A `knitr` + LaTeX dynamic report.

Giusi Moffa

Statistical bioinformatics, University of Regensburg

27 May 2014

Practical bioinformatics

p-values, one big topic for reproducible findings

So let's look at how we can get a dynamic and reproducible report with 'knitr', for getting in the news.

Recording the *R* session.

It may be useful since different versions in general will not produce identical results.

```
# Print R session info
toLatex(sessionInfo())
```

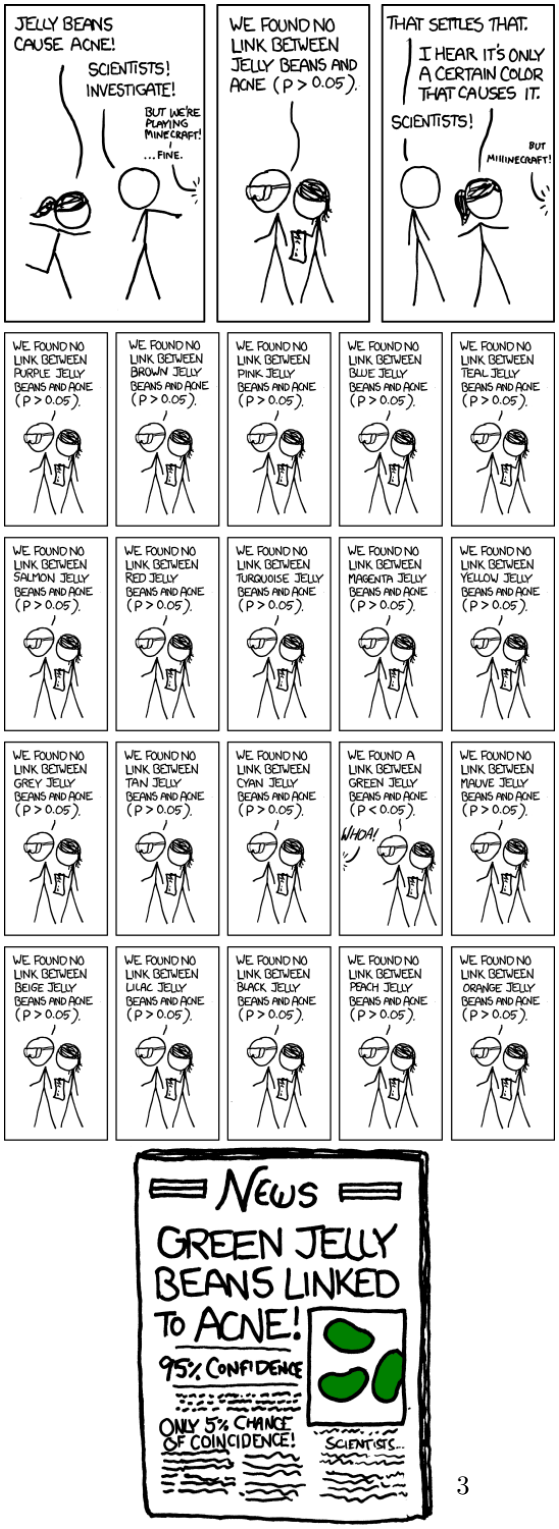
- R version 3.0.3 (2014-03-06), `x86_64-apple-darwin10.8.0`
- Locale: `en_GB.UTF-8/en_GB.UTF-8/en_GB.UTF-8/C/en_GB.UTF-8/en_GB.UTF-8`
- Base packages: base, datasets, graphics, grDevices, methods, stats, utils
- Other packages: knitr 1.5
- Loaded via a namespace (and not attached): evaluate 0.5.5, formatR 0.10, stringr 0.6.2, tools 3.0.3

It is a good habit also to set your working directory and check whether you are in the right place.

```
setwd("~/juicy/BioStatWork2012/mytex/journalClub/Reproducibility/repBioInfo")
getwd()  ## obtain the working directory

## [1] "/Users/giusimoffa/juicy/BioStatWork2012/mytex/journalClub/Reproducibility/repBioInfo
```

One old trick for getting in the news



Under the null-hypothesis p-values are expected to be uniformly distributed between 0 and 1. So if we have hundreds of crazy hypotheses we can try a fishing expedition for significance by testing all of them. On average 5

```
set.seed(31)
nColors <- 200
nObs <- 100
jellyBeans <- matrix(rnorm(nObs * nColors), ncol = nColors)
fishing4News <- apply(jellyBeans, 2, function(x) t.test(x)$p.value)
```

If we like simple numbers we can look at a table of summaries (here only for a small number of variables, for space constraints)

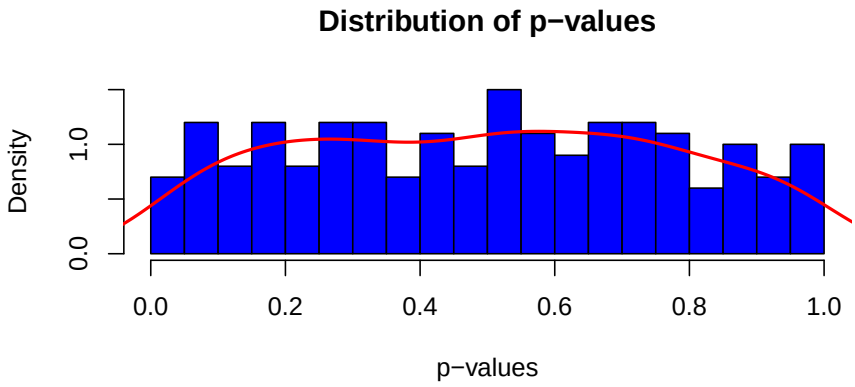
```
library(xtable)
xtable(summary(jellyBeans[,1:4]),
       caption="Some variables' summaries")
```

	V1	V2	V3	V4
1	Min. :-2.5408	Min. :-2.2187	Min. :-2.3257	Min. :-2.523
2	1st Qu.:-0.6512	1st Qu.:-0.6141	1st Qu.:-0.7047	1st Qu.:-0.425
3	Median :-0.0442	Median : 0.0846	Median :-0.1013	Median : 0.193
4	Mean :-0.0151	Mean : 0.0628	Mean :-0.0306	Mean : 0.113
5	3rd Qu.: 0.6460	3rd Qu.: 0.7606	3rd Qu.: 0.7086	3rd Qu.: 0.889
6	Max. : 2.3414	Max. : 2.2036	Max. : 2.6253	Max. : 2.159

Table 1: Some variables' summaries

A graph might be more pleasant. The distribution can be visualised through the 'hist' function and 'density' function.

```
hist(fishing4News, col=4, breaks=20, freq=FALSE,
     main="Distribution of p-values", xlab="p-values")
lines(density(fishing4News), col=2, lwd=2)
```



This way we get ‘r length(which(fishing4News¡.05))‘ significant results. This problem is well known in genomics and adressed by a plethora of methods for multiple correction. Transparency however is not always necessarily guaranteed, especially in in fields such as social science, and the issue is still a very much debated hot topic. (E.g. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "*p*-hacking" and the research hypothesis was posited ahead of time).

Correction for multiple testing

One way to account for multiple testing is by adopting the Benjamini-Hochberg correction. If we are "lucky" we can still get in the news after correcting for multiple testing, but we might need a rather larger number of crazy ideas.

```
set.seed(7)
nColors <- 10000
nObs <- 20
jellyBeans <- matrix(rnorm(nObs * nColors), ncol = nColors)
fishing4News <- apply(jellyBeans, 2, function(x) t.test(x)$p.value)
BHfishing <- p.adjust(fishing4News, "BH")
```

When testing for ‘r nColors‘ colours, and with a limited number (‘r nObs‘ in this case) of observations we still get one significant result. However only ‘r length(which(BHfishing ¡ .8))‘ are below .8.

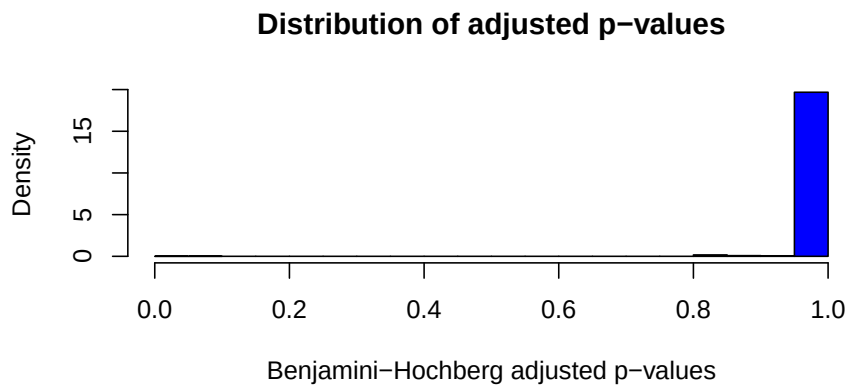
```
min(BHfishing)

## [1] 0.01483

sort(BHfishing)[1:5]

## [1] 0.01483 0.07893 0.83831 0.83831 0.83831
```

```
hist(BHfishing, col=4, breaks=20, freq=FALSE,
     main="Distribution of adjusted p-values",
     xlab="Benjamini-Hochberg adjusted p-values")
```



And if you really feel you need to, you can also save (and load) your entire workspace

```
save.image(file = "fishing.RData")
load(file = "fishing.RData")
```

Disclaimer

This report is freely available for the benefit of **science**, so that our steps on the way to the news can be checked by anybody who wishes to do so.