17 January 2011

# DISCOVERY: SOME STATISTICAL ASPECTS

D.R.Cox

*Nuffield College, Oxford, UK*

## 1. Formulation

- reference space
- nature of background noise
- nature of signal, point or distributed, accumulated counts or quantitative signals
- identification of uninteresting signals
- frequency of occurrence of signal
  - none or one
  - none or several, probably several, but not many
- single or multi-phase investigation

## A crucial distinction

Given many tests:

- Either there is a signal of interest or there is nothing but noise. Interest in probability of a false discovery.
- There are likely to be a modest number of signals: have we chosen the right ones? Interest in False discovery rate, i.e. proportion of those selected as signals that are in fact false (Benjamini and Hochberg, 1995)

## 2. Testing for presence of signal; simplest case

Tests at $n$ positions. Statistical independence. If there is no signal at position $j$ test statistic has a known distribution leading to a $p$-value, a tail area in a one-sided significance test. Gives a set $\{P_1, \ldots, P_n\}$. No deeper formulation!

Three formulations essentially equivalent:

- Schweder and Spjotvoll (Biometrika, 1982) studied the lower tail of distribution of $P$.
- transform to $Z = -\log P$
- transform to standard normal $T = \Phi^{-1}(1 - P)$

If no signal $P_j$ uniformly distributed; $Z_j$ unit exponentially distributed.

Define the order statistics

$$Z_{(1)} \leq Z_{(2)} \leq \ldots \leq Z_{(n)} = \max Z_j.$$

Plot ordered $Z$ against expected values. Under null hypothesis straight line of slope one. With signal one outlying point. With small number of signals small number of outlying points.

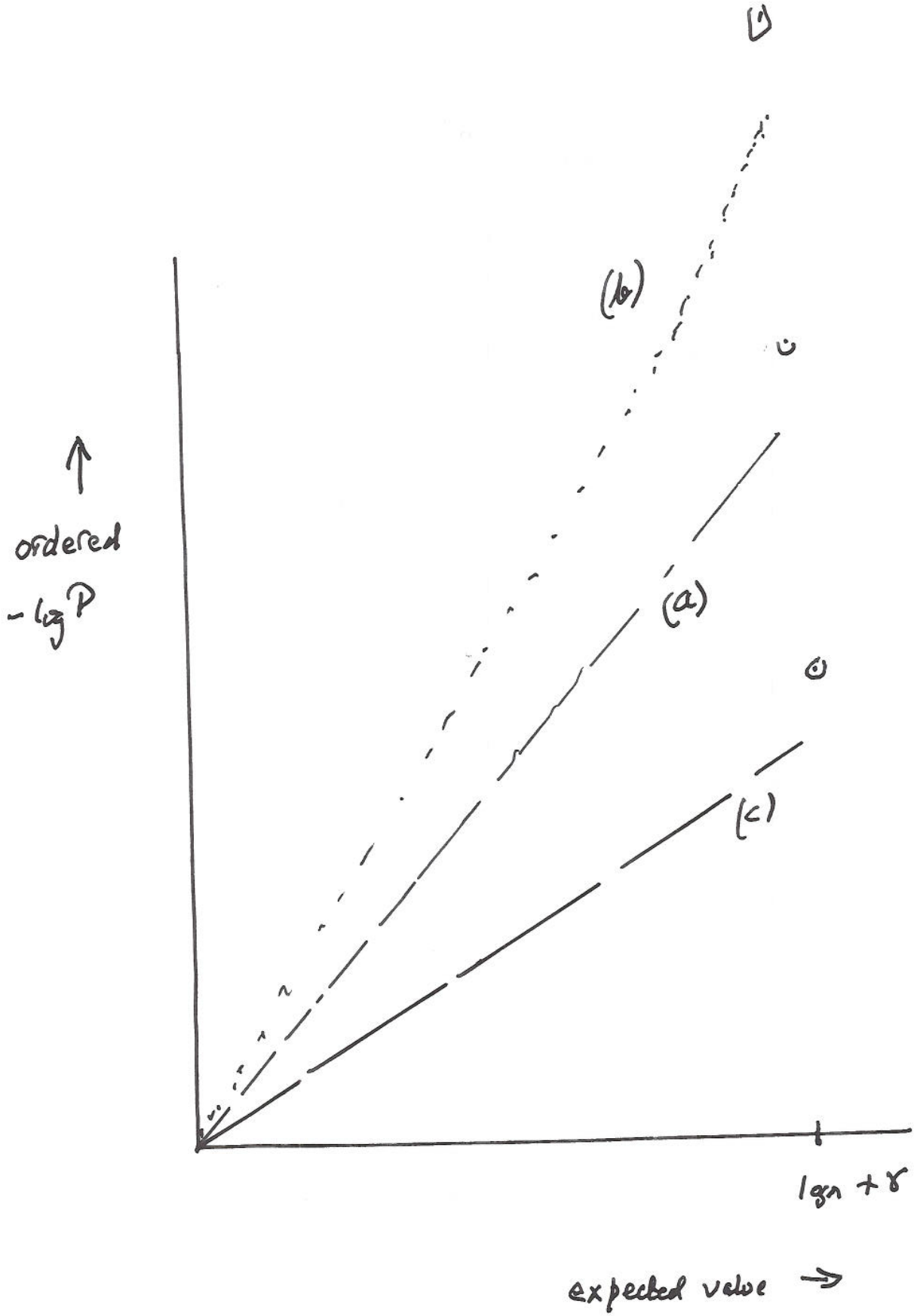
## 3. Renyi decomposition

Let $V_1, \ldots, V_n$ be independent and identically distributed in exponential distributions of mean one. Then

$$
\begin{aligned}
Z_{(1)} &= V_1/n, \\
Z_{(2)} &= V_1/n + V_2/(n-1), \\
&\cdot \\
&\cdot \\
&\cdot \\
Z_{(n-1)} &= V_1/n + V_2/(n-1) + \ldots + V_{n-1}/2, \\
Z_{(n)} &= V_1/n + V_2/(n-1) + \ldots + V_{n-1}/2 + V_n.
\end{aligned}
$$

This provides the basis for plotting the ordered $Z$ and for various tests.

Roughly log scale.




(a) standard case
(b) non-standard null distribution
(c) internal correlation

## 4. Simple test

The simplest procedure is to use the most significant value and this involves a direct selection allowance. The significance level attached to $\max(Z_j) = Z_{(n)} = z^*$ is

$$1 - (1 - e^{-z^*})^n = 1 - \exp(-ne^{-z^*})$$

approximately. The maximum has limiting Gumbel distribution, an extreme value distribution.

## 5. Some developments

Perhaps (Efron, 2010) the null distribution is wrong. Then $Z$ plot should be a smooth curve under null hypothesis. More secure to compare with trend from previous $m$ ordered values. Yields

$$\frac{Z_{(n)} - Z_{(n-1)}}{(2Z_{(n-1)} + Z_{(n-2)} + \ldots + Z_{(n-m+1)} - (m+1)Z_{(n-m)})/2}.$$

Omission of more higher values.

Under overall null hypothesis has standard variance-ratio, $F$, distribution with $(2, 2m)$ degrees of freedom. Alternatively discard one or two order statistics near the largest.

## Behaviour for wrong null distribution

Modification of Renyi decomposition

$$
\begin{aligned}
G(Z_{(1)}) &= V_1/n, \\
G(Z_{(2)}) &= V_1/n + V_2/(n-1), \\
&\vdots \\
G(Z_{(n-1)}) &= V_1/n + V_2/(n-1) + \ldots + V_{n-1}/2, \\
G(Z_{(n)}) &= V_1/n + V_2/(n-1) + \ldots + V_{n-1}/2 + V_n.
\end{aligned}
$$

Implication

## 6. Distributed signal

Suppose that the signal distributed across a small region. Simplest approach may be following in a one-sided version:

Transform individual tests to standard normality, $T_j = \Phi^{-1}(1 - P_j)$.

Write

$$T'_j = (T_j + aT_{j-1} + aT_{j-2} + \ldots)\sqrt{(1 - a^2)}.$$

The $Z$ plot from the $T'$ will be approximately a straight line of nonunit slope. That is, the previous discussion applies with an effective sample size $n'$, say, a function of $a$.

**Continuous version**

Take smoothing functions $h_S(u)$ and $h_B(u)$ for signal and background writing

$$Y_S(t) = \int h_S(u) dN(t-u), B(t) = \int h_B(u) dN(t-u),$$

and then

$$T_S(t) = \frac{Y_S(t) - B(t)}{\sqrt{\mathrm{var}(Y_S(t))}}$$

Finally

$$T_S^* = \mathrm{max}_t T_S(t).$$

Gumbel type approximation to the distribution of $-\log \Phi(-T_S^*)$.

## Recovery of information

These procedures do not use information contained in the position in the energy spectrum of nearby peaks.

This is independent of the probability plot if the underlying model is correct.

## Points of unequal precision

If the individual points are of very unequal precision some shrinkage for example by empirical Bayes may be helpful first.

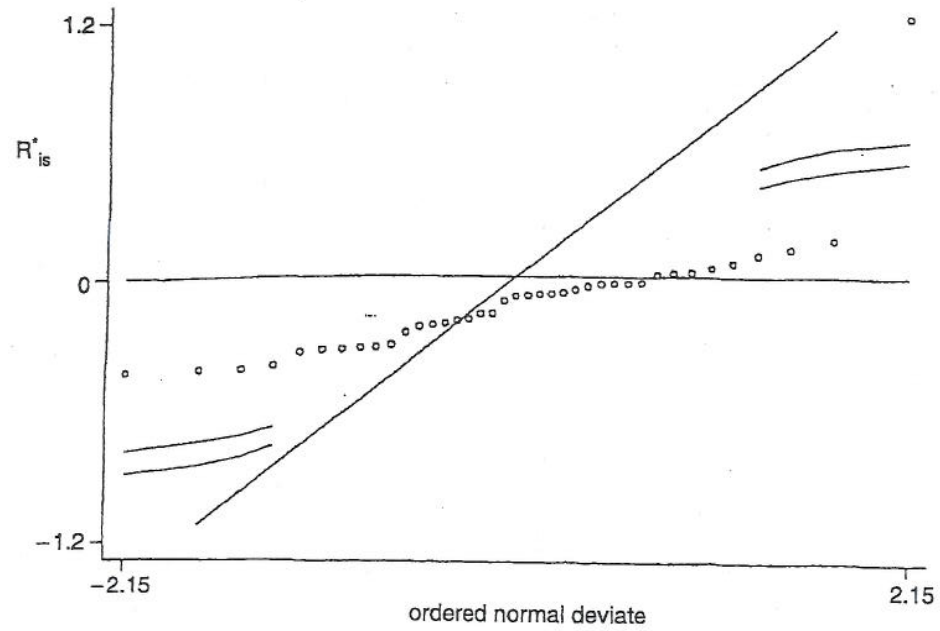



Fig. 4. Ordered normal plot of the log of the observed–expected ratio for 'Plumbers, fitters and heating engineers not elsewhere classified' in males ($R^*_{is}$), with guide-rails (set at $\alpha = 0.05$ and $\alpha = 0.01$).

The importance of examining all the disease and exposure plots side by side is further illustrated in Figure 4, which shows the occupational plot for 'Plumbers, fitters and heating engineers not elsewhere classified'. The anomalous point at the top of the plot is for pleural cancer. The observed to expected ratio for this job group was second highest on the pleural cancer line, but did not exceed the guide-lines. The anomalous estimate on the occupational plot (Figure 4), taken together with the magnitude of $\tilde{\sigma}_i$ for pleural cancer (Figure 2), is clearly indicative of a strong occupational effect.

## 7. Summary

Plot of the ordered $Z$ is helpful descriptively and is the basis for various formal tests. In particular

- null situation is a straight line of unit slope
- simplest alternative is one outlying point
- incorrect null distribution leads to a smooth curve
- internal correlation yields a straight line of slope different from one

When several nonnull hypotheses are present leads directly to FDR discussions: let $S$ denote the number of selected values of which $F$ are in fact null. Then the conditional false discovery rate is

$$E(F/S \mid S > 0).$$

## 8.Bayesian formulation

For Bayesian formulation need a prior probability for the existence of a signal at a particular point and a prior distribution for the magnitude of the signal under the alternative. Efron (2010) has given an empirical Bayesian analysis for situations in which appreciable numbers of null hypotheses are false and also covering the possibility that the null distribution is not the theoretically specified one.

Simpler and cruder approach effective for FDR problem (Cox and Wong, 2008) is as follows.

Under null hypothesis $T$ is standard normal. Under alternative with probability $\omega/n$ normal mean $\mu$ and variance one. With sufficient data, estimate parameters empirically.

With $n$ sites, each has prior probability $\omega/n$ of being nonnull. Thus the number of nonnull hypotheses has a Poisson distribution of mean $\omega$, so that $\omega = \log 2$ would give prior probability of no signal of $1/2$ and caution might require a much smaller value.

For a maximum Gaussian test statistic of $t^*$ out of $n$ tested the log odds for a real signal are

$$\log P(\text{realsignal})/P(\text{falsealarm}) = \log(\omega/n) + \mu(t^* - \mu/2).$$

Thus with $t^* = 5$ the log odds is maximal at $\mu = 5$ and with $\omega < 1$ takes values at most

At $\mu = 2$ log probability $8 + \log(\omega/n)$,

At $\mu = 5$ log probability $12.5 + \log(\omega/n)$,

At $\mu = 8$ log probability $8 + \log(\omega/n)$.

## SUMMARY

- many formalizations, physical and statistical
- detection of isolated very rare effects
- false discovery rate
- multi-stage process

# REFERENCES


Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate. A practical and powerful approach to multiple testing. *J. R. Statist. Soc.* B **57**, 289-300.

Cox, D.R. and Wong, M.Y. (2004). A simple procedure for the selection of significant effects. *J.R.Statist. Soc.* B **66**, 395-400.

Efron, B. (2010). *Large-scale inference.* Cambridge University Press.

Law, G., Cox, D.R., Maconochie, N., Simpson, J. , Roman, E. and Carpenter, L. (2001). Large tables. *Biostatistics* **2**, 163-171.

Schweder, T. and Spjotvoll, E. (1982). Plots of $P$-values to evaluate many tests simultaneously. *Biometrika* **69**, 493-502.