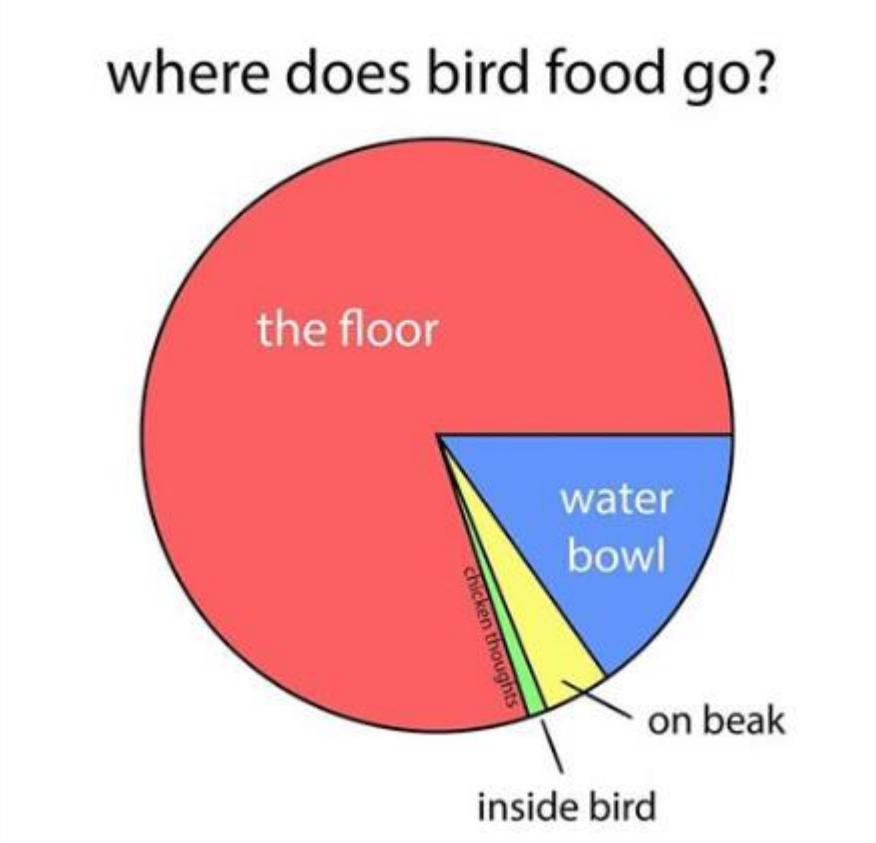
Goodies*

* Goodies related to animals, plants and numbers…



https://www.instagram.com/chickenthoughtsofficial

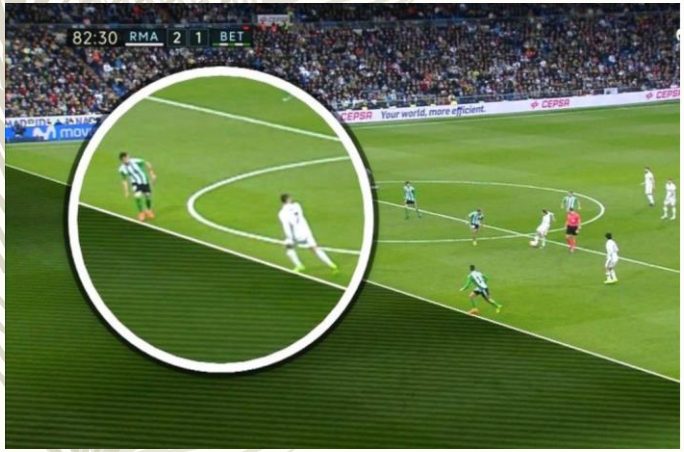
Os queijos ou *pie charts* talvez sejam os piores gráficos do mundo…

Axioma: toda a Informação contida num *pie chart* pode ser melhor transmitida através de outro tipo de gráfico! (mais à frente veremos exemplos…)

Sobre os erros… outra vêz… a ver se fica menos confuso que ontem

Hipótese nula (não há efeito, deixa jogar): não há fora de jogo

Realidade: não há fora de jogo
Arbitro marcou fora de jogo
É um falso positivo – erro de tipo I



https://cronaldodaily.com/2808/ronaldo-falsely-denied-brace-vs-real-betis-watch/

Tipo I: pensaram que havia um lobo, quando não havia!

Realidade: há fora de jogo
Arbitro não marcou for a de jogo
É um falso negativo – erro de tipo II

https://www.goal.com/en/news/offside-cristiano-ronaldo-goal-leaves-football-fans/9lnh8qqwuq9q1ogktzc7c3zgh

Tipo II: pensaram que não havia um lobo, quando havia!

A ECOLOGIA NUMÉRICA É IMPORTANTE

Já se tinham esquecido…?

Vejamos mais un(s) exemplos … !

(e-mail recebido **21 09 2018**) Olá Professor Tiago,

"O meu nome é **, sou aluna de PhD da Universidade de *** em fase final e preciso de fazer umas análises em R... O meu conhecimento em R é absolutamente zero e falei disso à *** que me recomendou que falasse consigo.
…centra-se na técnica de environmental DNA metabarcoding para detectar as espécies de peixe existentes no Rio *** (SE Asia). …amostras de água em vários pontos ao longo do rio, em duas épocas (wet e dry season. Por razões logísticas, em alguns locais as amostras de água foram apenas recolhidas numa das season) e em cada ponto de amostragem recolhi água à superfície (surface) e no fundo (deep). O meu objectivo é:

1) Perceber a diversidade de espécies existentes no Rio;
2) Perceber se existem diferenças estatisticamente significativas nas espécies detectadas entre seasons e entre surface and deep waters …

…foi-me dito que preciso de fazer análises de **site occupancy** e read counts … usando **ggplot**…o **teste chi-square** analisando a diversidade de espécies … uma vez que é impossível fazer uma **ANOVA de 2 factores** (Season - Wet e Dry; Depth Profile - Surface and Deep) já que em alguns sites (locais de amostragem) não foram recolhidas amostras de água nas duas season. … **multi-dimensional scaling**, para ser mais fácil visualizar os dados.

Tenho feito pesquisa e tentado "entender-me" com o R, mas confesso que não está nada fácil **e ajuda aqui é zero**."

(e-mail recebido **21 09 2019**) Olá, Professor Tiago!

Sou um estudante de doutoramento de Biologia e Ecologia das Alterações Globais na Faculdade de Ciências da Universidade de Lisboa orientado por ***

Recentemente submeti um artigo ao jornal ***Animal Behaviour***, no qual um dos revisores me aconselhou a utilizar modelos variados mistos para os meus dados. Ao longo das últimas duas semanas tenho lido sobre o tema. Estou atualmente a usar o pacote **sommer** do R para fazer estas análises (já que o pacote mais falado na literatura para fazer estas análises, o **ASReml**, é comercial), e julgo ter obtido algum sucesso. Mas estou com algumas dificuldades em interpretar a matriz de variância-covariância que a função me dá. Adicionalmente, também não sei se estou a fazer tudo corretamente, porque nunca antes tinha usado este pacote ou sequer feito modelos multivariados. Outra coisa que me está a incomodar é que já consegui estimar a correlação entre as variáveis de resposta do modelo, mas a função que uso não me permite calcular estimativas de erro destas correlações (a função *mmer* estima o erro-padrão das variâncias-covariâncias, mas não sei como replicar este erro-padrão para as correlações, ou se isto é sequer possível).

Resumindo, como sei que trabalha com R, gostava de saber se estava disposto a reunir-se comigo durante meia hora, ou uma hora, no seu gabinete para a semana para me ajudar a resolver estes problemas. Eu imagino que esteja ocupado com muitas outras coisas de momento, mas **foram os meus orientadores que me recomendaram perguntar-lhe** se me podia ajudar a desbloquear esta situação. Obrigado pela sua atenção :)

Taylor & Francis

The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: https://www.tandfonline.com/loi/utas20

The Impact of Results Blind Science Publishing on Statistical Consultation and Collaboration

Joseph J. Locascio

To cite this article: Joseph J. Locascio (2019) The Impact of Results Blind Science Publishing on Statistical Consultation and Collaboration, The American Statistician, 73:sup1, 346-351, DOI: 10.1080/00031305.2018.1505658

To link to this article: https://doi.org/10.1080/00031305.2018.1505658

will become increasingly needed as a very practical necessity, even if, regrettably, not for more elevated reasons.

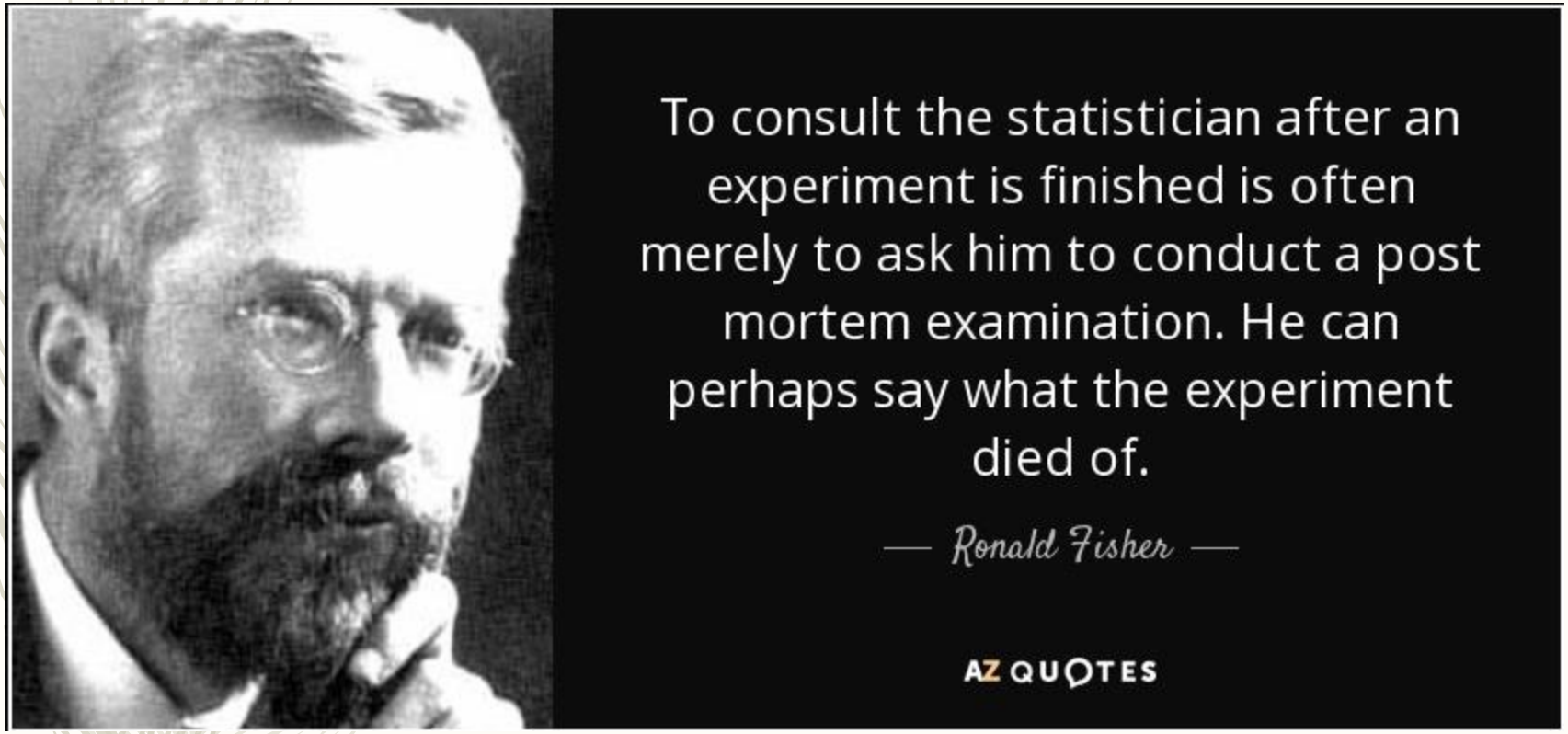
4.1. Statisticians Would Become More Involved in Research Studies

I am a member of a university based biostatistics consulting group, which provides statistical consultation and analysis assistance for studies conducted across research facilities throughout a large metropolitan area. In this role, it is more often the case than not, that I'm not merely asked to answer specific questions on an otherwise fully developed research study. Rather, oftentimes, the investigators have a somewhat vague idea for a study they wish to conduct, or are articulating it poorly, and I find myself trying to solve an equation in two unknowns. First, I need to clarify for myself, and for the investigators, what exactly the research question and hypotheses are, and then

4.2. Statisticians Would Become Involved in Studies Earlier

In my role as statistical consultant for research, I'm often asked for assistance very late in the game, so to speak. Someone has designed a study and collected a lot of data, and now wants to talk about statistical analysis. Or else, revisions to a manuscript submitted to a journal have been requested by reviewers including or especially regarding data analysis, and the investigators need the advice of a statistician and/or help in analysis. In these cases, oftentimes I feel what the study really needs is a new design and new data, more fitting to the purpose of the study, once I discover what that is, but it's too late for that. I try to be sensitive and do the best I can by recommending and conducting ad hoc statistical patch-ups, making clear the limitations of these methods, but these situations can be frustrating. Under RBME, investigators will know in a very practical sense that they

Ecologia Numérica - Aula Teórica 3 – 24-09-2018



https://www.azquotes.com/quote/97013

“Tipos” de estudos científicos

Unidisciplinares

Multidisciplinares

Interdisciplinares

https://www.researchgate.net/post/What_is_the_difference_between_interdisciplinary_and_multidisciplinary_research

"Tipos" de estudos científicos

- Descritivos vs. Experimentais
- Clássicos vs. Inovadores
- Fundamentais vs. Aplicados
- Importância regional vs. Importância global

Hipóteses sem dados não têm utilidade!

mas…

Dados sem hipóteses também não!

Que componentes deverá ter um programa de investigação?

Como podemos recolher os dados para responder à pergunta que queremos responder?

Onde está a informação nos dados para obter a resposta à pergunta que queremos responder?

> An approximate answer to the right question is worth a great deal more than a precise answer to the wrong question.
>
> John Tukey

introdução à análise de dados
o método científico

OBSERVAÇÕES
Padrões no espaço e/ou tempo

MODELOS
Explicações ou teorias

HIPÓTESES
Previsões baseadas no modelo

HIPÓTESE NULA (H_0)
Oposição lógica à hipótese de interesse

EXPERIÊNCIA
Teste da hipótese nula

INTERPRETAÇÃO

Uma decisão fraca

NÃO REJEITAR H_0
Rejeita a hipótese de interesse e modelo

Uma decisão forte

REJEITAR H0
Suporta a hipótese de interesse e modelo

introdução à análise de dados
o método científico

OBSERVAÇÃO
Um peixe salta fora de água

MODELO
Evitar a predação por peixes maiores

HIPÓTESE
O peixe irá saltar quando for adicionado um peixe maior

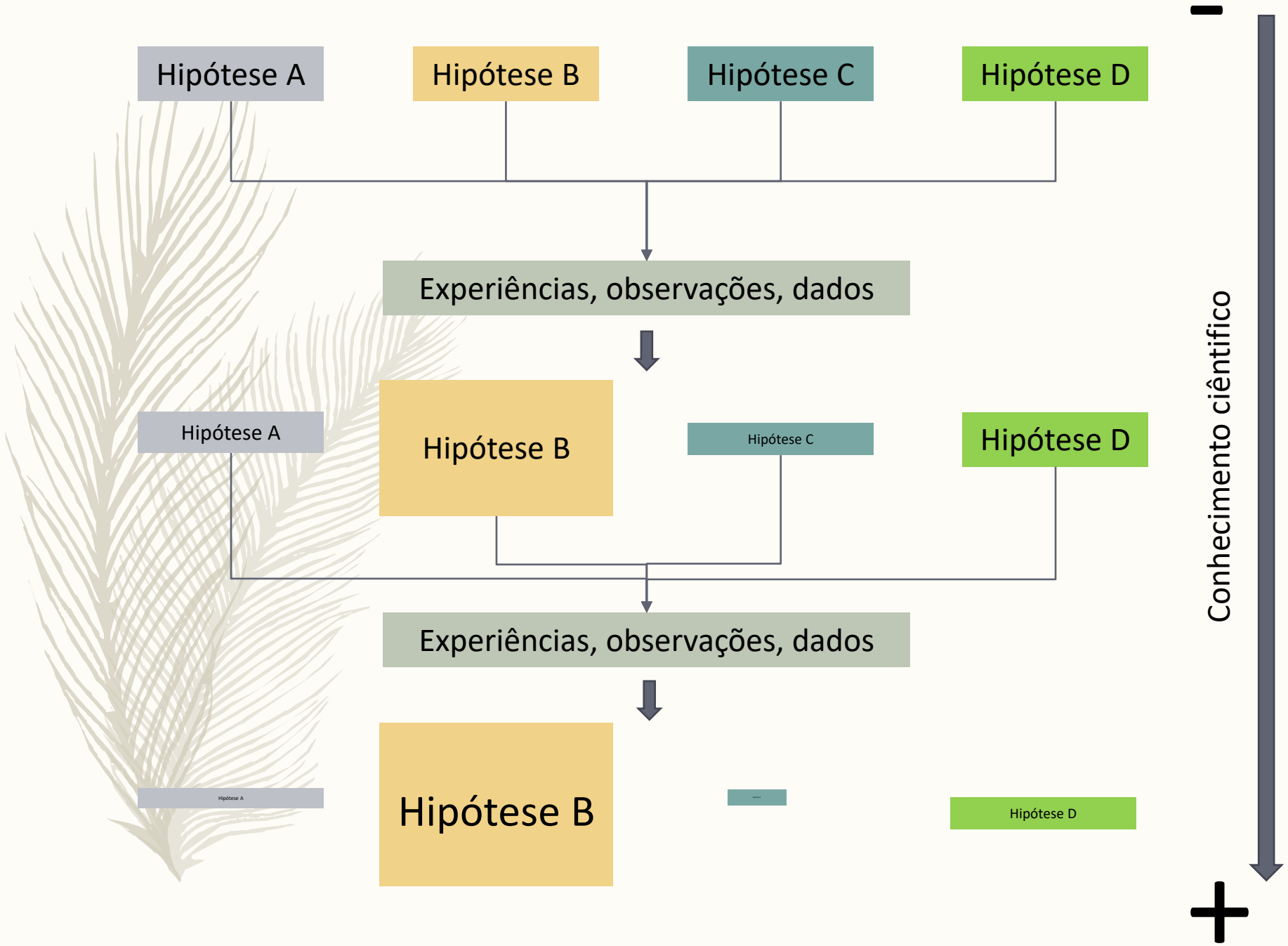
HIPÓTESE NULA (H_0)
Não há diferenças no comportamento quando é adicionado um peixe maior

EXPERIÊNCIA
Um conjunto de tanques com e sem predador

INTERPRETAÇÃO

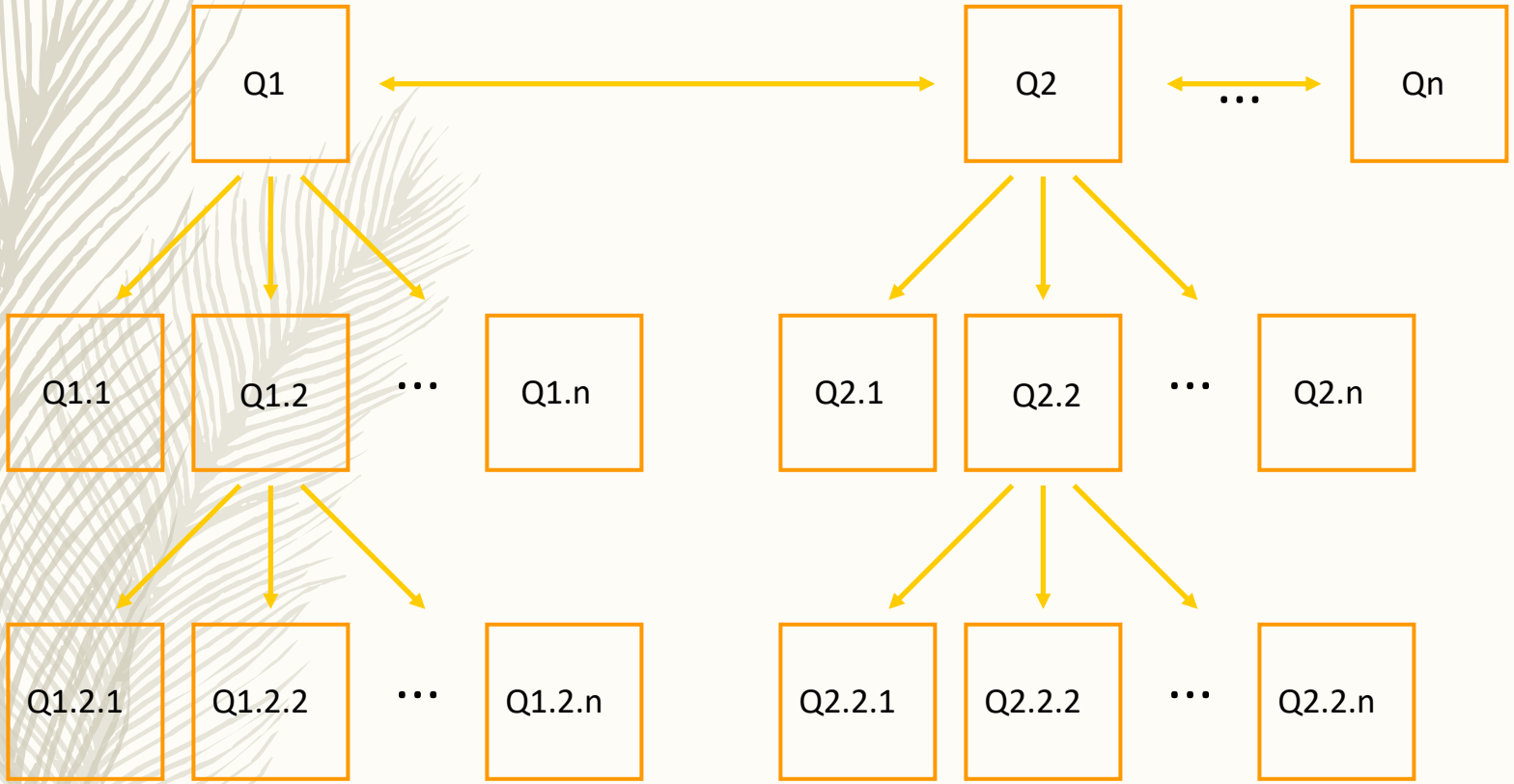
REJEITAR H_0
Admite-se a hipótese e o modelo

Output científico, novas questões


Hipótese A
Hipótese B
Hipótese C
Hipótese D
Experiências, observações, dados
Hipótese A
Hipótese B
Hipótese C
Hipótese D
Experiências, observações, dados
Hipótese B
Hipótese D
Conhecimento científico

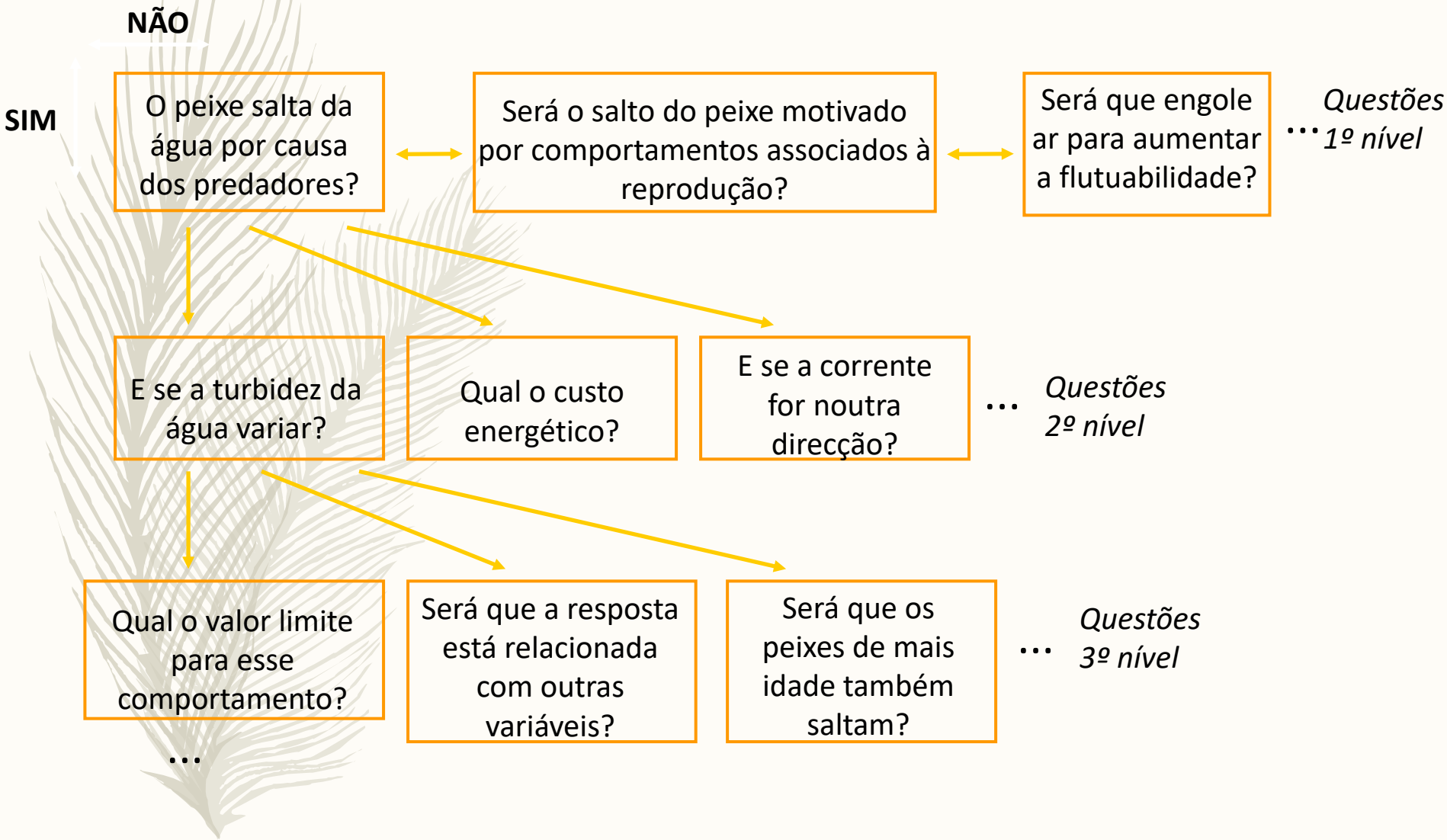
introdução à análise de dados
o método científico

Complementaridade, Comparações, Generalizações

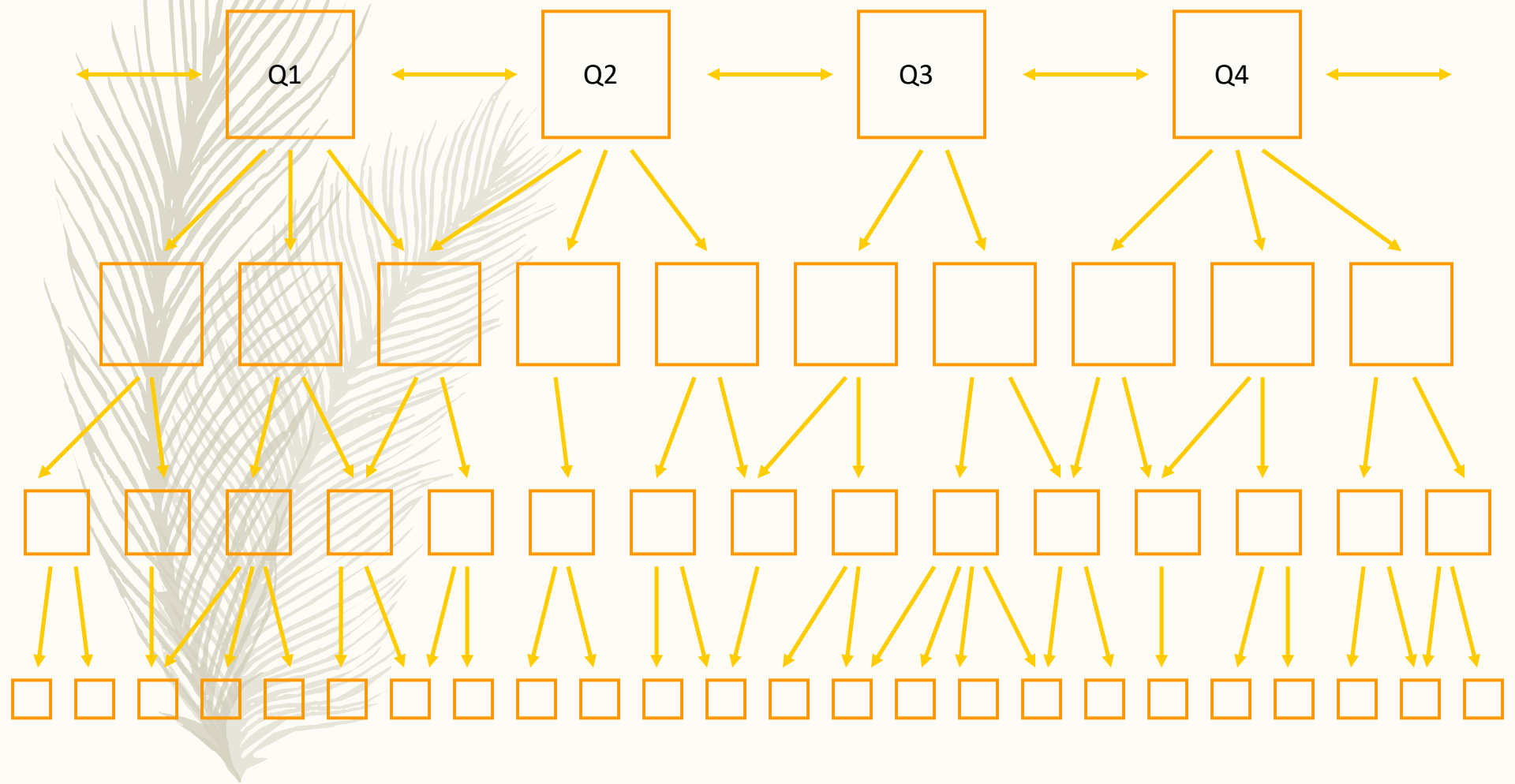


Progresso científico
Aumenta o conhecimento
Aumenta o detalhe
Inovação

introdução à análise de dados
o método científico

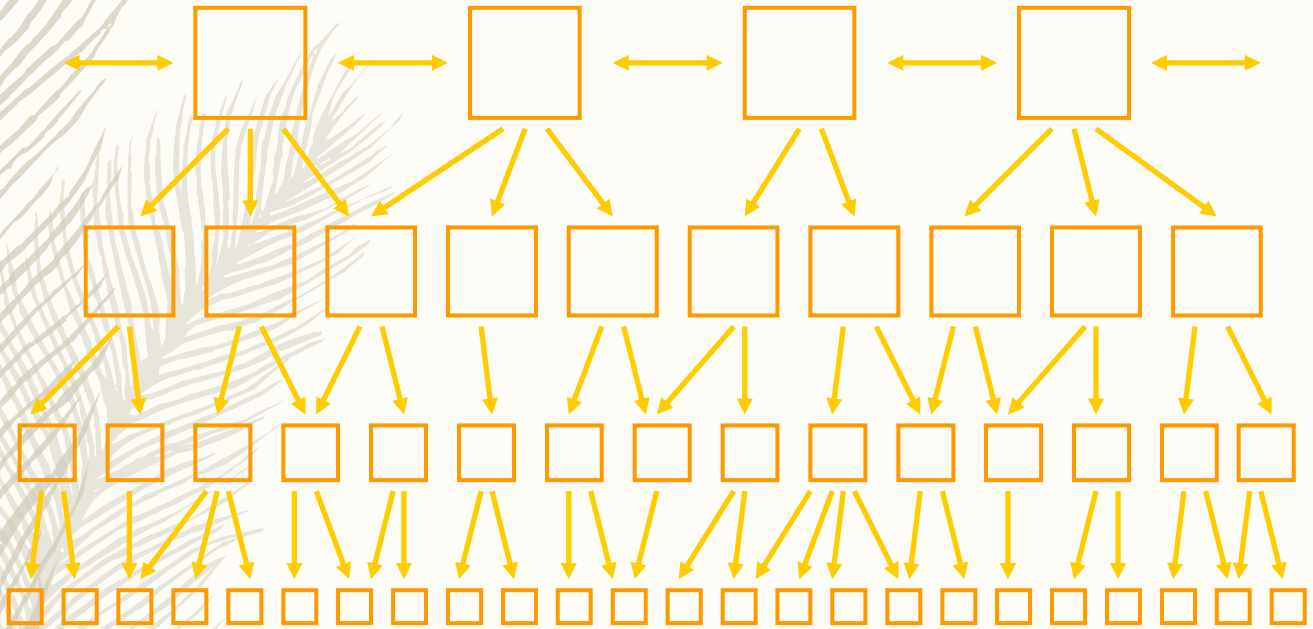


introdução à análise de dados
o método científico



introdução à análise de dados
o método científico

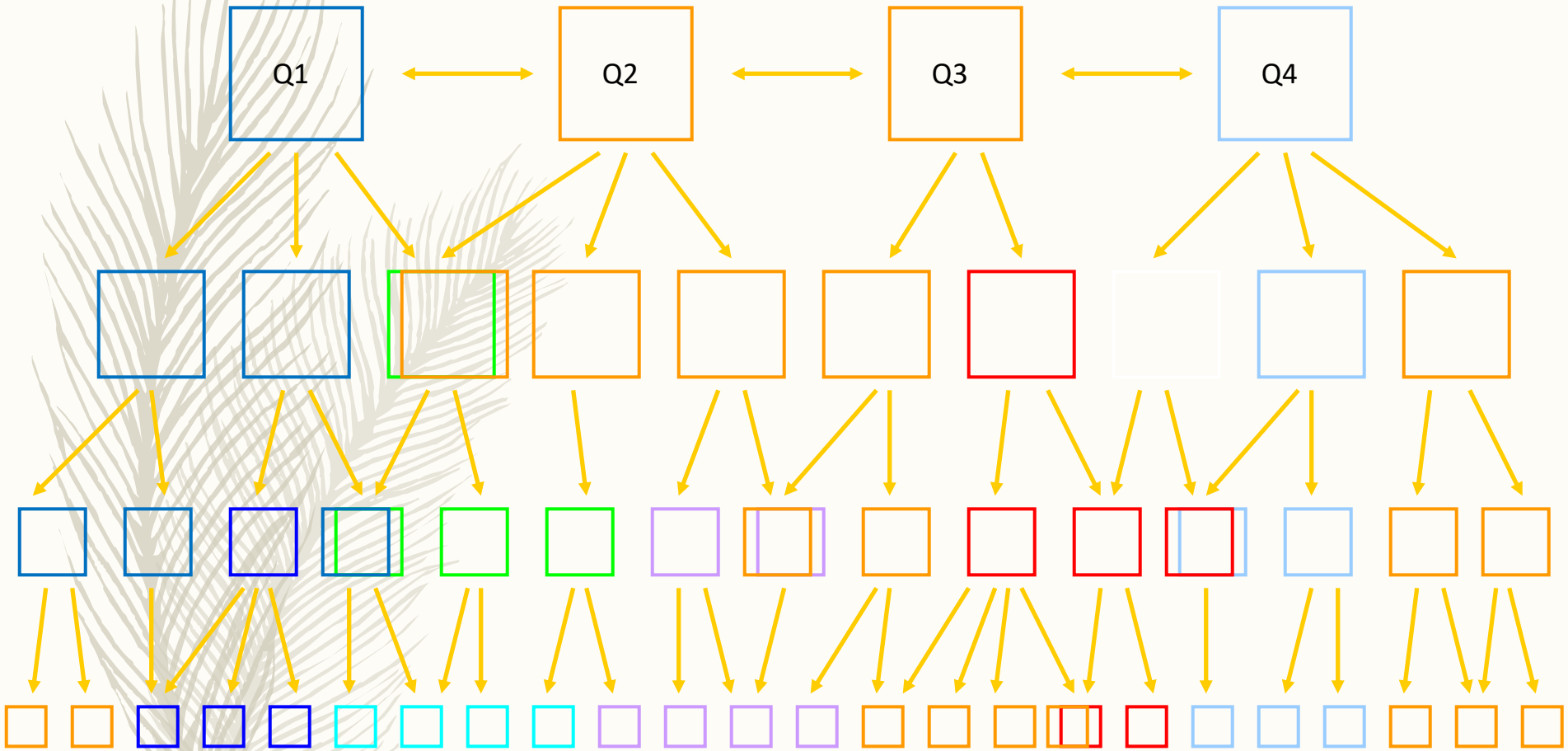
Conhecimento básico



Conhecimento avançado

De certo modo, com o progresso científico… o conhecimento avançado torna-se básico

introdução à análise de dados
o método científico



COMPLEMENTARIDADE, COLABORAÇÃO, COMPETIÇÃO, AVALIAÇÃO SÃO COMPONENTES ESSENCIAIS NO SEIO DA COMUNIDADE CIENTÍFICA

A ecologia numérica vive de dados

Hoje em dia é muito simples encontrar dados ecológicos, mesmo que não os recolhamos nós próprios.

Existem inúmeros recursos possíveis:

- Data Journals
  - Scientific Data
  - Biodiversity Data Journal
  - Large list of data journals here
- Repositórios online de dados – e.g. https://datadryad.org/, https://data.mendeley.com/ , etc.
- Repositórios específicos – e.g. LTER, Biotime, etc.
- Dados arquivados como suplementos de artigos

Por isso, não há desculpas para não analisar dados e praticar o que aprenderem nas aulas de Ecologia Numérica.

LTER Network Data Portal

HOME DATA TOOLS HELP LOGIN

enter search terms

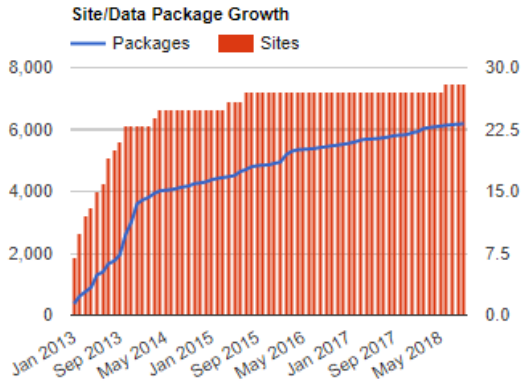
ADVANCED SEARCH

Welcome to the LTER Network Data Portal

Data are one of the most valuable products of the Long Term Ecological Research (LTER) Network. Data and metadata derived from publicly funded research in the U.S. LTER Network are made available online with as few restrictions as possible, on a non-discriminatory basis. In return, the LTER Network expects data users to *act ethically* by contacting the investigator prior to the use of data for publication.

The LTER Network Information System Data Portal contains ecological data packages contributed by past and present LTER sites. Please review the LTER Network Data Access Policy before downloading any data product. We request that you cite data sources in your published and unpublished works whenever possible. Digital object identifiers (DOI) are provided for each dataset to facilitate citation.

LTER Network scientists make every effort to release data in a timely fashion and with attention to accurate, well-designed and well-documented data. To understand data fully, please read the associated metadata and contact data providers if you have any questions. The LTER Network is not responsible for misinterpretation of data resulting from failure to consult metadata or data providers.



Site contributed data packages: **6180**

Total data packages: **42828**

https://portal.lternet.edu/nis/home.jsp

BioTIME

An ERC Funded Project

University of St Andrews | FOUNDED 1413

Sear

Home People Fieldwork BioTIME Database Publications Useful Links

We are assembling a database of biodiversity time series. Our intention is to make a database that can be accessed by anyone, with appropriate permissions by data owners. We invite you to contribute your data, and to help us explore the database and understand how biodiversity has changed worldwide.

https://synergy.st-andrews.ac.uk/biotime/biotime-database/

TPC: trabalho para casa

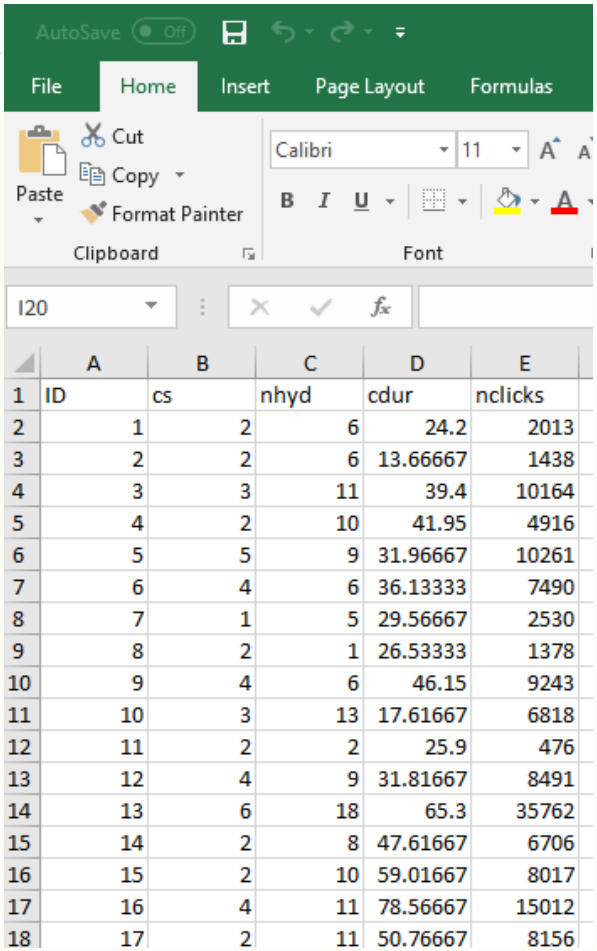
- Formular uma pergunta ecológica
- Recolher um conjunto de dados “ecológicos”, com um tamanho de amostra pelo menos igual a 30, idealmente maior que 50
- Registar (pelo menos) duas variáveis que possam ser comparadas, e uma variável que possa ser relacionada com as anteriores
- Exemplos:
  - selecionar 50 árvores. Recolher 2 folhas, uma numa posição mais baixa e outra numa posição mais alta, de cada árvore e o dap (diâmetro à altura do peito) de cada árvore
  - selecionar 50 plantas com flores. Medir a altura ao solo da planta. Selecionar a flor mais alta e a mais baixa. Contar quantos insetos há em cada uma das flores.
  - Selecionar 50 pombos. Registar se é macho ou fêmea. Registar se está só ou acompanhado. Andar em direção a cada pombo e registar a distância a que ele “para e olha” para avaliar o perigo.
  - Selecionar 50 formigas num carreiro, registar se vão para o ninho ou se afastam do mesmo, se tem ou não algo a ser transportado, a temperatura do ar e que distancia percorrem em 30 segundos (obviamente, tem de ser feito em dias diferentes… porquê?)

TPC: trabalho para casa

- Criar um ficheiro Excel com o seguinte nome:
  - 3 letras do primeiro nome + 3 letras do ultimo + número de aluno,
    - exemplo no meu caso, TiaMar19549.xlsx (ver no Fenix)
  - A primeira coluna vai-se chamar ID e conter os números 1 a n, em que n é o número de unidades de amostragem
  - Criar tantas colunas quantas variáveis recolhidas
  - O nome de cada variável deverá ter no máximo 5 letras (todas minúsculas)

TPC: trabalho para casa

TiaMar19549.xlsx



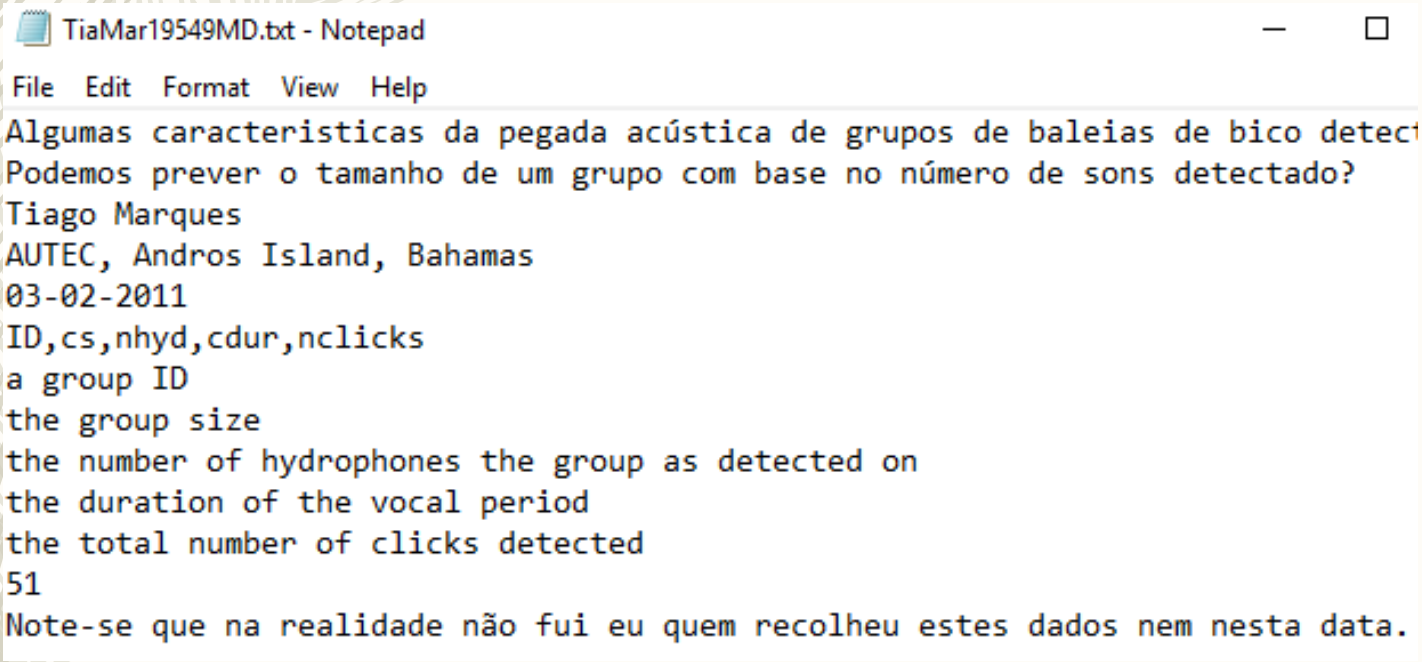
	A	B	C	D	E
1	ID	cs	nhyd	cdur	nclicks
2	1	2	6	24.2	2013
3	2	2	6	13.66667	1438
4	3	3	11	39.4	10164
5	4	2	10	41.95	4916
6	5	5	9	31.96667	10261
7	6	4	6	36.13333	7490
8	7	1	5	29.56667	2530
9	8	2	1	26.53333	1378
10	9	4	6	46.15	9243
11	10	3	13	17.61667	6818
12	11	2	2	25.9	476
13	12	4	9	31.81667	8491
14	13	6	18	65.3	35762
15	14	2	8	47.61667	6706
16	15	2	10	59.01667	8017
17	16	4	11	78.56667	15012
18	17	2	11	50.76667	8156

TPC: trabalho para casa

- Criar um ficheiro txt com **metadados** com o seguinte nome:
  - 3 letras do primeiro nome + 3 letras do ultimo + número de aluno + MD
  - exemplo no meu caso, TiaMar19549MD.txt (ver no Fenix)
- Linha 1: Descrição sumária dos dados
- Linha 2: Questão ecológica a responder
- Linha 3: Quem recolheu (1º nome + ultimo nome + número de aluno)
- Linha 4: Onde recolheu
- Linha 5: Data da recolha
- Linha 6: Nome das k variáveis recolhidas, separados por virgulas
- Linha 7 a 7+k-1: Descrição de cada variável
- Linha 7+k: número de observações
- Linha 7+k+1: comentários

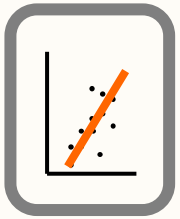
TPC: trabalho para casa

TiaMar19549MD.txt



Enviar ambos os ficheiros para tamarques@ciencias.ulisboa.pt (usar o tópico “dados”)

Tipos de variáveis e revisões sobre probabilidade



tipos de variáveis
revisões sobre probabilidades

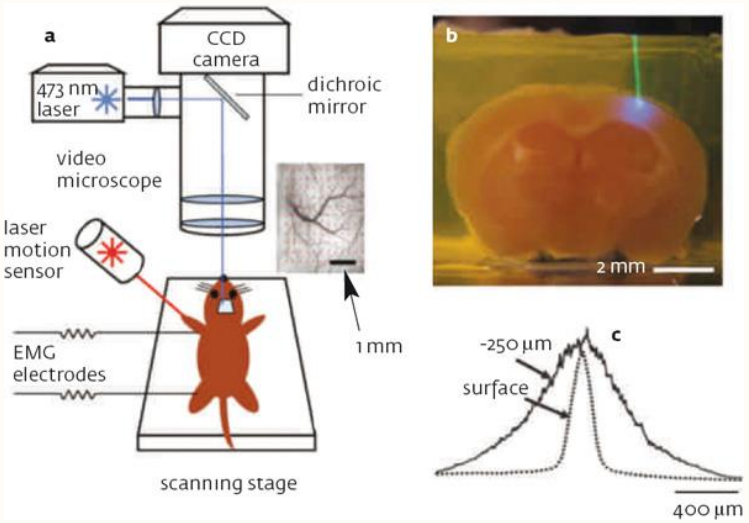
- Quais os tipos de variáveis nos estudos de ecologia?
- Qual a informação básica a obter sobre estas variáveis?
- Qual a utilidade das bases teóricas das probabilidades e estatística para a análise de dados?

- Que tipo de estudos se faz em ecologia?

tipos de variáveis
revisões sobre probabilidades

tipos de variáveis
revisões sobre probabilidades

tipos de variáveis
revisões sobre probabilidades



Escalas de medida

- Nominais
- Ordinais
- Intervalados
- Percentuais ou de razão

Escalas de medida

- Nominais
  e.g. espécie, sexo, cor

No R: fatores (factor)

- Ordinais
  - *pouco, médio, muito*
  - *muito menos, menos, igual, mais, muito mais*
- Intervalados
  e.g. temperatura, escalas circulares de tempo
- Percentuais ou de razão → Existe um zero absoluto!
  e.g. comprimento, peso, unidades de tempo, contagens

Stevens, S. 1946 On the theory of scales of measurement *Science* **103**:677-680

Dados discretos e contínuos

- Contínuos: *quando existe uma infinidade de valores possíveis entre quaisquer dois valores*
  e.g. comprimento

- Discretos: *quando existem valores impossíveis de obter entre duas medições*
  e.g. contagens

tipos de variáveis
revisões sobre probabilidades

escalas de razão, intervaladas ou ordinais → Contínuos ou Discretos

escalas nominais → Discretos

Exatidão, enviseamento e precisão

Accuracy – exatidão, fiabilidade, correcção, acurácia

Bias – viés, enviesamento

Precision – precisão

Termos usados no “Glossário Inglês-Português de Estatística” (disponivel nas refências e potencialmente util quando quizerem traduzir nomes de métodos, analises, etc)

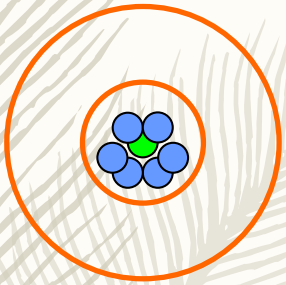
Accuracy is a qualitative term referring to whether there is agreement between a measurement made on an object and its true (target or reference) value.

Bias is a quantitative term describing the difference between the average of measurements made on the same object and its true value.

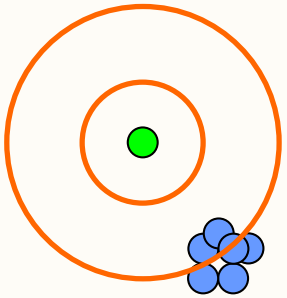
Enviesamento e Precisão (bias e precision)

- Enviesamento: *descreve a proximidade (ou mais concretamente a falta dela) entre uma medida de uma quantidade e o valor real.*
- Precisão: *é a proximidade entre sucessivas medidas a um mesmo item.*

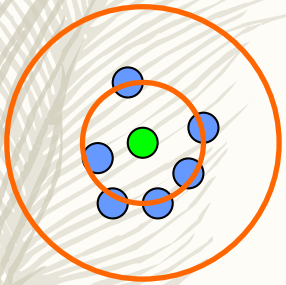
Enviesamento e Precisão



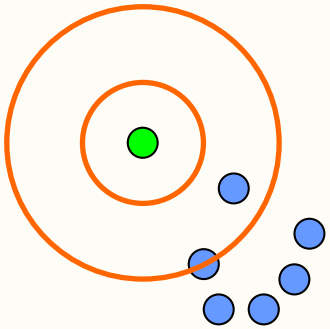
Não enviesado (correcto) e preciso



Enviesado (incorrecto) e preciso

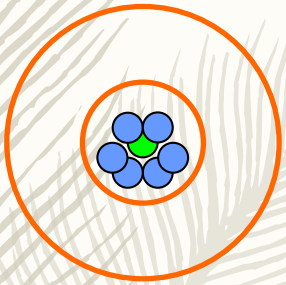


Não enviesado e pouco preciso

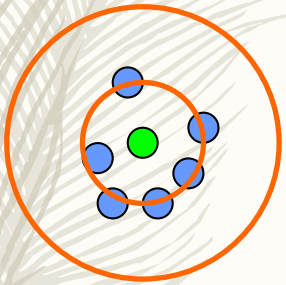


Enviesado e pouco preciso

Correcção e Precisão

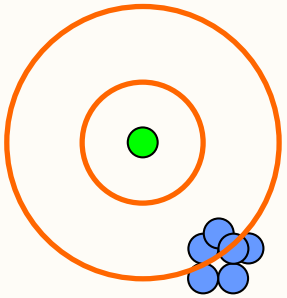


Correcto e preciso

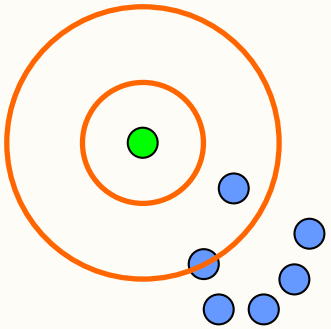


Correcto e pouco preciso

Estimativas enviesadas



Incorrecto e preciso



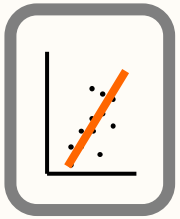
Incorrecto e pouco preciso

Números significativos

- Dados discretos: *não há dúvidas! A utilização de decimais não é apropriada*
  e.g. Contagens de organismos: 3 indivíduos, 27 indivíduos
  usar 3.0 e 27.0 seria errado.

- Dados contínuos: *são registados a um determinado nível de precisão e a utilização de diferentes números significativos tem as suas implicações*

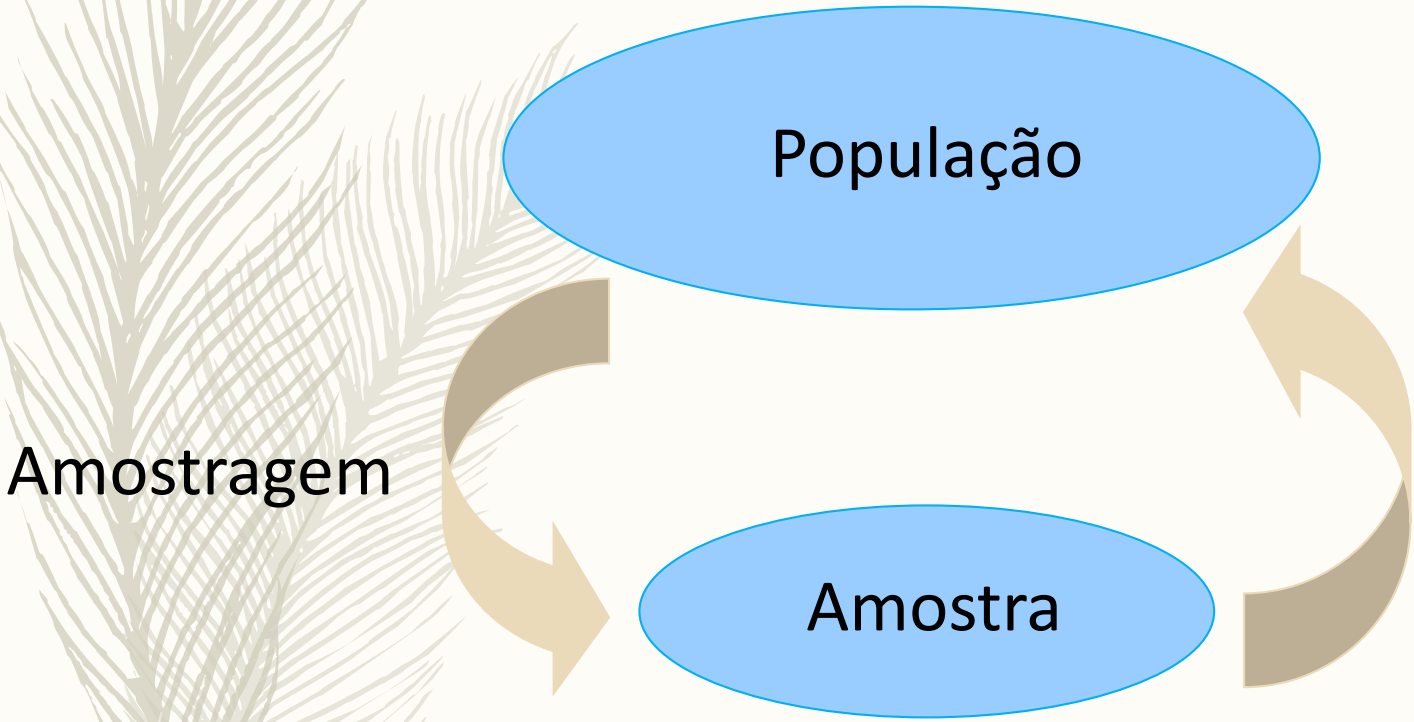
- *Eu meço 1.9 m… ou 1.899783457267362348764 m*

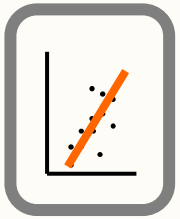


Um dos objectivos principais das análises estatísticas é fazer afirmações sobre uma qualquer população partindo de uma (pequena) amostra.

Uma quantidade tal como uma medida de tendência central ou de dispersão que caracteriza a população é denominada parâmetro.

Estimativas dos parâmetros são geralmente denominadas estatísticas.


População
Amostragem
Inferência
Amostra



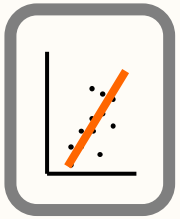
Conceitos básicos sobre amostragem

- População (população estatística)
- Unidade de amostragem
- Método de amostragem
- Amostra

Análise dos dados

- Antes de qualquer procedimento analítico mais elaborado deve proceder-se a uma análise exploratória dos dados

- Este tipo de análise permite-nos obter um maior conhecimento sobre os conjuntos de dados e identificar aspectos importantes para a selecção dos procedimentos a efectuar seguidamente



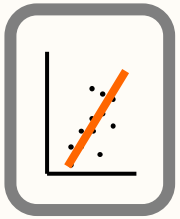
Análise exploratória de dados

- Geralmente baseada em estatísticas descritivas e representações gráficas
- As estatísticas descritivas mais frequentemente utilizadas são medidas de tendência central (e.g. média, moda, mediana) e de dispersão dos dados (e.g. variância, desvio padrão,etc.)

Distribuições de probabilidade

O que significa a probabilidade de um evento?

Embora sejam conceitos intuitivos para a generalidade das pessoas é necessário definir algumas regras.



Probabilidades

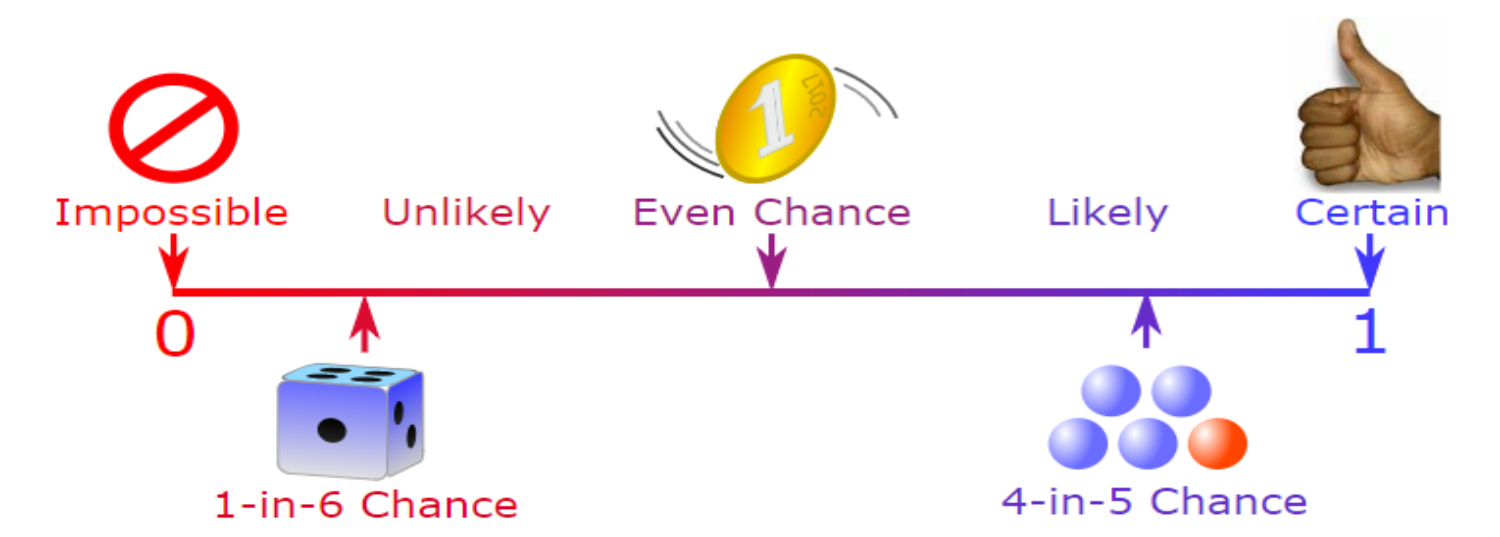
- A Probabilidade pode tomar valores entre 0 e 1
- Zero significa que esse evento é impossível
- Uma probabilidade de 1 significa que esse acontecimento é certo
- O que significa uma probabilidade intermédia?

*A probabilidade de chover amanhã é 0.25?!**

** Mas, de notar, se eu avaliar depois de amanhã, ou choveu ou não!*

Probabilidades

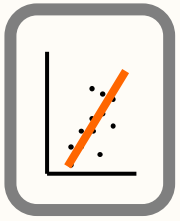
- A Probabilidade pode tomar valores entre 0 e 1



- O que significa uma probabilidade intermédia?

*A probabilidade de chover amanhã é 0.25?!**

** Mas, de notar, se eu avaliar depois de amanhã, ou choveu ou não!*



Notação e terminologia

- Designemos o evento por A. A probabilidade de um evento é geralmente escrita da seguinte forma

$$P(A) \text{ or } Pr(A)$$

- O complementar de determinado evento é $\bar{A}$ (tudo menos aquele evento).

$$P(\bar{A}) = 1 - P(A)$$

Probabilidades

- Uma probabilidade de 0.25 significa que será 3 vezes mais provável que não chova amanhã do que chova.

$$P(\text{não chover}) = 1 - P(\text{chover}) = 0.75$$

$$0.75/0.25 = 3$$

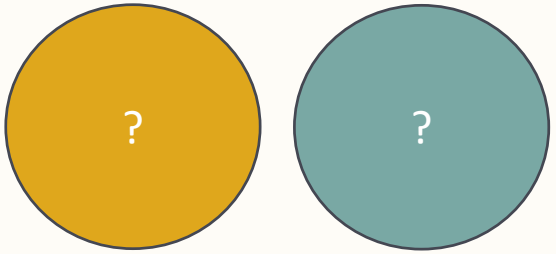
- Uma determinada probabilidade pode ser interpretada como uma proporção da concretização desse evento numa base temporal alargada.

A união de dois eventos consiste em tudo aquilo que estiver incluído em A ou B ou ambos.

Se

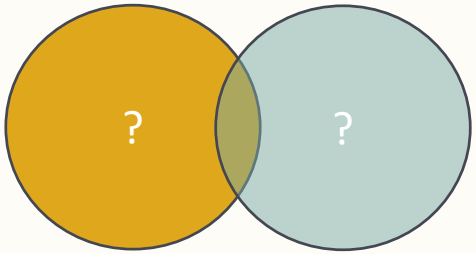
- *A = {chover amanhã}*
- *B = {chover amanhã e depois de amanhã}*
- *C = {3 peixes por arrasto}*
- *D = {4 ou 5 peixes por arrasto}*

Então



– *A∪B = {chover nos próximos dois dias}*

– *C∪D = {3 a 5 peixes por arrasto}*



$P\{A \cup B\} \neq P\{A\} + P\{B\}$,

$P\{C \cup D\} = P\{C\} + P\{D\}$,

porque apenas C e D são mutuamente exclusivos, enquanto que A e B se intersectam!