Introdução ao R e algumas aplicações para linguística com corpos

Exemplos de estatística

Diana Santos

Secção ibero-românica
Departamento de línguas, literaturas e culturas europeias
Faculdade de Letras, Universidade de Oslo

d.s.m.santos@ilos.uio.no

EBRaLC, 11-12 de Setembro de 2012

Interseção entre três disciplinas...

Este curso cruza três áreas de conhecimento

- informática
- estatística
- linguística

e pressupõe um mínimo de familiaridade com as três, em graus variados...
Uma coisa sobre que o curso **não** é: obter dados sobre corpos dentro do próprio R. Possível, mas não necessário.
O curso foi pensado para usar a estatística **depois** de obter esses dados no ambiente usado.

Casos em que um raciocínio estatístico faz sentido

1. Quando apenas temos acesso a uma amostra e queremos generalizar (e saber o grau de certeza ou confiança que podemos ter nesse processo)
2. Quando sabemos que temos ruído nas nossas medições (repetindo as ditas ajuda a aproximar-nos da verdade)
3. Quando queremos medir as relações e a possível dependência de vários fatores, numa mesma população ou amostra
4. Quando queremos comparar duas ou mais amostras: vêm da mesma população ou de populações diferentes?
5. Quando queremos detetar regularidades que não são visíveis a olho nu.

Conceitos essenciais I

- **probabilidade**: um número que quantifica a possibilidade, entre 0 e 1
- **distribuição ou densidade da probabilidade** (para casos contínuos)
- **distribuição cumulativa**: proporção de casos com valor $<= x$
- **surpresa**: diferença em relação às nossas expetativas: quanto mais surpreendente, menos provável que seja "por acaso"
- **quantis**: intervalos associados a uma dada percentagem (decis, percentis, quartis...)
- independência
- **significância**: uma medida de replicabilidade dos resultados, medida pelo erro de tipo I, *alpha*: rejeitar uma hipótese que é verdadeira. É dada pelo valor de p, geralmente exigido < 0.05...
- **hipótese nula**: o que vai ser posto à prova, e que se aceita se a probabilidade não for demasiado pequena

Conceitos essenciais II

- **poder de um teste**: probabilidade de rejeitar a hipótese nula quando esta é falsa $(1 - \beta)$ (β é a probabilidade de cometer um erro de tipo II, aceitar uma hipótese quando é falsa)
- **intervalos de confiança**: intervalo no qual se encontra a resposta certa com uma certa probabilidade

A disciplina da estatística

A observação empírica de muitos casos encontrou várias distribuições de probabilidade – funções matemáticas que os estatísticos sugerem ser apropriadas para descrever a realidade.
(No R existe uma família de funções associada a cada distribuição.)
A estatística é a arte de usar estas distribuições de probabilidade aplicadas ao mundo real, usando "estatísticas", ou seja, números ou conjuntos de números como "estimadores" da realidade.

Tipos de objetos no nosso discurso

- distribuição de probabilidade
- modelos estatísticos
- tabelas de contingência
- amostras emparelhadas

Tipos de informação

Assim como na língua nós temos adjetivos, substantivos, e verbos (que têm propriedades diferentes e se usam de maneiras diferentes), assim também nas linguagens de programação há tipos diferentes (com operações associadas diferentes), e na estatística há métodos diferentes para tipos de informação diferentes:

- categorias (como nomes de árvores, naipes, sexo, nacionalidade) – correspondendo a variáveis categóricas, tambem chamadas **nominais**
- categorias ordenadas (muito, pouco, nada; criança, jovem, adulto) – variáveis **ordinais**
- contínuas (é possível fazer aritmética, comparando as diferenças)
  - números arbitrários, como as temperaturas
  - números com sentido, em que o zero tem lógica: só neste caso é que se podem fazer razões (divisões) entre os números

A distribuição de Poisson

Usada para descrever a distribuição de quantidades (inteiras) – no tempo ou no espaço – independentes entre elas:

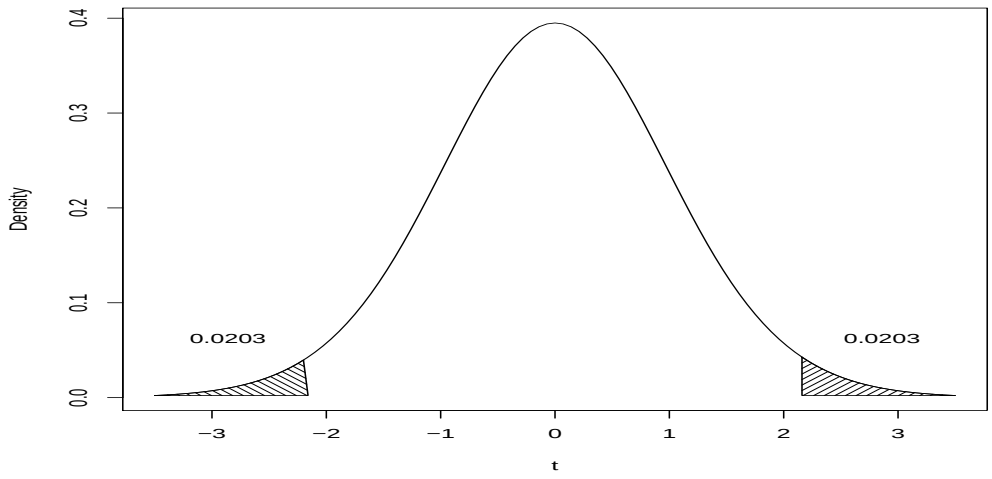
- número de assistentes numa aula
- número de casos de doença num mês
- número de sujidades num azulejo
- número de carros que passam numa ponte por dia

Sabemos quantas vezes um "fenómeno" aconteceu, mas não quantas vezes não aconteceu: número de suicídios, número de quedas de cavalo.

$$P(x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

Como se usa uma distribuição

Calcula-se a probabilidade de um dado resultado ser maior ou menor do que um dado valor. A probabilidade é o integral (ou a soma) dos valores à direita (ou à esquerda) desse valor.



Comandos no R

Em geral

```
dNOME(x,PARAMETROS)
qNOME(quantil,PARAMETROS)
pNOME(x,PARAMETROS)
rNOME(num,PARAMETROS)
```

```
dbinom(x,n,p)
qbinom(q,n,p)
pbinom(x,n,p)
rbinom(num,n,p)

dpois(x,lambda)
qpois(q,lambda)
ppois(x,lambda)
rpois(num,lambda)
```

Outras famílias: NORM, CHISQ, T, F, MULTINOM ...

RRRecreio

Entre no R. Experimente a distribuição de Poisson. Visualize para diferentes valores de λ.

Visualização no R

Uma das grandes vantagens desta linguagem, é que tem grandes capacidades gráficas e comandos "inteligentes".

```
plot()
boxplot()
hist()
assocplot()
barplot()
```

que conforme o tipo de argumentos produzem figuras diferentes, e que aceitam um número grande de parâmetros.

Para informáticos

Ao escrever funções no R, duas características interessantes:

1. definição dos valores por omissão na descrição da função

```
aminha<-function(x=3,y)
{ dist<- x-y
if (y>x) dist<-y-x
dist
}
```

2. passagem "vaga" de parâmetros

```
imprime<-function(tit="Em portugues")
{plot(xlab=tit,..)}
imprime(x=aleat)
```

Outras distribuições

F, χ^2, t, binomial, multinomial
As três primeiras são contínuas, as duas segundas são discretas.

RRRecreio (2)

Entre no R. Experimente outra distribuição. Visualize para diferentes valores dos parâmetros.

Voltemos à estatística... I

Desmistificando... a disciplina da estatística resume-se praticamente a estimar médias e variâncias, e a responder a perguntas baseada nestes conceitos, que, generalizando, se chamam momentos (de primeira ordem, de segunda, de terceira, etc.)

Centralidade Média dos sabem o que é... é uma medida que representa um conjunto de valores. Mediana e moda são outras medidas de centralidade.

Dispersão Variância é uma medida da variabilidade de um conjunto de valores. Outras medidas são o desvio padrão, o intervalo de variação (leque, "range"), o intervalo interquartis, o coeficiente de variação ...

Tendência Medida de assimetria (positiva ou negativa): pende mais para a direita ou para a esquerda?

Achatamento (Ou curtose) mede o grau de afilamento da curva relativamente à normal.

Voltemos à estatística... II

E relembrem – os estatísticos são platónicos! Uma coisa é os números que obtivemos, outra é a distribuição subjacente.
Por isso há sempre as medidas da distribuição, e as medidas da amostra, que podem ser estimadores das medidas da distribuição.
Lembrem-se: as medidas de uma amostra são para ser usadas como estimadoras da população.
Há dois tipos de estimadores: pelo método dos momentos, e pelo método da máxima verosimilhança (MLE - "maximum likelihood estimation").

RRRecreio (3)

Volte ao R. Leia alguns conjuntos de dados, e ausculte-os.

```
dim()
names()
summary()
```

Visualize-os. Altere aquilo que o R compreendeu mal.

RRRecreio (3)

Volte ao R. Leia alguns conjuntos de dados, e ausculte-os.

```
dim()
names()
summary()
```

Visualize-os. Altere aquilo que o R compreendeu mal.

```
condiv$novacoluna<-factor(condiv$decada)
```

Calcule as medidas de centralidade: média, mediana e moda, e as medidas de dispersão: desvio padrão e variância.

```
mean(), median(), sd(), var()
```

Folhas de registo I

- O medo: quantas palavras do campo semântico do medo encontradas em vários corpos ou subcorpos (Corpus NILC/São Carlos), correspondendo a géneros diferentes.
  Há diferença entre os géneros a este respeito?
- A cor: a mesma pergunta, mas agora feita ao CONDIV, que tem três temas, três décadas, e duas variantes.
  E há diferença entre autores no VERCIAL?
- Tamanho das traduções para inglês é maior do que dos originais em português? Tamanho das frases/sentenças no CorTrad jornalístico.
- Há palavras que são usadas mais numa variante do que noutra?
- Há correlação entre tempos usados por verbos?
- A forma de usar adjetivos pelas personagens da Jane Austen é diferente se forem homens ou mulheres? E se a conversa for entre um homem e uma mulher, ou com vários participantes?

Folhas de registo II

- É possível detetar se um dado tradutor é homem ou mulher? Ou se é traduzido ou não? Para cada texto, foi atribuído um conjunto de características/contagens de palavras gramaticais. E, com base nessas contagens, foi criado um classificador...
- Visualize a distribuição de 32 palavras neerlandesas acabadas em "likj" (Baayen 2008, página 234).
- Pesquise a distribuição de algumas palavras de cor mais frequentes no ConDiv.
- Investigue a distribuição dos 50 verbos mais frequentes de acordo com características de tempo e aspeto

A distribuição F

(Do nome de Fisher) A distribuição F é usada para comparar variâncias (é uma razão de variâncias).
Tem dois parâmetros: graus de liberdade da amostra do numerador, e da amostra do denominador.
Testa a homogeneidade da variância: Pode comparar duas amostras da mesma população, ou duas populações com a mesma variância.
É usada para verificar se as condições de um certo tipo de comparação – por exemplo ANOVA – estão preenchidas.

A distribuição χ^2

$\chi^2 = \sum_{i=1}^{N} \frac{(O_i - E_i)^2}{E_i}$
É usada em três situações diferentes

- quão apropriada uma certa expetativa é ("goodness of fit")
- se duas variáveis numa tabela de contingência são independentes, ou co-variam
- se uma população é homogénea em relação a uma variável

Nota: para poder usar χ^2, devemos estar a lidar com variáveis normais, observações independentes, e (todos os) valores esperados maiores de 5.

E a normal?

É a mais conhecida, devido à **lei dos grandes números**, que reza "mesmo que uma população não seja normal, as médias das suas amostras são-no".

- Mas a normal (ou gaussiana) não é aplicável à língua, pelo menos no caso do léxico, visto que estamos na situação de LNRE: "large numbers of rare events", ou seja cada palavra ou texto é muito rara (em muitos casos simplesmente não aparece), e por isso não se pode usar a aproximação normal.
- Doutra maneira: em relação ao vocabulário usado, as médias aumentam com o tamanho da amostra, mesmo que se aumente o texto, continuam a aumentar, e por isso não se pode assumir qualquer distribuição de médias.

O artigo de Dunning

Accurate Methods for the Statistics of Surprise and Coincidence

Em 1993 na revista *Computational Linguistics* teve uma repercussão enorme.
Ocorrências de palavras (da maior parte das palavras) são apropriadamente descritas por distribuições **binomiais**, e não normais.
A estimativa que se deve usar deve ser do tipo de máxima versosimilhança ("maximal likelihood"): G^2 é uma razão entre o máximo da hipótese que se testa e o máximo do espaço total, e $-2log\lambda$ é rapidamente aproximada por uma distribuição de χ^2.
Poisson e Fisher também são mencionados.

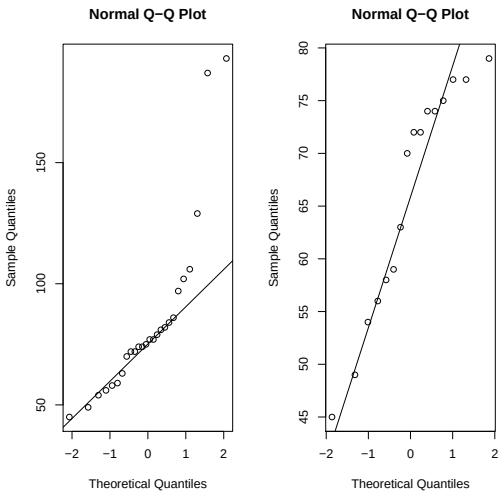
Distribuições paramétricas

Estas distribuições são paramétricas, no sentido de que, se conhecermos os parâmetros, sabemos o resto. Note-se que são **famílias** de distribuições, dependendo dos seus parâmetros.
Mas também existem outras formas de calcular a probabilidade, ou fazer testes, que não pressupõem a função conhecida – são não paramétricos.

Voltando à normal...

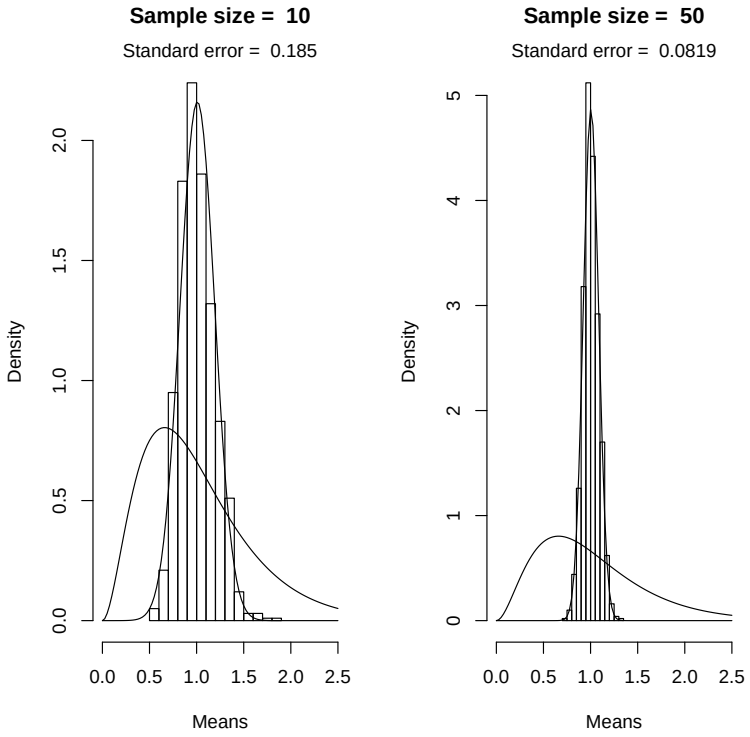
Mas e se estivermos na presença de um fenómeno normal? Como testar isso? Os testes dos quantis dão-nos uma ideia:

```
par(mfrow=c(1,2))
qqnorm(vot01)
qqline(vot01)
qqnorm(vot01[vot01<80])
qqline(vot01[vot01<80])
```



Observando o teorema do limite central

```
par(mfrow=c(1,2))
source("central.limit")
central.limit(10)
central.limit(50)
```



Perguntas que se podem fazer

Os métodos estatísticos apenas produzem números – nunca respondem diretamente a questões linguísticas. É a interpretação linguística que lhes dá sentido. (Stefan Evert, 2006)

- caracterizar a população através de amostras (estimação)
- averiguar se a diferença entre duas amostras é por acaso (teste de hipóteses)
- averiguar se duas ou mais características estão relacionadas (estudo da correlação)

Unidades

A unidade de amostragem (textos de um corpo) é quase sempre diferente da unidade de medida (por exemplo palavras, construções sintáticas, frases).
Ou seja, frases seguidas estão relacionadas, palavras seguidas ou próximas estão relacionadas. É o que se chama em inglês "burstiness" .
Este é um dos principais problemas da aplicação de métodos estatísticos à língua, já que os métodos estatísticos pressupõem uma amostra aleatória.

Testar se as diferenças são significativas

Por outras palavras, testar se é muito pouco provável que as diferenças sejam devidas simplesmente à sorte (entre amostras diferentes de um mesmo material).
Quanto menos provável for obter essas diferenças, mais provável é que elas se devam a uma diferença real (e não causada pela amostragem).
Testa-se entre medidas feitas a **amostras diferentes**.

- contagens em géneros diferentes
- contagens em variantes diferentes
- contagens em autores diferentes

```
prop.test(PROPORCAO1,PROPORCAO2)
```

Quando há várias medidas (ou fatores) que podem fazer sentido

Tem de se fazer uma análise multivariada, ou seja, tem de se variar todas as possibilidades, e investigar se existem dependências entre os vários fatores. Muito comum é ANOVA, para fatores categóricos e não numéricos.

Testar independência

Falta de correlação significa independência. Mas correlação não significa dependência no sentido de causalidade. Testa-se entre duas propriedades medidas às mesmas unidades:

- cor de cabelo e cor de olhos
- sexo e tipo de adjetivos
- fumador e doente de cancro do pulmão

```
chisq.test(TABELA)
```

Paradoxo de Simpson

Em conjunto (amontoado/"pooled") há uma direção de associação (marginal), mas, se virmos cada caso, temos associações condicionais na direção inversa.
Dados[1]: número de criminosos condenados à morte (ou não) na Florida (1976-1987), em termos da raça da vítima e da raça do defensor (branca/negra). Temos pois uma tabela de contingência 2x2x2.
Ignorando a raça da vítima, defensores negros são melhores a defender do que defensores brancos. Entrando em conta com a raça do assassino ("controlando a raça mantendo-a fixa"), a conclusão é inversa.
Tentativa de explicação: há uma dependência muito forte entre a raça do defensor e da vítima. A sentença de morte é muito mais frequente quando a vítima é branca.

[1] Exemplo de Agresti 1996. Embora chamado paradoxo de Simpson, foi primeiro documentado por Yule.

RRRecreio (4)

Explore o paradoxo de Simpson, calculando as tabelas intermédias e a tabela total.
Familiarize-se com as funções

`table, xtabs, prop.table`

Visualizar o material

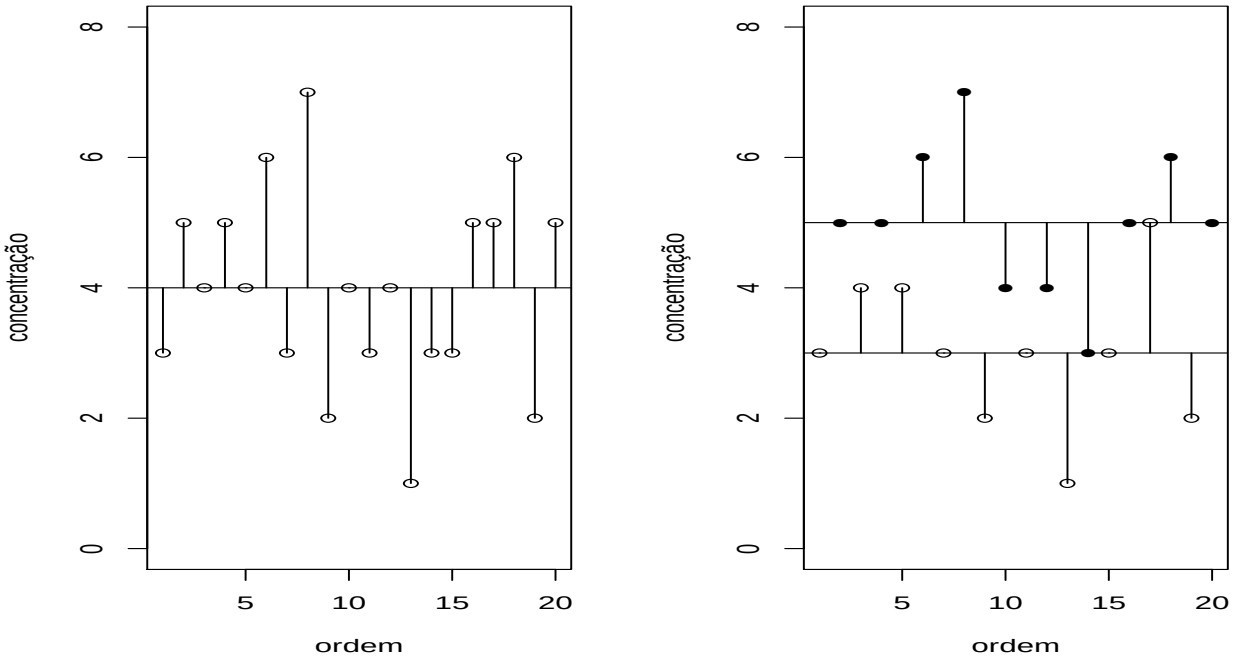
Muitas vezes temos muitos números, e muita informação, mas não sabemos o que fazer com ela. Por exemplo em relação a tempo: podemos experimentar com vários tamanhos de período, e com a identificação da data de publicação ou da data de nascimento do autor.
No caso dos dados do projeto vercial, muitos visualizações diferentes podem e devem ser tentadas.

ANOVA: dependência de vários fatores

ANOVA (de "analysis of variance") usa-se para testar se as médias relativas a contextos diferentes são dependentes desses contextos/fatores, ou se correspondem apenas a variação das medidas.

- ANOVA funciona comparando as várias variâncias: considerando médias diferentes, ou a média global.
- Para o método ser aplicável, é preciso que as variâncias das subpopulações sejam relativamente semelhantes.
- A tabela da ANOVA indica a percentagem de variância explicada pelos diferentes fatores. Idealmente, a variável deve ser normalmente distribuída em cada subpopulação.
- ANOVA é uma generalização do teste t, e usa a distribuição F ($F = t^2$).

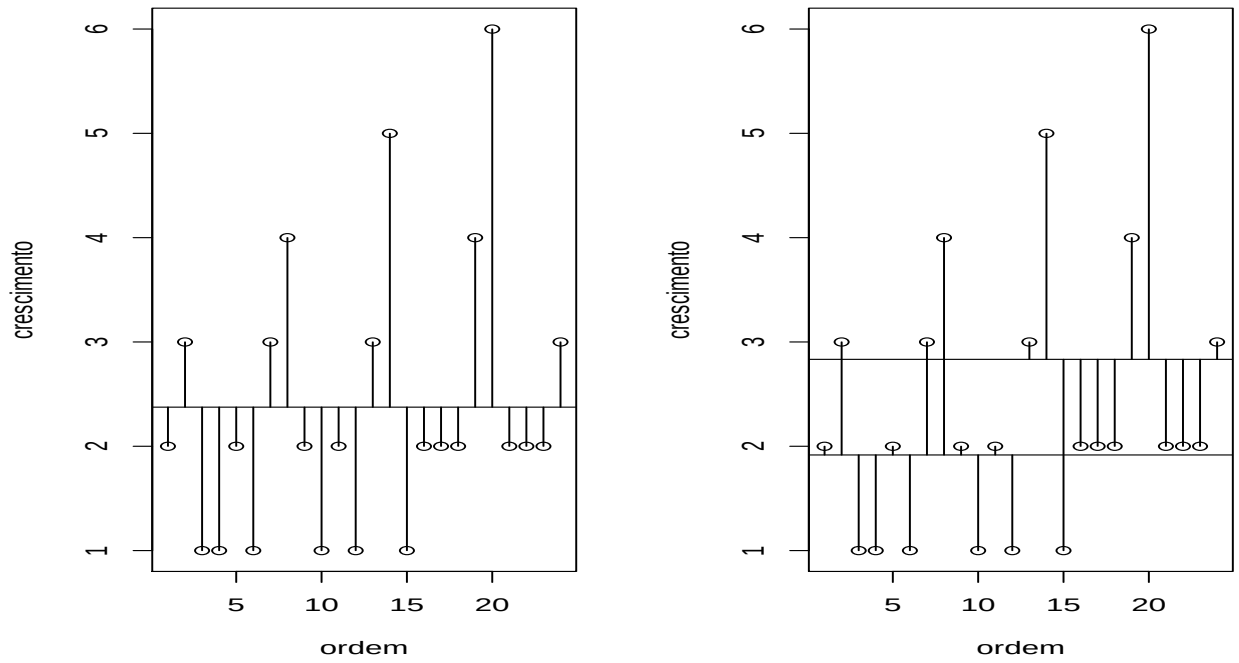
Exemplo detalhado de ANOVA (1)



Exemplo detalhado de ANOVA (1): veredito

```
> summary(aov(ozono~jardim))
            Df Sum Sq Mean Sq F value Pr(>F)
jardim       1     20 20.0000      15 0.001115 **
Residuals   18     24  1.3333
```

Exemplo detalhado de ANOVA (2)



Exemplo detalhado de ANOVA (2): veredito

```
> summary(aov(Growth~Photoperiod)
             Df Sum Sq Mean Sq F value Pr(>F)
Photoperiod   3   7.12   2.375   1.462  0.255
Residuals    20  32.50   1.625
```

Alternativas a ANOVA para contagens ou proporções

No caso de as variáveis não serem normais, e se referirem a contagens (Poisson) ou proporções (binomial), é preciso transformar as contagens noutra função, e fazer o que se chama regressão logística (de logaritmo). No R, usa-se os modelos lineares generalizados

```
glm( A ~ B, poisson)
glm( A ~ B, quasipoisson)
glm( A ~ B, binomial)
```


Últimas palavras

Método estatístico, segundo Cramer

- recolha de dados (obtenção da amostra dada a população)
- descrição desses mesmos dados
- análise do que eles mostram
- previsão

Muito ficou por contar...

1. Existe uma classe de métodos, os exploratórios, que apenas pode aqui ser ilustrada mas que corresponde muito provavelmente ao grosso dos métodos usados em corpos...
   O seu objetivo é descobrir padrões, associações, e tendências em grandes quantidades de material.
2. Métodos não paramétricos: usando simulação e métodos numéricos, têm a vantagem de não pressupor nada sobre as funções subjacentes...
3. Crítica de modelos estatísticos: como desenvolver um modelo através de refinamento ou de comparação...
4. Modelos geométricos da língua

Resumo da estatística

Três tipos de técnicas

1. Estatística descritiva: descrição do material
2. Teste de hipóteses: de independência ou homogeneidade, de diferença ou concordância com uma distribuição teórica, ou entre duas populações
3. Exploração de dados

Resumo do R

1. funções estatísticas para cada um dos casos
2. visualização
3. contato com o exterior

Obrigada pela atenção!