E depois?

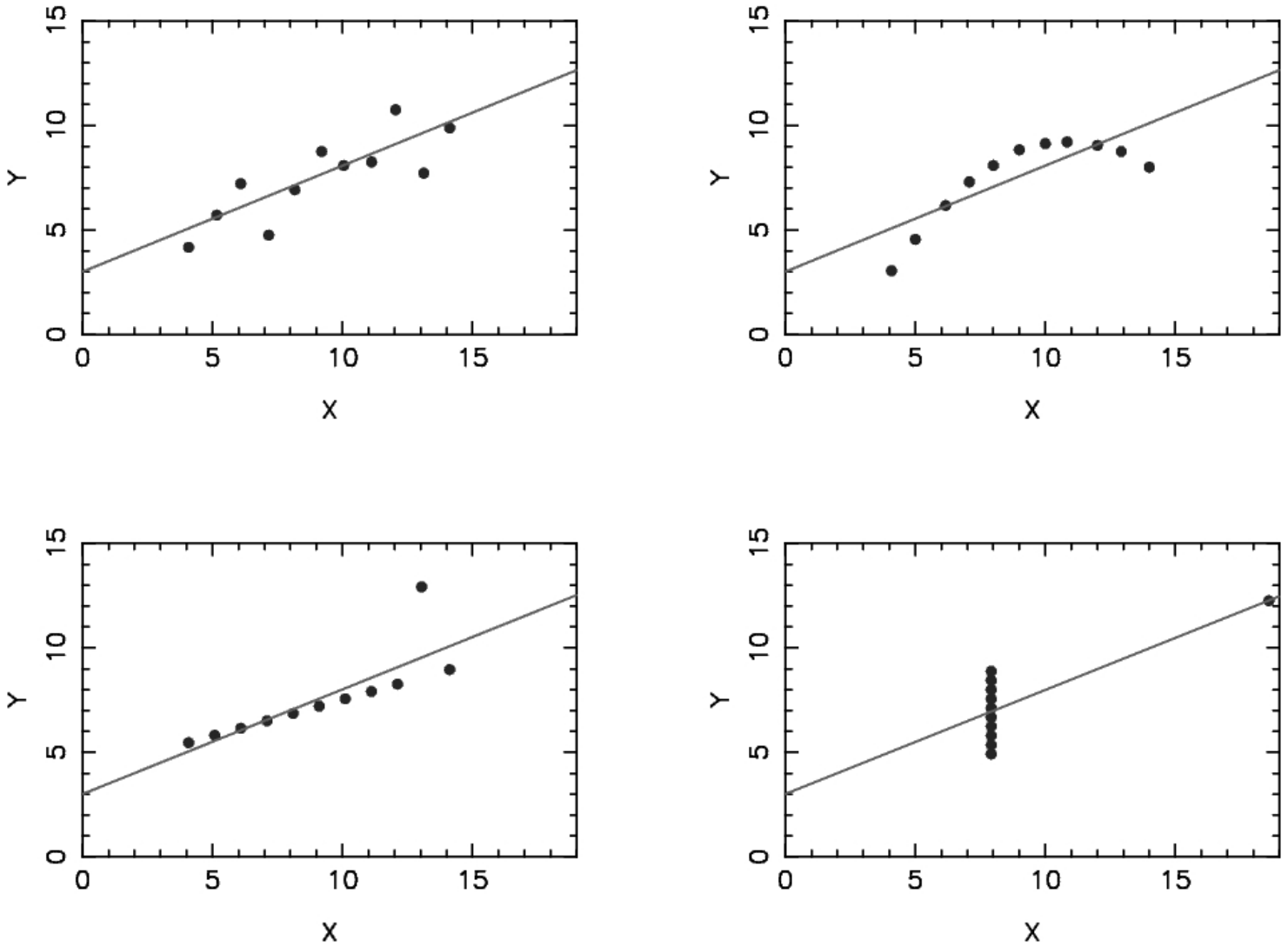
- Quando temos sucesso em demonstrar a existência de uma correlação, o próximo passo é perguntar o que essa correlação significa, ou seja, que lei relaciona as duas variáveis.

- É muito comum nesses casos fazer uma *regressão linear* aplicando o *método dos mínimos quadrados*. É essencial perceber que nesse ponto estamos falando de ajuste de modelos. A distinção entre isso e *teste de hipóteses* deve ser clara.

- Antes de discutirmos esse ponto, é necessário que se façam várias considerações (o que se fará em mais detalhe na seção 6.2, como por exemplo:
  1. Existem grandezas melhores a se minimizar do que os quadrados dos desvios?
  2. Quais são os erros nos parâmetros da regressão?
  3. Porque a relação deve ser linear?
  4. E a mais crucial de todas: O que estamos tentando encontrar? Se existe uma correlação entre x e y queremos encontrar $x(y)$ ou $y(x)$? Em geral os coeficientes encontrados são completamente diferentes.

E depois?

Como um argumento contra a aplicação cega de testes de correlação e ajustes lineares considere o famoso quarteto de Anscombe (1973) mostrado na figura abaixo:



Quarteto de Anscombe: quatro conjuntos de 11 pares (X_i, Y_i) fictícios, cada qual com os mesmos $(\overline{X}, \overline{Y})$, coeficientes de regressão, resíduos em Y e erros nas inclinações.

E depois?

- O ponto que Anscombe queria fazer é que gráficos são essenciais para uma boa análise estatística.
- Por outro lado, os exemplos ilustram outros pontos: A regra do dedão e a distinção entre *independência* dos pontos e *correlação*. Em mais de um dos conjuntos de dados de Anscombe os dados são claramente não-independentes, ainda que não apresentem nenhuma *correlação* formal.
- O painel *b*) é um caso onde um ajuste linear é de qualidade indiferente, embora a escolha da relação 'correta' entre X e Y resultaria num ajuste perfeito.
- *Independente* significa $prob(X, Y) = prob(X)prob(Y)$ ou $prob(X|Y) = prob(X)$; enquanto que X *correlacionado* com Y significa $prob(X, Y) \neq prob(X)prob(Y)$, numa forma que $r \neq 0$.
- É perfeitamente possível ter-se $prob(X, Y) \neq prob(X)prob(Y)$ e $r = 0$. O exemplo clássico são pontos distribuídos com as faixas na bandeira do Reino Unido.

Análise de Componentes Principais (PCA)

- Análise de componentes principais é o 'buscador de correlações' definitivo quando muitas variáveis estão presentes. Dados N objetos com n parâmetros, como encontrar o que está correlacionado com o que?

- PCA é uma técnica de uma família de algoritmos (estatística multivariada) desenhada para esse tipo de situação.

- A tarefa é a seguinte: dado um conjunto de N objetos com n variáveis medidas x_n encontrar um conjunto de ξ_n variáveis que são ortogonais (independentes), cada uma combinação linear das variáveis originais:

$$\xi_i = \sum_{j=1}^{n} a_{ij} x_j$$

  com valores de a_{ij} tais que o menor número de novas variáveis contém a maior fração da variância total. As ξ_i são as *componentes principais*

- Se a maioria da variância envolver umas poucas variáveis então os dados podem ser descritos de forma mais simplificada.

Análise de Componentes Principais (PCA)

- Seguindo Francis & Willis (1999) o PCA pode ser descrito algebricamente, através de matrizes de covariança ou de forma geométrica.

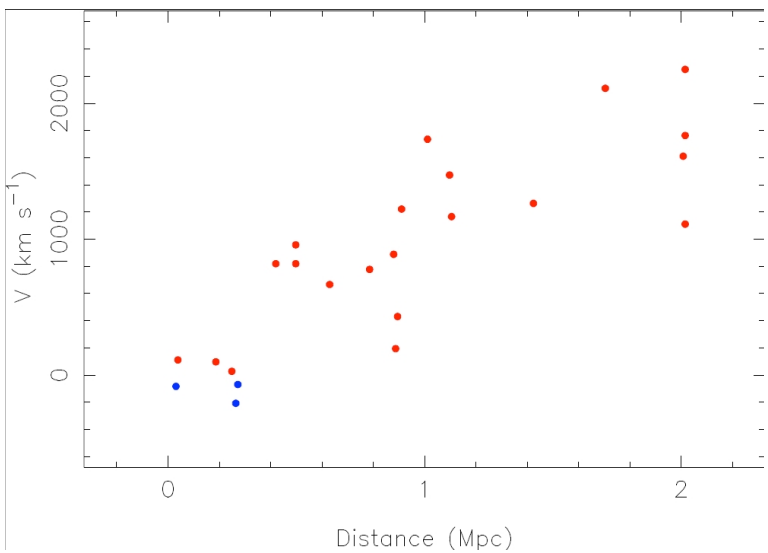
- Na forma geométrica, considerando os N objetos, esses formam um nuvem no espaço n-dimensional. Se dois ou mais parâmetros estão correlacionados, então a nuvem se elonga numa dada direção desse espaço.

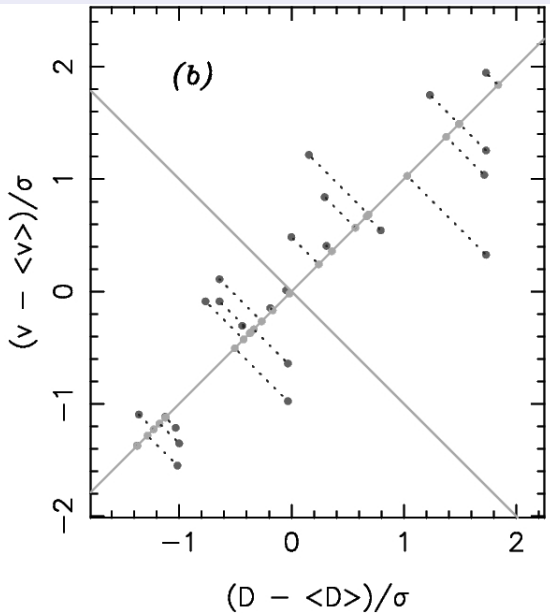
- O PCA identifica essas direções de extensão dos dados e as usa com uma sequencia de eixos. Sequencial no sentido de que a primeira é encontrada através da minimização da some dos quadrados dos desvios. Essa direção forma a primeira componente ou *autovetor 1* e é a direção individual que contém a maior variância.

- Nesse ponto repete-se a operação, agora no hiperplano $n-1$ dimensional ortogonal à primeira componente. O procedimento pode ser repetido até achar-se a n-ézima componente.

Análise de Componentes Principais (PCA)

Exemplo W&J P. 87





Análise de Componentes Principais (PCA)

- Consideremos agora a forma matricial. No processo de PCA o método usual é construir a matriz de erro. No caso das duas variáveis: $a(1,1) = \sum d'^2$, $a(1,2) = a(2,1) \sum v'd'$ e $a(2,2) = \sum v'^2$.

- Procura-se então a transformação que faz os termos cruzados desaparecerem, o que é simples em termos matriciais.

- Determinam-se os autovalores e autovetores da matriz e com eles forma-se a matriz transposta T, para transformação de variáveis e rotação. No caso uma rotação que diagonaliza a matriz de erro.

- Como nota final, nada disso é 100% automático. Deve haver uma preocupação com a linearidade das variáveis, com pesos e a presença de dados espúrios. Ao final não deve haver correlação entre PC1 e PC2.

W&J Cap 5: Teste de hipóteses

- Sempre que comparamos um conjunto de dados com outro ou com um modelo estamos *testando hipóteses*. Eles são consistentes um com o outro ou não?

- Do ponto de vista frequentista todo o processo de *inferência estatística* pode ser visco como *teste de hipótese* seguido de *determinação de parâmetros* (Cap. 6). No ponto de vista Bayesiano esses dois passos podem ser resumidos a um apenas: o método bayesiano de escolha de modelos (Cap. 7), que veremos em cursos mais avançados.

- Frequentistas trabalham a partir de distribuição probabilística de uma dada estatística. Bayesianos trabalham a partir da distribuição probabilística de uma hipótese ou modelo. Claramente é o segundo que queremos fazer mas em ocasiões em que métodos clássicos tem de ser usados (e.g. não há um modelo disponível). Esse capítulo se concentra basicamente nesses casos.

Teste de hipóteses

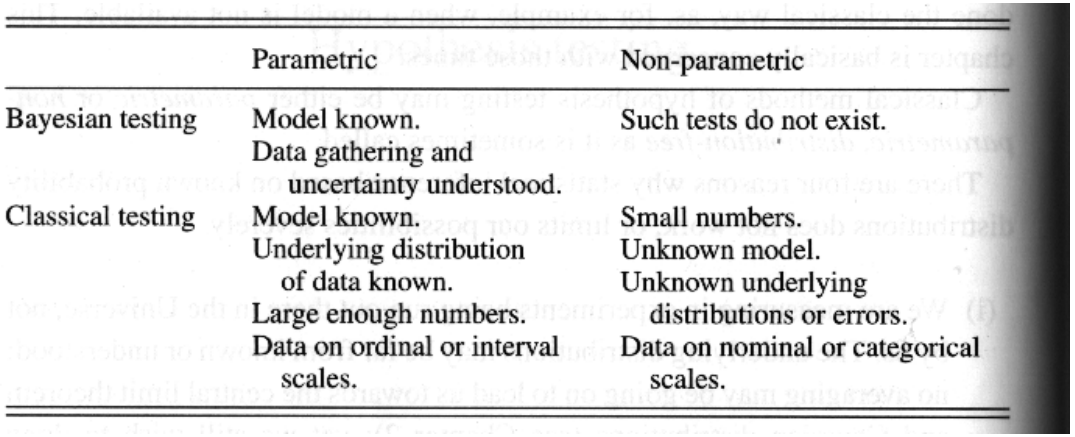
- Métodos clássicos podem ser *paramétricos* ou *não-paramétricos, livres de distribuições*
- Existem quatro razões pelas quais a inferência estatística baseada em distribuições conhecidas não funcionam, ou limita grandemente as possibilidades.
  - Nossos experimentos foram rodados pelo Universo e não nós, de modo que as distribuições subjacentes, muitas vezes, não são conhecidas e ainda assim queremos fazer inferências. Nesses caso é necessário o uso de *estatísticas não-paramétricas*, que são métodos que não requerem conhecimentos sobre as distribuições.
  - Temos que lidar com amostras muito pequenas, tais como $N = 3$. Técnicas não paramétricas conseguem lidar com isso.
  - O intervalo de escalas observáveis é dado na tabela 1.3. Cada escala tem definição propriedades formais, bem como operações admissíveis. Basta dizer que o uso de escalas outras que não numéricas requer, na maioria dos casos, o uso de métodos não paramétricos.

Table 1.3 *Measurement scales*

Scale type	Also called	Description: example
Nominal/categorical	Binned	**gender:** male / female
Ordinal/ranking	Ordered	**army ranks:** private, corporal, sergeant, ...
Ratio	Numerical/ measures	uniformly calibrated scale with zero point: **temperature in kelvin**
Interval	Numerical/ measures	uniformly calibrated scale: **time**, whose beginning (and end) we do not know, but which we arbitrarily 'zero-point' in many ways to give it ratio scales, e.g. Gregorian calendar, Julian Day, UTC, GMT

  - Outros usam esses métodos para fazer inferências e precisamos entender o que eles estão fazendo.

Teste de hipóteses

- 

	Parametric	Non-parametric
Bayesian testing	Model known. Data gathering and uncertainty understood.	Such tests do not exist.
Classical testing	Model known. Underlying distribution of data known. Large enough numbers. Data on ordinal or interval scales.	Small numbers. Unknown model. Unknown underlying distributions or errors. Data on nominal or categorical scales.

- Embora a não existência de métodos bayesianos não-paramétricos pareça óbvia, isso não é 100% correto. Se os dados forem compreendido a ponto de ser possível modelar seu processo de medida é possível usar o caminho bayesiano para achar soluções.

- Os testes clássicos envolvem 'rejeitar a hipótese nula'. Ou seja, rejeitar ao invés de aceitar a hipótese num dados nível de significância. A hipótese rejeitada pode ser uma sobre a qual não temos o menor interesse. É um *processo de eliminação*.

Metodologia do teste de hipóteses clássico

- O teste de hipóteses, tal qual como formulado por Neyman e Pearson segue os seguintes passos:
  1. Encontre duas hipóteses possíveis e excludentes, cada uma associada com uma *ação terminal*:
     H_0, a *hipótese nula* ou hipótese do não-efeito, usualmente formulada para ser rejeitada e
     H_1, uma alternativa ou *hipótese de trabalho*.
  2. Especifique de antemão o *nível de significância* α; escolha um teste que *a*) aproxima as condições e *b*) encontra o que é preciso; obtenha a *distribuição amostral* e a *região de rejeição*, cuja área é uma fração α da área total da distribuição amostral.
  3. Faça o teste; rejeite H_0 se o teste resultar num valor para a estatística cuja probabilidade de ocorrência sob H_0 (chamado usualmente de p) é $\leq \alpha$.

- É importante enfatizar o segundo item. É vital determinar o nível de significância antes de fazer o teste! Você deve estar preparado para chegar à conclusão indicada pelos dados. *Não existe teste de hipótese inconclusivo!*

Metodologia do teste de hipóteses clássico

- Existem dois tipos de erro nesse processo:de tipo I e II.
  - Erros de *Tipo I* ocorrem quando H_0 é, de fato, verdadeiro, e esse é rejeitado. Ou seja, sua probabilidade do ocorrência é α.
    Falso positivo
  - Erros de *Tipo II* ocorrem quando H_0 é falso e falha-se em rejeita-lo. Sua probabilidade de ocorrência é β, onde β e α não se correlacionam de modo direto. A *força* de um teste é a probabilidade de rejeitar um H_0 falso, ou $1 - \beta$.
    Falso negativo

Metodologia do teste de hipóteses clássico

- A distribuição probabilística da estatística teste, ou seja, a distribuição de frequência dos valores do teste incluindo todos os valores possíveis do testes sob H_0, cuja área total é uma unidade.

- A probabilidade da ocorrência de um valor da estatística do teste *na região da rejeição* é menor do que p, por definição; mas onde a região de rejeição se encontra depende de H_1

- Se H_1 indicar *direção*, então existe apenas uma região de rejeição e o teste é unilateral (*one-tailed*); se nenhuma direção for indicada, a região de rejeição é composta pelas duas extremidades da distribuição e portando é trata-se de um teste de bilateral (*two-tailed*).

- Esse é o único uso que se faz de H_1. O procedimento de teste apenas nos convence a aceitar H_1 se esse for a única alternativa a H_0. Atenção: é da natureza humana pensar que o seu H_1 é a única alternativa a H_0.

Metodologia do teste de hipóteses clássico

- Se não temos um H_1 bem definido então usar um dado valor de α faz menos sentido. Um modo de fazer esses teste com mais nuances pode ser calcular o valor de p, notar que ele é desconfortavelmente pequeno e sentir-se provocado a buscar alternativas. Não há nada de errado com isso, contanto que não se associem probabilidades com a rejeição de H_0.

- A tabela abaixo esclarece as noções de erros do tipo I e II. Suponha que o valor crítico de nossa estatística é, por exemplo, t_c e suponha ainda que a probabilidade da chance de excedermos o valor de t_c sob H_0 é α.

- O procedimento então é: computar a estatística T, comparar com t_c e rejeitar H_0 (ou seja aceitar H_1) caso $T > t_c$:

	H_0 verdadeiro	H_1 verdadeiro
$T \geq t_c$	A: erro Tipo I	B: correto
$T \leq t_c$	C: correto	D: erro Tipo II

- A soma das probabilidades das colunas dessa tabela ($A + C$ ou $B + D$) devem somar 1, mas não há relação entre as linhas, ou seja entre as taxas de erros do tipo I e II.

- A probabilidade de B é a força do teste. Geralmente existe um compromisso entre a 'força' e a taxa de erros tipo I que deve ser feito para hipóteses alternativas H_1 específicas. Fim da aula 6