Chapter 7

Randomization

1 Fisher on randomization

Fisher (1966, Chapter III) discussed in detail the analysis of an experiment on plant growth made by Charles Darwin. After mentioning a number of possible systematic effects due to the placement of plants within pots, Fisher commented (page 44):

> Randomisation properly carried out, in which each pair of plants are assigned their positions independently at random, ensures that the estimates of error will take proper care of all such causes of different growth rates, and relieves the experimenter from the anxiety of considering and estimating the innumerable causes by which his data may be disturbed. The one flaw in Darwin's procedure was the absence of randomisation.

Later in the same chapter (page 45) he asserted that

> . . . the physical act of randomisation, which, as has been shown, is necessary for the validity of any test of significance, affords the means, in respect of any particular body of data, of examining

> the wider hypothesis in which no normality of distribution is implied.

He then proceeded to describe what is now often called a "Fisher randomization test", which uses only the probabilities supplied by the randomization. After some "tedious" calculations he then commented that the tail probability he had obtained was "very nearly equivalent to that obtained using the t test with the hypothesis of a normally distributed population".

In other words, the randomization is needed to turn some systematic effects into random noise and apparently the probabilistic analysis based only on the randomization leads to conclusions similar to those obtained under a model with normal errors.

Fisher's view has not been accepted by all statisticians. Debabrata Basu was a notable strong critic. For his insightful critique read Chapters XIV and XV of Ghosh (1988), a collection of some of his papers and conference talks.

2 Shoes: a paired comparison

Shoes data from Box et al. (1978, Section 4.2):

> ... measurements of the amount of wear of the soles of shoes worn by 10 boys. The shoe soles were made of two different synthetic materials, A and B. ... the experiments were run in pairs. Each boy wore a special pair of shoes, the sole of one shoe having been made with A and the sole of the other with B. The decision as to whether the left or the right sole was made with A or B was determined by the flip of a coin.
>
> ... the variability among the boys has been eliminated. ... by working with the 10 differences $B - A$ most of this boy-to-boy variation could be eliminated. An experimental design of this kind is called a *randomized paired comparison* design

I have coded the coin tosses as ±1, with +1 meaning "apply treatment A to the left sole" and −1 meaning "apply treatment A to the right sole". With that convention, the difference between the B and A treatments for boy$_i$ equals $y_i = x_i * d_i$, where d_i denotes the difference "wear for right foot minus wear for left foot".

```
shoes <- read.table("shoes.data",sep="\t",header=T)
shoes$coin <- 2 * (shoes$foot == "L") - 1 # +1 for left
shoes$BA.diff <- shoes$B - shoes$A
shoes$RL.diff <- shoes$BA.diff * shoes$coin
shoes

##       A    B foot.A coin BA.diff RL.diff
## 1  13.2 14.0      L    1     0.8     0.8
## 2   8.2  8.8      L    1     0.6     0.6
## 3  10.9 11.2      R   -1     0.3    -0.3
## 4  14.3 14.2      L    1    -0.1    -0.1
## 5  10.7 11.8      R   -1     1.1    -1.1
## 6   6.6  6.4      L    1    -0.2    -0.2
## 7   9.5  9.8      L    1     0.3     0.3
## 8  10.8 11.3      L    1     0.5     0.5
## 9   8.8  9.3      R   -1     0.5    -0.5
## 10 13.3 13.6      L    1     0.3     0.3
```

Presumably the experimenter was seeking to learn which treatment was better (= less wear). BHH later revealed that material B was cheaper. So perhaps the experimenter was looking for evidence than material A did produce significantly less wear. That hypothesis suggests the use of a one-sided t-test: Is the difference `shoes$BA.diff` significantly large (positive)?

Let me start with the analysis based on the assumption that the differences y_i are independent $N(\delta, \sigma^2)$ random variables. The estimate of σ^2 is $\widehat{\sigma}^2 = \sum_i (y_i - \overline{y})^2/9$ and the t-statistic is $tstat = \sqrt{10}\,\overline{y}/\widehat{\sigma}$. As I am feeling lazy I'll let **R** do all the work:

```
summary(lm(shoes$BA.diff ~ 1))

##
## Call:
## lm(formula = shoes$BA.diff ~ 1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -0.610 -0.110 -0.010  0.165  0.690
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.4100     0.1224   3.349  0.00854 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3872 on 9 degrees of freedom
```

The observed value of the t-statistics has a one-sided p-value of about 0.004. Material A really does seem to be doing better.

3 The randomization distribution

The randomization is represented by a vector x in $\{-1,+1\}^{10}$, which, by construction, takes each of the 1024 possible values with probability 2^{-10}.

Under the null hypothesis that the two treatments actually have the same (random?) effect on the amount of wear, the random variable x_i should have no influence on $d_i = \texttt{shoes\$RL.diff}_i$. That is, x and d should be independent. Under the null hypothesis, the observed values of y_i is just $\pm d_i$, where d_i is the random value generated by the pecularities of each boy's behavior and the sign is attached independently of d_i.

```r
# Create a 10 by 2^{10} matrix representing all possible
# outcomes for tosses of 10 coins
rand <- coin(10)        # defined in cointoss.R
yrand <- rand * shoes$RL.diff # 1024 values for different x's
# My function tstat() calculates the t -statistic
# for each possible y:
Tstats <- apply(yrand,2,tstat)
```

According to Fisher, the spread in the 1024 t-statistics (one for each possible realization of x) should look like the spread in 1024 observations from the t_9 distribution.

A plot of the sorted values of `Tstats` against the corresponding quantiles of the t_9 distribution,

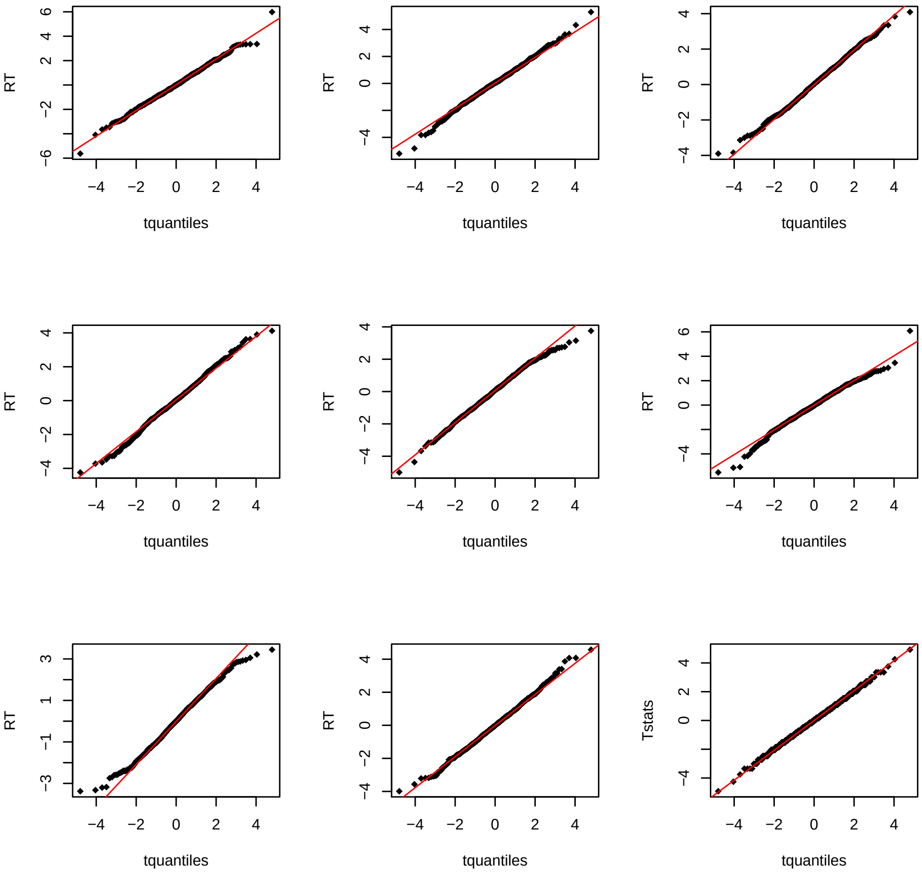
```r
# quantiles at 1024 p's of t_9:
tquantiles <- qt(ppoints(1024),df=9)
```

would be almost a straight line if the values in `Tstats` were spread out like a t_9 distribution.

For the sake of comparison, I include some quantile plots for samples taken from the t_9 distribution (at least according to what **R** regards as a random sample):

```
old.par <- par(no.readonly = T)
par(mfrow = c(3,3))
t9 <- function(p){  qt(p, df = 9) }

for (ii in 1:8){
        RT <- rt(1024,df=9) # sample of size 1024 from t_9
        qqplot(tquantiles,RT,pch=18,new=T)
        qqline(RT,distribution = t9,col="red")
}
qqplot(tquantiles,Tstats,pch=18,new=T)
qqline(Tstats,distribution = t9,col="red")
```



```
par(old.par)
```

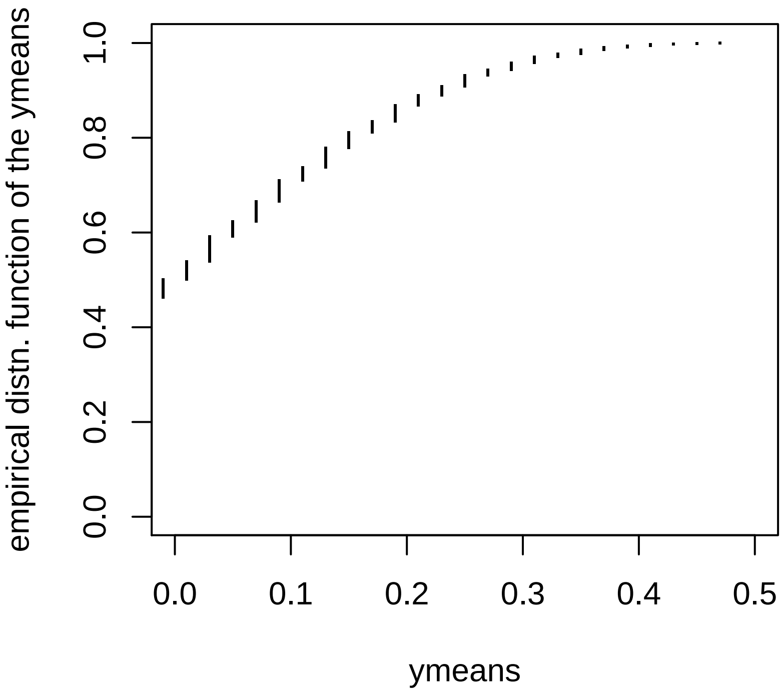
Are you impressed?

The randomization test looks at where the observed value of $\overline{y}$ (namely, 0.41) sits amongst the set of all possible $\overline{y}$ generated by different realizations of x:

```
ymeans <- sort(apply(yrand,2,mean))
print(ymeans[1015:1024])

##  [1] 0.39 0.39 0.39 0.41 0.41 0.41 0.41 0.43 0.45 0.47

plot(ymeans, (1:1024)/1024,pch=".",xlim=c(0,0.5),
        ylab="empirical distn. function of the ymeans")
```



The interpretation is complicated slightly by the ties in the `ymeans` values. BHH page 100 decided to count half the ties at 0.41 as being greater than the observed value, which led to a p-value $5/1024 \approx 0.005$, impressively close to the p-value calculated via the normal approximation.

4 Theoretical justification of t-approximation

Several authors have tried to justify Fisher's assertion about the approximate t_{n-1}-distribution under the null hypothesis for

$$T = \frac{\sqrt{n}\overline{y}}{\sqrt{\sum_{i\le n}(y_i - \overline{y})^2/(n-1)}}$$

when the y_i's are generated by a randomization: $y_i = x_i d_i$. I have found the paper by Box and Andersen (1955) the most helpful.

The idea is that the random variable

$$B = \frac{T^2}{n-1+T^2} = \frac{n\overline{y}^2}{\sum_{i\le n} y_i^2}$$

would have a beta$(1/2, (n-1)/2)$ distribution if the y_i's were independent $N(0,\sigma^2)$'s. That beta distribution has expected value $1/n$ and variance $3/(n^2+2n)$.

If $y_i = x_i d_i$, with the d_i's being treated as constants, then

$$B = n^{-1}\left(\sum\nolimits_i x_i f_i\right)^2 \qquad \text{where } f_i = d_i/\sqrt{\sum\nolimits_{j\le n} d_j^2},$$

which (under the randomization distribution) has expected value $1/n$ and a variance that is close to $3/(n^2+2n)$ if the f_i's are "not too extreme".

References

Box, G. E. P. and S. L. Andersen (1955). Permutation theory in the derivation of robust criteria and the study of departures from assumption. *Journal of the Royal Statistical Society, Series B 17*, 1–34.

Box, G. E. P., W. G. Hunter, and J. S. Hunter (1978). *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building.* New York: Wiley.

Fisher, R. A. (1966). *The Design of Experiments* (8th ed.). Haffner. First edition 1935. Republished in 1991 by Oxford as part of a collection of three of Fisher's books, under the title "Statistical Methods, Experimental Design, and Scientific Inference".

Ghosh, J. K. (Ed.) (1988). *Statistical Information and Likelihood: A Collection of Critical Essays by Dr. D. Basu*, Volume 45 of *Lecture Notes in Statistics*. Springer-Verlag.