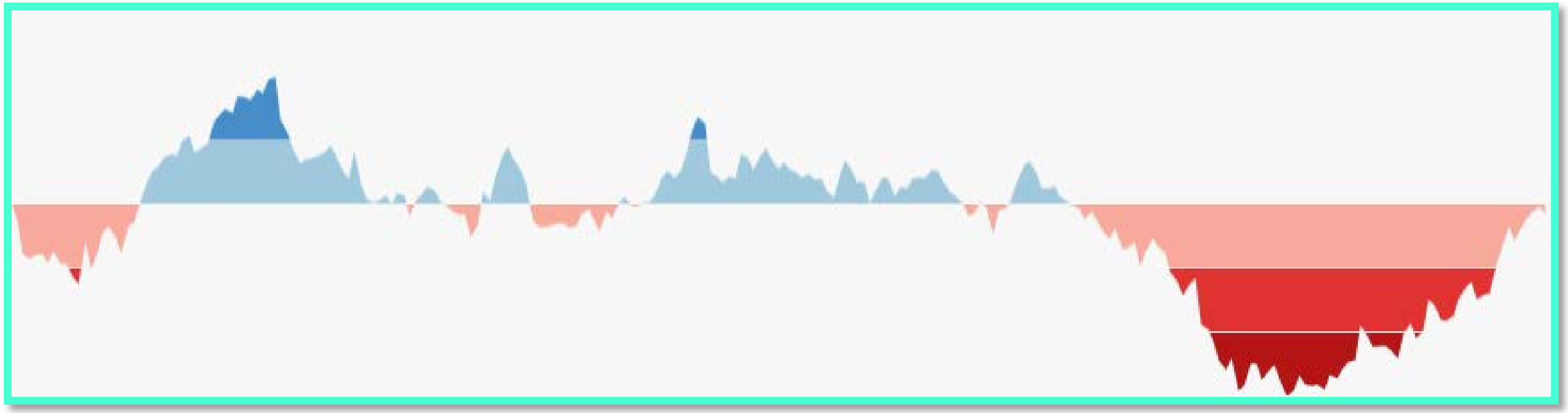
how can many lines be displayed on the screen simultaneously in a perceivable and meaningful way?

jlborges@fe.up.pt

goal - difference between increases versus decreases compared to the starting value to stand out distinctly.

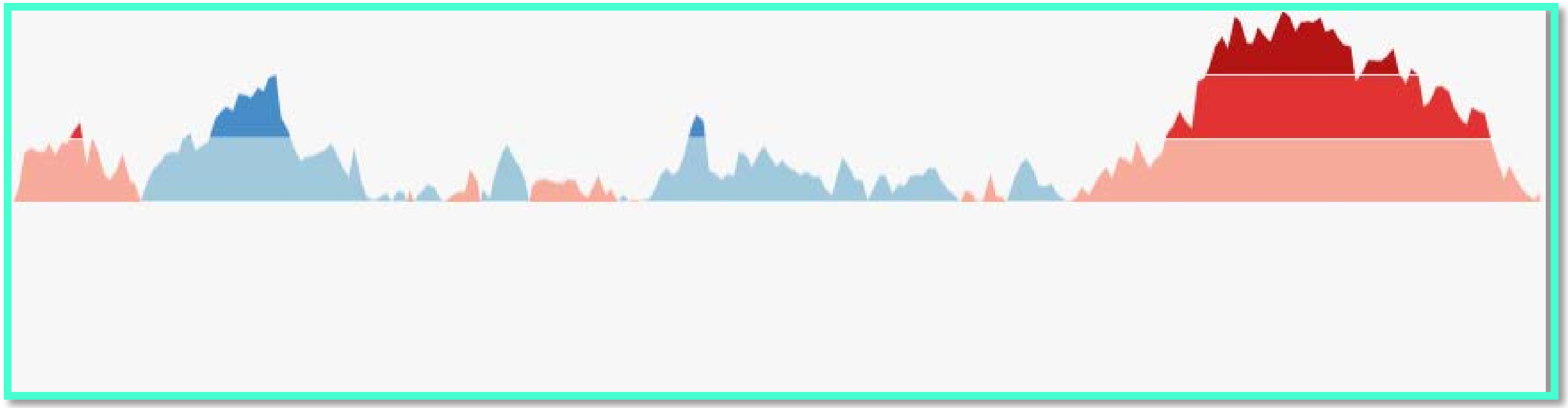
Blue - positive differences
Red - negative differences

magnitude within either positive or negative values, encoded as differences in color intensity



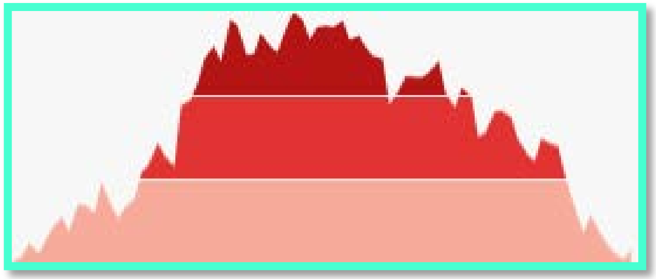
jlborges@fe.up.pt

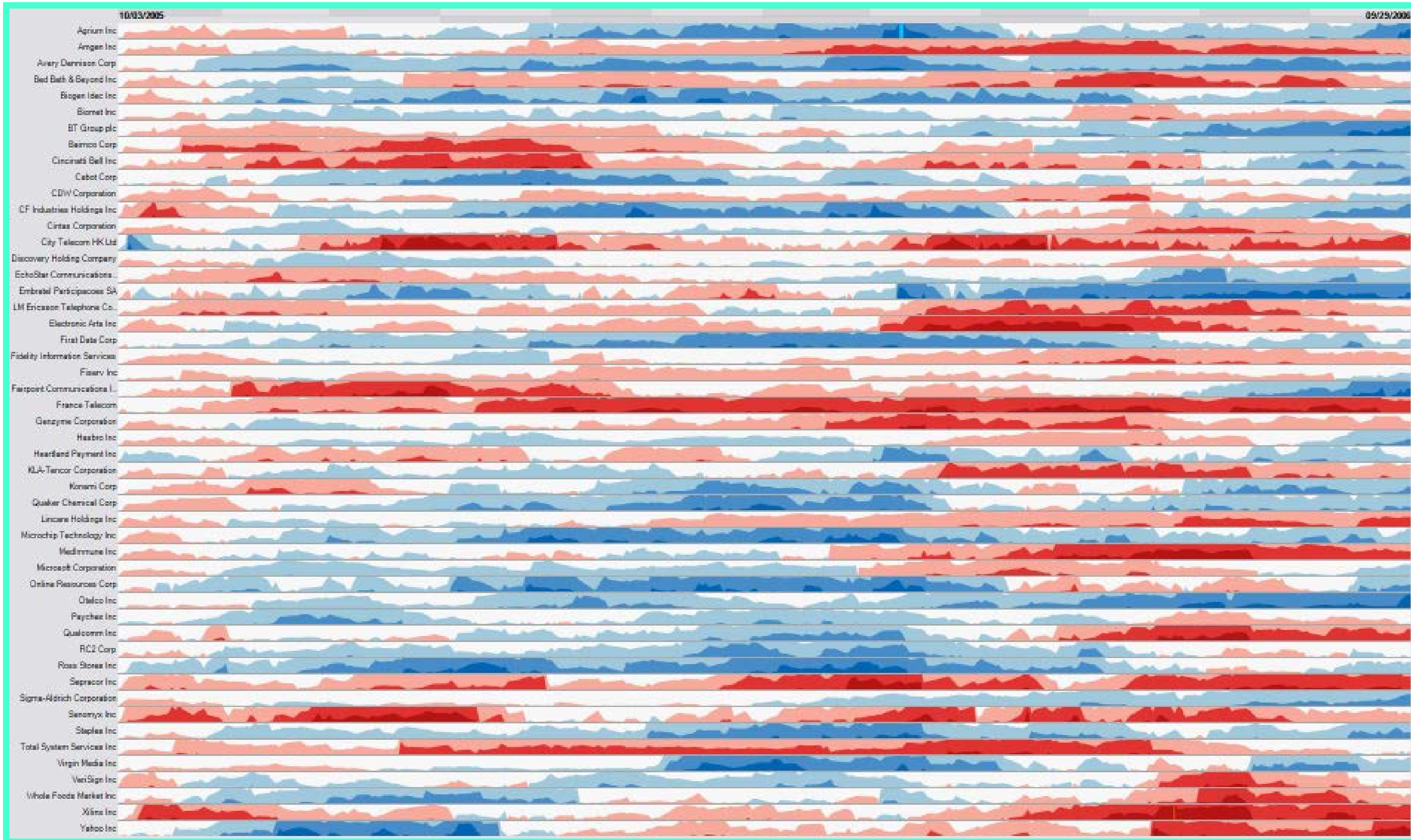
How can we save vertical space?



How can we save even more vertical space?

Collapse the color bands to display the values in less vertical space




10/03/2005
09/29/2006
Agrium Inc
Amgen Inc
Avery Dennison Corp
Bed Bath & Beyond Inc
Biogen Idec Inc
Biomet Inc
BT Group plc
Baimco Corp
Cincinatti Bell Inc
Cabot Corp
CDW Corporation
CF Industries Holdings Inc
Cintas Corporation
City Telecom HK Ltd
Discovery Holding Company
EchoStar Communications...
Embratel Participacoes SA
LM Ericsson Telephone Co...
Electronic Arts Inc
First Data Corp
Fidelity Information Services
Fiserv Inc
Fairpoint Communications I...
France Telecom
Genzyme Corporation
Hasbro Inc
Heartland Payment Inc
KLA-Tencor Corporation
Konami Corp
Quaker Chemical Corp
Lincare Holdings Inc
Microchip Technology Inc
MedImmune Inc
Microsoft Corporation
Online Resources Corp
Otelco Inc
Paychex Inc
Qualcomm Inc
RC2 Corp
Ross Stores Inc
Sepracor Inc
Sigma-Aldrich Corporation
Senomyx Inc
Staples Inc
Total System Services Inc
Virgin Media Inc
VeriSign Inc
Whole Foods Market Inc
Xilinx Inc
Yahoo Inc

Resultados de Aprendizagem

- Identificar experiências aleatórias que podem ser representadas por uma distribuição
  - Uniforme
  - Exponencial negativa
  - Normal

- Calcular probabilidades de acontecimentos de experiências aleatórias que podem ser representadas por uma distribuição
  - Uniforme
  - Exponencial negativa
  - Normal

- Saber verificar graficamente se uma variável tem as características de uma distribuição Normal (*Normal Probability Plots*)