U.PORTO
FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO

# Capítulo 1

# Introdução

AMG, JFO (v8 – 2017)
adaptado de: *Estatística*,
Rui Campos Guimarães,
José A. Sarsfield Cabral

Slide 1.-1

## Conteúdo

**1.1 Estatística** . . . . . **1-1**
**1.2 População e Amostra** . . . . . **1-3**
**1.3 A Estatística e o Método Científico** . . . . . **1-4**
**1.4 Método de Análise Estatística** . . . . . **1-5**
1.4.1 Objectivo da análise e definição da população . . . . . 1-5
1.4.2 Amostragem . . . . . 1-5
1.4.3 Recolha de dados . . . . . 1-6
1.4.4 Análise dos dados . . . . . 1-7
1.4.5 Estabelecimento de inferências sobre a população . . . . . 1-7
**1.5 "Six-Steps Statistical Investigation"** . . . . . **1-8**
1.5.1 Exemplo: "Recruiting Organ Donors" . . . . . 1-8
1.5.2 "Four Pillars of Statistical Inference" . . . . . 1-10
**1.6 Caso de Estudo: "Learning About Lottery Strategies"** . . . . . **1-10**

Slide 1.0

### Resultados de aprendizagem

- Conhecer o objecto da estatística
- Distinguir os conceitos de DADOS e OBSERVAÇÕES
- Distinguir os conceitos de POPULAÇÃO e AMOSTRA
- Conhecer o método de análise estatística ou o "Six-Steps of a Statistical Investigation"

Slide 1.1

## 1.1 Estatística

### O que é a Estatística?

Statistics is the science of designing studies or experiments, collecting data and modeling/analyzing data for the purpose of decision making and scientific discovery when the available information is both limited and variable. *Statistics is the science of Learning from Data.*

(An Introduction to Statistical Methods, Ott and Lyman)

- O que é **a Estatística**? É uma forma de raciocínio e um conjunto de técnicas e métodos desenvolvidos para nos ajudar a compreender e a lidar com a incerteza.
- O que são **Estatísticas**? São valores calculados a partir de conjuntos de Dados ("Data").
- O que são **Dados**? Dados são os valores recolhidos juntamente com o seu contexto.

Slide 1.2

**Principais dificuldades**

- Ao contrário de outras disciplinas da área da Matemática (Análises, Álgebra, ...) não há uma resposta única ou certa (p.e., dois Estatísticos podem testemunhar num tribunal em lados contrários).
- A afirmação anterior leva a que algumas pessoas pensem que podem provar qualquer afirmação com Estatística, o que não é verdade. Tal deve-se a erros (p.e., utilização da técnica ou do método errado, a má interpretação de resultados, ...), que às vezes são intencionais.

**Importância da Estatística**

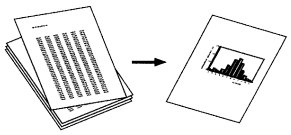
- Actualmente a quantidade de dados e informação recolhida é enorme ("Big Data"), desde o que se compra no supermercado aos clicks em páginas da internet. O que fazer com tantos dados e informação?
- A Estatística está na base das técnicas e métodos que lidam com grandes quantidades de informação e que permitem extrair conhecimento a partir dessa informação.
- Existem também técnicas estatísticas específicas para situações onde a quantidade de dados disponível é pequena ("small samples").

Slide 1.3

## Exemplo

Um gestor de um banco dispõe de uma lista de 550 saldos de contas à ordem seleccionadas ao acaso, numa determinada data, entre as contas de profissionais liberais clientes das suas agências nacionais.

(i) Mais útil do que dispor dos dados em bruto (isto é, da lista de 550 saldos) é ter a informação devidamente classificada (saldos negativos, saldos incluídos nos intervalos 0–100 €, 100–200 €, etc.) e representada, por exemplo, num gráfico:



(ii) E se o gestor do banco pretender, a partir do conjunto de 550 saldos de que dispõe, retirar conclusões acerca da forma como se distribuem os saldos relativos ao conjunto das contas à ordem de todos os profissionais liberais clientes do banco?

Slide 1.4

**Ramos da Estatística:**

**Estatística Descritiva** – Sintetizar e representar de forma compreensível a informação contida num conjunto de dados, situação (i) do exemplo anterior

*(tabelas, gráficos e cálculo de estatísticas)*

**Inferência Estatística** – Caracterizar o todo (a *população*) a partir da análise de um conjunto limitado de dados (uma *amostra*), situação (ii) do exemplo anterior

*(amostra* $\longrightarrow$ *população)*

**Aplicações:**

- Avaliação do risco da introdução de alimentos geneticamente manipulados ou de um novo medicamento.
- Previsão do número de novos casos de infecções por HIV por região ou o número de clientes aderentes a uma promoção de um supermercado.
- Perceber como o nível de desemprego está relacionado com o controlo ambiental ou se a vitamina C realmente previne doenças.

Slide 1.5

### A Estatística e a Variabilidade

- Uma resposta eficaz às aplicações anteriores depende largamente da nossa capacidade em entender e lidar com um conceito fundamental em Estatística: a *variabilidade*
- A *variabilidade* está naturalmente presente em todos os dados que recolhemos, já que estes estão sujeitos a imprecisões e erros de medida e quando recolhidos em alturas diferentes sofrem variações
- Isto leva a que os dados em que nos baseamos para tomar decisões constituam uma imagem imperfeita do mundo real.
- Por estas razões, a *variabilidade* está no centro da Estatística e como lidar com ela é o principal desafio da Estatística e a chave para se conseguir extrair conhecimento a partir de dados.

$\Rightarrow$ A Estatística tem assim um papel importante na compreensão de um mundo complexo e repleto de incerteza e variabilidade.

Slide 1.6

## 1.2 População e Amostra

**Terminologia**

**Observações (ou objectos)** – conjunto de itens ou indivíduos que são estudados numa análise estatística

**Variável(is)** – característica(s) da observação que é medida ou contada

**Dados** – conjunto dos valores associados a uma ou mais variáveis

**Definições operacionais (ou contexto)** – dados só fazem sentido se as variáveis tiverem definições operacionais, que sejam aceites por todos os envolvidos na análise estatística

**Tipos de Variáveis**

**Quantitativas** – expressas em escalas numéricas, podendo-se realizar operações aritméticas com os valores de variáveis (p.e., altura, peso, distâncias, ...)

**Categóricas (ou Qualitativas)** – expressas em categorias, sobre as quais não faz sentido realizar operações aritméticas (p.e., raça, cor, ...)

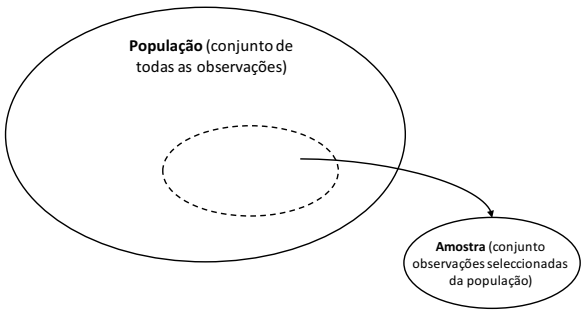
**População (ou universo)** – conjunto de todas observações em análise

**Amostra** – subconjunto de observações seleccionadas de uma população

Slide 1.7

### População e Amostra

**Análise estatística**
Estudo que incide sobre um conjunto de observações com uma determinada característica comum que varia em quantidade e/ou qualidade




Uma análise estatística tanto pode incidir sobre a toda a *população* ou apenas sobre uma pequena parte dela (uma *amostra*)

Slide 1.8

**Exemplo (cont.)**

Um gestor de um banco dispõe de uma lista de 550 saldos de contas à ordem seleccionadas ao acaso, numa determinada data, entre as contas de profissionais liberais clientes das suas agências nacionais.

- Objectos sobre os quais incide a análise?
- Característica analisada?
- População?
- Amostra?

**Objectos** – contas à ordem de todos os profissionais liberais clientes do banco

**Característica** – saldo registado numa determinada data

**População** – conjunto dos saldos das contas à ordem de todos os profissionais liberais clientes do banco, na data em causa

**Amostra** – conjunto de 550 saldos seleccionados

Slide 1.9

**Comentários**

- frequentemente os termos *população* e *amostra* são usados referindo-se indistintamente aos objectos sobre os quais incide a análise estatística e aos dados que medem a característica analisada
- os conceitos de *população* e *amostra* variam em função do objectivo da análise estatística
- os conceitos de *população* e *amostra* podem ser generalizados a mais do que uma dimensão, quando interessa estudar *simultaneamente* várias características bem como as relações de associação entre elas
  - cada elemento da população ou da amostra $\longrightarrow$ vector de dados
  - populações $\longrightarrow$ populações conjuntas
- as populações podem ser finitas ou infinitas; populações finitas de dimensão muito elevada podem ser tratadas como infinitas

Slide 1.10

**Porquê usar amostras ... quando se pretende inferir as características de uma população?**

- população *infinita*
- *custo* excessivo do processo de recolha e tratamento de dados para toda a população
- *tempo* excessivo para o processo de recolha e tratamento de dados para toda a população
- a informação é recolhida por *métodos destrutivos*
- *inacessibilidade* de alguns elementos da população

Slide 1.11

# 1.3 A Estatística e o Método Científico

**Evidência Estatística vs. Evidência "anedótica"**
Quantas vezes ouviram afirmações do tipo:

- "Não vacinem os vossos filhos. Eu vacinei o meu e ele ficou autista."
- "Eu não corro porque o pai do meu amigo correu durante toda a sua vida e faleceu aos 46 anos vítima de ataque cardíaco."
- "Os jovens não devem ser autorizados a guiar porque toda a gente sabe que eles distraem-se facilmente."
- "Sabe-se que fumar provoca cancro, mas a avó de um conhecido viveu até aos 105 anos e fumava um maço de cigarros por dia."

$\Rightarrow$ Conclusões científicas não se devem basear em evidências "anedóticas" como histórias pessoais ou casos extremos para justificar conclusões genéricas, mas sim em evidências estatísticas obtidas a partir de Dados ("Data").

Slide 1.12

- A Estatística é a ciência de produzir os dados adequados para responder a uma determinada "questão de investigação", analisar os dados recolhidos e retirar as conclusões apropriadas com base nesses dados.

- De seguida apresentam-se duas metodologias alternativas apresentadas por autores diferentes capazes de garantir quer o rigor de resultados quer a capacidade de os reproduzir.
  - *Método de Análise Estatística*, Guimarães e Sarsfield Cabral, "Estatística", Verlag Dashöfer Portugal
  - "*Six-Steps Statistical Investigation*", Nathan Tintle, *et al.*, "Introduction to Statistical Investigations", Wiley

  Obs.: Realcem-se as semelhanças entre as duas metodologias apesar das diferenças na terminologia e no número de passos.

Slide 1.13

# 1.4 Método de Análise Estatística

Metodologia mais tradicional adaptada de *Estatística*, Guimarães e Sarsfield Cabral, Verlag Dashöfer Portugal

Fases do Método de Análise Estatística

1. Estabelecimento do *objectivo da análise* a efectuar e definição da(s) população(ões) correspondente(s)
2. Concepção de um procedimento adequado para a selecção de uma ou mais amostras – *amostragem*
3. *Recolha* de dados
4. *Análise* dos dados
5. Estabelecimento de *inferências* acerca da população (com base na informação amostral)

*Nota*: existem interacções entre as várias fases, com ciclos de execução

Slide 1.14

## 1.4.1 Objectivo da análise e definição da população

- O rigor colocado nesta fase tem implicações no esforço a despender nas fases seguintes (nomeadamente, em tempo e dinheiro) e na qualidade das soluções que daí advêm
- Comparações de características envolvem habitualmente mais do que uma população

**Exemplo**
Se o objectivo da análise for o de estabelecer uma comparação entre as alturas dos portugueses adultos e dos suecos adultos, é óbvio que estarão envolvidas duas populações (o conjunto das alturas de uns e o das alturas dos outros)

Slide 1.15

## 1.4.2 Amostragem

**Amostragem probabilística**

- é possível calcular a probabilidade de cada elemento da população ser incluído na amostra
- permite medir o rigor das inferências

**amostragem aleatória** – quando todos os elementos da população têm igual probabilidade de integrar a amostra

*(permite evitar o enviesamento de selecção)*

**Amostragem não probabilística**

**amostragem por conveniência** – amostra seleccionada de modo a ser *conveniente* para o analista

**amostragem subjectiva** – os elementos da amostra são seleccionados com base num critério *subjectivo* de representatividade

Slide 1.16

### Exemplos

**População** – intenções de voto dos eleitores de uma cidade

**Amostra** – intenções de voto de um grupo de eleitores seleccionados ao acaso a partir da lista telefónica correspondente

**Enviesamento de selecção** – sub-representação das intenções de votos dos eleitores pertencentes a estratos sociais mais baixos

- Teste da aceitabilidade de uma nova cerveja pelos consumidores, efectuado com base nas opiniões recolhidas entre um conjunto de empregados da empresa cervejeira, seleccionados ao acaso
- Análise da clareza de um texto didáctico, com base na opinião de um conjunto de alunos que o autor julga serem típicos da população estudantil à qual se dirige

Slide 1.17

### Representatividade de amostras

**Objectivo da amostragem** – garantir a representatividade da(s) amostra(s) a recolher de forma a correctamente analisar a característica em estudo

**Como garantir a representatividade de uma amostra?**

- controlando os *factores externos* que possam influenciar a característica em estudo

  (*e.g.*, *género* e *idade* em entrevistas e inquéritos, *temperatura* em experiências químicas, ...)
- recolha de amostras com *tamanho adequado* para permitir retirar conclusões sobre a(s) população(ões) com base em processos de inferência estatística

  (mais tarde estudaremos métodos que permitem determinar a dimensão da(s) amostra(s) a recolher)

Slide 1.18

## 1.4.3 Recolha de dados

**Fontes:**

**primárias –** dados obtidos directamente pelo analista ou pela sua organização, por *observação* (exame directo e posterior registo de características) ou por *questionários*

**secundárias –** dados compilados ou publicados por outra organização (agências governamentais, empresas especializadas em consultas de mercado, ...)

**Processos:**

**experimentais –** exerce-se um controlo directo sobre os factores que potencialmente afectam a característica em análise

**observacionais –** não se controlam esses factores (forma mais frequente nas análises no âmbito da gestão, da economia ou das ciências sociais – questões éticas/financeiras)

Slide 1.19

### Exemplos

- Para estudar o efeito poluente de uma fábrica sobre a água do rio no qual são descarregados os seus efluentes, foram efectuadas medições da quantidade de oxigénio dissolvido na água, para um conjunto de amostras recolhidas a jusante da fábrica. Metade das amostras foi recolhida no fim de um dia em que a fábrica laborou; e a outra metade, no fim de um dia em que a fábrica se encontrou encerrada
- No âmbito de um estudo de tráfego num túnel rodoviário, procurou analisar-se a relação entre a densidade de tráfego no túnel e a velocidade média de circulação. O estudo foi efectuado com base num conjunto de medições simultâneas incidindo sobre valores da densidade e da velocidade observados ao longo de um mês
- Leitura regular e registo da temperatura ambiente em diferentes divisões de um edifício
- Recolha de dados relativos aos planos de investimento de empresas de um determinado sector, através da realização de entrevistas aos seus directores-gerais

Slide 1.20

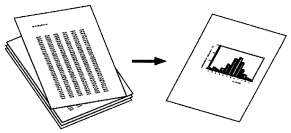
### 1.4.4 Análise dos dados

- sintetizar a informação contida nos dados
- detecção de eventuais erros no processo de recolha de dados

$\Longrightarrow$ *Estatística descritiva* (próximo capítulo)

**Exemplo**

- Um gestor de um banco dispõe de uma lista de 550 saldos de contas à ordem seleccionadas ao acaso, numa determinada data, entre as contas de profissionais liberais clientes das suas agências nacionais.
- Classificar devidamente a informação (saldos negativos, saldos incluídos nos intervalos 0–100 €, 100–200 €, etc.) e representá-la, por exemplo, através de um gráfico

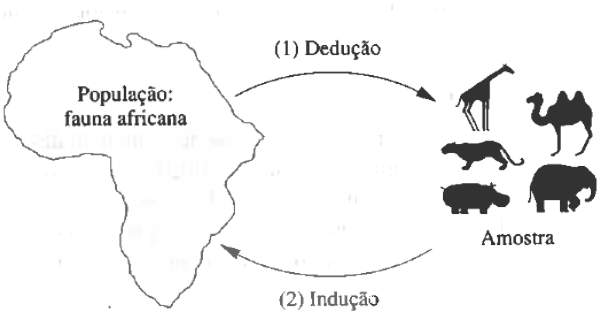


Slide 1.21

### 1.4.5 Estabelecimento de inferências sobre a população

- Com base na informação contida numa (ou mais) amostra(s) retirar *conclusões* relativas a uma (ou mais) população(ões)
- associar-lhes um *grau de credibilidade* ou certeza

$\Longrightarrow$ *Estatística indutiva* (amostra $\longrightarrow$ população)




Slide 1.22

**Exemplos**

- A partir de uma amostra aleatória constituída pelas alturas de 120 portugueses adultos do sexo masculino, pretende-se estimar qual a proporção de alturas de indivíduos naquelas condições que têm um valor superior a 1.70m
- Com base num estudo de mercado, estima-se que um determinado produto atingirá, no próximo ano, uma quota de 35% do mercado nacional. Não faz sentido perguntar se esta estimativa será verdadeira ou falsa. Interessará é saber qual o rigor e a credibilidade que se lhe devem atribuir.

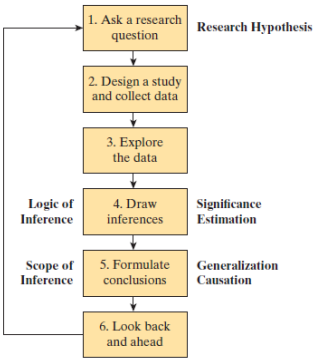
  Por exemplo, para dimensionar a capacidade de produção para o próximo ano, será razoável admitir como praticamente certo que a quota de mercado se situará entre 33% e 35%? Ou dever-se-á ser mais cauteloso, alargando aquele intervalo?

Slide 1.23

## 1.5 "Six-Steps Statistical Investigation"

Metodologia apresentada em "*Introduction to Statistical Investigations*", Nathan Tintle, *et al.*, Wiley, com maior ênfase colocado quer na definição da "research question" quer no estabelecimento da lógica e no âmbito da inferência.

Six-Step Statistical Investigation Method




Slide 1.24

**"Six-Steps Method"**

1. *Ask a research question* that can be addressed by collecting data. These questions often involve comparing groups, asking whether something affects something else, or assessing people's opinions.
2. *Design a study and collect data.* This step involves selecting the people or objects to be studied, deciding how to gather relevant data on them, and carrying out this data collection in a careful, systematic manner.
3. *Explore the data*, looking for patterns related to your research question as well as unexpected outcomes that might point to additional questions to pursue.
4. *Draw inferences* beyond the data by determining whether any findings in your data reflect a genuine tendency and estimating the size of that tendency.
5. *Formulate conclusions* that consider the scope of the inference made in Step 4. To what underlying process or larger group can these conclusions be generalized? Is a cause-and-effect conclusion warranted?
6. *Look back and ahead* to point out limitations of the study and suggest new studies that could be performed to build on the findings of the study.

Slide 1.25

### 1.5.1 Exemplo: "Recruiting Organ Donors"

- While a majority of people approve of organ donation in principle, far less than that actually sign up when getting a driver's license.
- Different states (and different countries) have different recruiting methods.
- Do these different methods result in different sign-up rates?

Obs. Nos Estados Unidos é habitual a opção entre ser ou não ser dador de orgãos ser feita no momento que se pede a carta de condução, já que não existe nem cartão de cidadão nem bilhete de identidade.

***Step 1.*** **Ask a Research Question**

- In general: Is there a method that will increase the likelihood that a person agrees to become an organ donor.
- More specifically: Does the default option presented to driver's license applicants influence the likelihood of someone becoming an organ donor?

Slide 1.26

***Step 2.*** **Design a study and collect data**

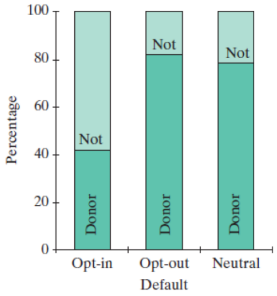
- The researchers decided to recruit various participants and ask them to pretend to apply for a new driver's license.
- The participants did not know in advance that different options were given for the donor question, or even that this issue was the main focus of the study.

- Some of the participants were forced to make a choice of becoming a donor or not, without being given a default option (the "neutral" group, Michigan's current practice).
- Other participants were told that the default option was not to be a donor but that they could choose to become a donor if they wished (the "opt-in" group, Michigan's past practice).
- The remaining participants were told that the default option was to be a donor but that they could choose not to become a donor if they wished (the "opt-out" group, some countries use this practice).

Slide 1.27

***Step 3.*** **Explore the data**

- 23 of 55 (41.8%) participants in the opt-in group agreed to become organ donors
- 41 of 50 (82.0%) participants in the opt-out group agreed to become organ donors
- 44 of the 56 (78.6%) participants in the neutral group agreed to become organ donors




Slide 1.28

***Step 4.*** **Draw inferences beyond the data**

- Using methods that you will learn in this course, the researchers analyzed whether the observed differences between the groups was large enough to indicate that the default option had a genuine effect.
- In particular, they reported strong evidence that the neutral and opt-out versions do lead to a higher chance of agreeing to become a donor, as compared to the opt-in version currently used in many states.
- In fact, they could be quite confident that the neutral version increases the chances that a person agrees to become a donor by between 20 and 54 percentage points, a difference large enough to save thousands of lives per year in the United States.

Slide 1.29

***Step 5.*** **Formulate conclusions**

- Based on the analysis of the data and the design of the study, the researchers concluded that the neutral version causes an increase in the proportion who agree to become donors over the opt-in.
- But because the participants in the study were volunteers recruited from various general interest Internet bulletin boards, generalizing conclusions beyond these participants is only legitimate if they are representative of a larger group of people. (The authors believed their sample included a "broad range of demographics.")

Slide 1.30

***Step 6.*** **Look back and ahead**

- One limitation of the study is that participants were asked to imagine how they would respond, which might not mirror how people would actually respond in such a situation.
- A new study might look at people's actual responses to questions about organ donation or could monitor donor rates for states that adopt a new policy.
- Researchers could also examine whether presenting educational material on organ donation might increase people's willingness to donate.
- Another improvement would be to include participants from wider demographic groups than these volunteers.

Slide 1.31

### 1.5.2 "Four Pillars of Statistical Inference"

**"Four Pillars of Statistical Inference"**

1. *Significance*: How *strong* is the evidence of an effect? You will learn how to provide a measure of the strength of the evidence provided by the data. For example, how strong is the evidence that the neutral and opt-out versions increase the chance of agreeing to become an organ donor, as compared to the opt-in version?
2. *Estimation*: What is the *size* of the effect? You will learn how to estimate how different two groups are. For example, how much larger (if at all) are the chances someone agrees to donate organs when asked with the neutral version instead of the opt-in version?
3. *Generalization*: How *broadly* do the conclusions apply? You will learn to consider what larger group of individuals you believe your conclusions can be applied to. For example, what are the characteristics of the individuals who participated in the organ donation study, and how are they similar or different than typical drivers?
4. *Causation*: Can we say what *caused* the observed difference? You will learn whether you can legitimately identify what caused the observed difference. For example, can you conclude that the way the researchers asked the organ donation question was the cause of the observed differences in proportions of donors?

Slide 1.32

## 1.6 Caso de Estudo: "Learning About Lottery Strategies"

- Forms of lotto are played world-wide and many people have theories about how to make money at the game
- We will examine a particular lotto game, to see whether it might be possible to play it profitably
- The game we will look at is the daily Pick-3 lottery run by the state of New Jersey in the USA

Slide 1.33

**Playing "Pick-3" Lotto**

- Each player selects a three digit number between 000 and 999
- A winning number is selected by independently picking three digits between 0 and 9 at random
- All players who hold the winning numbers split the prize money for the game; the size of the prize depends on the number of players who choose the winning numbers

**Choosing a Winning Strategy**

- The results of the games (winning number and winning amount) are publicly available
- Does this data contain information which will enable us to choose a profitable strategy for this game?
- We will use the results of 254 consecutive games to look for a profitable strategy (`http://www.state.nj.us/lottery/games/1-6_pick3.shtml`)

Slide 1.34

**Winning Numbers and Amounts**

(810, $190.0), (156, $120.5), (140, $285.5), (542, $184.0), (507, $384.5),
(972, $324.5), (431, $114.0), (981, $506.5), (865, $290.0), (499, $869.5),
(020, $668.5), (123, $83.0), (356, $188.0), (015, $449.0), (011, $289.5),
(160, $212.0), (507, $466.0), (779, $548.5), (286, $260.0), (268, $300.5),
(698, $556.5), (640, $371.5), (136, $112.5), (854, $254.5), (069, $368.0),
(199, $510.0), (413, $102.0), (192, $206.5), (602, $261.5), (987, $361.0),
(112, $167.5), (245, $187.0), (174, $146.5), (913, $205.0), (828, $348.5),
(539, $283.5), (434, $447.0), (357, $102.5), (178, $219.0), (198, $292.5),
(406, $343.0), (079, $332.5), (034, $532.5), (089, $445.5), (257, $127.0),
(662, $557.5), (524, $203.5), (809, $373.5), (527, $142.0), (257, $230.5),
(008, $482.5), (446, $512.5), (440, $330.0), (781, $273.0), (615, $171.0),
(231, $178.0), (580, $463.5), (987, $476.0), (391, $290.0), (267, $176.0),
(808, $195.0), (258, $159.5), (479, $296.0), (516, $177.5), (964, $406.0),
(742, $182.0), (537, $164.5), (275, $137.0), (112, $191.0), (230, $298.0),
(310, $110.0), (335, $353.0), (238, $192.5), (294, $308.5), (854, $287.0),
(309, $203.5), (026, $377.5), (960, $211.5), (200, $342.0), (604, $259.0),
(841, $231.0), (659, $348.0), (735, $159.0), (105, $130.5), (254, $176.0),
(117, $128.5), (751, $159.0), (781, $290.0), (937, $335.0), (020, $514.0),
(348, $191.0), (653, $304.5), (410, $167.0), (468, $257.0), (077, $640.0),

(921, $142.0), (314, $146.0), (683, $356.0), (000, $96.0), (963, $295.0),
...

Slide 1.35

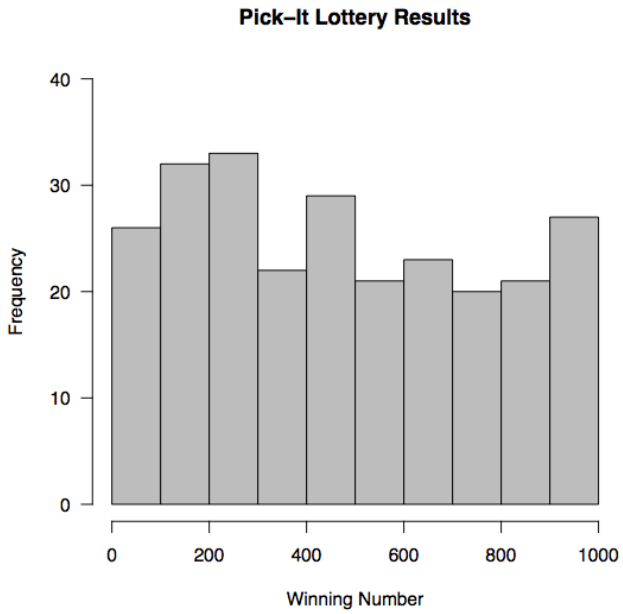
**How Visualisation Helps**

- Humans can really only make sense of three or four numbers at a time
- By representing the values in a graphical form we make it easier to handle large numbers of values
- Using graphs should make it possible to learn more about this data

**Choosing Good Numbers**

- One approach to making money at "Pick-3" is to try to select numbers which are more likely to win
- Since we have data on the winning numbers we can look at the distribution of the winning numbers and see whether some ranges of values are more like to produce a winner than others
- One way to do this is to produce a *histogram* of the winning numbers

Slide 1.36




Slide 1.37

**Analysis**

- It looks there tend to be more winners in the region from 100 to 300 than in other regions
- This suggests that we might be best to choose numbers in this range as they are more likely to be drawn as winners than other values
- We have to be careful about making this kind of judgement because the winning numbers are chosen randomly and we can expect to sometimes see clusters of winning number
- To judge the significance of what we see in the histogram we must resort to formal statistical theory

Slide 1.38

**Statistical Variability**

- There are 254 values; we would expect the number of values in each cell (column) to be approximately:

$$254 \times \frac{1}{10} = 25.4$$

- Statistical theory $^{(*)}$ tells us that, if the winning ticket is chosen randomly, the level of variability in each cell is:

$$\sqrt{254 \times \frac{1}{10} \times \frac{9}{10}} = 4.78 \quad ^{(**)}$$

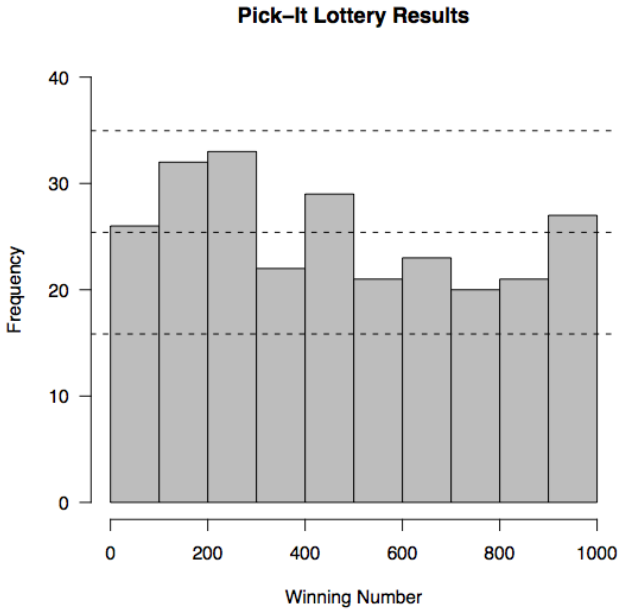
- Discrepancies of up to twice this amount can be attributed to "random variability" $^{(***)}$

$^{(*)}$ see chapter about confidence intervals for further details

$^{(**)}$ see the binomial distribution for further details

(***) random variability means by chance

Slide 1.39

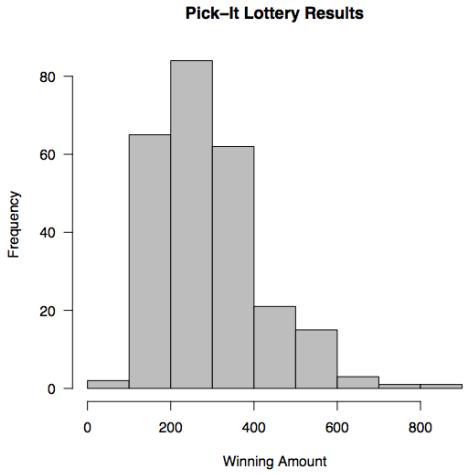



Slide 1.40

## "Pick-3" Lotto – Initial Conclusions

- There is no reason to believe that the winning values are anything other than random
- This says that we have no reason to believe that any particular value is more likely to win than any other
- Since we can't choose values which are more likely to win than others, we might instead see if we can influence the amount won

Slide 1.41

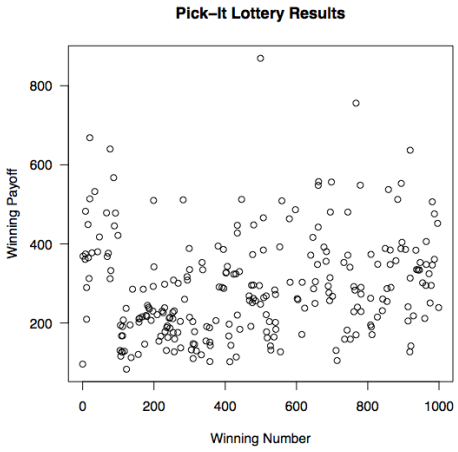



Slide 1.42

**Winning Amounts**

- The histogram shows that there is a wide range of amounts won in the game
- This suggests that it *might* be possible to choose the numbers which win larger amounts
- To do thus we need to see if there is some relationship between ticket number and winning amount
- A scatter plot is the natural way to look for such a relationship

Slide 1.43

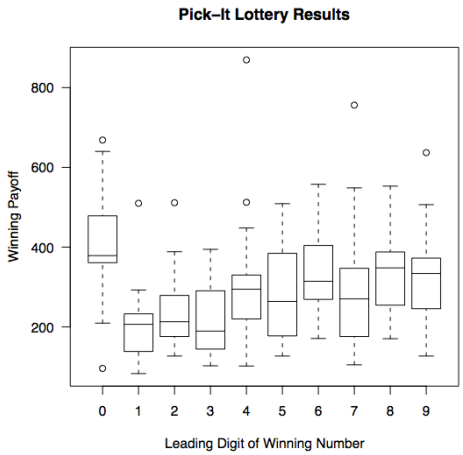



Slide 1.44

**Scatterplot Features**

- The winning amounts in a band to the left of the plot appear to generally be higher than those in the rest of the plot
- We can investigate this further by separating the values into groups according to the first digit of the ticket number and drawing *box plots* for each group

Slide 1.45

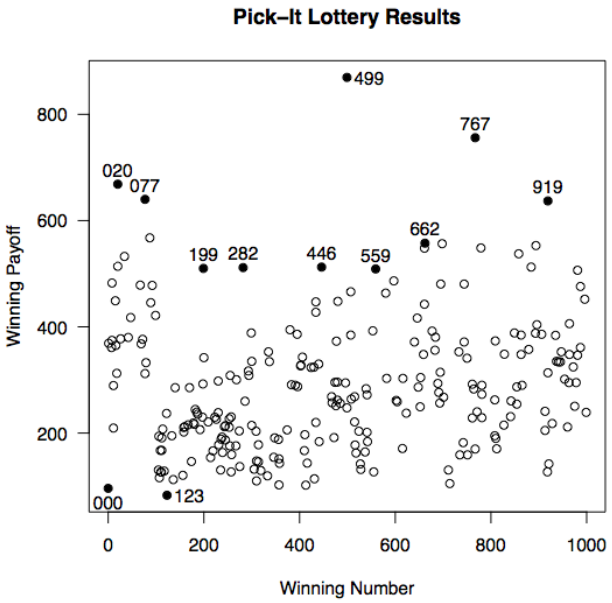



Slide 1.46

## "Pick-3" Lotto – Conclusions

- Tickets with a leading zero digit clearly tend to produce larger winnings
- It is also apparent that there are some very large and some very small winning amounts
- Probably, it is interesting to identify the ticket numbers corresponding to these extremes

Slide 1.47




Slide 1.48

### Choosing a "Pick-3" Lotto Strategy

- Choose numbers which are less likely to be chosen by other players
- Then, when you win, you will tend to win more
- Possible ways to choose:
  - Choose a number with a leading zero
  - Choose a number with repeated digits
  - Avoid "obvious" numbers
    (e.g., 000, 123, 246, . . . )

Slide 1.49