

Resultados do projeto “Estatística descritiva de um banco de dados de pacientes pediátricos com tumores do sistema nervoso central”

Um caderno aberto de pesquisa

Francisco Hélder Cavalcante Félix, Centro Pediátrico do Câncer - Hospital Infantil Albert Sabin

Abstract: Um banco de dados de pacientes pediátricos com tumores do sistema nervoso central diagnosticados em um grande hospital estadual foi criado pelos autores. Os resultados e análises foram colocados num repositório do serviço GitHub. O código das análises foi checado com o serviço de integração contínua em nuvem Travis Ci e o resultado final publicado neste arquivo. Este arquivo traz tabelas, gráficos e texto mostrando os resultados do projeto. Foi elaborado em Rmarkdown, utilizando a linguagem de marcação simplificada Markdown com “pedaços” de código da linguagem estatística R entremeados. O arquivo foi avaliado pelo pacote rmarkdown e compilado para o formato pdf neste texto. Devido à integração contínua, pode ser atualizado em tempo real, enquanto dados novos são acrescentados.

Keywords: tumores do sistema nervoso central, cancerologia pediátrica, estatística descritiva, rmarkdown, integração contínua, ciência aberta

January 21, 2018

Introdução:

A ciência aberta baseia-se principalmente na capacidade de divulgar (compartilhar) eletronicamente as informações coletadas (dados brutos) e produzidas (análises e seus resultados) de um projeto de pesquisa através da internet. Dessa forma, 2 consequências advêm imediatamente: 1 - Transparência da informação e do processo científicos, inclusive para públicos não técnicos. 2 - Capacidade irrestrita de comentários, tanto por especialistas (análogo à *revisão por pares*), quanto por não especialistas (que poderíamos chamar de *revisão cidadã*).

O observador arguto já pode levantar a questão de que o controle de comentários numa plataforma é dos controladores daquele serviço, ou seja, é possível criar um canal de comunicação tipo “ciência aberta”, porém unidirecional, sem recepção de comentários (ou pior, é possível censurá-los). Independentemente disso, os comentários a uma publicação livremente disponível na internet podem ser publicados em qualquer canal sem relação com o canal original e referenciado ao primeiro. Ou seja, não há como verdadeiramente censurar comentários a uma publicação livre na internet.

Outra característica importante da ciência aberta é a capacidade de *reuso* de informações, o que pode ser entendido como a principal utilidade social da ciência aberta. Esta característica simples tem o potencial de otimizar a produção científica a nível global. Bastaria isso para justificar a implementação em larga escala da ciência aberta. Outros benefícios podem ser descritos de forma ilimitada.

Esta é a implementação de ciência aberta que criei, baseando-me largamente em projetos já existentes. Trata-se de um caderno de pesquisa aberto, armazenado num repositório remoto para o programa Git (existem vários), gerado através de um serviço de integração contínua (CI) em nuvem (vários idem) e com a estrutura de um pacote da linguagem estatística R, usada para as análises. *Não se trata de um pacote verdadeiro*, apesar de ter um diretório de código R e um arquivo de definições DESCRIPTION. O objetivo desse mimetismo é facilitar as análises numa plataforma de CI. Um pacote de R é um programa com funções utilizáveis. Não é isso que este(s) caderno(s) é(são). Assim, propositadamente deixo de fora partes imprescindíveis de um pacote, como o NAMESPACE e os manuais.

Na seção a seguir, são mostrados resultados de análises estatísticas concernentes a este caderno aberto de pesquisa em particular. Todos os dados pertinentes a seres humanos são adequadamente desidentificados.

Análises:

```
require(pander)
require(survival)
snc<-read.csv('../data/snc.csv')
attach(snc)
```

```
barplot(summary(as.factor(snc$sex)),names.arg=c("masculino","feminino"),
xlab="Sexo",width=0.5,xlim=c(0,1.7),space=0.5,col=0)
```

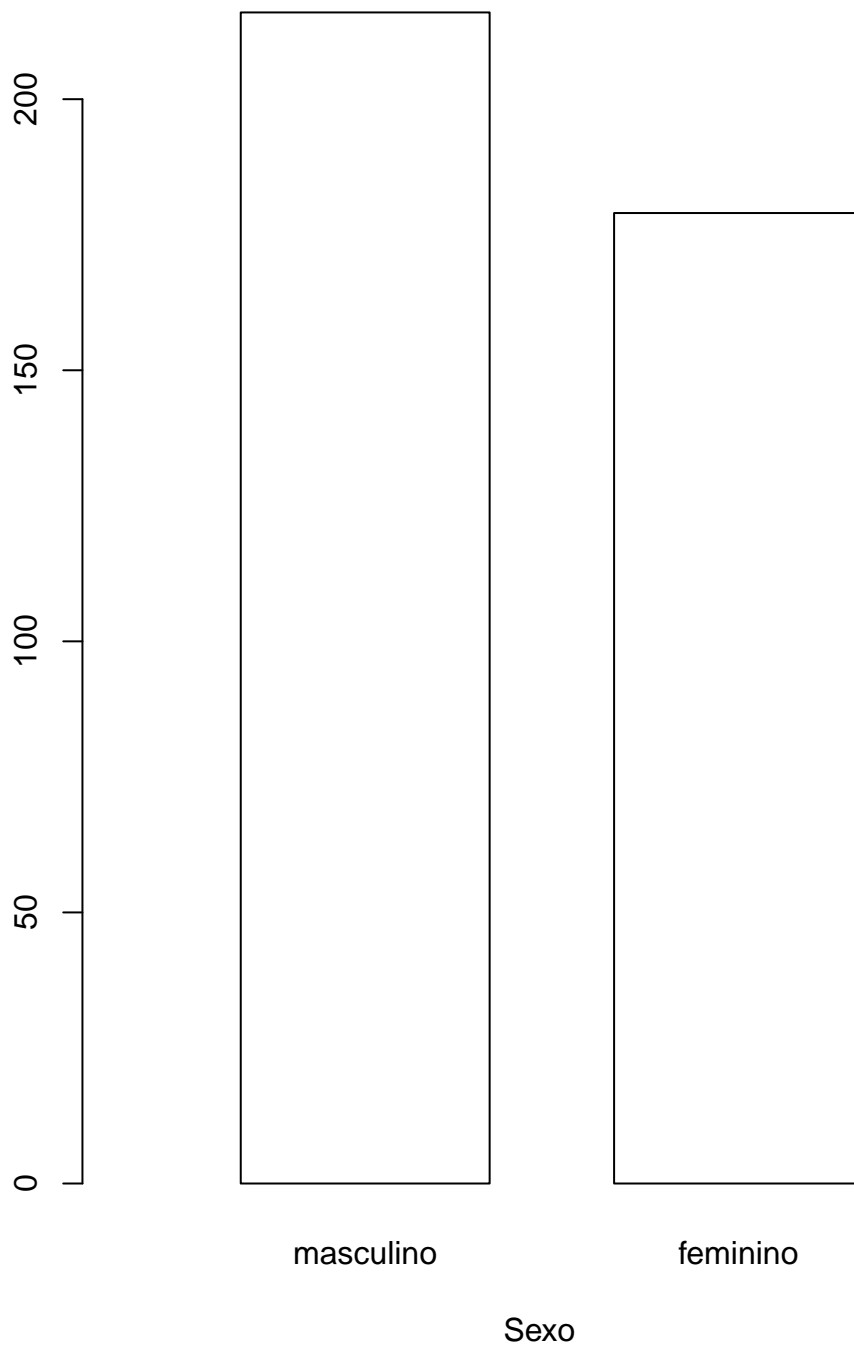


Figura 1: número de pacientes segundo o sexo.

```
boxplot(snc$age/365.25,xlab="Idade (anos)",boxwex=0.6,staplewex=0.4,  
frame.plot=F)
```



Figura 2: idade dos pacientes ao diagnóstico.

Min. : 1.0 , 1st Qu.: 99.5 , Median :198.0 , Mean :198.0 , 3rd Qu.:296.5 , Max. :395.0 , NA, Min. :0.0000 , 1st Qu.:0.0000 , Median :0.0000 , Mean :0.4532 , 3rd Qu.:1.0000 , Max. :1.0000 , NA, Min. : 28 , 1st Qu.:1609 , Median :2658 , Mean :2839 , 3rd Qu.:4012 , Max. :6546 , NA, Min. : 6 , 1st Qu.: 225 , Median : 629 , Mean :1079 , 3rd Qu.:1626 , Max. :5620 , NA, Min. :0.0000 , 1st Qu.:0.0000 , Median :1.0000 , Mean :0.6835 , 3rd

Qu.:1.0000 , Max. :3.0000 , NA, Unknown :105 , Meduloblastoma classico : 54 , Astrocitoma pilocítico : 30 , Astrocitoma difuso/fibrilar: 26 , Ependimoma : 26 , Ependimoma anaplasico : 17 , (Other) :137 , Min. :1.000 , 1st Qu.:2.000 , Median :3.000 , Mean :2.741 , 3rd Qu.:4.000 , Max. :4.000 , NA's :105 , Cerebelo :116 , DIPG : 64 , Diencefalo : 63 , Gliomatose : 31 , IV_Vent : 28 , Ventriculo Lat: 16 , (Other) : 77 , Fortaleza:131 , Caucaia : 21 , Itapipoca: 11 , Caninde : 7 , Maracanaú: 7 , Itapaje : 6 , (Other) :212 , Min. :0.000 , 1st Qu.:0.000 , Median :0.500 , Mean :1.159 , 3rd Qu.:2.000 , Max. :4.000 , NA's :23 , Min. : 0.00 , 1st Qu.: 0.00 , Median : 50.00 , Mean : 50.13 , 3rd Qu.: 90.00 , Max. :100.00 , NA's :24 , Min. : 0.000 , 1st Qu.: 0.000 , Median : 1.000 , Mean : 3.706 , 3rd Qu.: 4.000 , Max. :27.000 , NA, Min. : 0.0 , 1st Qu.: 0.0 , Median : 5.0 , Mean : 226.3 , 3rd Qu.: 5.4 , Max. :9640.0 , NA, Min. :0.000 , 1st Qu.:0.000 , Median :1.000 , Mean :1.111 , 3rd Qu.:2.000 , Max. :2.000 , NA, Min. : 0.000 , 1st Qu.: 0.000 , Median : 0.000 , Mean : 1.775 , 3rd Qu.: 0.000 , Max. :25.000 , NA, Min. : 0.0 , 1st Qu.: 0.0 , Median : 0.0 , Mean : 156.3 , 3rd Qu.: 0.0 , Max. :7200.0 , NA, Min. :0.0000 , 1st Qu.:0.0000 , Median :0.0000 , Mean :0.1744 , 3rd Qu.:0.0000 , Max. :2.0000 , NA's :28 , Min. : 0.0000 , 1st Qu.: 0.0000 , Median : 0.0000 , Mean : 0.6177 , 3rd Qu.: 0.0000 , Max. :27.0000 , NA, Min. :0.00000 , 1st Qu.:0.00000 , Median :0.00000 , Mean :0.09114 , 3rd Qu.:0.00000 , Max. :9.00000 , NA, Min. : 0.00000 , 1st Qu.: 0.00000 , Median : 0.00000 , Mean : 0.03333 , 3rd Qu.: 0.00000 , Max. :13.00000 , NA's :5

Dados e código para replicação estão disponíveis no repositório do GitHub do projeto