# Estatística Q.B.

## Statistics *Quantum Satis*

Dinis PESTANA[1,2,3]
**Acta Med Port 2013 Sep-Oct;26(5):499-504**

**RESUMO**

A estatística é um instrumento de eleição na transformação de informação em conhecimento, pois a informação é sempre limitada, e o objectivo é inferir da amostra para a população que pretende representar. Mas a banalização do uso de *software* estatístico — que dá sempre respostas, quer as questões colocadas sejam ou não adequadas — e o abuso da estatística para conferir um lustro científico ao que muitas vezes é treta, tem contribuído para uma desconfiança justificada sobre a investigação científica. Decerto Lord Rutherford, se fosse vivo, já não recomendaria que não se usasse estatística. Mas recomendaria que apenas se use "estatística q.b.", pois se com poucos dados eventualmente não se pode estabelecer nada, com demasiados dados pode concluir-se significância de tudo, seja ou não efectivamente relevante. A medicina baseada na evidência corre por isso o risco de, ao meta analisar cada vez mais dados, acabar por 'provar' resultados desprovidos de real sentido.

**Palavras-chave:** Interpretação Estatística de Dados; Meta-Análises; Probabilidade; Técnicas de Investigação.

**ABSTRACT**

Statistics is a privileged tool in building knowledge from information, since the purpose is to extract from a sample limited information conclusions to the whole population. The pervasive use of statistical software (that always provides an answer, the question being adequate or not), and the absence of statistics to confer a scientific flavour to so much bad science, has had a pernicious effect on some disbelief on statistical research. Would Lord Rutherford be alive today, it is almost certain that he would not condemn the use of statistics in research, as he did in the dawn of the 20th century. But he would indeed urge everyone to use statistics *quantum satis*, since to use bad data, too many data, and statistics to enquire on irrelevant questions, is a source of bad science, namely because with too many data we can establish statistical significance of irrelevant results. This is an important point that addicts of evidence based medicine should be aware of, since the meta analysis of two many data will inevitably establish senseless results.

**Keywords:** Data Interpretation, Statistical; Investigative Techniques; Meta-Analysis; Probability.

### Estatística — uma Panorâmica Breve

No último quartel do século XIX Galton escreveu ditirambos sobre a capacidade de a estatística decifrar a complexidade do mundo ("[...] *a Estatística é uma disciplina cheia de beleza e interesse. Quando a Estatísticas não é tratadas por brutos, sendo pelo contrário trabalhada com métodos avançados, e interpretada com argúcia, tem um poder extraordinário para dilucidar questões complexas. É a única caixa de ferramentas capaz de abrir uma fenda por onde penetrar o formidável amontoado de dificuldades que barram o caminho do progresso a todos os que querem avançar nas ciências humanas*."), qual canivete suíço intelectual capaz das mais variadas tarefas.

Galton foi de facto um impulsionador da estatística inferencial (fundou o primeiro laboratório de estatística, sob direção de Karl Pearson). O próprio Galton percebeu que a estatística permitia "medir" uma variável difícil ou cara de observar usando uma variável facilmente acessível que com ela estivesse correlacionada. E Karl Pearson, que já nos finais do século XIX tinha proposto em *The Grammar of Science* uma mudança de paradigma, complementando a tradicional pesquisa de causalidade com a investigação da associação estatística, inventou o teste do qui-quadrado, usável como teste de ajustamento de um modelo e também na análise de tabelas de contingência, formalizou a correlação linear, investigou modelos mais gerais do que o modelo 'normal', descrevendo uma família de curvas com assimetria e achatamento não nulos, que inclui betas, gamas (de que a qui-quadrado é um caso especial) e *F*.

A estatística tornou-se rapidamente o instrumento de implementação da metodologia da investigação nas ciências experimentais, pois a indução é sempre uma amplificação de observações limitadas de uma amostra para toda a população, em muitos casos virtualmente ilimitada, em que haverá sempre que lidar com alguma incerteza (e se possível domesticá-la, usando probabilidade e valores esperados). A associação de estatística e medicina cedo começou a produzir frutos, e é interessante notar que já em 1892 Machado de Assis, no capítulo CLXIV de Quincas Borba, põe Sofia a olhar, numa livraria, para a secção de livros de anatomia e de estatística.

A estatística teve um salto qualitativo interessante quando um jovem químico, William Gossett, publicou na "Biometrika" de 1908 uma análise "On the probable error of the mean", em que deduziu a *t* de Student (o pseudónimo

1. Departamento de Estatística e Investigação Operacional. Centro de Estatística e Aplicações. Faculdade de Ciências. Universidade de Lisboa (CEAUL). Lisboa. Portugal.
2. Centro de Filosofia das Ciências. Universidade de Lisboa (CFCUL). Lisboa. Portugal.
3. Instituto de Investigação Científica Bento da Rocha Cabral. Lisboa. Portugal.
Recebido: 29 de Setembro de 2012 - Aceite: 07 de Maio de 2013 

que usou para assinar a publicação) e mostrou a sua utilidade em estudos inferenciais sobre o valor médio de uma população normal, ou na comparação de valores médios de duas normais. Nesse trabalho introduziu também algumas ideias importantes sobre simulação, uma área que se tornou verdadeiramente importante quando o uso de computadores se banalizou, dando origem a desenvolvimentos notáveis na estatística computacional.

Anote-se que os resultados de Student são exactos para pequenas amostras normais; e que, para grandes amostras não normais, podem em condições geralmente válidas ser usados como aproximações.

Claro que em ciências da vida, sempre complexas, é frequente ser necessário comparar mais do que dois valores médios. Fisher mostrou que comparações dois a dois usando os resultados de Student era um disparate, pois as decisões em estatística são tomadas sob risco de erro, e o erro combinado aumenta enormemente. Inventou por isso a Análise da Variância (análise no sentido químico, pois trata-se de uma decomposição da variância em parcelas independentes, entre grupos e dentro de cada grupo), divulgada no seu livro "Statistical Methods for Research Workers" (1925); e na década que se seguiu inventou as bases do Planeamento de Experiências, uma forma de recolha de dados mais sofisticada, que alterou indelevelmente o conceito de experimentação científica.

No período inicial de desenvolvimento da estatística era comum dispor apenas de amostras de pequena dimensão. Se não soubermos se os dados provêm de uma população normal, os testes *t* de Student, $X^2$ e *F* deixam de ser adequados. Como em todas as populações não normais os estimadores do valor médio e da variância são dependentes, a dedução de testes exactos é em geral um quebra-cabeças. Por isso se desenvolveu uma área da estatística em que nada se pressupõe sobre a população, a estatística não-paramétrica, que usa contagens, relações de ordem e *ranks*. Foi nesta área que se desenvolveram testes, como o de Kolmogorov-Smirnov, para avaliar se os dados provinham de uma determinada população (nomeadamente normal), ou se duas amostras podiam ser consideradas como provenientes da mesma população. A área não paramétrica é muito inventiva, e para além de testes sobre localização e sobre escala, desenvolveu testes sobre assimetria, alteração de padrões, homogeneidade, independência, aleatoriedade, e muitos outros.

A banalização do uso de computadores trouxe profundas alterações à estatística e à sua prática. A capacidade que um computador tem para gerar números que imitam o acaso pôde ser aproveitada para 'fingir' a realidade, e na estatística computacional podemos sempre fazer o milagre da multiplicação dos pães e dos peixes do evangelho, mas com algo mais imaterial, claro: números. Esta capacidade de aumentar computacionalmente amostras, replicar amostras, extrair subamostras com ou sem replicação, permitiu que se desenvolvessem novos métodos de trabalho, por vezes com nomes humorísticos, como o *bootstrap* de Efron, um nome inspirado nas míticas aventuras do barão de Munchausen, um admirável mentiroso compulsivo, que entre outras façanhas contava que se tinha salvo de morrer afogado puxando-se do fundo de um lago pelos atacadores das botas até chegar à superfície.

Foi também possível, com as capacidades de cálculo e gráficas dos computadores, alterar dramaticamente a análise inicial dos dados, usando uma análise exploratória de dados, mais arrojada do que a tradicional estatística descritiva. Por outro lado, o desenvolvimento de *software* estatístico, cada vez mais amigável, permitiu a banalização do uso da estatística — ainda que muitas vezes de forma atropelada, porque executar ordens inadequadas costuma dar mau resultado, e o computador é bom a executar ordens, mas não a avaliar se as ordens são ou não apropriadas. Não consigo simpatizar com a ideia de ensinar estatística a médicos a partir de *outputs* de computadores (a médicos e a quaisquer outros, mas no caso de médicos é uma perversão mais perigosa).

A escassez de dados foi em parte mitigada pela arte da simulação. Mas o desenvolvimento de aplicações computacionais permitiu também a aquisição de dados em grande escala, a tal ponto que foi necessário desenvolver uma nova área na charneira da estatística e da computação, *data mining*, para procurar a informação que vale ouro no meio de carradas de ganga de informação que é mais ruído do que sinal.

Apesar de actualmente em geral não haver escassez de dados, há exceções, nomeadamente em Medicina. No caso de doenças raras, por exemplo, há poucos dados, e dispersos. Diversos grupos de investigadores podem ter amostras pequenas, sendo a análise estatística inconclusiva, sendo porém possível que uma meta análise dos resultados venha a ser conclusiva. Pode mesmo haver evidência controversa, haver estudos que apontam para a significância de um determinado resultado enquanto outros advogam e contrário, e também neste caso uma meta análise pode eventualmente harmonizar os resultados.

**Ter muitas ideias e deitar quase todas fora**

Com muitas variantes, atribui-se a Linus Pauling o comentário "*If you want to have good ideas you must have many ideas. Most of them will be wrong, and what you have to learn is which ones to throw away*." Em certo sentido, ilustra a Lei dos Grandes Números (com muitas ideias, ainda que muitas vão pela borda fora porque a probabilidade de resistirem ao aparato crítico a que devem ser submetidas seja baixa, o número esperado de sucessos acaba por ser considerável).

Não é disparatado dizer que a teoria dos testes de hipóteses foi desenvolvida para deitar deitar fora ideias sem interesse, falsear hipóteses que devem ser descartadas. O antepassado dos testes de hipóteses é um trabalho deveras interessante de John Arbuthnott,[1] em que começa por constatar que em 72 anos consecutivos de registos de nascimentos na sua paróquia nasceram, ano a ano, mais rapazes do que raparigas. Arbuthnott comenta que se alguém em 72 lançamentos de uma moeda ao ar verifica-

se que em cada um deles, consistentemente, o resultado era 'cara', só por imbecilidade manteria a ideia que haveria igual probabilidade de sair cara ou coroa. Da mesma forma, seria completamente inadequado admitir, face à referida observação, que é igualmente provável nascer um rapaz ou uma rapariga. (Daí Arbuthnott conclui que a Divina Providência faz nascer mais rapazes do que raparigas porque os rapazes morrem mais, e assim se garante que na idade adulta cada senhora possa ter o consolo de encontrar um par masculino, mas esta é se calhar uma das tais ideias que Pauling acha que devem ser deitadas fora.)

Os historiadores da estatística consideram que esse trabalho desenvolve algumas ideias fundamentais da teoria dos testes de hipóteses. Arbuthnott quer investigar se nascem mais rapazes do que raparigas (hipótese alternativa). Nega essa hipótese alternativa, tomando como hipótese nula que a probabilidade de nascer macho é igual à probabilidade de nascer fêmea = 1/2. Como a probabilidade de observar em 72 ocasiões consecutivas que nascem mais rapazes do que raparigas, condicionalmente à hipótese nula ser verdadeira é apenas $p = 0{,}5^{72} \approx 2{,}11758 \times 10^{-22}$, considera que é tolice não rejeitar a hipótese nula (passando implicitamente a considerar mais credível a alternativa de que a probabilidade de nascer um macho é maior do que a de nascer uma fêmea) [*].

Aquele 'valor de prova' (*p-value*) *p* pode ser encarado como uma avaliação da estranheza que causa o resultado observado da estatística de teste, considerando que a hipótese nula é verdadeira. Quanto maior a "informação", que é proporcional a -ln(*p*), e consequentemente muito grande quando *p* assume valores próximos de 0, maior é essa surpresa, e mais plausível parece rejeitar a hipótese nula por esta não se coadunar com os factos, pois contra factos não há argumentos.

Uma caricatura do procedimento científico é então: se quer tentar fundamentar uma novidade $H_A$, veja se pode rejeitar a sua negação $H_0$, por a probabilidade *p* de uma cauda da distribuição da estatística de teste delimitada pelo que foi observado, condicionalmente à validade de $H_0$, ser surpreendentemente baixa.

É por isso que faz sentido dizer que rejeitar a hipótese nula $H_0$ é uma decisão forte (progresso no conhecimento), manter $H_O$ é uma decisão fraca — muitas vezes, as amostras são insuficientes para rejeitar $H_0$. De facto, num passado não muito remoto, o *p-value* ser significativo era meio caminho andado para publicação, e por outro lado a obtenção de um *p-value* 'grande', interpretado como afinal os dados se compatibilizarem com o expresso na hipótese nula, não era considerado publicável (o que levanta grandes problemas de sínteses meta analíticas poderem ser fortemente enviesadas devido a estudos que não provavam significância não terem sido publicados).

**Acumulação de Evidência Estatística**

A meta análise, que começou a ser desenvolvida no último quartel do século XX, tem a ambição de tirar uma resultante de evidência apresentada no relato de diversas investigações sobre a mesma questão, conseguindo eventualmente que a acumulação de informação leve, afinal, à rejeição da hipótese nula (frisamos de novo que em geral a hipótese nula é um 'grau zero' do conhecimento, que gostaríamos de rejeitar, porque a hipótese alternativa $H_A$ é que representa progresso no conhecimento).

Neste sentido, é uma técnica de recurso. A Amostragem e o Planeamento de Experiências são as duas disciplinas complementares que se ocupam da obtenção dos dados apropriados para levar a cabo determinada investigação, e desenvolveram resultados sobre o tamanho que as amostras devem ter, dada a variabilidade estimada, para que a inferência seja feita com um determinado grau de precisão, com uma probabilidade elevada pré-fixada. Infelizmente mesmo os rudimentos de amostragem são em geral mal conhecidos ou ignorados por exigirem gastos de recursos que as equipas de investigação não estão dispostas a fazer. Por outro lado, mesmo que se tenha uma ideia corretíssima da quantidade de dados que deveriam ser recolhidos, há circunstâncias em que tal não é possível. Ninguém vai infetar doentes com Ebola apenas porque os que tem ao dispor são escassos, e a teoria da amostragem aconselhava a que houvesse mais 43 doentes do que os que efetivamente existem.

Mas com o correr do tempo essa equipa e outras equipas vão colecionando dados, e se tiverem a informação e formação adequadas, talvez partilhem os dados na Cochrane Collaboration — e é de esperar que a evidência acumulada um dia venha a ser a adequada para tomar decisões sob risco, mas com o risco controlado ao nível que se deseja.

Assim exposto, parece que a meta análise é uma área que todos devíamos abraçar com entusiasmo. É, sem dúvida, uma área cheia de interesse, mas que nas mãos de entusiastas acaba por virar o feitiço contra o feiticeiro.

Atente-se no seguinte exemplo: sejam dois caracteres qualitativos *A* (com níveis $A_1$ e $A_2$) e *B* (com níveis $B_1$ e $B_2$) — por exemplo sexo e percepção extra-sensorial, numa investigação que pretende verificar se a capacidade de adivinhar o naipe de cartas de jogar que outrem está a virar. $A_1$ representa sexo feminino, $A_2$ sexo masculino, $B_1$ que identificou corretamente pelo menos o naipe de 5 em 8

[*] Tem havido ideias curiosas sobre o que determina o sexo. McManus, na excelente e premiada monografia *Right Hand, Left Hand*, conta que houve a convicção de que o bebé seria de sexo masculino se o esperma proviesse do testículo direito e feminino se proviesse do testículo esquerdo, o que levava alguns homens a amarrar o que não correspondia às suas aspirações. Nem sempre dava o resultado pretendido, decerto porque não tinham suportado a dor necessária para atingir a recompensa que desejavam. Atualmente, a maior parte das pessoas continua a ter a convicção de que é igualmente provável nascer um bebé de sexo masculino ou de sexo feminino, por pensarem que no momento de formação de células sexuais, por meiose, são formados tantos espermatozóides X quantos Y — mas entre a formação e o uso muita coisa pode acontecer, sabe-se por exemplo que existem espermatozóides assassinos, cuja função é eliminar outros espermatozóides, e consequentemente o equilíbrio inicial entre X e Y é alterado. Actualmente, em Portugal, a estimativa da probabilidade de nascer um rapaz é 0,516.

cartas, $B_2$ o contrário (uma investigação que penso que não serve para nada, senão para exemplificar algumas tolices). Suponha-se que os dados obtidos são os que se seguem:

| | $A_1$ | $A_2$ |
|---|---|---|
| $B_1$ | 54 | 41 |
| $B_2$ | 32 | 21 |

Nesta situação é usual recorrer ao teste de independência do qui-quadrado (admitindo que é uma tabela de contingência de margens livres, decorrente de uma classificação cruzada de uma amostra em que não houve fixação prévia de quantos homens e mulheres eram solicitados para colaborar no estudo).

Numa tabela 2x2 com valores observados $o_{ij}$, $i,j$=1,2, que genericamente representamos

| | $A_1$ | $A_2$ |
|---|---|---|
| $B_1$ | a | b |
| $B_2$ | c | d |

o valor observado da estatística de teste

$$\chi^2_{22} = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} = \frac{n(ad - bc)^2}{(a + b)(a + c)(b + d)(c + d)}$$

uma vez que os valores esperados $e_{ij}$ sob a hipótese nula de independência são obtidos dividindo o produto das margens pela dimensão $n=a+b+c+d$ da amostra. Assim, no caso do exemplo acima, o valor observado seria $\chi^2_{22}$ (obs.) = 0,1746675, muito inferior ao valor crítico $\chi^2_{1,0.95}$ = 3,83, geralmente usado como valor a partir do qual se rejeita a hipótese nula, correndo o risco de ao decidir com essa regra haver uma probabilidade de 5% de rejeitar erradamente a hipótese nula. O valor de prova é $p$ = 0,676, ou seja, sendo a hipótese nula verdadeira há uma probabilidade muito grande de o valor obtido da estatística de teste ser superior ao valor observado 0,175, não há razões para rejeitar a hipótese nula.

Mas por outro lado se com mais observações obtivermos a tabela

| | $A_1$ | $A_2$ |
|---|---|---|
| $B_1$ | 100x54 | 100x41 |
| $B_2$ | 100x32 | 100x21 |

o valor observado da estatística de teste é $\chi^2_{22}$ (obs.) = 17,46675, 100 vezes maior, muito maior do que o valor crítico $\chi^2_{1,0.95}$ = 3,83.

Apesar de a proporção de homens e mulheres que identificaram corretamente os naipes de pelo menos 5 de 8 cartas ser a mesma, o valor de prova é agora $p$ = 0,000029 (a probabilidade de observar um resultado pior do que o observado é inferior a três em cem mil), e a hipótese nula deve ser rejeitada.

Objectar-me-ão que com uma amostra 100 vezes maior é natural ser agora evidente que deve haver rejeição, quando antes não havia evidência suficiente para rejeitar. De acordo, mas não é isso que está em causa: o que pretendo ilustrar com este exemplo é que TUDO acaba por ser rejeitado, se se fizer crescer desmedidamente a amostra. Significância estatística e relevância real não são a mesma coisa. Amostras excessivas são tão inadequadas como amostras demasiado escassas.

Jessica Utts[2] escreveu um notável trabalho sobre o uso de meta análise em estudos de percepção extra-sensorial, mostrando que do ponto de vista estatístico não há dúvida que há mais identificação correta de naipes, sob condições experimentais fortemente blindadas, do que seria de esperar se a identificação fosse feita apenas com uma tiragem ao acaso (por exemplo, gerando números pseudo-aleatórios entre 0 e 1, e escolhendo copas se saísse um número em [0,0.25], ouros se saísse um número em (0.25,0.5], paus se saísse em (0.5, 0.75], e espadas se saísse em (0.75,1]. Não é surpreendente, uma vez que está em causa uma amostra com dezenas de milhares de dados (e permito-me juntar uma explicação suplementar mais profunda: num passeio aleatório simétrico, o tempo de espera para retorno a equilíbrio é infinito, pelo que basta haver um desequilíbrio favorável nos dados iniciais (em que porventura a blindagem nem era perfeita e pode ter havido "batota", intencional ou não) para esse desequilíbrio perdurar, parecendo surpreendente.

Gosto de meta análise, recomendo que se tente usar informação — mas com contenção, e com espírito crítico. E o espírito crítico leva-nos também a reconhecer que ao usar dados alheios podemos estar a usar dados de baixa qualidade. Recordo-me de ajudar um amigo médico a estudar dados sobre natrémia (variável resposta) e tempo de intervenção cirúrgica em operações ao útero por laparascopia, em que a duração das operações fornecidas por diversas equipas iam da precisão de minutos e segundos ao caso extremo de operações que duravam 15 minutos, ou 20 minutos, ou 25 minutos, nunca valores diferentes destes. Por outro lado, tem-se verificado frequentemente que dados fornecidos por outrem são excessivamente aleatórios (distribuição equilibradas dos algarismos, enquanto a lei de Benford-Zipf mostra que em muitos fenómenos naturais a ocorrência de algarismos é hiperbolicamente decrescente), levando a suspeitar que os dados foram ajeitados *a posteriori*, ou foram simplesmente inventados.

Recomendo de qualquer modo a coleção de artigos de Egger e Davey Smith,[3-8] já um pouco antigos, mas muito esclarecedores. Os livros de Hartung et al[9] e de Kulinskaya et al[10] podem servir de introdução a quem queira aprofundar a questão.

**Usos e abusos da estatística**

Sou decerto suspeito, mas creio que o elogio da estatística feita por Galton é completamente justo. Como tudo, a estatística tem grandezas e misérias. É fácil apresentar exemplos da beleza e utilidade da estatística. Creio que não nos damos conta, no dia-a-dia, da estatística que está por detrás da ergonomia do que nos rodeia, da simplicidade com que realizamos tarefas 'simples' como usar uma fita em que se deposita uma gota de sangue para avaliar a glicemia — no aparelho que uso, o que é de facto medido é o ângulo de refração da luz no meu sangue, que é convertido em unidades de glicemia usando regressão estatística. Esta arte de medir uma coisa por outra é um feito fantástico, mas não é só para isso que serve a regressão. Permite por exemplo determinar qual é a melhor relação dose/resposta de medicamentos, ou auxiliar um cirurgião maxilo-facial a reconstruir um rosto destruído, determinando com grande precisão o ângulo de abertura que deve proporcionar aos maxilares reconstruídos, por forma a que o doente possa depois articular bem as palavras (fechar bem a boca) ou comer sopa com comodidade (abrir de forma a usar uma colher de sopa — mas não confundir com a colher do arroz, ou não o deixar a só comer sopa com a colher de café.)

A análise discriminante permite-nos com pouca informação — por exemplo, medições várias de um dente — ter uma ideia em geral correcta se é de homem ou de mulher; e a análise química da racemização de ácido aspártico na dentina permite ter uma ideia aproximada da idade de um cadáver desconhecido. Técnicas diversas permitem uma redução inteligente da dimensionalidade de dados multivariados, fazendo sobressair a informação que de facto é relevante, por exemplo ajudando médicos da área desportiva a corrigir posturas ou a selecionar atletas com mais potencial para alta competição. É também a análise de dados que permite treinar para o ouro olímpico, por exemplo em corridas com barreiras analisar o tamanho do passo, para o corrigir porventura uns escassos milímetros, para garantir que o atleta chega regularmente ao intervalo de confiança em que tem que fazer a chamada para passar a barreira sem bater nela, o que na melhor das hipóteses o atrasaria, e na pior o levaria a cair.

Amostragem e planeamento de experiências não têm a expressão que mereceriam na preparação de cientistas experimentais. Inúmeros trabalhos são baseados em 'amostras de conveniência', que deviam antes ser chamadas amostras inconvenientes. De facto, se representam a população de que foram extraídas, será por mero acaso. Quando vir num trabalho uma frase do estilo "este estudo usou 46 voluntários..." pode imediatamente suspeitar que as boas regras de amostragem não foram aplicadas para determinar o tamanho que a amostra deveria ter, e que também não é representativa porque tem características especiais — por exemplo, os estudos da equipa de Kinsey sobre sexualidade humana, usando muitos universitários, que tendem a mostrar-se mais desinibidos e mesmo exibicionistas, e prisioneiros, que decerto têm vida sexual diferente da do comum da população, seriamente compromete as conclusões que relataram. Decerto ninguém duvida que se a mesma pergunta, "deve o presidente da República marcar eleições antecipadas" a voluntários num jantar com o presidente do PS, ou num jantar com o presidente do PSD, aquele 'D' vai fazer uma enorme diferença na percentagem de sins — e se usarmos qualquer dessas percentagens para estimar a proporção de portugueses que desejam eleições antecipadas essa estimativa deve ser fortemente enviesada. E se a pergunta for feita a uma amostra de conveniência de empregados do Metro, ou a estudantes da Faculdade de Direito de Lisboa, a estimativa que se obtém não terá decerto mais qualidade (será mais próxima do verdadeiro valor, por serem amostras mais heterogéneas, mas continuam a ser uma coleção de dados insuficientemente representativa). Replicar pequenas amostras para obter grandes amostras — por exemplo, dividir uma emissão de esperma em 1000 'amostras', como se escrevia num relatório que li, para analisar a hipermetilação de bases como possível indicador de infertilidade, nada tem que ver com a independência que os dados devem ter para serem representativos. Encontrei trabalhos em que se analisam os dados correspondentes à evolução de todos os doentes de um hospital num determinado período, para fazer uma estatística que se pretende inferencial — mas faz algum sentido inferir do todo para o todo? Ou haverá a ingenuidade de pensar que todos os doentes de um hospital são representativos de doentes com a mesma patologia em outros hospitais, no que se refere à evolução da doença, quando os tratamentos usados são quase certamente diferentes (até no mesmo hospital — num estudo retrospectivo sobre tratamento de grávidas com o síndroma HELLP as 46 doentes que tinham sido tratadas tinham recebido fármacos para desenvolver os pulmões do feto antes da IMG, mas os médicos tinham usado dois fármacos (não por razões experimentais, mas por preferência de cada médico, ou por ser o que havia disponível), e não havia duas doentes, nas 46 tratadas, que tivessem tido a administração de doses iguais durante períodos iguais. É interessante analisar estes dados, nomeadamente para planear adequadamente experiências futuras, mas usá-los como uma amostra para trabalho inferencial é ingenuidade.

Os estudos retrospectivos muitas vezes devem-se a não ter havido um planeamento experimental adequado. Quando ouço falar de 'experiências' educativas, fico sempre de pé atrás, porque duvido que seja uma verdadeira experiência, no sentido em que os resultados de um tratamento a que são sujeitos os elementos de um grupo experimental são comparados com os de um grupo de controlo. A ausência de um grupo de controlo é algo que já não espero ver em Medicina, onde há um severo controlo ético, que tanta falta faz em outras ciências. Nem sempre foi assim, com resultados trágicos. É célebre a história da 'operação porta-cava', destinada a dar maior qualidade de vida a doentes com patologias hepáticas, que foi largamente feita sem haver um controle adequado. Quando se fez o controlo retrospectivo, tarde e más horas, alguns médicos portugueses com um sentido de humor apurado passaram

a chamar-lhe 'operação porta à cova', contou-me o Professor Nuno Grande.

Penso que um indivíduo com uma unha partida pode auto-receitar-se um antibiótico, mas se calhar era melhor usar uma tesoura ou uma lima de unhas. Um dos maiores abusos da estatística é ser usada por quem não percebe os fundamentos, até porque acha que tem o melhor dos *softwares* (filosofia similar à de usar um antibiótico como panaceia universal) para analisar os dados — esquecendo que o grande defeito das virtudes do computador é responder, seja a pergunta adequada ou não. Felizmente um grupo de brincalhões de Harvard começou há alguns anos a conceder prémios IgNobel, troçando (ocasionalmente sem razão, como o futuro veio a demonstrar) de trabalhos particularmente disparatados. Publicam também uns *Annals of Improbable Research*, que deviam ser lidos e meditados por qualquer trabalhador científico, para se auto-criticar e para se defender das aldrabices e ciência da treta cuja fertilidade imparável invade todas as áreas.

Inspirado neles, a resolução do velho problema: o que apareceu primeiro, o ovo ou a galinha?

Projeto de investigação: dois investigadores telefonam quase simultaneamente para cada um de 100 aviários escolhidos ao acaso, um deles encomendando uma galinha e o outro uma dúzia de ovos, para serem entregues na mesma morada (para não haver enviesamento, vão alternando de aviário para aviário a ordem em que as duas encomendas são feitas).

Resultado da análise de dados: não há dúvida, em 97% dos casos os ovos chegaram antes.

Medite o leitor se não passou já por trabalhos com aparência de investigação feita com todos os requisitos, mas tão pateta e patética como esta.

### Recomendação Final

É muito citada a frase de Lord Rutherford "*Se a sua experiência precisar de Estatística, faça uma experiência melhor*". Creio que um homem inteligente como Lord Rutherford, se praticasse investigação atualmente, diria antes: "*Controle tudo o que for possível, o que não for possível controlar aleatorize, evite confundimento e enviesamento — e depois use a estatística q.b., mas não mais do que essa. Nunca use os dados para confirmar o que esses mesmos dados sugeriram, e não use repetidamente os mesmos dados, não só porque se deixa influenciar por eles, como também porque não há dados perfeitos, e usando os mesmos dados vai ter enviesamento no mesmo sentido. Não tente correlacionar tudo com tudo, e sobretudo não confunda associação com causalidade, lembre-se da loucura do trabalho que 'demonstrava' que o número de crimes nas cidades inglesas, no século XIX, se devia a haver mais pastores na igreja anglicana — quando apenas são dois fenómenos que cresceram ambos com o aumento populacional. Ou tenha sempre presente o texto humorístico de Graham Greene (May we Borrow your Husband?), do quarentão virgem que nunca tinha praticado sexo porque o médico dos alunos da* public school *em que estudara tinha uma teoria: as relações sexuais causam cancro, pois 100% das pessoas que morreram de cancro ou praticaram relações sexuais ou são filhas de pessoas que praticaram relações sexuais.*"

E Lord Rutherford não continuaria com recomendações porque esgotaria o tempo e o espaço disponível; e sobretudo porque sabia que não é preciso repisar quando se tem leitores inteligentes.

### CONFLITO DE INTERESSES

O autor declara que não houve conflito de interesses na realização deste trabalho.

### FONTES DE FINANCIAMENTO

A investigação do autor é financiada por fundos nacionais através da FCT - Fundação para a Ciência e a Tecnologia, projecto PEst-OE/MAT/UI0006/2011.

## REFERÊNCIAS

1. Arbuthnott J. An argument for divine providence, taken from the constant regularity observed in the birth of both sexes. Philos Trans. 1710;27:186-90.
2. Utts J. Replication and meta-analysis in parapsychology. Stat Sci. 1991;6:363-403.
3. Egger M, Davey Smith G. Meta-Analysis — potentials and promise. BMJ. 1997;315:1371-4.
4. Egger M, Davey Smith G. Meta-Analysis — principles and procedures. BMJ. 1997;315:1533-7.
5. Egger M, Davey Smith G. Meta-Analysis — beyond the grand mean? BMJ. 1997;315:1610-4.
6. Egger M, Davey Smith G. Meta-Analysis — bias in location and selection of studies. BMJ. 1998;316:61-6.
7. Egger M, Schneider M, Davey Smith G. Meta-Analysis — spurious precision? Meta-analysis of observational studies. BMJ. 1998;316:140-4.
8. Davey Smith G, Egger M. Meta-Analysis — unresolved issues and future developments. BMJ. 1998;316:221-5.
9. Hartung J, Knapp G, Sinha BK. Statistical Meta-Analysis with Applications. New York: Wiley; 2008.
10. Kulinskaya E, Morgenthaler S, Staudte RG. Meta Analysis. A Guide to Calibrating and Combining Statistical Evidence. Chichester: Wiley; 2008.