Practical 2: Hardy Weinberg Equilibrium and multiple testing

Jerome Goudet and Bruce Weir

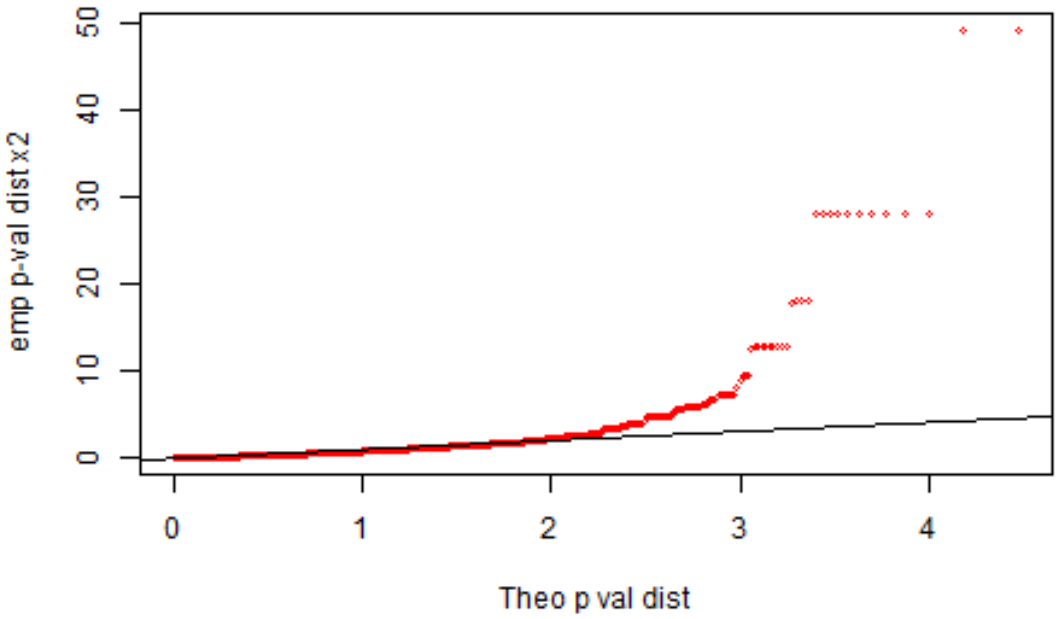
2021-07-15

```
library(gaston)
library(hierfstat)
library(JGTeach)
```

HW χ^2 tests

1. With the `pan` bed object you have crated in the previous practical, calculate for each locus its inbreeding coefficient ; use it to test whether the loci are in Hardy Weinberg Equilibrium ; plot $-log10$ of these p-values against their expectations under the null hypothesis.

```
# F=1-Ho/He
pan<-ms2bed("https://www2.unil.ch/popgen/teaching/SISGData/pan.txt")
inb.coeff<-1-pan@snps$hz/2/pan@p/(1-pan@p)
#nb inds
ni<-dim(pan)[1]
x2<-ni*inb.coeff^2
p.val.x2<-pchisq(x2,df=1,lower=FALSE)
nl<-dim(pan)[2]
#theo dist p.val under null
theo.pval<-1:nl/nl
plot(-log10(theo.pval),-log10(sort(p.val.x2)),
     col="red",cex=0.5,xlab="Theo p val dist",
     ylab="emp p-val dist x2");abline(c(0,1))
```

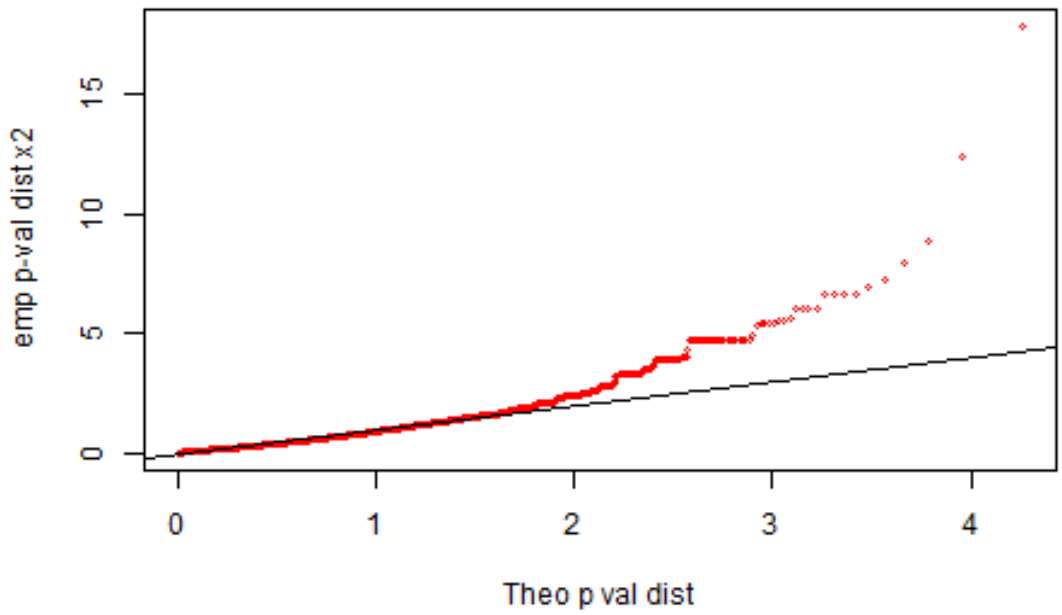


Is it what you would have expected ?

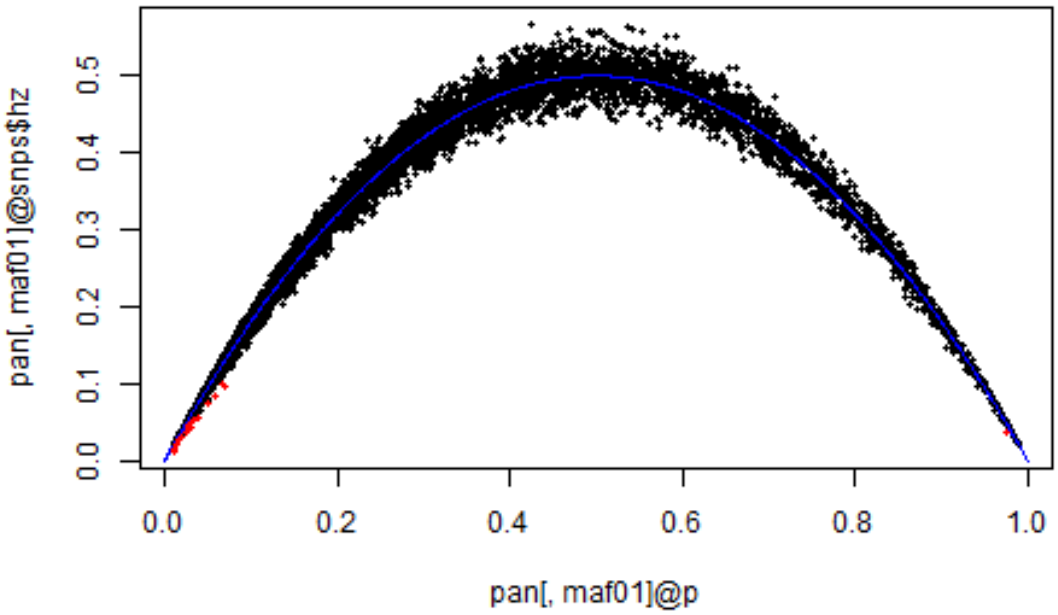
```
No, we would have expected no more
significant tests than expected by chance,
i.e, values along the diagonal of the graph
```

2. redo the same but for loci with minor allele count of at least 10 (maf ≥ 0.01 ; this filtering "rule" is often used in genomic analysis). Identify, using e.g. a different color, the loci that are not in HWE according to this test on the plot of observed heterozygosity against allele frequencies. Do you see a pattern ?

```
x<-0:1000/1000
maf01<-which(pan@snps$maf>=0.01)
nl<-length(maf01)
theo.pval<-1:nl/nl
plot(-log10(theo.pval),-log10(sort(p.val.x2[maf01])),
     col="red",cex=0.5,xlab="Theo p val dist",
     ylab="emp p-val dist x2");abline(c(0,1))
```




```
plot(pan[,maf01]@p,pan[,maf01]@snps$hz,col="black",pch=16,cex=0.6)
lines(x,2*x*(1-x),col="blue")
outliers<-which(-log10(p.val.x2[maf01])>4)
points(pan[,maf01][,outliers]@p,pan[,maf01][,outliers]@snps$hz,col="red",pch=16,cex=0.6)
```




```
filtering on MAC (minor allele count) at 10 improves a bit,
but not much. All of the outliers are in the tail of
p distribution, for high or low freq.
```

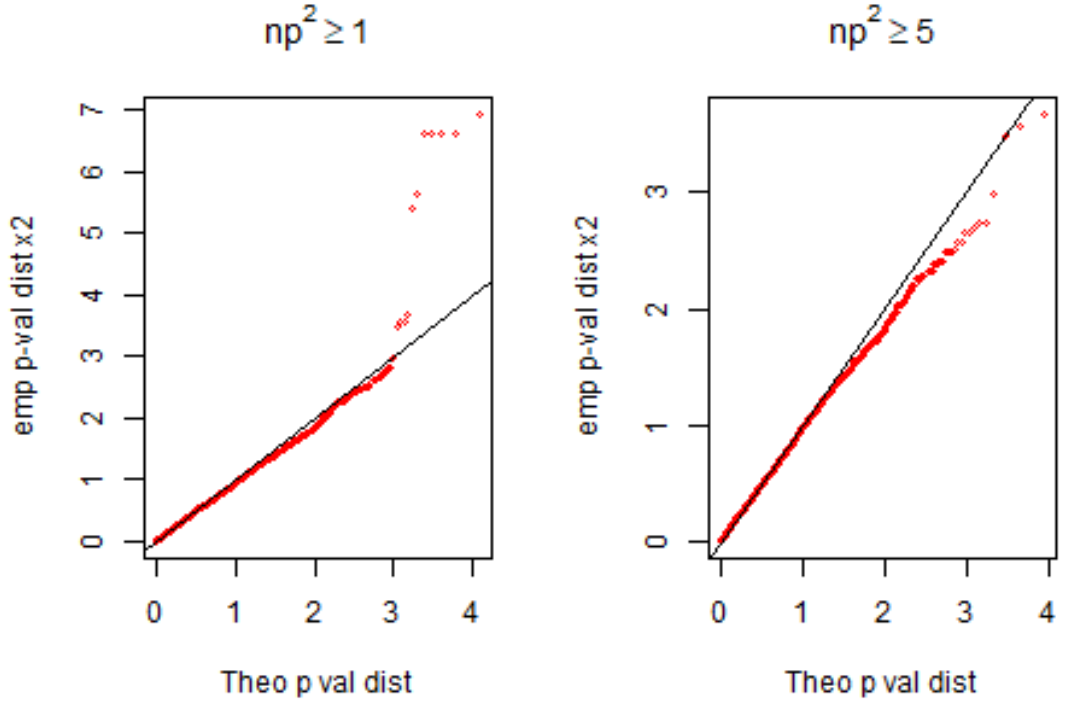
— A rule often stated for the validity of a χ^2-test is

> The minimum expected number per cell is 1 (5) [the proportion of cells with expected counts lower than 5 should not exceed 20%]

Is this rule working here ?

```
par(mfrow=c(1,2))
#what frequency leads to np^2==1 e.g. p=(1/n)^0.5
#nb inds
ni<-dim(pan)[1]
xi<-1
mafn1<-which(pan@snps$maf>=(xi/ni)^.5)
nl<-length(mafn1)
theo.pval<-1:nl/nl
plot(-log10(theo.pval),-log10(sort(p.val.x2[mafn1])),
     col="red",cex=0.5,xlab="Theo p val dist",
     ylab="emp p-val dist x2",main=expression(np^2>=1));abline(c(0,1))

#what frequency leads to np^2==5
xi<-5
mafn5<-which(pan@snps$maf>=(xi/ni)^.5)
nl<-length(mafn5)
theo.pval<-1:nl/nl
plot(-log10(theo.pval),-log10(sort(p.val.x2[mafn5])),
```

```
     col="red",cex=0.5,xlab="Theo p val dist",
     ylab="emp p-val dist x2",main=expression(np^2>=5));abline(c(0,1))
```




```
par(mfrow=c(1,1))
```

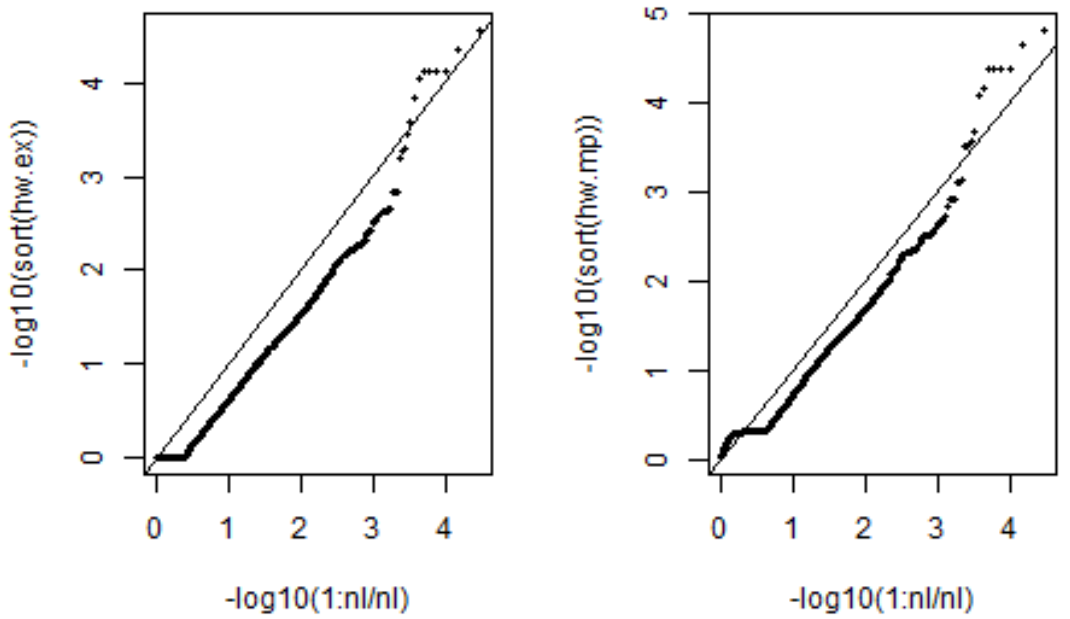
```
in Q2 we use np as our expected, but in fact it is np^2,
the expected number of rare homozygotes.
While things are better when np^2=1, we still have outliers,
and it is only with np^2=5 that we remove most of them.

We are thus loosing very many loci for no reasons
when using the chi2 test
```

HW exact tests

3. Now load (install it if not done yet) the `HardyWeinberg` library, and use the function `HWExactStats` to obtain the exact p-values for these loci, using first the argument `midp` set to FALSE and then set to TRUE.

```
library(HardyWeinberg)
```

```
hw.ex<-HWExactStats(cbind(pan@snps$N0,pan@snps$N1,pan@snps$N2),midp=FALSE)
hw.mp<-HWExactStats(cbind(pan@snps$N0,pan@snps$N1,pan@snps$N2),midp=TRUE)
nl<-length(hw.mp)
par(mfrow=c(1,2))
plot(-log10(1:nl/nl),-log10(sort(hw.ex)),cex=0.6,pch=16);abline(c(0,1))
plot(-log10(1:nl/nl),-log10(sort(hw.mp)),cex=0.6,pch=16);abline(c(0,1))
```




```
par(mfrow=c(1,1))
```

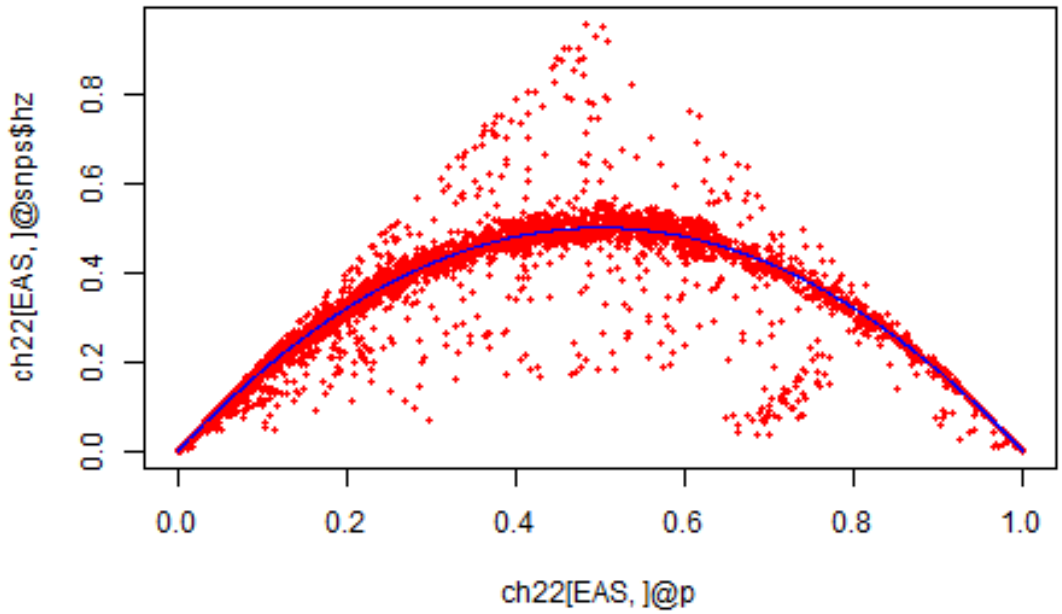
```
Using expact p values or exact mid p values improve things
drastically, as we don't have outlier loci anymore
-> use mid p-value (or exact test, not much difference) for QC
```

4. using the East Asian samples from the 1000 genome project, plot the SNPs heterozygosity against the frequency of the alternate allele for chr22 :0-20M. Add to the plot a line showing the expected heterozygosity under Hardy Weinberg Equilibrium ; Discuss the resulting figures, compared to what we saw with the simulated data.

```
ch22<-read.VCF("chr22_Mb0_20.recode.vcf.gz")
```

```
## ped stats and snps stats have been set.
## 'p' has been set.
## 'mu' and 'sigma' have been set.
```

```
samp.desc.file<-"https://www2.unil.ch/popgen/teaching/SISG18/integrated_call_samples_v3.20130502.ALL
samp.desc<-read.table(samp.desc.file,header=TRUE)
EAS<-which(samp.desc$super_pop=="EAS")
plot(ch22[EAS,]@p,ch22[EAS,]@snps$hz,col="red",pch=16,cex=0.6)
lines(x,2*x*(1-x),col="blue")
```



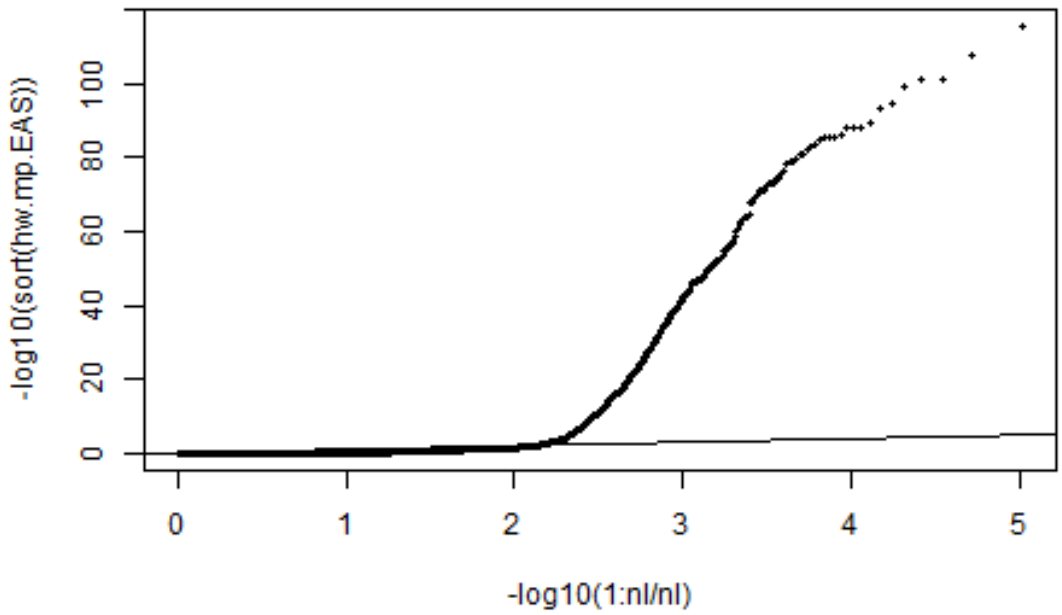

```
when we look at  EAS samples hz vs p,
we see a LOT of points far away from the blue line,
despite sample size being around 500. In particular
many points along the edges of a triangle.
Probably issues with SNP callin
(even though this is supposed to be a highly curated data set!)
```

5. Test for Hardy Weinberg using the exact mid-pvalue test for all these loci.

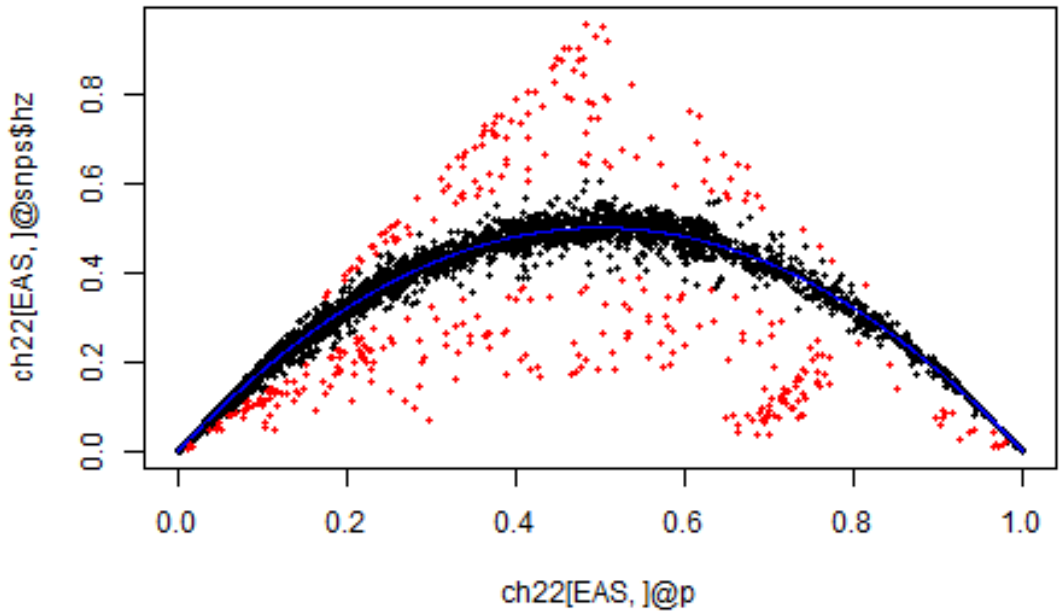
```
hw.mp.EAS<-HWExactStats(cbind(ch22[EAS,]@snps$N0,ch22[EAS,]@snps$N1,ch22[EAS,]@snps$N2),midp=TRUE)
```

6. plot the p-values of the previous tests against their expectation under the null (Rather than the p-values, -log10 of the p-values is more informative). Identify on the plot of heterozygosity against allele frequencies the loci not conforming to HWE, and discuss the results.

```
nl<-length(hw.mp.EAS)
plot(-log10(1:nl/nl),-log10(sort(hw.mp.EAS)),cex=0.6,pch=16);abline(c(0,1))
```




```
outliers<-which(-log10(hw.mp.EAS)>6)
plot(ch22[EAS,]@p,ch22[EAS,]@snps$hz,col="black",pch=16,cex=0.6)
lines(x,2*x*(1-x),col="blue")
points(ch22[EAS,outliers]@p,ch22[EAS,outliers]@snps$hz,col="red",pch=16,cex=0.6)
```




```
by filtering out on mid p-values, we managed to remove most outliers,
even in the tails of allelic frequencies distribution. Job done!
Note still that we may remove these loci for pop structure analysis etc,
but we might have very interesting biological signal there, so must investigate
```

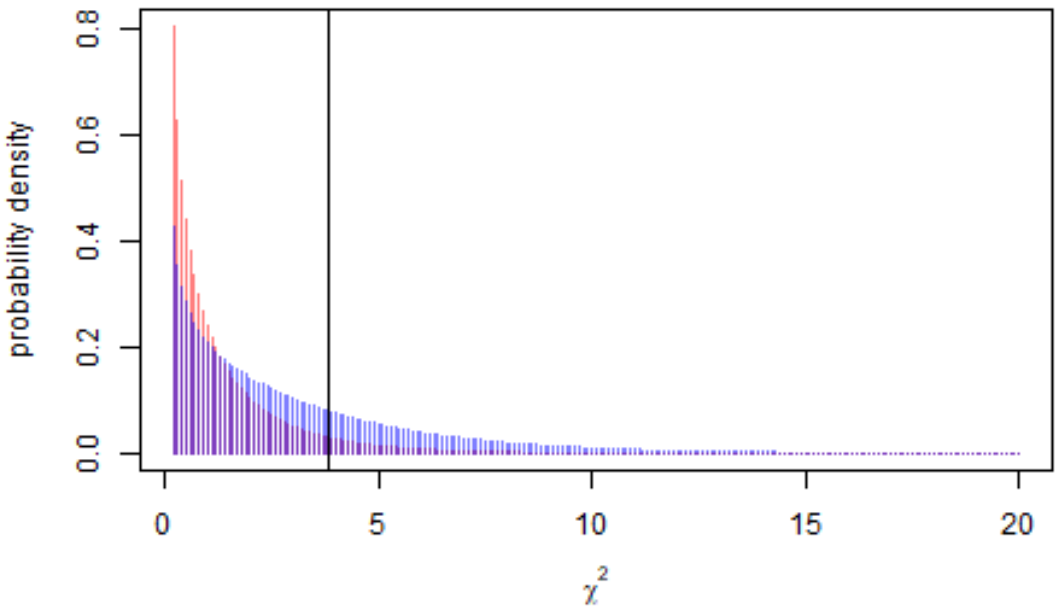
Power to detect HW departure [optional]

7. [optional] How likely are we to detect departure from HW if $f = 0.125$ with a sample of 100 individuals ?

```r
ni<-100
f<-0.125
pchisq(qchisq(0.95,df=1),df=1,ncp=ni*f^2,lower=FALSE)
```

```
## [1] 0.239527
```

```r
#density of chisq with ncp nf2

x<-seq(0.2,20,0.1)
plot(x,dchisq(x,df=1),type="h",col="#FF000080",
     xlab=expression(chi^2),ylab="probability density") #chisq prob dens
lines(x,dchisq(x,df=1,ncp=ni*f^2),type="h",col="#0000FF80")
abline(v=qchisq(0.95,df=1)) # 95th centile of chisq dist
```



— How many individuals would be needed to detect departures from HW when $f = 0.125$ with a power of 0.8 ?

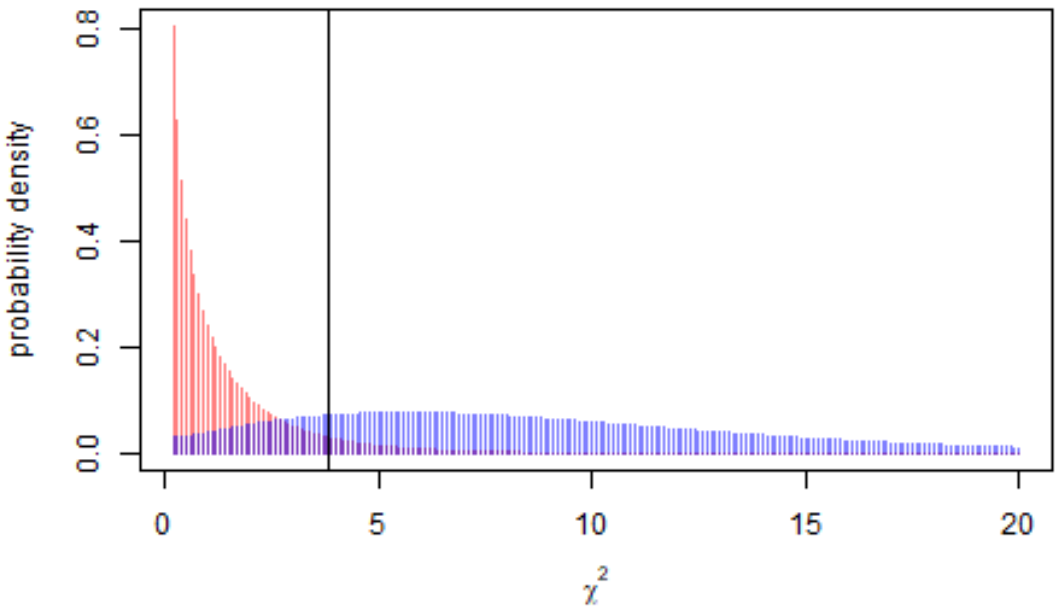
```
ns<-1:10*100
round(pchisq(qchisq(0.95,df=1),df=1,ncp=ns*f^2,lower=FALSE),digits=3)
```

```
##  [1] 0.240 0.424 0.581 0.705 0.798 0.865 0.911 0.942 0.963 0.977
```

```
ni<-500
f<-0.125
pchisq(qchisq(0.95,df=1),df=1,ncp=ni*f^2,lower=FALSE)
```

```
## [1] 0.7981762
```

```
plot(x,dchisq(x,df=1),type="h",col="#FF000080",
     xlab=expression(chi^2),ylab="probability density")
lines(x,dchisq(x,df=1,ncp=ni*f^2),type="h",col="#0000FF80")
#density of chisq with ncp nf2
abline(v=qchisq(0.95,df=1)) # 95th centile of chisq dist
```




```
Power to detect departure from HW with a single SNP,
even with fairly large f, is low,
and we need around 500 individuals to have a 80% chance
of finding significant departure with f=0.125 for a single SNP
```