MPEI 2018-2019

Aula 26

Teorema do Limite Central e aplicações

Exercícios de consolidação

- Sendo $X_1, X_2, \ldots$ i.i.d com média finita μ e variância finita σ^2:
- E
- e $M_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$

- $E[M_n] = ?$
- $Var[M_n] = ?$

- $P(|M_n - E[M_n]| \geq \epsilon) \leq ???$
- $P(|M_n - E[M_n]| \geq \epsilon) \leq \frac{Var(M_n)}{\epsilon^2}$

- M_n converge em probabilidade para μ

Exercícios de consolidação

- 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:

$$X_i = \begin{cases} 1, & se\ gosta \\ 0, & se\ não\ gosta \end{cases}$$

  - $M_n = \frac{X_1+X_2+\cdots+X_n}{n}$ 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 ?

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

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 ?

- Comecemos com alguns exemplos …

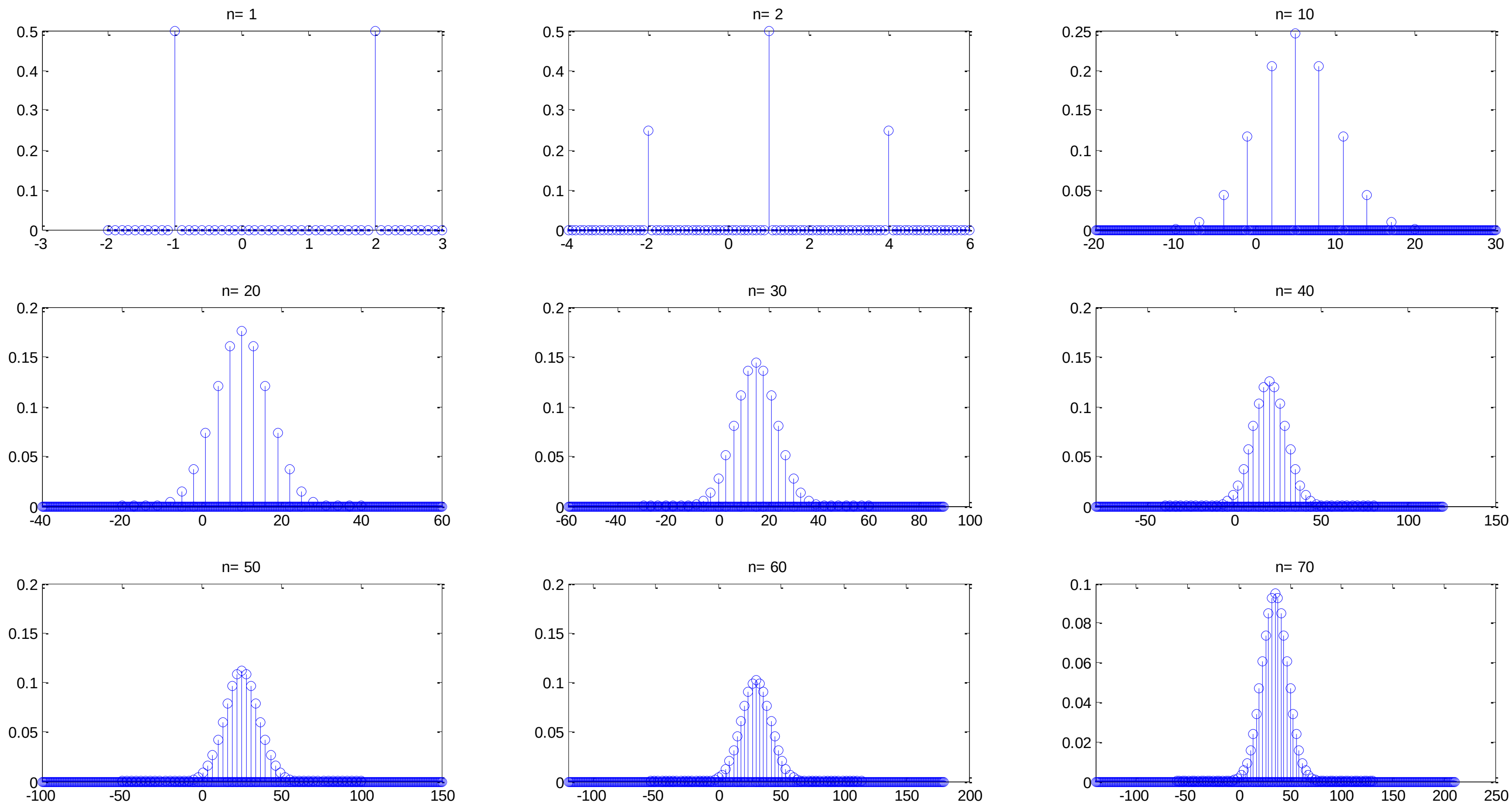
Exemplo 1

- 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 ?

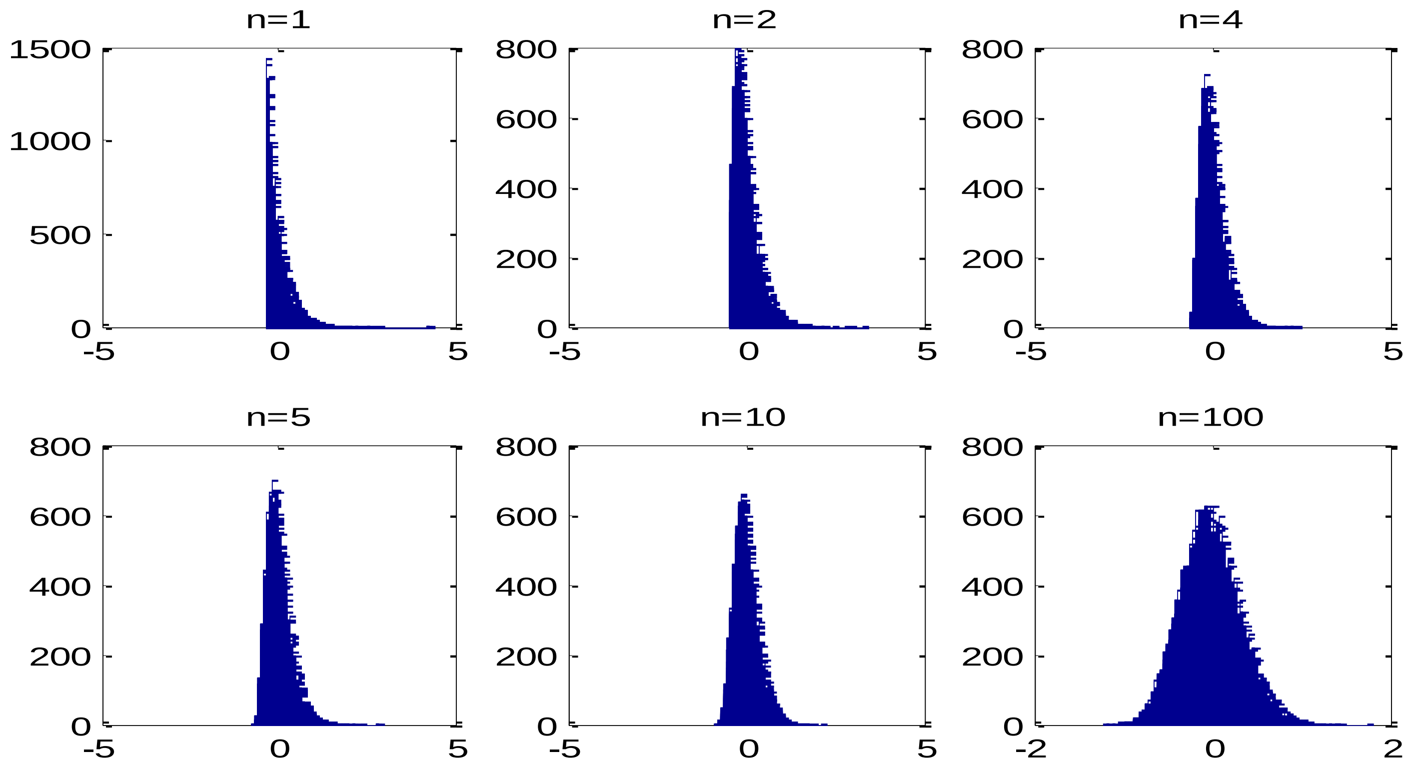
Continuando o jogo

- Recorrendo a simulação em Matlab...



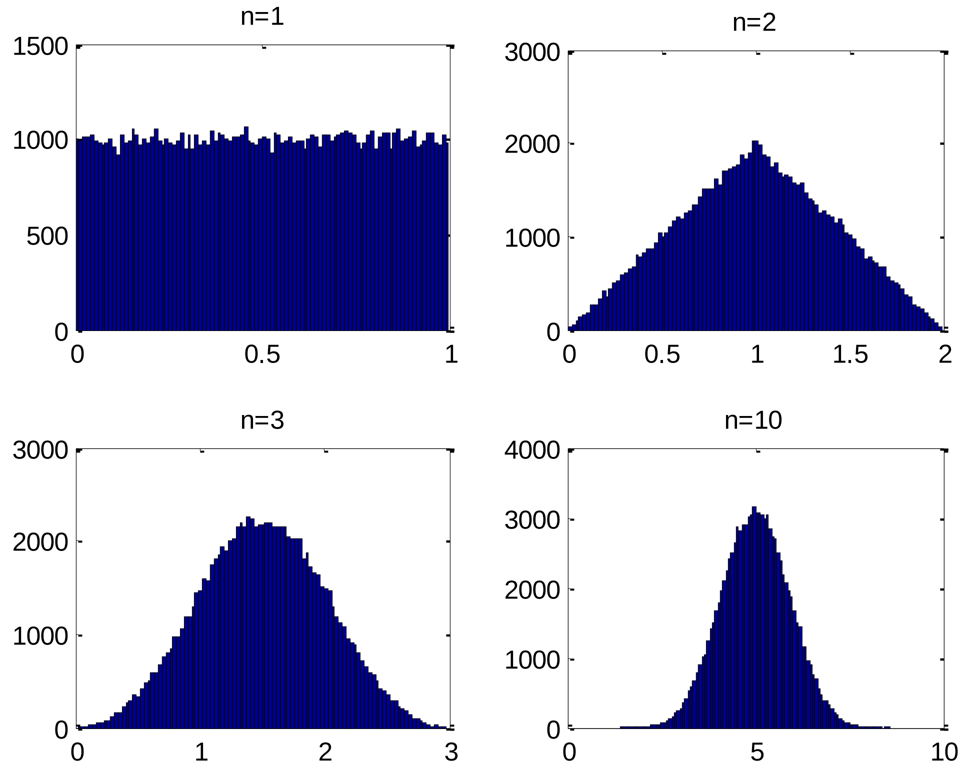
E se tivermos outras distribuições iniciais ?

- Exponencial: y=-log(rand(1,len))./lambda



Outro exemplo

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



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 online

- **Central Limit Theorem for the Continuous Uniform Distribution**

- http://demonstrations.wolfram.com/CentralLimitTheoremForTheContinuousUniformDistribution/

- This Demonstration illustrates the central limit theorem for the continuous uniform distribution on an interval.

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.

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:
  - $X_1, X_2, \ldots$ variáveis aleatória I.I.D.
  - X_i com distribuição F e $E[X_i] = \mu$ e $Var(X_i) = \sigma^2$
    - μ e σ^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:

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

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

Aplicando à média

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

- Pelo TLC temos

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

A distribuição da média das 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

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

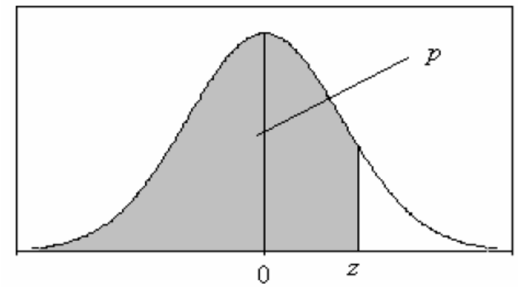
Resolução

- 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 $N(0,1)$:
- $P(S_{100} > 600) = P\left(Z_{100} > \frac{600-650}{25}\right)$
- $= P\left(Z_{100} > \frac{600-650}{25}\right) = \boldsymbol{P(Z_{100} > -2)}$

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

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

1,8	0,96407	0,96485
1,9	0,97128	0,97193
2,0	0,97725	0,97778
2,1	0,98214	0,98257
2,2	0,98610	0,98645

=0,97725

Em Matlab

- Obter $\Phi(2)$

```
z=2
m=0
sigma=1
p = cdf('Normal',z,m,sigma)
    >>  0.9772
```

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(S100 > s=600)
m=650              % média de S100
sigma=25           % desvio padrão de S100

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

Retomemos o nosso inquérito futebolístico …

- Relembremos o problema:
- 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 5 %**

- Considere:
  - Resultado de um inquérito à pessoa i:

$$X_i = \begin{cases} 1, & se\ gosta \\ 0, & se\ não\ gosta \end{cases}$$

  - $M_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$ fracção de "gosta" na amostras

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 $\mathrm{P}\left(\left|\frac{S_n - \boldsymbol{nf}}{n}\right| \leq 0{,}05\right)$
- 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

$$p(1-p) \leq 1/4 \qquad (\Rightarrow \sigma = 1/2)$$

- $P(|M_n - f| \leq 0{,}05) \quad \leq P(|Z| \leq 0{,}1\sqrt{n})$

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

- $P(|Z| \leq 0{,}1\sqrt{n})$
- $= P(-0{,}1\sqrt{n} \leq Z \leq 0{,}1\sqrt{n})$
- $= F_{N(0,1)}(0{,}1\sqrt{n}) - F_{N(0,1)}(-0{,}1\sqrt{n})$

- 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,
- $\sqrt{n} \geq (19{,}6)^2 \Rightarrow n \geq 384{,}16$

- $n = 385$ é o número mínimo que procurávamos

Para mais informação

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