Permutation methods

Conditional inference: Fisher's exact test

From Efron and Hastie (2016) Section 4.3, page 46

The following Table

```
tab = matrix(c(9,7,12,17), ncol=2)
colnames(tab) = c("success","failure")
rownames(tab) = c("new","old")
addmargins(tab)
```

```
##     success failure Sum
## new       9      12  21
## old       7      17  24
## Sum      16      29  45
```

shows the results of a randomized trial on 45 ulcer patients, comparing new and old surgical treatments.

Was the new surgery significantly better?

Fisher argued for carrying out the hypothesis test conditional on the marginals of the table $(16, 29, 21, 24)$. With the marginals fixed, the number y in the upper left cell determines the other three cells by subtraction.

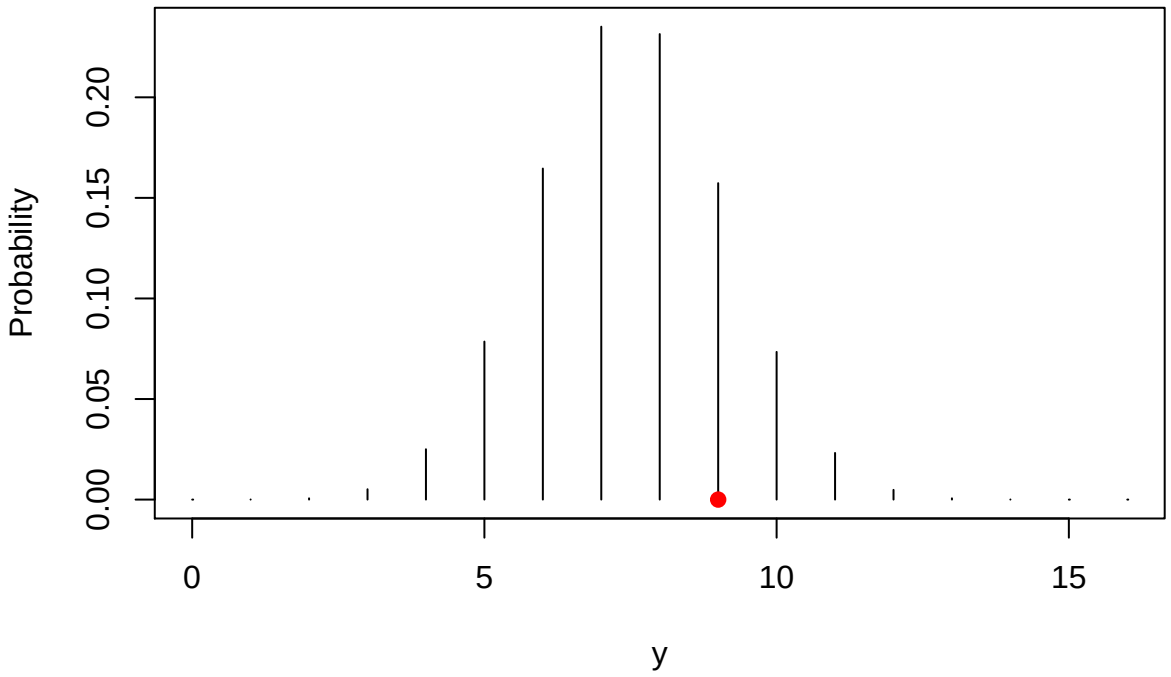
We need only test whether the number $y = 9$ is too big under the null hypothesis of no treatment difference, instead of trying to test the numbers in all four cells.

	success	failure	
new	Y	$N_{new} - Y$	N_{new}
old	$S - Y$	$N_{old} - S + Y$	N_{old}
	S	$N_{new} + N_{old} - S$	$N_{new} + N_{old}$

Under the null hypothesis, the random variable $Y|(S, N_{new}, N_{old})$ follows the Hypergeometric Distribution with

$$\Pr(Y = y|S, N_{new}, N_{old}) = \frac{\binom{N_{old}}{S-Y}\binom{N_{new}}{Y}}{\binom{N_{new}+N_{old}}{S}}$$

```
plot(0:16,dhyper(0:16, 21, 24, 16), type="h", xlab="y", ylab="Probability")
points(9,0,col=2,pch=19)
```



In the surgery example, the difference is not significant: the p-value of the Fisher's exact test is

```
fisher.test(tab, alternative="greater")$p.value
```

```
## [1] 0.2594465
```

Permutation and randomization

From Efron and Hastie (2016) Section 4.4, pages 49-51

Consider the comparison between the 47 ALL and 25 AML patients in the gene 136 leukemia example.

The two-sample t-statistic had value $t = 3.13$, with two-sided significance level 0.0025 according to a Student-t null distribution with 70 degrees of freedom. All of this depended on the Gaussian, or normal, assumptions.

As an alternative significance-level calculation, Fisher suggested using permutations of the 72 data points. The 72 values are randomly divided into disjoint sets of size 47 and 25, and the two-sample t-statistic is recomputed. This is done some large number B times, yielding permutation t-values $t^{*1}, t^{*2}, \ldots, t^{*B}$, where $t^{*1} = t$ is the identity permutation.

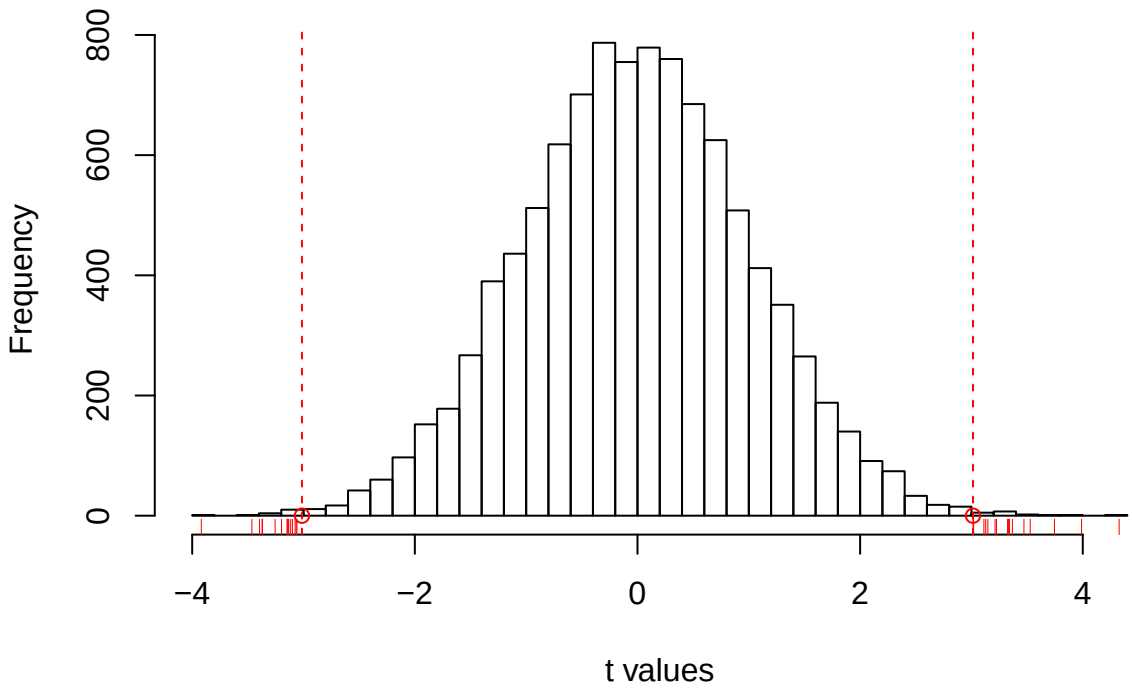
The two-sided permutation significance level for the original value t is then the proportion of the in absolute value

$$\frac{\sum_{b=1}^{B} I\{|t^{*b}| \geq |t|\}}{B}$$

We reproduce Figure 4.3 of Efron and Hastie (2016). Figure 4.3 shows the histogram of $B = 10000$ t^{*b} values for the gene 136 of the Leukemia data: 33 of these exceeded t in absolute value, yielding significance level 0.0033 against the null hypothesis of no ALL/AML difference, remarkably close to the normal-theory significance level 0.0025.

```
leukemia <- read.csv("http://web.stanford.edu/~hastie/CASI_files/DATA/leukemia_big.csv")
x <- (substr(names(leukemia),1,3)=="AML")*1
Y <- leukemia[136,]
tstat = vector()
tstat[1] = t.test(Y[x==1],Y[x==0], var.equal =T)$stat
```

```
B = 10000
set.seed(123)
for (b in 2:B){
Yperm = sample(Y)
tstat[b] =  t.test(Yperm[x==1],Yperm[x==0], var.equal =T)$stat
}
hist(tstat, 50, main="", xlab="t values")
points(tstat[1], 0, col=2); abline(v=tstat[1], col=2, lty=2)
points(-tstat[1], 0, col=2); abline(v=-tstat[1], col=2, lty=2)
rug(tstat[abs(tstat)>= abs(tstat[1])], col=2)
```




```
mean( abs(tstat) >= abs(tstat[1]) )
```

```
## [1] 0.0033
```

Why should we believe the permutation significance level? Fisher provided two arguments.

1. Suppose we assume as a null hypothesis that the $n = 72$ observed measurements are an i.i.d. sample obtained from the *same* distribution $f_\mu(y)$

$$Y_i \overset{\text{i.i.d.}}{\sim} f_\mu(y), \quad i = 1, \ldots, n \tag{1}$$

   There is no normal assumption here, say that f_μ is $\mathcal{N}(\mu, \sigma^2)$. Let $(y_{(1)}, y_{(2)}, \ldots, y_{(n)})$ indicate the order statistic of $(y_1, \ldots, y_n)$, i.e., the 72 numbers ordered from smallest to largest, with their AML or ALL labels removed. Then it can be shown that all $72!/(42!25!)$ ways of obtaining $(y_1, \ldots, y_n)$ by dividing $(y_{(1)}, y_{(2)}, \ldots, y_{(n)})$ into disjoint subsets of sizes 47 and 25 are equally likely under null hypothesis (1). A small value of the permutation significance level indicates that the actual division of AML/ALL measurements was not random, but rather resulted from negation of the null hypothesis (1). This might be considered an example of Fisher's logic of inductive inference, where the conclusion "should be obvious to all." It is certainly an example of conditional inference, now with conditioning used to avoid specific assumptions about the sampling density .

2. In experimental situations, Fisher forcefully argued for *randomization*, that is for randomly assigning the experimental units to the possible treatment groups. Most famously, in a clinical trial comparing drug A with drug B, each patient should be randomly assigned to A or B. Randomization greatly strengthens the conclusions of a permutation test. In the AML/ALL gene 136 situation, where randomization wasn't feasible, we wind up almost certain that the AML group has systematically larger numbers, but cannot be certain that it is the different disease states causing the difference. Perhaps the AML patients are older, or heavier, or have more of some other characteristic affecting gene 136. Experimental randomization almost guarantees that age, weight, etc., will be well-balanced between the treatment groups. Fisher's RCT (randomized clinical trial) was and is the gold standard for statistical inference in medical trials.

Permutation testing is frequentistic: a statistician following the procedure has 5% chance of rejecting a valid null hypothesis at level 0.05, etc.

Randomization inference is somewhat different, amounting to a kind of forced frequentism, with the statistician imposing his or her preferred probability mechanism upon the data. Permutation methods are enjoying a healthy computer-age revival, in contexts far beyond Fisher's original justification for the t-test.

Westfall & Young max$T$ and min$P$

From Goeman and Solari (2014) Multiple Hypothesis Testing in Genomics. Statistics in Medicine 2014 33:1946-78, Section 4.4

Instead of making assumptions on the dependence structure of the p-values, it is also possible to adapt the procedure to the dependence that is observed in the data by replacing the unknown true null distribution with a permutation null distribution. In this way, the conservativeness associated with probability inequalities such as Boole's or Simes' can be avoided, and instead, the multiple testing procedure can adapt to the true null distribution of the p-values. The permutation-based FWER-controlling procedures of Westfall & Young are often more powerful than the methods of Holm, Hommel etc. and have even been shown to be asymptotically optimal for a broad class of correlation structures.

In the methods of Westfall & Young, permutations are used to find the α-quantile of the distribution of the minimum p-value of the m_0 true hypotheses. This same quantile is bounded by α/m by Bonferroni inequality, but this bound is often conservative. The methods of Westfall & Young use permutations to obtain a more accurate threshold, which is usually larger than α/m.

Two variants of their method were designed by Westfall & Young: the max$T$ and min$P$ methods. The max$T$ method uses the raw p-values directly as input for the method, whereas the min$P$ method transforms these p-values to (raw) permutation p-values first.

Practically, the max$T$ method of Westfall & Young starts by making k permuted data sets and calculating all m raw p-values for each permuted data set. Let us say we store the results in an $m \times k$ matrix $\mathbf{P}$. We find the k minimal p-values along each column to obtain the permutation distribution of the minimum p-value out of m. The α-quantile $\tilde{\alpha}_0$ of this distribution is the permutation-based critical value, and Westfall & Young reject all hypotheses for which the p-value in the original data set is strictly smaller than $\tilde{\alpha}_0$.

Next, we may continue from this result in the same step-down way in which Holm's method continues on Bonferroni's. In the next step, we may remove from the matrix $\mathbf{P}$ all rows corresponding to the hypotheses rejected in the first step and recalculate the k minimal p-values and their α-quantile $\tilde{\alpha}_1$. Removal of some hypotheses may have increased the values of some of the minimal p-values, so that possibly $\tilde{\alpha}_1 > \tilde{\alpha}_0$.

We may now reject any additional hypotheses that have p-values below the new quantile $\tilde{\alpha}_1$. The process of removing rows of rejected hypotheses from $\mathbf{P}$ and recalculating the α-quantile of the minimal p-values is repeated until any step fails to result in additional rejections, or until all hypotheses have been rejected, just like in Holm's procedure.

Westfall & Young's min$P$ method is similar to their maxT method, except that instead of the matrix $\mathbf{P}$ of raw p-values, it uses a matrix $\tilde{\mathbf{P}}$ with per-hypothesis permutation p-values.

The i, jth element of $\tilde{\mathbf{P}}$ is given by

$$\tilde{p}_{ij} = \frac{\#\{l : p_{il} \leq p_{ij}\}}{k}$$

where p_{ij} is the i, jth element of $\mathbf{P}$. This matrix $\tilde{\mathbf{P}}$ is subsequently used in the same way as $\mathbf{P}$ in the max$T$ procedure. Because permutation p-values take only a limited number of values, the matrix $\tilde{\mathbf{P}}$ will always contain many tied values, which is an important practical impediment for the min$P$ method, as we will see in the succeeding text.

The number of permutations is always an issue with permutation-based multiple testing. In data with a small sample size, this number is necessarily restricted, but this ceases to be an issue already for moderate data sets. Although it would be best always to use the collection of all possible permutations, this is often computationally not feasible, so a collection of randomly generated permutations is often used. Additional randomness is introduced in this way, which makes rejections and adjusted p-values random, especially with only few random permutations. The minimum number of permutations required depends on the method, on the the α-level and on the presence of randomness in the permutations.

- The max$T$ method requires fewest permutations and can work well with only $1/\alpha$ permutations, whatever the value of m, if p-values are continuous and all permutations can be enumerated. With random permutations, a few more permutations are recommended to suppress randomness in the results, but a number of 1000 permutations is usually quite sufficient at $\alpha = 0.05$, whatever m.
- The min$P$ method requires many more permutations. Because of the discreteness of the permutation p-values, the minimum observed p-value will be equal to the minimum possible p-value for most of the permuted data sets unless the number of permutations is very large, resulting in zero power for the method. For the min$P$ procedure, therefore, we recommend to use m/α permutations as an absolute minimum, but preferably many more. Such numbers of permutations are computationally prohibitive for typical values of m.

A gain in power for Westfall & Young's maxT method relative to Holm or Hommel can be expected especially if strong positive correlations between p-values exist. The permutation method will adapt to the correlation structure found in the data and does not have to take any worst case into account. A gain in power may also occur if the raw p-values are conservative. Permutation testing does not use the fact that p-values of true hypotheses are uniformly distributed but adapts to the actual p-value distribution just as it adapts to the true correlation structure. Use of Westfall & Young therefore does not require blind faith in the accuracy of the asymptotics underlying the raw p-values. Where methods that are not permutation-based become conservative or anti-conservative with the underlying raw p-values, Westfall & Young can even work with invalid and possibly anti-conservative p-values calculated from an incorrect model and produce correct FWER control on the basis of such p-values. Although this sounds wonderful, it is still sensible to be careful with this, because p-values from invalid models tend to be especially wild for probes for which the model fits badly, rather than for probes with an interesting effect. For this reason, the power of a Westfall & Young max$T$ procedure based on such p-values can be disappointing. The min$P$ variant of the Westfall & Young procedure partially mends this by working on the per-probe permutation p-values instead of the raw p-values, guaranteeing a uniform distribution of the input p-values for the method.

Software for the Westfall & Young procedures is somewhat limited. For some hypothesis tests, the max$T$ and min$P$ procedures are implemented in the `mt.maxT` and `mt.minP` procedures in the `multtest` package in R.

Golub data

We illustrate the implementation of theWestfall & Young max$T$ procedure on a microarray data set from Golub et al. We first load the data that are in the `multtest` R package. The golub data set is a 38×3051 matrix. Each column reports the expression level for $m = 3051$ genes. The first 27 rows correspond to patients with leukemia of type ALL, the last 11 rows correspond to patients with leukemia of type AML.

Our goal is to find genes that have a differential expression between these two conditions.

```
library(multtest)
data(golub)

Y <- t(golub)
colnames(Y) <- golub.gnames[,3]
X <- golub.cl
m <- ncol(Y)
n <- nrow(Y)

alpha <- 0.05
```

Therefore, for each gene we perform a two-sample t-test and we store the permutation p-values in the $(k \times m)$ matrix $\mathbf{P}$ (1st row: identity permutation):

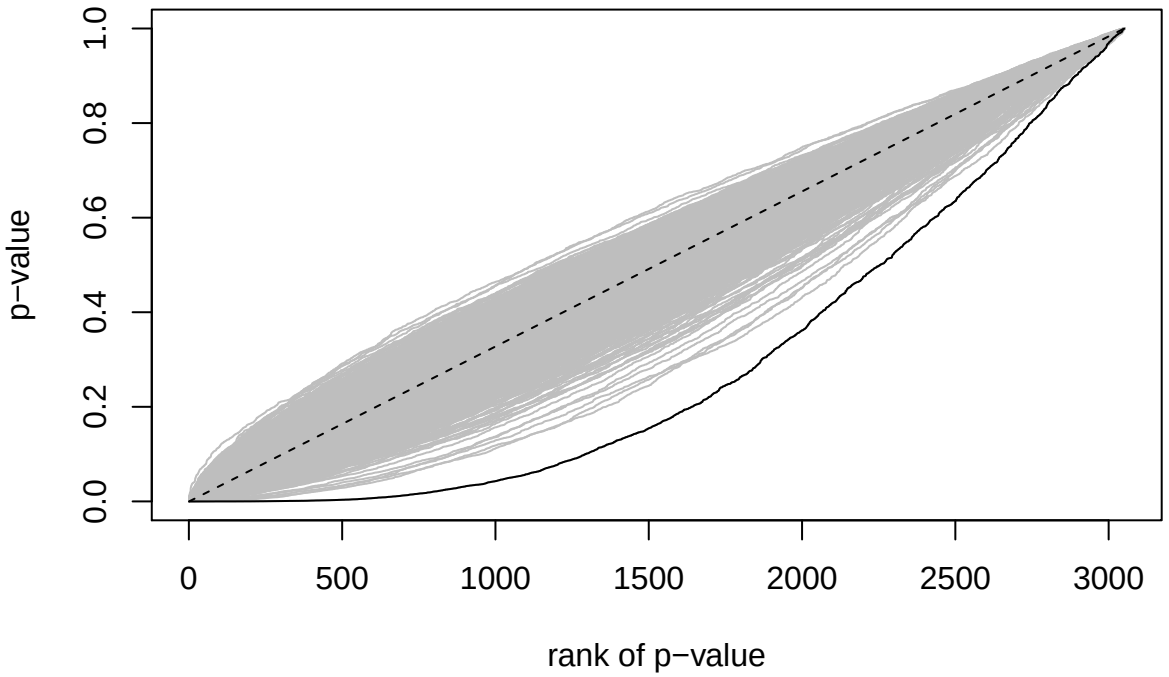
```
k <- 1000
set.seed(123)

S <- sapply(1:m, function(i)
  mt.sample.teststat( Y[,i], X, test="t.equalvar",B=k) )
P <- 2*(1-pt(abs(S),df=n-2))

p.raw <- P[1,]
```

Permutation null distribution: p-value curves for the observed and permuted data sets

```
matplot(apply(P,1,sort), type="l", lty=1, col="gray", xlab="rank of p-value", ylab="p-value")
lines(sort(p.raw))
lines(c(1,m), c(0,1), lty="dashed")
```



Westfall and Young max$T$ method: single-step critical value

```
minP <- apply(P,1,min)
tildealpha0 <- quantile(minP,alpha)
tildealpha0
```

```
##           5%
## 2.014186e-05
```

```
# number of rejections
sum(p.raw < tildealpha0)
```

```
## [1] 104
```

```
# number of rejections: Bonferroni
sum(p.adjust(p.raw,"hommel") <= alpha)
```

```
## [1] 98
```

Westfall and Young max$T$ method: step-down critical value

```
reject <- rep(F,length(p.raw))
ready <- FALSE
while (!ready) {
  minP <- apply(P[,!reject,drop=FALSE], 1, min)
  tildealpha <- quantile(minP, alpha)
  newreject <- (p.raw <  tildealpha) & (!reject)
  reject <- reject | newreject
  ready <- !any(newreject)
}
tildealpha
```

```
##           5%
## 2.299729e-05
```

```
# number of rejections
sum(p.raw < tildealpha)
```

```
## [1] 107
```

```
# number of rejections: Holm
p.holm <- p.adjust(p.raw,"holm")
```

Westfall and Young max$T$ method by using `multtest` R package

```
set.seed(123)
res.maxT <- mt.maxT(t(Y),X,test="t.equalvar",side="abs",B=k)
```

```
## b=10 b=20    b=30    b=40    b=50    b=60    b=70    b=80    b=90    b=100
## b=110    b=120   b=130   b=140   b=150   b=160   b=170   b=180   b=190   b=200
## b=210    b=220   b=230   b=240   b=250   b=260   b=270   b=280   b=290   b=300
## b=310    b=320   b=330   b=340   b=350   b=360   b=370   b=380   b=390   b=400
## b=410    b=420   b=430   b=440   b=450   b=460   b=470   b=480   b=490   b=500
## b=510    b=520   b=530   b=540   b=550   b=560   b=570   b=580   b=590   b=600
## b=610    b=620   b=630   b=640   b=650   b=660   b=670   b=680   b=690   b=700
## b=710    b=720   b=730   b=740   b=750   b=760   b=770   b=780   b=790   b=800
## b=810    b=820   b=830   b=840   b=850   b=860   b=870   b=880   b=890   b=900
## b=910    b=920   b=930   b=940   b=950   b=960   b=970   b=980   b=990   b=1000
```

```
head(res.maxT)
```

```
##                 index  teststat  rawp  adjp
## M27891_at         829 10.255974 0.001 0.001
## D88422_at         378  8.448676 0.001 0.001
## X95735_at        2124  8.166010 0.001 0.001
## M23197_at         808  7.981284 0.001 0.001
```

```
## U22376_cds2_s_at  2489 -7.855191 0.001 0.001
## HG1612-HT1612_at   394 -7.836707 0.001 0.001
```

```
p.maxt <- res.maxT$adjp[order(res.maxT$index)]

allp <- cbind(p.raw, p.holm, p.maxt)
mt.plot(allp, plottype="rvsa", proc=c("rawp","Holm","maxT"), leg=c(alpha,m), lty=1:ncol(allp), col=1:nc
```