Simple examples simulation from ‘Why significant variables aren’t automatically good predictors.’ (2015)

This is an R Markdown Notebook. When you execute code within the notebook, the results appear beneath the code.

Set up the variabile sets

```
vs2.a=matrix(rep(c(50, 61), 9), 2, 9, byrow=T)/999
vs2.b=matrix(c(c(5, 26), rep(67, 7), c(26, 5), rep(67, 7)),
                2, 9, byrow=T)/1000
options(digits = 2)
vs2.a
```

```
##       [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9]
## [1,] 0.050 0.061 0.050 0.061 0.050 0.061 0.050 0.061 0.050
## [2,] 0.061 0.050 0.061 0.050 0.061 0.050 0.061 0.050 0.061
```

```
rowSums(vs2.a)
```

```
## [1] 0.49 0.51
```

```
vs2.b
```

```
##       [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9]
## [1,] 0.005 0.026 0.067 0.067 0.067 0.067 0.067 0.067 0.067
## [2,] 0.026 0.005 0.067 0.067 0.067 0.067 0.067 0.067 0.067
```

```
rowSums(vs2.b)
```

```
## [1] 0.5 0.5
```

Normalizing step

```
vs2.a[1,]=vs2.a[1,]/sum(vs2.a[1,])
vs2.a[2,]=vs2.a[2,]/sum(vs2.a[2,])
vs2.b[1,]=vs2.b[1,]/sum(vs2.b[1,])
vs2.b[2,]=vs2.b[2,]/sum(vs2.b[2,])
vs2.a
```

```
##      [,1]  [,2] [,3]  [,4] [,5]  [,6] [,7]  [,8] [,9]
## [1,] 0.10 0.123 0.10 0.123 0.10 0.123 0.10 0.123 0.10
## [2,] 0.12 0.099 0.12 0.099 0.12 0.099 0.12 0.099 0.12
```

```
rowSums(vs2.a)
```

```
## [1] 1 1
```

```
vs2.b
```

```
##       [,1]  [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9]
## [1,] 0.010 0.052 0.13 0.13 0.13 0.13 0.13 0.13 0.13
## [2,] 0.052 0.010 0.13 0.13 0.13 0.13 0.13 0.13 0.13
```

```
rowSums(vs2.b)
```

```
## [1] 1 1
```

Bayes rate

Variable set 1

```
sum(pmax(vs2.a[1,], vs2.a[2,]))/2
```

```
## [1] 0.55
```

Variable set 2

```
sum(pmax(vs2.b[1,], vs2.b[2,]))/2
```

```
## [1] 0.52
```

Running the simulation

```
BB=10000

chi2.a=rep(0, BB)
chi2.b=rep(0, BB)
I2.a=rep(0, BB)
I2.b=rep(0, BB)

for(i in 1:BB){

    tab.a =cbind(rmultinom(1, 500, vs2.a[1,]),
                 rmultinom(1, 500, vs2.a[2,]))

    tab.b =cbind(rmultinom(1, 500, vs2.b[1,]),
                 rmultinom(1, 500, vs2.b[2,]))

    chi2.a[i]=chisq.test(tab.a)$p.value
    chi2.b[i]=chisq.test(tab.b)$p.value

    I2.a[i]=500*sum((tab.a[,1]/500-tab.a[,2]/500)^2)/2
    I2.b[i]=500*sum((tab.b[,1]/500-tab.b[,2]/500)^2)/2


}
```

Simulations results

Median I scores

```
options(digits=3)
median(I2.a)
```

```
## [1] 1.88
```

```r
median(I2.b)
```

```
## [1] 1.68
```

Median p-value

```r
options(digits=3)
median(chi2.a)
```

```
## [1] 0.0314
```

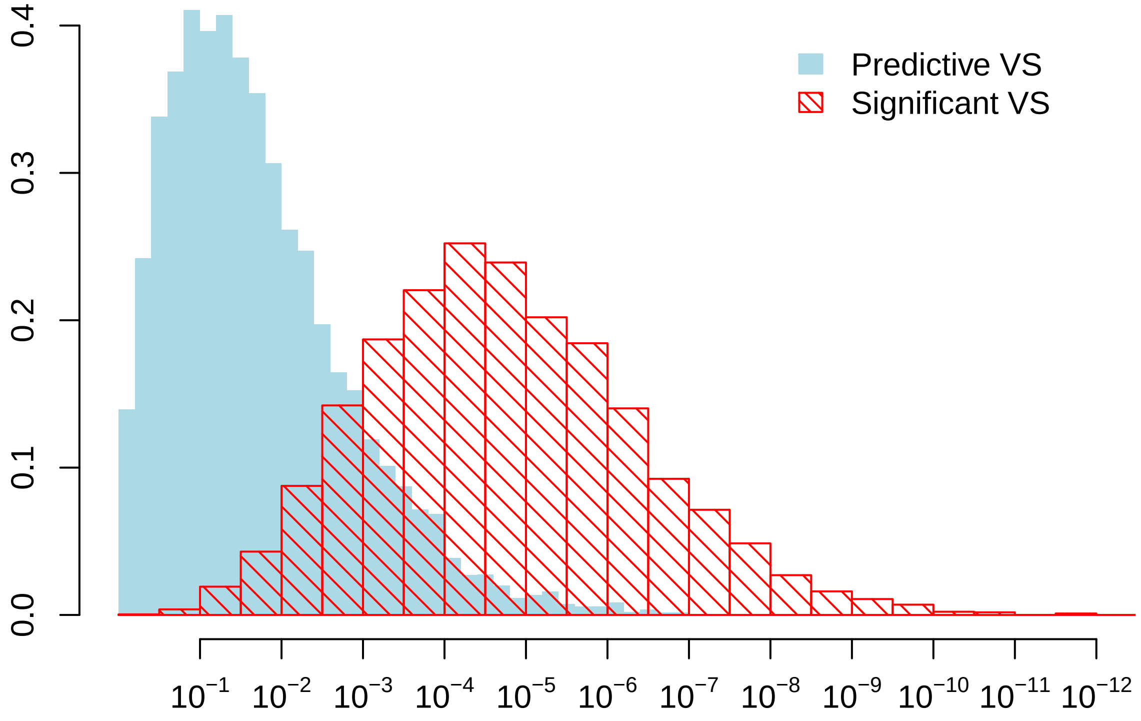
```r
median(chi2.b)
```

```
## [1] 2.54e-05
```

```r
par(mfrow=c(1,1), mar=c(4.5,4,1,1), font.main=1, cex.main=1)

hist(-log(chi2.a)/log(10), breaks=30, freq=F,
     #density=6, lwd=1.5,
     col="light blue",
     border=0,
     xlim=c(0,12), main="",
     xlab="", ylab="", xaxt="n")
ticks <- seq(1, 12, by=1)
labels <- sapply(ticks, function(i) as.expression(bquote(10^ .(-i))))
axis(1, at=1:12, labels=labels)

#abline(v=qchisq(1-0.01, 2), lwd=2)
hist(-log(chi2.b)/log(10), breaks=40, freq=F, density=12, lwd=2, angle=-45,
col=rgb(1,0,0),
xlim=c(0,7), main="",
     xlab="",ylab="",
     add=T
     )
#abline(v=qchisq(1-0.01, 2), lwd=2)
legend(8, 0.4, c("Predictive VS", "Significant VS"), fill=c("light blue", "red"), border=c("light blue"
```


0.0
0.1
0.2
0.3
0.4
Predictive VS
Significant VS
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{-7}
10^{-8}
10^{-9}
10^{-10}
10^{-11}
10^{-12}