

# Variáveis Aleatórias em Situações Limite

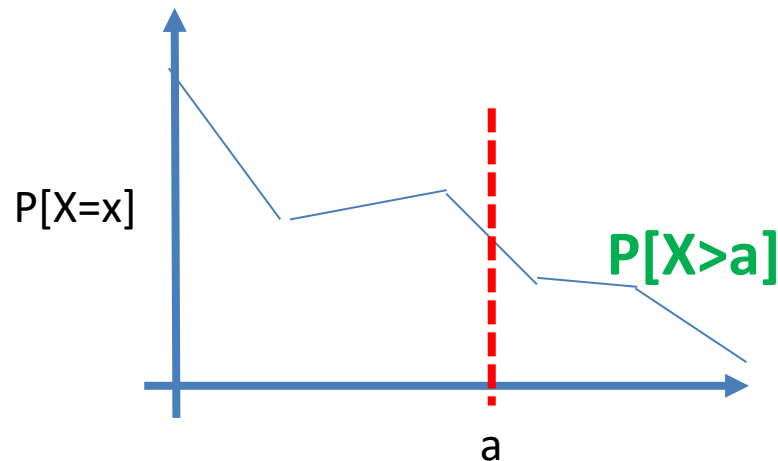
Desigualdades de Markov e Chebyshev

Lei dos Grandes Números

Teorema do Limite Central

# Motivação

- Vimos que  $E[X]$  dá informação sobre o valor médio de uma variável aleatória  $X$ 
  - Muito útil para muitos problemas
- No entanto, em diversas situações interessa-nos saber a probabilidade para valores distantes de  $E[X]$ 
  - Ou seja, o que acontece na “cauda” (tail) da distribuição



# Motivação

- Também nos interessam situações limite
- Por exemplo, saber **o que acontece quando  $n$  tende para infinito** ao valor esperado e variância da Média de  $n$  medições
- Outro exemplo, muito importante:
  - Ao fazermos centenas de milhar ou milhões de experiências **nas nossas simulações** (teoria frequencista) **estamos de facto a garantir boas estimativas das probabilidades ?**

# Média e variância da Média

- Se criarmos a v.a. relativa à média de  $n$  variáveis IID  $X_i$ ,  
$$\mathbf{M}_n = \frac{S_n}{n}$$
- assumindo  $E[X_i] = \mu$  e  $Var(X_i) = \sigma^2$ , teremos :

$$E[\mathbf{M}_n] = E\left[\frac{S_n}{n}\right] = \frac{\sum_i E[X_i]}{n} = E[X_i] = \mu$$

- $\mathbf{Var}[\mathbf{M}_n] = Var\left(\frac{S_n}{n}\right) = \frac{1}{n^2} \frac{\sum_i Var[X_i]}{1} = \frac{Var(X_i)}{n} = \frac{\sigma^2}{n}$
- À medida que se aumenta o número de experiências vai diminuindo a variância da estimativa da média

# Questões

- Quanto provável é termos a média das amostras superior a um determinado valor?
  - Exemplo: o dobro do valor esperado
- Quanto próxima a média obtida com as amostras fica do valor médio ?
- Qual a distribuição da média para valores de  $n$  muito grandes ?

# Desigualdades de Markov e Chebyshev

- Os dois teoremas que apresentaremos de seguida, sem muita preocupação com demonstrações, **permitem estabelecer facilmente majorantes** para probabilidades de certas classes de acontecimentos
  - partindo apenas do conhecimento da média e variância de uma variável aleatória
  - Mais informação, por exemplo, na secção 3.8 do livro de F. Vaz

# Questão 1

- Probabilidade de termos valores superiores a um determinado valor ?
- Exemplo:
- A média das classificações numa turma é 15,2.
- Será que conseguimos determinar um limite superior para probabilidade de um dos alunos ter nota igual ou superior a 17 ?

# Desigualdade de Markov

- Seja  $X$  uma variável aleatória **não negativa**
- Pela Desigualdade de Markov:

$$P(X \geq a) \leq \frac{E[X]}{a}, \quad \forall a > 0$$

- Esta desigualdade dá-nos um limite superior para a probabilidade de a função  $X$  ser maior ou igual a um determinado valor
- Qual o valor de  $P$  com  $a = E[X]$  ?
  - $E a < E(X)$  ?
  - $E a > E(X)$  ?



# Desigualdade de Markov

- Demonstração:

- $E[X] = ?$

- $= \int_0^a x f_X(x) dx + \int_a^\infty x f_X(x) dx \geq$

- $\geq \int_a^\infty x f_X(x) dx \geq \int_a^\infty a f_X(x) dx =$

- $\geq a P[X \geq a]$

- Logo:  $P[X \geq a] \leq \frac{E[X]}{a}$

# Exemplo (continuação)

- A média da altura de uma população é 1,65 m.
- Qual o limite superior de probabilidade de um indivíduo ultrapassar os 2 metros ?
- $P(X \geq 2) \leq \frac{1,65}{2} = 0,825$
- Limite não muito útil ou significativo !

# Exemplo 2

- A média das classificações numa turma é 15,2.
- Qual o limite superior de probabilidade de um dos alunos ter nota igual ou superior a 17 ?
- $P(X \geq 17) \leq \frac{15,2}{17} = 0.8941$
- E superior a 19?
- $P(X \geq 19) \leq \frac{15,2}{19} = 0.8$

## Questão 2

- Quão provável é a **diferença entre a variável e o seu valor esperado** ser superior/inferior a um determinado valor ?

- Isto é  $P(|X - E[X]| \geq a) = ?$

Ou  $P(|X - E[X]| < a) = ?$

- Exemplo: Probabilidade de os valores diferirem da média mais que 2 desvios padrão ?

# Desigualdade de Chebyshev

- Pela Desigualdade de Chebyshev temos:

- $P(|X - E[X]| \geq a) \leq \frac{\text{Var}(X)}{a^2}$

- Ou, em alternativa:

- $P(|X - E[X]| < a) \geq 1 - \frac{\text{Var}(X)}{a^2}$

# Desigualdade de Chebyshev

- Demonstração:
- Define-se  $D^2 = (X - E[X])^2$
- É óbvio que  $D^2 \geq 0$  e  $D^2 \geq a^2 \Leftrightarrow |D| \geq a$
- Aplicando a Desigualdade de Markov
- $P(|D| \geq a) = P(D^2 \geq a^2)$
- $P(D^2 \geq a^2) \leq \frac{E[(X - E[X])^2]}{a^2}$
- $P(D^2 \geq a^2) \leq \frac{\text{Var}(X)}{a^2} ; P((X - E[X])^2 \geq a^2) \leq \frac{\text{Var}(X)}{a^2}$
- Assume-se  $E[X]$  e  $\text{Var}(X)$  finitos

# Desigualdade de Chebyshev

- Se expressarmos  $a$  em função do desvio padrão, fazendo  $a = h\sigma$ , teremos:
- $P(|X - E[X]| \geq h\sigma) \leq \frac{\sigma^2}{(h\sigma)^2} = \frac{1}{h^2}$
- Ou seja: a probabilidade de obter um valor que dista da média de  $h$  desvios padrão ou mais é menor ou igual a  $\frac{1}{h^2}$ 
  - Exemplos:
    - $h=1 \Rightarrow P \leq 1$
    - $h=2 \Rightarrow P \leq \frac{1}{4}$

Valores com pouco precisão

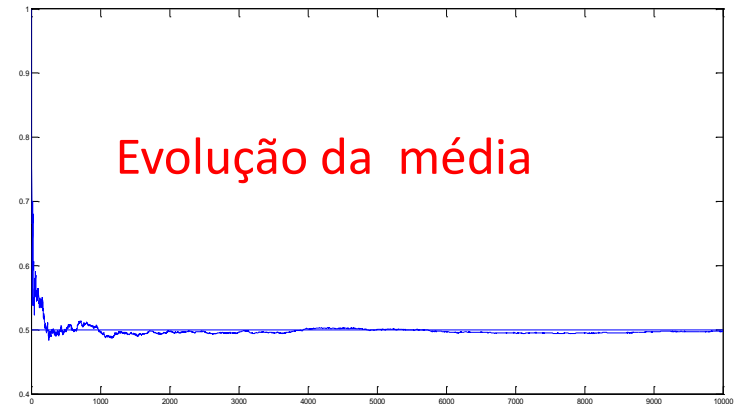
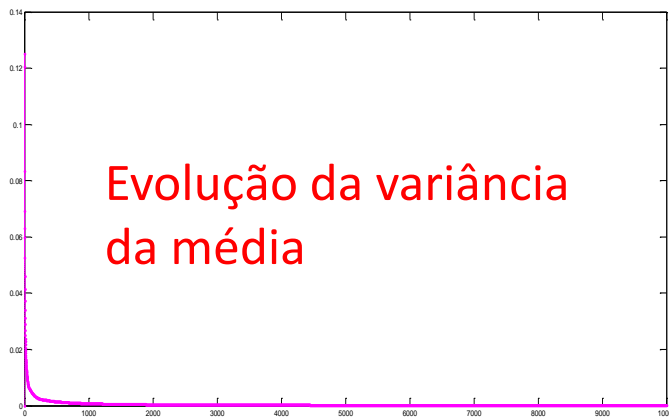
# Questão 3

- Ao fazermos centenas de milhar ou milhões de experiências nas nossas simulações (teoria frequencista) estamos de facto a garantir boas estimativas das probabilidades ?



# Voltando à média de $n$ variáveis aleatórias ...

- Como vimos, **a variância da média das estimativas** tende para 0 à medida que  $n$  aumenta



- O que se pode interpretar como a probabilidade da média das amostras se aproximar do valor médio ser cada vez maior, aproximando-se de 1

# Voltando à média de $n$ variáveis aleatórias ...

- Qual a **probabilidade da média das amostras se aproximar do valor médio** (a menos de  $\epsilon$ ) ?
  - Ou seja:  $P(|M_n - E[M_n]| < \epsilon)$
- Recorrendo à Desigualdade de Chebyshev temos:
- $$P(|M_n - E[M_n]| \geq \epsilon) \leq \frac{\text{Var}(M_n)}{\epsilon^2}$$
- $$P(|M_n - E[M_n]| \geq \epsilon) \leq \frac{\frac{\sigma^2}{n}}{\epsilon^2}$$
- $$P(|M_n - E[M_n]| < \epsilon) \geq 1 - \frac{\sigma^2}{n \epsilon^2}$$

# Lei fraca dos grandes números

- Passando ao limite a última expressão teremos:

$$\lim_{n \rightarrow \infty} P(|M_n - E[M_n]| < \epsilon) = 1$$

- Resultado que é conhecido por **Lei Fraca dos Grandes Números**

# Leis dos grandes números (LGN)

- Existe um segundo enunciado (fora dos objectivos de MPEI), a **lei forte dos grandes números**, que afirma:

$$P\left(\lim_{n \rightarrow \infty} M_n = \mu\right) = 1$$

- A **Lei Fraca dos Grandes Números** afirma que para um valor de  $n$  suficientemente elevado a média das amostras estará muito próxima do valor esperado

- Enquanto que a lei forte garante que é certo que o limite para que tende a média (das amostras) é o valor esperado

# L. G. N. e definição frequencista

- Consideremos uma **sequência de experiências aleatórias independentes e repetidas**
- e seja  $I_j$  uma variável aleatória indicadora da ocorrência do evento A na experiência de ordem j
  - [1 significa que A ocorreu]
- O número total de ocorrências de A nas n experiências será:

$$N_n = I_1 + I_2 + \cdots + I_n$$

# L. G. N. e definição frequencista

- Como a frequência relativa de A é

$$f_A(n) = \frac{(I_1 + I_2 + \dots + I_n)}{n}$$

- $f_A$  é a média das amostras das variáveis aleatórias  $I_i$

- Então (pelas duas leis dos grandes números):

$$\lim_{n \rightarrow \infty} P(|f_A(n) - p(A)| < \epsilon) = 1$$

e

$$P[\lim_{n \rightarrow \infty} f_A(n) = p(A)] = 1$$

- Permitindo-nos dizer que **a frequência relativa é uma boa estimativa da probabilidade**



# Um pouco de História (para terminar esta parte)

- 1713: Lei fraca descrita por Jacob **Bernoulli**
- 1835: **Poisson** chama-lhe “La Loi des Grands Nombres”
  - Lei dos Grandes Números em Francês
- 1909: Émile Borel desenvolve a Lei forte para variáveis de Bernoulli
- 1928: Andrei Nikolaevich **Kolmogorov** prova a Lei forte no caso geral

Qual a distribuição de  $M_n$  para valores de  $n$  muito grandes ?



# Questão

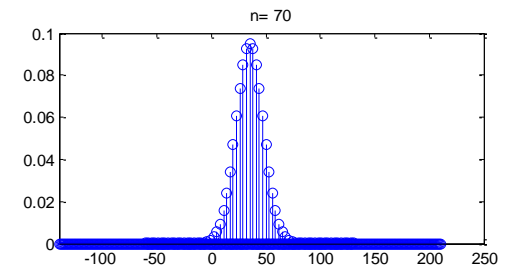
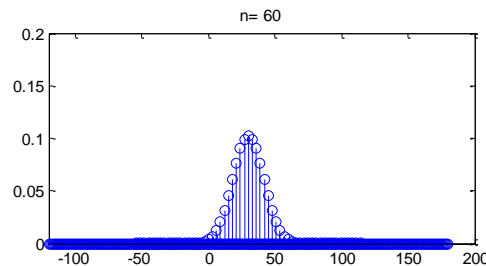
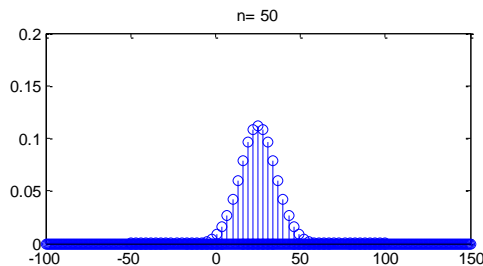
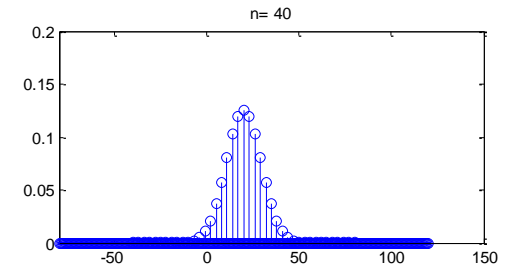
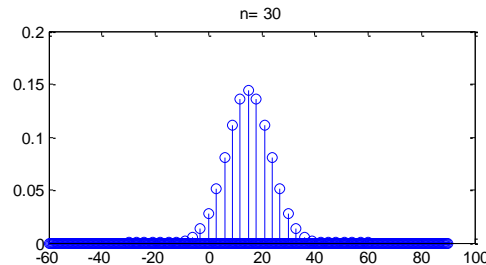
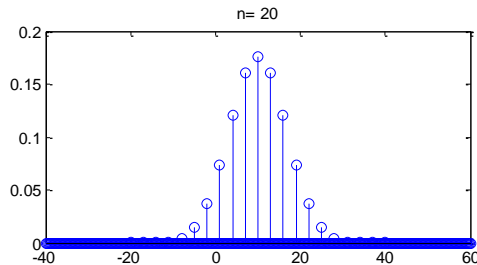
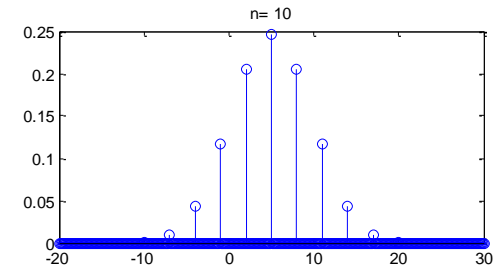
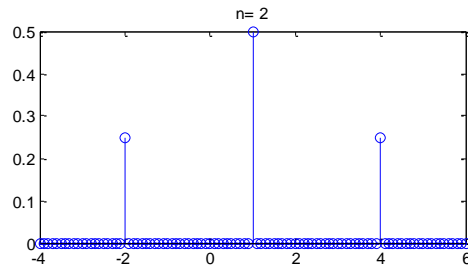
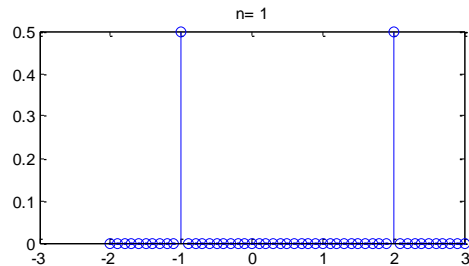
- Já vimos o comportamento limite da média de uma sequência de variáveis aleatórias
- Conseguimos avançar mais e dizer alguma coisa quanto à distribuição ?
- Começemos com alguns exemplos ...

# Exemplo 1

- Consideremos um jogo em lançamos uma moeda ao ar e **perdemos 1 Euro se sair CARA** e **ganhámos 2 Euros se sair COROA**
- A moeda é honesta e existe independência entre as jogadas
- Como se comporta a distribuição com as jogadas ?

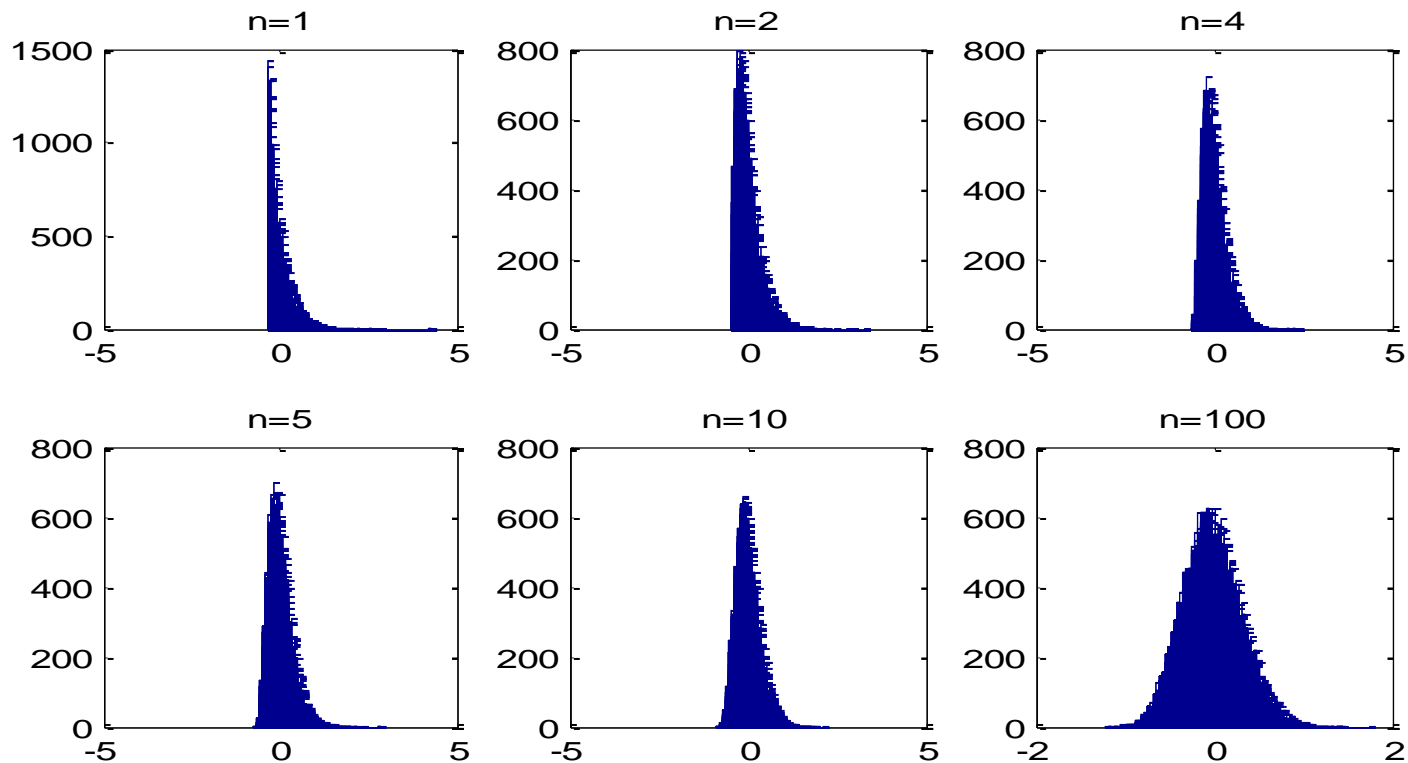
# Continuando o jogo

- Recorrendo a simulação em Matlab...



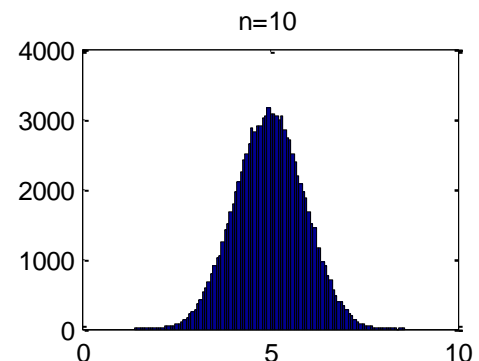
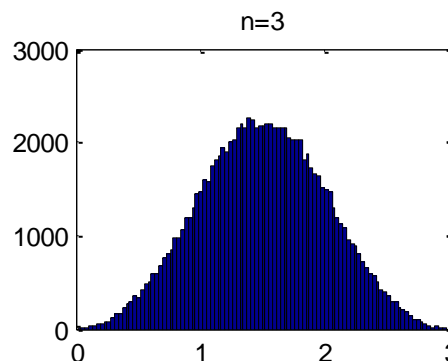
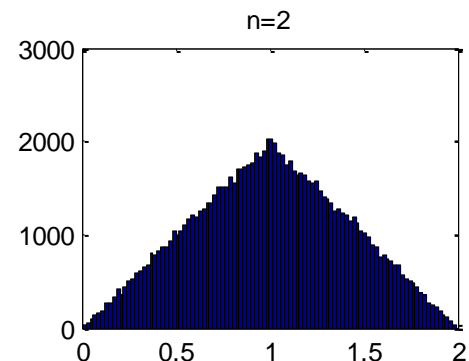
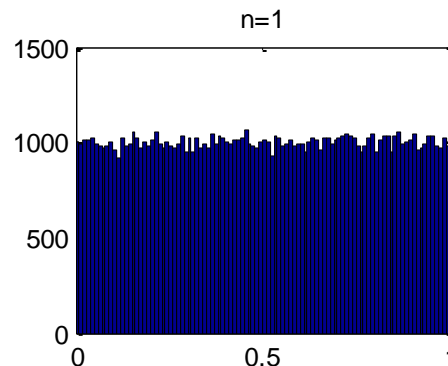
# E se tivermos outras distribuições iniciais ?

- Exponencial:  $y = -\log(\text{rand}(1, \text{len})). / \text{lambda}$



# Outro exemplo

- **Usando geração** de números aleatórios:
- Geradas 10 sequências de números aleatórios com distribuição uniforme no intervalo  $[0,1]$  e somadas ...



# Teorema do Limite Central

- Nos exemplos, para valores grandes de  $n$ , temos sempre uma distribuição com a forma da Gaussiana
- De facto **demonstra-se que a soma de variáveis i.i.d. tende para uma distribuição normal quando o número de variáveis é grande**
  - Teorema do Limite Central
- A média é, como já vimos, igual à das variáveis originais

# De uma forma mais formal

- Sendo:
  - $X_1, X_2, \dots$  **variáveis aleatórias I.I.D.**
  - $X_i$  com distribuição  $F$ ,  $E[X_i] = \mu$  e  $Var(X_i) = \sigma^2$ 
    - $\mu$  e  $\sigma^2$  finitos
  - $S_n$  a soma das  $n$  primeiras **variáveis**
  - $Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$  v.a. de média nula e variância unitária

- O Teorema do Limite Central afirma:

$$\lim_{n \rightarrow \infty} P(Z_n \leq z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-\frac{x^2}{2}} dx$$

- Isto é, a função de distribuição de  **$Z_n$  tende para a distribuição de uma variável Normal normalizada  $N(0, 1)$**

# Aplicando à média ( $M_n$ )

- Fazendo  $M_n = \frac{1}{n} \sum_{i=1}^n X_i$
- Pelo TLC temos

$$M_n \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{quando } n \rightarrow \infty$$

A distribuição da **média de  $n$  variáveis i.i.d. tende para a distribuição normal com parâmetros  $\mu$  e  $\frac{\sigma^2}{n}$**



# Teorema do Limite Central

- O Teorema do Limite Central é a **razão da importância da distribuição Normal/Gaussiana**
  - É um **resultado extremamente importante e abre caminho a muitas aplicações**
- “Formulação qualitativa”:

Coisas que são o resultado da soma de muitos pequenos efeitos tendem a ser Gaussianas

# Demos online

- Wolfram Demonstrations Project : **The Central Limit Theorem**

The central limit theorem states that the sampling **distribution of the sample mean approaches a normal distribution as the size of the sample grows.**

This means that the histogram of the means of many samples should approach a bell-shaped curve.

Each sample consists of 200 pseudorandom numbers between 0 and 100, inclusive.

- <http://demonstrations.wolfram.com/TheCentralLimitTheorem/>

# Demos

- **Central Limit Theorem Applied to Samples of Different Sizes and Ranges**
- <http://demonstrations.wolfram.com/CentralLimitTheoremAppliedToSamplesOfDifferentSizesAndRanges/>
- This Demonstration shows the applicability of the central limit theorem (CLT) to the means of samples of random integer or real numbers having random ranges.
- It allows the user to generate such datasets and plot the histogram of their means.
- Superimposed on the histogram is the normal (Gaussian) distribution function that gives the theoretical distribution of these sample means.
- Also shown for comparison are the numeric values of the mean and standard deviation, both of the theoretical distribution and of the generated data.

# Exemplo de aplicação do TLC

- Suponha que as despesas feitas por cada cliente de um restaurante são variáveis aleatórias I.I.D. com média 6.5 Euros e desvio padrão 2.5 Euros.
- Estime a probabilidade de os primeiros 100 clientes gastarem um total superior a 600 Euros

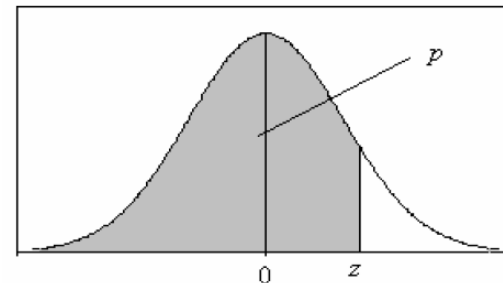
# Resolução

- Consideremos  $S_{100} = X_1 + X_2 + \cdots + X_{100}$
- Como  $E[S_{100}] = 100\mu = 650$
- e  $n\sigma^2 = 625$
- Teremos  $Z_{100} = \frac{S_{100} - 650}{25}$
- Como pelo TLC  $Z_{100}$  segue um lei  $N(0,1)$ :
- $P(S_{100} > 600) = P\left(Z_{100} > \frac{600 - 650}{25}\right)$
- $= P\left(Z_{100} > \frac{600 - 650}{25}\right) = \mathbf{P(Z_{100} > -2)}$

# Calc. probabilidades na $N(0,1)$

- $P(Z_{100} > -2)$  ?
- Como se obtém ?
- Existem valores tabelados de  

$$P(Z \leq z) = \Phi(z)$$



– Exemplo:

<http://www.professores.uff.br/patricia/images/stories/arquivos/TabelaNormal.pdf>

- $P(Z_{100} > -2) = 1 - \Phi(-2)$
- $= 1 - (1 - \Phi(2)) = \Phi(2) =$

1,7	0,96445	0,96407
1,8	0,96485	0,96447
1,9	0,96485	0,96447
2,0	0,97725	0,97778
2,1	0,98214	0,98257
2,2	0,98610	0,98645
2,3	0,98956	0,98979

$=0,97725$

# Em Matlab

- Obter  $\Phi(2)$

z=2

m=0

sigma=1

p = cdf('Normal',z,m,sigma)

>> 0.9772

OU

p= normcdf(z,m,sigma)

Nota: usa Statistics Toolbox

# Em Matlab

- Com ferramentas como Matlab não é necessário estar a efectuar a normalização
- Aplicando directamente a  $S_{100}$ :

```
s=600           % pq queremos  $P(S_{100} > s=600)$   
m=650           % média de  $S_{100}$   
sigma=25        % desvio padrão de  $S_{100}$ 
```

```
p = 1- cdf('Normal',600,m,sigma)  
>>> 0.9772
```

OU

```
p = 1- normcdf(600,m,sigma)
```



# Exercício - Inquérito futebolístico

- $f$ : fracção da população que gosta de futebol
- Queremos fazer uma sondagem/inquérito a  $n$  pessoas
- Quantas pessoas devemos inquirir para ter uma confiança (probabilidade) de 95% de que **não cometemos um erro superior a 1 %**
- Considere:
  - Resultado de um inquérito à pessoa  $i$ :
$$X_i = \begin{cases} 1, & \text{se gosta} \\ 0, & \text{se não gosta} \end{cases}$$
  - $M_n = \frac{X_1 + X_2 + \dots + X_n}{n}$  fracção de “gosta” na amostras

# Resolução

- Sugestões ?
- Uma das formas (veremos outra) é usando a Desigualdade de Chebyshev ...
- O que diz a desigualdade ?
- $$P(|M_n - E[M_n]| \geq \epsilon) \leq \frac{Var(M_n)}{\epsilon^2}$$

# O que sabemos ?

- $\epsilon = ?$
- $\epsilon = 0.01$
- $Var(M_n) = ?$
- $Var(M_n) = \frac{Var(X_i)}{n}$

$$Var(X_i) = ?$$

- Todas as  $X_i$  são v. a. de Bernoulli
  - Mas não sabemos  $p$  (o inquérito é para estimar isso)
- Para o nosso caso é útil o valor máximo de  $Var(X_i)$  . Qual esse valor ?
- $Var(X_i) = p(1 - p) \leq \frac{1}{4}$

# Voltando à desigualdade

- Substituindo temos:

- $$P(|M_n - E[M_n]| \geq 0,01) \leq \frac{\frac{1}{4}n}{0,01^2} = \frac{1}{4 n 10^{-4}}$$

- Como queremos  $P(\quad) \leq 0,05$

- $$\frac{1}{4 n 10^{-4}} \leq 0,05$$

- $n = ?$

- $n \geq 50\,000$  (valor conservador)

E se  $\epsilon = 0,05$  ?

- $P(|M_n - E[M_n]| \geq 0,05) \leq \frac{1}{4 n (0,05)^2}$
- Obtendo-se  $n$  de:
- $\frac{1}{4 n (0,05)^2} \leq 0,05$
- $n \geq 2000$

# Discussão

- Problemas com os valores de  $n$  que obtivemos:
  1. São muito grandes
  2. Baseiam-se numa desigualdade que apenas pode dar um majorante/minorante
    - E não um valor “exacto”
- Veremos de seguida que se pode fazer melhores estimativas de  $n$ 
  - Mas para isso precisamos saber mais sobre a distribuição de  $M_n$

# Resolução usando TLC

- Pretendemos  $P(|M_n - f| \leq 0,05) \geq 0,95$
- O evento que nos interessa calcular a probabilidade é  $|M_n - f| \leq 0,05$
- Pretendemos portanto  $P\left(\left|\frac{S_n - \textcolor{teal}{n}f}{n}\right| \leq 0,05\right)$
- Como  $Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$  manipulamos para obter  $\sqrt{n}\sigma$  no denominador, obtendo
- $P\left(\left|\frac{S_n - nf}{\sqrt{n}\sigma}\right| \leq \frac{0,05\sqrt{n}}{\sigma}\right)$



# Resolução usando TLC (cont.)

- Como  $Z_n$  tende para  $N(0,1)$
- Teremos:
- $P(|M_n - f| \leq 0,05) \approx P(|Z| \leq 0,05 \frac{\sqrt{n}}{\sigma})$
- E usando majorante para a variância
$$p(1 - p) \leq 1/4 \quad (=> \sigma = 1/2)$$
- $P(|M_n - f| \leq 0,05) \leq P(|Z| \leq 0,1\sqrt{n})$

$$P(|Z| \leq 0.1\sqrt{n}) ?$$

- $P(|Z| \leq 0.1\sqrt{n})$
- $= P(-0.1\sqrt{n} \leq Z \leq 0.1\sqrt{n})$
- $= F_{N(0,1)}(0.1\sqrt{n}) - F_{N(0,1)}(-0.1\sqrt{n})$
- Para permitir usar tabelas, coloquemos em função de  $Q(z) = 1 - F_{N(0,1)}(z)$ 
  - Sabe-se também que  $F_{N(0,1)}(-z) = Q(z)$
- $= 1 - Q(0.1\sqrt{n}) - Q(0.1\sqrt{n})$
- $= 1 - 2 Q(0.1\sqrt{n})$

# Terminando...

- $1 - 2 Q(0,1\sqrt{n})$  terá de ser  $\geq 0,95$
- $1 - 2 Q(0,1\sqrt{n}) \geq 0,95$
- $\Rightarrow Q(0,1\sqrt{n}) \geq \mathbf{0,025}$
- $\Rightarrow 0,1\sqrt{n} \geq \mathbf{1,96}$  por consulta a tabela
- Resolvendo em ordem a  $n$  temos, finalmente,
- $\sqrt{n} \geq (1,96)^2 \Rightarrow n \geq 384,16$
- $n = \mathbf{385}$  é o número mínimo que procurávamos

# Para mais informação

- Capítulo 5, “Somas de variáveis aleatórias”, do livro de F. Vaz