

## 第八部分：評審幾篇論文

本部分是個習題。附上四篇論文，請你用 referee 的角度來讀它。你的態度是「儘量找麻煩」，因為你預備建議主編「這文章不能要」。因此你要挑出所有的錯來。

爲省空間，我們拿掉所有的圖。

### 第一篇 (此文的摘要和緒論會用作習題)

#### Testing for Activation in Data from FMRI Experiments

*Abstract:* The traditional method for processing functional magnetic resonance imaging (FMRI) data is based on a voxel-wise, general linear model. For experiments conducted using a block design, where periods of activation are interspersed with periods of rest, a haemodynamic response function (HRF) is convolved with the design function and, for each voxel, the convolution is regressed on prewhitened data. An initial analysis of the data often involves computing voxel-wise two-sample t-tests, which avoids a direct specification of the HRF. Assuming only the length of the haemodynamic delay is known, scans acquired in transition periods between activation and rest are omitted, and the two-sample t-test is used to compare mean levels during activation versus mean levels during rest. However, the validity of the two-sample t-test is based on the assumption that the data are Gaussian with equal variances. In this article, we consider the Wilcoxon rank test as well as modified versions of the classical t-test that correct for departures from these assumptions. The relative performance of the tests are assessed by applying them to simulated data and comparing their size and power; one of the modified tests (the CW test) is shown to be superior.

*Key words:* Excess kurtosis, haemodynamic response function, Shapiro-Wilk test, skewness, two-sample t-test, Welch test, Wilcoxon Rank test.

### 1. Introduction

Functional Magnetic Resonance Imaging (FMRI) is a non-invasive method that produces a time sequence of images of a subject's brain that are sensitive to changes in blood oxygenation caused by neural activation. The vast majority of analytical techniques that are applied to FMRI data assume the transfer function between neural activation and subsequent changes in blood oxygenation, the haemodynamic response function (HRF), is known fully *and* the data follow the Gaussian distribution. In this article, we consider the analysis of FMRI data collected in one of two states, called "activation" and "rest," based on two-sample tests. From knowledge of the length of the haemodynamic delay, measurements during the transition period between activation and rest can be omitted. The validity of the classical two-sample t-test is based on the assumption that the activation data and the rest data are Gaussian with equal variances. In this article, we propose

use of a modified two-sample test for FMRI data that allows for departures from this assumption. We study three competing tests. One is the Welch test (Welch, 1937), which is a modification of two-sample t-test that allows unequal covariances. A second competitor is the Cressie-Whitford (CW) test (Cressie and Whitford, 1986) that can be used with non-Gaussian data. The third competitor is the Wilcoxon rank (WR) test (Wilcoxon, 1945). In what follows, we compare the classical t-test with the Welch, CW, and WR tests for FMRI data based on a block design, where the blocks alternate between periods of activation and rest.

The next section describes the physiological background and physical processes used in FMRI and the most common methods used to process FMRI data; it also defines the four two-sample tests (including the classical two-sample t-test) that are compared in Section 4. Section 3 discusses the application of the two-sample tests for FMRI data and describes the methods used to identify and quantify departures from Gaussianity for each voxel. The size and power of the four tests are compared in Section 4 using a simulation study of FMRI data, from which recommendations are given. Section 5 contains discussion and conclusions.

## 2. FMRI Experiments

### 2.1 Some physiology

All neuronal activation is linked to an increase in oxygen consumption, causing a local increase in the blood flow. The body's response is to supply more oxygen than is required for the neuronal activity. Due to the different magnetic properties of oxygenated and de-oxygenated blood, the excess oxygenated blood that circulates during neuronal activation alters the magnetic properties of the venous blood, resulting in the so-called *blood oxygenation level dependent* (BOLD) signal. FMRI produces a sequence of brain images that is sensitive to changes in the BOLD signal.

In a classical FMRI experiment, the subject is scanned every few seconds to obtain an image of the brain; the subject is exposed to an experimental stimulus in some time periods, and is in a rest state during the remaining time periods. The stimulus can either be applied for brief periods in rapid, possibly random succession (an “event-related” experimental design, Josephs *et al.*, 1997), or for longer periods with interspersed rest periods (a “block” experimental design, Frackowiak *et al.*, 1997). In this paper, we focus on FMRI experiments conducted using a block experimental design.

Even though neuronal activation occurs immediately after exposure to the experimental stimulus, the vascular response evolves more slowly, resulting in the BOLD signal. The temporal relationship between neuronal activation and the observed BOLD signal is called the haemodynamic response. To model the haemodynamic response, it is common to convolve the experimental design with a so-called haemodynamic response function (HRF). Poisson, gamma, and Gaussian distributions are used widely as HRFs (Friston *et al.*, 1994).

The region of the brain where there is neural activation is found by regressing the

observed fMRI data on the expected BOLD signal, obtained as a convolution of the experimental design with the HRF. Of course, this depends on a well-specified HRF.

## 2.2 fMRI data

Observed fMRI data are four-dimensional, in space and time. At each time point, a three-dimensional image of the brain is acquired, called a volume. Each volume consists of voxels, and each voxel has an associated one-dimensional time series of observed signal intensities.

The most common approach to the analysis of fMRI data is to consider the voxels independently. A widely-used approach assumes a general linear model (GLM) for the voxel-wise time series (Friston *et al.*, 1995). For example, after various preprocessing steps, including prewhitening to achieve approximately independent errors, a two-sample test statistic is computed for each voxel where the two samples correspond to activation data and rest data. A voxel is declared to be significant if the test statistic exceeds some threshold. The distribution theory associated with this approach is based on the assumption of Gaussianity of the observed data and the proper specification of the HRF leading to the expected BOLD signal.

For initial data analysis, it is enough for us to know the length of the haemodynamic *delay* between neural activation and changes in the BOLD signal (Bandettini *et al.*, 1993). This knowledge is used to omit scans acquired in transition periods between possibly “activated” BOLD signals and “resting” BOLD signals. The delay between the neural activation and changes in the BOLD signal depends on many different factors; the type of stimuli, the duration of each stimulus, and the brain activation regions can all effect the length of the delay. Empirical studies have proposed methods for estimating HRFs that can adapt to different experimental designs. By using the block designs described in Section 2.1 and deleting transition data in our preliminary analysis, we have a sample of data acquired under activation and a second sample of data acquired under rest. In the next section, we describe four possible two-sample tests that might be used to test for the presence of activation at each voxel.

## 2.3 Two-sample tests

The null hypothesis of no difference between the means of two populations can be investigated with appropriate two-sample tests. In what follows, we summarize the four tests to be compared where, under activation the voxel data are  $\mathbf{F}Y_a = \{Y_i\}_{i \in \mathbf{F}A}$  and, under rest the voxel data are  $\mathbf{F}Y_r = \{Y_j\}_{j \in \mathbf{F}R}$ ; here  $\mathbf{F}A$  and  $\mathbf{F}R$  denote the activation and rest acquisition times, respectively.

### The classical two-sample t-test

The classical two-sample t-test assumes:

(A1) Observations  $\mathbf{F}Y_a$  and  $\mathbf{F}Y_r$  are uncorrelated.

(A2) The observations within each of  $\mathbf{F}Y_a$  and  $\mathbf{F}Y_r$  have identical Gaussian distributions; that is,

$$\mathbf{F}Y_f \sim \text{Gau}(\mu_f * \mathbf{F}\mathbf{1}, \sigma_f^2 * \mathbf{F}\mathbf{I}); \quad f \in \{a, r\}.$$

(A3)  $\sigma_a^2 = \sigma_r^2$ .

To test the hypothesis:

$$H_0: \mu_a \leq \mu_r \quad \text{versus} \quad H_1: \mu_a > \mu_r, \quad (2.1)$$

the classical two-sample t-test uses test statistic,

$$T \equiv \frac{\bar{Y}_a - \bar{Y}_r}{\sqrt{\left(\frac{1}{n_a} + \frac{1}{n_r}\right) \left(\frac{(n_a-1)s_a^2 + (n_r-1)s_r^2}{n_a+n_r-2}\right)}}, \quad (2.2)$$

with

$$\bar{Y}_f = \frac{1}{n_f} \sum_{i \in \mathbf{F}} Y_i \quad \text{and} \quad s_f^2 = \frac{\sum_{i \in \mathbf{F}} (Y_i - \bar{Y}_f)^2}{n_f - 1}; \quad f \in \{a, r\},$$

where  $\mathbf{F}$  is the set of activation times  $\mathbf{F}A$  (rest times  $\mathbf{F}R$ ) if  $f = a$  ( $f = r$ ), and  $n_a$  ( $n_r$ ) is the number of the observations in the sample  $\mathbf{F}Y_a$  ( $\mathbf{F}Y_r$ ).

If Assumptions (A1), (A2), and (A3) are satisfied, the classical two-sample t-test with significance level  $\alpha$  is:

$$\begin{aligned} &\text{Accept } H_0 \text{ if } T < t_d(1 - \alpha) \\ &\text{Accept } H_1 \text{ otherwise,} \end{aligned}$$

where  $t_d(1 - \alpha)$  is the  $100(1 - \alpha)$  percentile of the  $t$  distribution on  $d = n_a + n_r - 2$  degrees of freedom.

### The Welch test

The Welch test (Welch, 1937) is used to test the same hypotheses (2.1), but it assumes only (A1) and (A2); that is, it is possible that  $\sigma_a^2 \neq \sigma_r^2$ . Welch (1937) has shown that under the null hypothesis  $H_0$ , the test statistic

$$T^* \equiv \frac{\bar{Y}_a - \bar{Y}_r}{\sqrt{\frac{s_a^2}{n_a} + \frac{s_r^2}{n_r}}} \quad (2.3)$$

has approximately a  $t$  distribution with

$$e \equiv \frac{\left(\frac{\sigma_a^2}{n_a} + \frac{\sigma_r^2}{n_r}\right)}{\left(\frac{\sigma_a^4}{n_a^2(n_a-1)} + \frac{\sigma_r^4}{n_r^2(n_r-1)}\right)} \quad (2.4)$$

degrees of freedom. In practice, the population variances  $\sigma_a^2$ ,  $\sigma_r^2$  in (2.4) are estimated from data using sample variances  $s_a^2$ ,  $s_r^2$ . The Welch test with significance level  $\alpha$  is:

$$\text{Accept } H_0 \text{ if } T^* < t_e(1 - \alpha)$$

Accept  $H_1$  otherwise,

where the cut-off value  $t_e(1 - \alpha)$  is based on fractional degrees of freedom and is obtained by interpolation of the  $t_d(1 - \alpha)$  cut-off levels based on the nearest integers  $d$  to  $e$ .

### The CW test

The CW test (Cressie and Whitford, 1986) also tests hypotheses (2.1), but makes only Assumption (A1); that is, it is possible that the data are non-Gaussian with unequal variances. To account for this, we use the same statistic  $T^*$  given by (2.3) as Welch, but modify its null distribution according to the skewnesses  $\alpha_{3a}$ ,  $\alpha_{3r}$  and the excess kurtoses  $\alpha_{4a}$ ,  $\alpha_{4r}$  of the non-Gaussian activation and rest distributions, respectively.

By calculating the Cornish-Fisher expansion of  $T^*$ , Cressie and Whitford (1986) show that under Assumption (A1) and  $H_0$ , the distribution of  $T^*$  is approximately that of the random variable,

$$V = U + \frac{\frac{\alpha_{3a}\sigma_a^3}{n_a^2} - \frac{\alpha_{3r}\sigma_r^3}{n_r^2}}{6\left(\frac{\sigma_a^2}{n_a} + \frac{\sigma_r^2}{n_r}\right)^{3/2}}(U^2 - 1) - \frac{\frac{\alpha_{3a}\sigma_a^3}{n_a^2} - \frac{\alpha_{3r}\sigma_r^3}{n_r^2}}{2\left(\frac{\sigma_a^2}{n_a} + \frac{\sigma_r^2}{n_r}\right)^{3/2}}U^2 - \frac{1}{2}gUZ, \quad (2.5)$$

where  $U, Z$  are i.i.d.  $N(0, 1)$  and

$$g \equiv \left\{ \frac{\frac{\sigma_a^4}{n_a^3}(\alpha_{4a} + 2) + \frac{\sigma_r^4}{n_r^3}(\alpha_{4r} + 2) - \left(\frac{\alpha_{3a}\sigma_a^3}{n_a^2} - \frac{\alpha_{3r}\sigma_r^3}{n_r^2}\right)^2}{\left(\frac{\sigma_a^2}{n_a} + \frac{\sigma_r^2}{n_r}\right)^2} - \frac{\left(\frac{\alpha_{3a}\sigma_a^3}{n_a^2} - \frac{\alpha_{3r}\sigma_r^3}{n_r^2}\right)^2}{\left(\frac{\sigma_a^2}{n_a} + \frac{\sigma_r^2}{n_r}\right)^3} \right\}^{1/2}. \quad (2.6)$$

The CW test with significance level  $\alpha$  is

Accept  $H_0$  if  $T^* < v(1 - \alpha)$

Accept  $H_1$  otherwise,

where  $v(1 - \alpha)$  is the  $100(1 - \alpha)$  percentile of the distribution of  $V$ , obtained by simulation. As for the Welch test, the population moments in (2.5) and (2.6) are estimated from data using sample versions; see Section 3.3.

### The Wilcoxon Rank (WR) Test

The WR test (Wilcoxon, 1945) makes only assumption (A1), as does the CW test. In addition, it assumes that the distribution function  $F(y)$  of the observations  $\mathbf{FY}_r$  is continuous and the distribution function of the observations  $\mathbf{FY}_a$  is  $F(y - \delta)$ , for  $\delta \in \mathbb{R}$ . Then the WR statistic tests the hypotheses,

$$H_0 : \delta \leq 0 \text{ versus } H_1 : \delta > 0. \quad (2.7)$$

In order to test (2.7), the WR test sums the ranks of each of the  $\mathbf{FY}_a$  values in the combined sample of  $N = n_a + n_r$  data consisting of the  $\mathbf{FY}_a$  and  $\mathbf{FY}_r$  values ordered from smallest to largest. Let  $R_i$  denote the rank of  $Y_i$ ;  $i \in \mathbf{FA}$ . The test statistic for the WR test is

$$W = \sum_{i \in \mathbf{FA}} R_i.$$

An exact  $p$ -value is then computed based on the null distribution ( $\delta = 0$ ) of  $W$ , which is obtained by considering all possible  $N!$  permutations of ranks of the  $\mathbf{FY}_a$  and  $\mathbf{FY}_r$ .

However, this is computationally demanding for large  $n_a$  and  $n_r$ . For large  $n_a$  and  $n_r$ , we approximate the distribution of the centered and scaled version of  $W$ ,

$$W^* = \frac{W - .5 - n_a(n_a + n_r + 1)/2}{\sqrt{n_a n_r (n_a + n_r + 1)/12}},$$

with a standard normal (Hollander and Wolfe, 1999). Hence the WR test with significance level  $\alpha$  is:

$$\begin{aligned} &\text{Accept } H_0 \text{ if } W^* < z(1 - \alpha) \\ &\text{Accept } H_1 \text{ otherwise,} \end{aligned}$$

where  $z(1 - \alpha)$  is the  $100(1 - \alpha)$  percentile of the Gaussian distribution with zero mean and unit standard deviation.

### 3. Methods of Analysis and Comparisons

In this section, we continue to consider inference based on a single generic voxel. Simultaneous inference involving all voxels is considered in Section 4.

#### 3.1 Application of Two-Sample Tests to FMRI Data

Let  $\mathbf{T}$  be the set of acquisition times of the observed intensities associated with the given voxel. Assuming the subject was exposed to only one type of neural activation,  $\mathbf{T}$  can be divided into three groups: the time points  $\mathbf{FA}$  where activation of the BOLD signal is expected, the time points  $\mathbf{FR}$  during which the BOLD signal is expected to be in a rest state, and the time points  $\mathbf{B}$  corresponding to the transition periods between the activation and the rest times. An example of such a division of time points is illustrated in Figure ???. In the two-sample tests considered in this article, one sample corresponds to  $\mathbf{FA}$  and other sample corresponds to  $\mathbf{FR}$ ; intensities corresponding to  $\mathbf{B}$  are omitted from further analysis.

Figure 1 about here

Consider the two-sample tests of  $H_0$  versus  $H_1$  given in Section 2. For a given voxel and a given test, accepting the alternative hypothesis  $H_1$  means that the associated voxel is declared to be activated by the experimental stimulus.

#### 3.2 Simulated FMRI data

Six datasets were obtained from 3 healthy volunteers (1 female, 2 males) using a 1.5T Signa scanner. The data were collected under rest conditions; that is, the subjects were not exposed to any stimulus during the experiment and they were instructed to relax in the scanner with their eyes closed. One such rest dataset was obtained from the first male subject (30 years old), two rest datasets were obtained from the second male subject (27

years old), and three rest datasets were obtained from the female (30 years old). Each dataset consisted of 200 volumes, every observed volume contained 28 slices, and each slice had 64x64 voxels. These datasets were preprocessed for motion correction and prewhitened to make the time series uncorrelated (using the software FEAT, which is part of the FSL package; see Smith *et al.*, 2001).

We created activation datasets by essentially adding a signal having *known magnitude and location* of the activation to each preprocessed rest dataset. The signal component was calibrated against an image acquired from a previous unrelated visual-activation fMRI experiment; see Figure ?? for an example. By applying the signal in the locations acquired from a previous visual experiment, we avoided the possibility of applying the signal near so-called default regions (regions which show decreased neuronal activity during the activation of the stimulus) and their confounding effects on the simulated signal. The activation datasets alternated blocks of 10 time points of rest with 10 time points of activation. The average peak-signal change, defined as a ratio between the average of the intensities under the activation and the average of the intensities measured during the rest periods for the most activated voxel, was set to be 3%. Each dataset contains 200 time points; the three sets of time points  $\mathbf{FA}$ ,  $\mathbf{FR}$ , and  $\mathbf{B}$  were obtained assuming a haemodynamic delay of 3 time periods, resulting in  $n_a = 70$  and  $n_r = 73$ .

Figure 2 about here

### 3.3 Violations of equal variances and Gaussianity assumptions

Several methods were used to assess the degree of departure of the activation datasets from (A2) and (A3). Consider a generic voxel and recall from Section 1 that  $\mathbf{FY}_a = \{Y_i\}_{i \in \mathbf{FA}}$  make up the so-called “activated” sample and  $\mathbf{FY}_r = \{Y_j\}_{j \in \mathbf{FR}}$  make up the “rest” sample.

To investigate the violation of Assumption (A3) given in Section 2, thereby allowing  $\sigma_r^2 \neq \sigma_a^2$ , we computed the sample variances for  $\mathbf{FY}_a$  and  $\mathbf{FY}_r$  for each voxel in each activation dataset. The pairs of sample variances of active and rest samples for all voxels that are located in subject’s brain (out of all  $64 \times 64 \times 28 = 114,688$  voxels, only 2,2340 of them were located in subject’s brain) are plotted in Figure ??; the 45-degree line corresponding to equal variances is superimposed. In all panels, and especially in 3(c), we see some points far from the diagonal, which suggests that the assumption of homogeneity is violated for three voxels. A formal F-test ( $\alpha = 0.05$ ) of equal variances detected 1,225 out of 22,340 (5.5%) brain voxels to have significantly different sample variances, and visual inspection of these voxels indicated no spatial pattern. This indicates that, overall, unequal variances may not be a serious problem for these fMRI data.

Figure 3 about here

To investigate departures from Gaussianity, Assumption (A2), we computed the sample skewness and sample excess kurtosis for  $\mathbf{FY}_a$  and  $\mathbf{FY}_r$ , for all six activation datasets. For

the activation sample these are:

$$\hat{\alpha}_{3a} = \frac{\sqrt{n_a} \sum_{i \in \mathbf{FA}} (Y_i - \bar{Y}_a)^3}{\{\sum_{i \in \mathbf{FA}} (Y_i - \bar{Y}_a)^2\}^{3/2}},$$

$$\hat{\alpha}_{4a} = \frac{n_a \sum_{i \in \mathbf{FA}} (Y_i - \bar{Y}_a)^4}{\{\sum_{i \in \mathbf{FA}} (Y_i - \bar{Y}_a)^2\}^2} - 3,$$

and likewise we computed  $\hat{\alpha}_{3r}$  and  $\hat{\alpha}_{4r}$  for the rest sample.

To illustrate graphically the relationship between skewness and kurtosis, we chose one activation dataset. The pairs  $(\hat{\alpha}_{3a}, \hat{\alpha}_{4a})$  for the 22,340 brain voxels from one activation dataset are plotted on the left panel of Figure ??, and the pairs  $(\hat{\alpha}_{3r}, \hat{\alpha}_{4r})$  are plotted on the right panel. For Gaussian data, the plotted pairs should be very close to the origin. In Figure ??, we observe strong departures from zero skewness and zero excess kurtosis in both panels. Thus, we might expect an improvement in hypotheses testing for activation using the CW test or the WR test over the classical two-sample t-test or the Welch test.

Figure 4 about here

More formally, we calculated the Shapiro-Wilk test (e.g., Royston, 1982) for normality ( $\alpha = .05$ ) for each voxel and rest/activation combination. For the dataset used in Figure ??, Table ?? summarizes the number (out of 22,430) of brain voxels that were significantly non-Gaussian. About 12% of activated samples and about 11% of rest samples were declared significant by the Shapiro-Wilk test; if the samples were Gaussian, we would expect only 5% to be declared significant. More than 20% of voxels were declared significant in at least one of the activated or rest samples.

Table 1: Brain-voxels declared significant using Shapiro-Wilk test ( $\alpha = .05$ ), based on one of the six datasets.

		Activated samples		Total
		Significant	Not significant	
Rest samples	Significant	647	2095	2742 (12.3%)
	Not significant	1774	17824	19598
Total		2421 (10.8%)	19919	22340

The spatial distribution of the voxels declared significant is shown in Figure ??; while they are distributed fairly homogeneously between regions of the brain, there is some indication that, within a region, they can clump together.

Figure 5 about here

#### 4. Results

All four two-sample tests were used to test for activation in each voxel. We obtained p-values as in Section 2 where the p-value for the CW test was obtained from simulation of the random variable given by (2.5) and that for the WR test was obtained from the standard normal approximation to  $W^*$ .

Because of the multiple hypotheses being tested (one for each brain voxel), the voxels declared as active were obtained by comparing the p-values with  $\alpha^* \equiv \alpha / \{\# \text{ of brain voxels}\}$  with  $\alpha = .05$ . This is the voxelwise Bonferroni-adjusted level of significance based on an overall level of significance of  $\alpha = .05$ . Voxels with p-values less than or equal to  $\alpha^*$  were pronounced active. Because the activation pattern of each dataset was known, we can estimate and compare the sizes and powers of the two-sample tests.

Let  $A$  denote the set of voxels to which an activation signal has been added and  $R$  the set of voxels with no added activation. Let  $\mathcal{A}_{\text{right}}$  denote the voxels in  $A$  declared to be active, and let  $\mathcal{A}_{\text{WRONG}}$  denote the voxels in  $A$  not declared active. All voxels from category  $R$  can be similarly divided into  $\mathcal{R}_{\text{right}}$ , those non-activated voxels not declared active, and  $\mathcal{R}_{\text{WRONG}}$ , those non-activated voxels which were declared active.

The achieved size of each test was estimated by

$$\hat{\alpha} \equiv (|\mathcal{R}_{\text{WRONG}}| / |\mathcal{R}|),$$

where  $|\mathcal{C}| \equiv \#$  voxels in the region  $\mathcal{C}$  of the brain. The quantity  $\hat{\alpha}$  is also called the false-positive rate and should be comparable to the desired familywise level of significance  $\alpha$  ( $= .05$ ). If  $\hat{\alpha} < \alpha$ , the test is conservative. The power of each test was estimated by

$$\hat{\pi} \equiv (|\mathcal{A}_{\text{right}}| / |\mathcal{A}|),$$

which is the true-positive rate.

Table ?? lists the estimated sizes and powers of each test for all six simulated FMRI datasets. All four tests were consistently very conservative, with the Wilcoxon test being the most conservative. The classical t-test and Welch test had equivalent power, which was consistently greater than that of the Wilcoxon test. The CW test was the most powerful test, uniformly over the six datasets.

Table 2: Estimated size and power of the four two-sample tests for the six datasets.

Dataset	TEST							
	Classical t-test		Welch		CW		WR	
	$\hat{\alpha}$	$\hat{\pi}$	$\hat{\alpha}$	$\hat{\pi}$	$\hat{\alpha}$	$\hat{\pi}$	$\hat{\alpha}$	$\hat{\pi}$
1	.496E-4	.289	.496E-4	.288	.992E-4	.305	0	.277
2	0	.208	7.466E-4	.208	19.927E-4	.224	4.479E-4	.200
3	0	.221	0	.217	0	.235	0	.202
4	.583E-4	.233	.583E-4	.233	1.750E-4	.253	.583E-4	.224
5	0	.205	0	.205	0	.223	0	.188
6	0	.239	0	.237	27.527E-4	.251	1.101E-4	.227

Table ?? gives a more detailed comparison of the classical t-test and the CW test for one of the datasets. While 626 out of 2,173 activated brain voxels were correctly detected as significant by both tests, 37 additional activation voxels were correctly detected by the CW test that were not identified by the classical t-test. Only one activation voxel was identified by the classical t-test that was missed by the CW test.

Table 3: Comparison of the performance of the CW test and the classical t-test, based on one of the six datasets

			CW test			
			Voxels from $\mathcal{A}$		Voxels from $\mathcal{R}$	
			$\mathcal{A}_{\text{right}}$	$\mathcal{A}_{\text{wrong}}$	$\mathcal{R}_{\text{right}}$	$\mathcal{R}_{\text{wrong}}$
Classical t-test	Voxels from $\mathcal{A}$	$\mathcal{A}_{\text{right}}$	626	1	.	.
		$\mathcal{A}_{\text{wrong}}$	37	1509	.	.
	Voxels from $\mathcal{R}$	$\mathcal{R}_{\text{right}}$	.	.	20165	1
		$\mathcal{R}_{\text{wrong}}$	.	.	0	1

## 5. Discussion and Conclusions

While the results were obtained from only one type of scanner, the 1.5T Signa GE, and with fMRI data for three subjects, they show that fMRI data can exhibit both unequal variances and non-Gaussianity. Using the Shapiro-Wilk test, more than 20% of voxels in the dataset were declared significant in one or both of the rest or activated samples. We believe that more powerful scanners will lead to data that are even more non-Gaussian, since their finer spatial resolution involves less averaging of the response.

The Welch test is valid for unequal variances but when non-Gaussianity is suspected, the CW test accounts for both. The WR test is a nonparametric analog of the classical t-test. In the six datasets studied in Section 3, non-Gaussianity was a bigger problem than unequal variances. The results in Section 4 showed that the CW test performed better than the other three tests. These results suggest that the CW test should replace any standard use of the classical parametric or nonparametric two-sample tests based on fMRI data.

## Acknowledgement

This research was supported by the Office of Naval Research under grants N00014-99-1-0214 and N00014-02-1-0052 and by the National Science Foundation Grant DMS-0406026. The authors would like to thank Antonio Algaze and Petra Schmalbrock for providing the fMRI data and the members of FMRI, Oxford UK for initial consultation about simulating activation fMRI datasets. Perceptive comments by the referees led to improvements in the exposition and strengthening of our conclusions.

---

**References**

- Bandettini, P. A., Jesmanowicz A., Wong, E. C. and Hyde, J. S. (1993). Processing strategies for time-course data sets in functional MRI of the human brain. *Magnetic Resonance in Medicine* **30**, 161-173.
- Cressie, N. and Whitford, H. J. (1986). How to use the two sample t-test. *Biometrical Journal* **28**, 131-148.
- Frackowiak, R. S. J., Friston, K. J., Frith, C. D., Dolan, R. J. and Mazziotta, J. C. (1997). *Human Brain Function*. Academic Press.
- Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J. B., Frith, C. D. and Frackowiak, R. S. J. (1995). Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping* **2**, 189-210.
- Friston, K. J., Jezzard, P. and Turner, R. (1994). Analysis of functional MRI time-series. *Human Brain Mapping* **2**, 69-78.
- Hollander and Wolfe, (\*\*\*\*\* add initials \*\*\*\*\*) (1999). *Nonparametric Statistical Methods, 2nd edn*. John Wiley and Sons.
- Josephs, O., Turner, R. and Friston, K. J. (1997). Event-related fMRI. *Human Brain Mapping* **5**, 243-248.
- Royston, P. (1982). An extension of Shapiro and Wilk's *W* test for normality to large samples. *Applied Statistics* **31**, 115-124.
- Smith, S. M., Bannister, P., Beckmann, C., Brady, M., Clare, S., Flitney, D., Hansen, P., Jenkinson, J., Lebovici, D., Ripley, B., Woolrich, M. and Zhang, Y. (2001). FSL: New tools for functional and structural brain image analysis. *NeuroImage* **13**, S249.
- Welch, B. L. (1937). The significance of the difference between two means when the population variances are unequal. *Biometrika* **29**, 350-362.
- Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics* **1**, 80-83.

**第二篇** (此文的摘要和緒論曾用作習題)**An Evaluation of Multiple Behavioral Risk Factors for Cancer in a Working Class, Multi-Ethnic Population**

*Abstract:* Behavioral risk factors for cancer tend to cluster within individuals, which can compound risk beyond that associated with the individual risk factors alone. There has been increasing attention paid to the prevalence of multiple risk factors (MRF) for cancer, and to the importance of designing interventions that help individuals reduce their risks across multiple behaviors simultaneously. The purpose of this paper is to develop methodology to identify an optimal linear combination of multiple risk factors (score function) which would facilitate evaluation of cancer interventions.

*Key words:* Community based research, conditional logistic regression, multiple risk factors, random effects.

**1. Introduction**

Despite the considerable biomedical advances of the last half-century, facilitating improvement in lifestyle behaviors remains the most efficacious population-level strategy for reducing cancer risk. Estimates vary, but suggest that over fifty percent of new cancer cases and up to one-third of cancer mortality could be prevented through improvements in health behavior practices (American Cancer Society, 2004; Doll and Peto, 1981). A 19 percent decline in the rate at which new cancer cases occur, and a 29 percent decline in the rate of cancer deaths, could potentially be achieved by 2015, if prevention efforts were heightened and behavior change sustained. This would translate to the prevention of approximately 100,000 cancer cases and 60,000 cancer deaths each year, by the year 2015 (National Cancer Policy Board and Institute of Medicine, 2003).

There is ample epidemiological evidence for the consideration of red meat consumption, physical activity, and folic acid intake in cancer prevention efforts. Regular physical activity lowers the risk of cancers of the colon, breast, and possibly prostate (Colditz, Cannuscio, and Frazier, 1997; Friedenreich and Rohan, (1995).). An additional 30 percent of cancer deaths can be attributed to adult diet (Anonymous, 1996); higher intake of red meat has been associated with increased risk of colon (Sandhu, White and McPherson, 2001) and prostate cancers (Michaud, Augustsson, Rimm, Stampfer, Willett, and Giovannucci 2001). Associated with both physical inactivity and diet is obesity, which may account for between 25-30 percent of cancers of the colon, breast (postmenopausal), endometrium, kidney, and esophagus (Vainio and Bianchini, 2002). Folic acid is protective against colon cancer (Giovannucci, Stampfer, Colditz, Hunter, Fuchs, Rosner, Speizer, and Willett, 1998); long-term multi-vitamin use, in particular has been found to reduce risk for colon cancer, likely because of its folic acid content (Giovannucci *et al.* 1998).

The risk for many diseases, including colon cancer, is associated with multiple behavioral risk factors (MRF); these behaviors are highly interrelated and tend to cluster within individuals. For example, those who eat high-fat diets are also more likely to be sedentary, suggesting that the behaviors may be mutually reinforcing (see e.g., Emmons, Marcus, Linnan, Rossi, and Abrams, 1999). Change in one behavioral risk factor thus may serve as a stimulus or gateway for change in the other health behaviors (see e.g., Emmons *et al.* 1999), and there are overarching behavioral principles and intervention frameworks that guide behavior change efforts across risk factors.

Consequently, to facilitate population-level reductions in cancer risk, it may be inefficient to target discrete behavioral risk factors, when similar principles might be applied simultaneously to multiple behaviors (Institute of Medicine, 2000).

The literature provides little consensus as to the most appropriate analytic strategy for evaluating the efficacy of MRF interventions; most studies have analyzed the various outcomes independently or by creating a simplistic sum (e.g., 1 RF + 1 RF = 2RFs) (see e.g., Prochaska and Sallis 2004; Campbell, James, Hudson, Carr, Jackson, Oakes, Demissie, Farrell, and Tessaro, 2004). This could be problematic, because the use of separate analytic strategies may result in improper inferences regarding the effect of an MRF intervention because of correlation among the factors. Such strategies may overlook the clustering effect brought about by the agglomeration of multiple behavioral risk factors and have been criticized as being too simplistic. The purpose of this paper is to develop a methodology to identify an optimal linear combination of multiple behavioral risk factors (MRF score function) for cancer that would best facilitate evaluation of an MRF cancer intervention.

## 2. Methods

### 2.1 Study design

The data analyzed in this paper are from the Harvard Cancer Prevention Program Project (HCPPP) Healthy Directions, which is composed of two randomized controlled trials, one in health centers (HC) (Emmons, Stoddard, Gutheil, Suarez, Lobb, and Fletcher 2003), and another in small businesses (SB) (Hunt, Stoddard, Barbeau, Wallace, and Sorensen 2003). The overarching goal of the HCPPP was to create a new generation of cancer prevention interventions that would be effective among working class, multi-ethnic populations. Together, the two arms of the trial were successful in enrolling a sub-population of the multi-ethnic working class population in eastern Massachusetts. The study aims and sampling strategies are published in greater detail elsewhere (Emmons *et al.*, 2003; Hunt *et al.*, 2003).

### 2.2 Health centers

Healthy Directions-HC (Emmons *et al.*) was a randomized controlled trial conducted in collaboration with a large health care delivery system, comprised of 14 multi-specialty medical group practices that serve over 270,000 patients. Ten of the fourteen health centers were invited to participate in this study, and all agreed. Health center served as the unit of randomization and intervention. Briefly, patients who resided in low income, multi-ethnic neighborhoods (defined using census block-groups that were predominantly working class, impoverished, or with low levels of education) were identified and approached for participation through their health center. Individuals identified through geocoding to be residents in the target neighborhoods were deemed eligible if they met the following criteria: (1) being 18-75 years old, (2) having a well-care or follow-up visit scheduled with a participating provider, (3) being able to speak and read either English or Spanish, (4) not having cancer at the time of enrollment, (5) not being employed by the participating health centers, (6) not being employed by a worksite participating in the companion small business study, and (7) providing consent to participate in the randomized study. All providers practicing in the Internal Medicine Departments of the health centers were approached for permission to recruit from among their patient pools. Provider participation averaged 83% across sites (range 50%-100%; 97 clinicians). Patients scheduled for appointments with the participating providers and in the eligible age range were identified through the automated central appointment system. Study staff attempted to recruit 8,963 potentially eligible candidates; 2,547 (28%) individuals were

unreachable. Among the 6,414 potential subjects reached, 867 (14%) were ineligible, 3,330 (52%) refused, and 2,219 (35%) were enrolled. Assuming that 14% of those not reached were also ineligible, the response rate is 29% of those assumed eligible. The cohort recruited at baseline was contacted by telephone after the intervention period to complete a follow-up survey. Of the 2,219 who completed the baseline survey (n=1088 intervention condition; n=1131 control condition), 1,954 (88%) completed the follow-up survey. The follow-up response rate was equivalent across conditions.

### 2.3 Small business

The Healthy Directions-SB study (Hunt *et al.*, 2003) was a randomized controlled trial in which the worksite was the unit of randomization and intervention. Worksites were identified using the Dun and Bradstreet database to locate small businesses with Standard Industrial Classification (SIC) codes 20-39 (manufacturing industries) and employing between 50-150 employees. Additional inclusion criteria included: (1) employing a multi-ethnic population (defined as 25% of workers being first-or second-generation immigrants or people of color), (2) having a turnover rate of less than 20% in the previous year, (3) being autonomous in decision-making power to participate in a study, and (4) agreeing to be randomly assigned to the intervention condition. One hundred thirty-three (133) companies met the eligibility criteria, and of these, 26 agreed to participate (Barbeau, Wallace, Lederman, Lightman, Stoddard and Sorensen 2004).

Data were collected using interviewer-administered surveys among individuals who were permanent employees and worked 20 hours or more per week. On site interviews were administered on company time in the language (either English, Spanish, Portuguese, or Vietnamese) preferred by respondents. Two cross-sectional samples were collected, one at baseline in which 1,740 participants from 26 worksites completed the survey (response rate 84%). The second sample was collected at follow-up 1,408 participants in 24 worksites (during the course of the intervention two worksites dropped out, one intervention and one control) with a response rate of 77%. 974 participants (518 in control worksites and 456 in intervention worksites) completed both the baseline and follow-up surveys forming the embedded cohort used in this analysis.

### 2.4 Data and analysis

The goals of the intervention were to: (1) increase fruit and vegetable intake, (2) decrease red meat consumption, (3) increase physical activity levels, and (4) increase daily multivitamin usage. The following variables assess the individual risk factors measured on a continuous scale: number of servings of fruit and vegetables per day, number of servings of red meat consumed per week (RM), and hours of moderate or vigorous physical activity per week (PA). The fourth measure is a binary variable indicating use of a multi-vitamin on 6 or 7 days per week (MV). In order to keep all variables on an equivalent time scale, we created a new variable for fruit and vegetable consumption that calculated the amount of fruits and vegetables consumed in one week (FV) by multiplying the current measure of fruit and vegetable intake by seven. The continuous variables (FV, RM, PA) were standardized using the formula in Equation 2.1;

$$STV = \frac{V - P_{05}}{P_{95} - P_{05}}, \quad (2.1)$$

where V are the original values for the continuous variables (FV, RM, PA),  $P_{05}$  and  $P_{95}$  are the fifth and ninety-fifth percentile values respectively for a given variable and STV are the new standardized variables (STFV, STRM, and STPA respectively). Standardization was implemented for consistency (to make a one unit change in one variable similar to a one unit change in another) and interpretability. The 5<sup>th</sup> and 95<sup>th</sup> percentiles were used to minimize the influence of outliers.

For the purposes of identifying an optimal linear combination that would show an intervention effect we restricted our sample to only those subjects who received the intervention, responded to both the baseline and follow-up surveys, and have complete data for the four risk behaviors. As opposed to the usual situation of observing how the covariate vector or a linear combination of the covariate vector will change because of treatment, the idea here is to determine how the covariate vector or the linear combination will predict the intervention status. This is similar in spirit to a matched case-control analysis.

A popular method for the analysis of longitudinal data with a dichotomous outcome is a mixed effects logistic regression model. A mixed effects logistic regression model with a logit link will have the form:

$$\log \left[ \frac{\text{pr}(Y_{ij} = 1)}{1 - \text{pr}(Y_{ij} = 1)} \right] = a_i + \beta' X_{ij}. \quad (2.2)$$

Here  $Y_{ij}$ ,  $i = 1, \dots, n$ ,  $j = 1, 2$ , denotes the indicator of intervention time (i.e. pre-intervention  $Y_{i1} = 0$  and post-intervention  $Y_{i2} = 1$ ),  $X_{ij}$  is the covariate vector, and  $a_i$  is a random cluster effect. The subscript  $i$  is an indicator for individual and the subscript  $j$  is an indicator for time. Each individual subject  $i$  is a cluster of two sets of observations, pre-intervention and post-intervention. The random effect variable  $a_i$  can be thought of as measuring an individual's demographic characteristics (i.e., age, gender, race). In our analysis, we want to control for an individual's specific demographic characteristics, therefore, we treat the random effect variable  $a_i$  as a nuisance parameter and condition it out of the model. We can condition them out by using the conditional likelihood based on the fact that  $Y_{i1} + Y_{i2} = 1$ . We are left with a conditional logistic regression model. These types of models are often used to analyze matched case-control studies, where the outcome of interest is whether a subject is a case or control.

In this framework we intend to model

$$\text{logit}(\text{Pr}(Y_{ij} = 1|a_i)) = \beta' X_{ij}, \quad (2.3)$$

where an optimal linear combination, or the best score, will be  $\hat{\beta}' X$ .

We set up our data as if it came from a 1:1 matched case-control study; each individual is a cluster of two observations, one "case" and one "control". One observation is pre-intervention ("control"/baseline) and the second observation is post-intervention ("case"/follow-up). At each time point (pre and post-intervention) each subject has a vector (containing STFV, STRM, STPA, and MV) of covariates.

For matched case-control studies with one case per matched set, the likelihood function for the conditional logistic regression reduces to the partial likelihood of the Cox model for the continuous time scale (Hosmer and Lemeshow 1998). We created dummy survival times so that all cases have the same event time and the corresponding controls are censored at a later time. We used Proc PHREG in SAS<sup>1</sup> to fit the conditional logistic regression model by forming a stratum for each matched set (individual id number). This allowed us to obtain estimates for  $\hat{\beta}$ .

### 3. Results

Using the combined Health Center and Small Business data from the Healthy Directions baseline and follow-up surveys on the 1,209 study participants that received the intervention, we found an optimal score function for the four risk factors:

$$\text{score} = 1.05 * \text{STFV} + 1.70 * \text{MV} + 0.25 * \text{STPA} - 1.35 * \text{STRM}. \quad (3.1)$$

The score is a summary measure of the health behaviors of a subject based on these four factors. From this score, we can see that increasing the number of fruits and vegetables consumed per week,

taking a multivitamin six or more days a week, increasing the amount of physical activity done in a week, and/or decreasing the amount of red meat consumed in a week will increase the score for a subject which in turn means an overall improvement in health behaviors. The dynamics of the score are consistent with the goals of the intervention. A participant can increase their health behavior score by changing one risk factor, or combinations of the four risk factors in a manner consistent with the goals of the intervention.

We believe that these factors not only have individual effects, but that some factors may also have compounding effects. This belief is based on previous evidence of the interrelationships seen in modifying behavioral risk factors (see e.g., Emmons *et al.*, 2004; Butterfield *et al.*, 2004). Therefore, we looked for significant interactions between the four variables. Table ?? shows the analysis of maximum likelihood estimates for our final model. In our final score function (see Equation 3.2), we multiply the effects (parameter estimates) by 100 to increase the range of the scores as well as to simplify interpretation.

Table 1: Analysis of Maximum Likelihood Estimates

Standardized Variable	Parameter Estimate	Standard Error	P-Value
STFV	0.576	0.303	0.0570
MV	2.008	0.2078	<.0001
STPA	0.232	0.193	0.2294
STRM	-1.515	0.343	<.0001
STFV*STRM	1.229	0.565	0.0296
MV*STRM	-0.707	0.343	0.0392

$$\begin{aligned} \text{score} = & 57.6 * STFV + 200.8 * MV + 23.2 * STPA - 151.5 * STRM \\ & + 122.9 * STFV * STRM - 70.7 * MV * STRM \end{aligned} \quad (3.2)$$

There was a significant interaction between the amount of fruits and vegetables consumed per week and the amount of red meat consumed per week, suggesting that changing both behaviors simultaneously is better than changing either behavior alone, but the effect of changing both behaviors is not equal to the sum of the individual changes on the MRF score. There was also a significant interaction between multivitamin usage more than six times a week and the amount of red meat consumed per week, suggesting that changing either behavior alone is good, but changing both behaviors simultaneously will result in an even larger increase on the MRF score.

Table 2: Examples of changes in individual risk factor measures and resulting MRF score

	FV	MV	PA	RM	MRF Score
Baseline values	20	0	4	5	-30.23
<i>Case 1: Optimal values at final</i>					
Final values	35	1	10	1	247.46
change	+15	+1	+6	-4	+277.69
<i>Case 2: Improves only FV</i>					
Final values	35	0	4	5	17.39
change	+15	0	0	0	+47.62
<i>Case 3: Improves only MV</i>					
Final values	20	1	4	5	135.22
change	0	+1	0	0	+165.45
<i>Case 4: Improves only PA</i>					
Final values	20	0	10	5	-19.09
change	0	0	+6	0	+11.14
<i>Case 5: Improves Only RM</i>					
Final values	20	0	4	1	14.64
change	0	0	0	-4	+44.87
<i>Case 6: Improves FV and RM</i>					
Final values	35	0	4	1	72.82
change	+15	0	0	-4	+103.05

Table ?? displays a few examples of how a change in an individual risk factor from the baseline case to the optimal case will change the score. If we consider the first row of Table ?? to be a baseline value in which a subject consumes 20 servings of fruits and vegetables per week, does not take a multivitamin six or more days a week, has four hours of physical activity per week, and consumes five servings of red meat per week (the average values for study subjects at baseline, meeting only the recommend level of physical activity), the standardized values would be 0.32, 0.32, and 0.5 respectively. Therefore the score for a subject at baseline would be

$$score = 57.6*0.32 + 200.8*0 + 23.3*0.32 - 151.5*0.5 + 122.9*0.32*0.5 - 70.7*0*0.5 = -30.2. \quad (3.3)$$

We can consider an arbitrary optimal case as a subject who consumes 35 servings of fruits and vegetables per week (or five a day), takes a multivitamin 6 or more days a week, engages in 10 hours of physical activity per week, and eats one serving of red meat per week (meeting and/or exceeding all of the recommended levels). Table ?? shows the effects of these changes on the score from the baseline case to the optimal case for each variable alone and the effects of combinations of two and three variables. Figure ?? compares our final model (MRF, Equation 3.2) with a main effects model (a model without interactions) showing that the main effects model can both overestimate and underestimate scores predicted from the MRF model due to the absence of the two significant interactions.

Table 3: Score changes with one, two, and three variable changes

Variables Changed	Score Change
FV	47.62
MV	165.45
PA	11.14
RM	44.87
FV + MV	213.07
FV + PA	58.76
FV + RM	72.82
MV + PA	176.59
MV + RM	238.60
PA + RM	56.00
FV + MV + PA	224.21
FV + MV + RM	266.55
FV + PA + RM	83.96
MV + PA + RM	249.73

Although we used only those subjects that received the intervention to develop the score, the score is generalizable to the entire study population. It was created, and is most useful for, the purpose of comparing the subjects that received the intervention to those that received usual care, because it provides a summary measure of the health behaviors of a subject on all intervention risk factors pre and post-intervention. There were 1,297 subjects that received usual care and took both the baseline and follow-up surveys. These subjects can be considered controls for the effect of the intervention. Figure ?? shows box plots of score comparing baseline and follow-up for subjects that received the intervention compared to those that received usual care. In the intervention group, the mean score at baseline was 48.1, while the mean score at follow up was 104.3. In the usual care group the mean score at baseline was 40.4, and the mean score at follow-up was 53.2. The mean change in score for the usual care group was 12.8, while the mean change in score for the intervention group was 56.2. There was a statistically significant difference in the mean change in score from baseline to follow-up when comparing the usual care group to the intervention group ( $p < 0.001$ ). The intervention group showed greater improvements in score at follow-up proving the intervention quite successful.

#### 4. Discussion

Increasing attention has been paid to multiple risk factor interventions, across a range of disease outcomes, both because adverse behavioral risk factors tend to cluster within individuals and because of recognition of the utility of facilitating change across multiple risk behaviors. However, most MRF studies to date have used individual risk factor methods to analyze intervention effects (see e.g., Prochaska and Sallis (2004); Campbell *et al.*, 2004). As shown in Figure ??, the main effects model both over-estimates (e.g., FV & PA & RM) and under-estimates (e.g., MV & PA & RM) the scores predicted from the MRF model, depending on the combination of variables and the degree of change for a given participant in the intervention. Thus, such analytic models may compromise determinations of the efficacy of a MRF intervention. We were successful in modeling a linear combination of behavioral risk factors including interactions between risk factors, an effort that represents an advance over the existing methods for analyzing MRF intervention efficacy.

To illustrate, note that in our final model there are two interaction terms. One between the amount of fruits and vegetables consumed per week and the amount of red meat consumed per week, and another between multivitamin usage more than six times a week and the amount of red

meat consumed per week. Looking at Table ??, we can see that with all the other variables held constant, a change in fruit and vegetable consumption alone from 20 to 35 servings per week will increase the score by 47.62, and a decrease in red meat alone from 5 to 1 servings per week will increase the score by 44.87. However, because of the interaction term, if both variables are changed by the amounts indicated the score would increase by 72.82, which because of the interaction is a smaller than 92.49, the sum of the individual changes. Similarly, if a subject begins to take a multivitamin daily the score will increase by 165.45, and if they decrease red meat from 5 to 1 serving per week the score will increase by 44.87. However, if a participant begins to take a multivitamin daily and decreases red meat consumption by 4 servings per week the score will increase by 238.60, a larger increase than 210.32 that you would get by adding 165.45 from taking a multivitamin daily and 44.87 by decreasing red meat consumption. Cluster effects are not captured by main effects models and are an advantage of this method.

There are some limitations to the method proposed here, namely that the score function depends on the efficacy of the intervention to determine variable weighting. For example, if the intervention was most effective at increasing multivitamin use, the weight (coefficient) for the multivitamin use variable would be largest in magnitude, whereas if the intervention was least effective in changing the participants' physical activity patterns, the weight (coefficient) for the physical activity variable would be the smallest in magnitude. In some cases then, the weights may be a proxy for the amount of participant effort necessary to change the health behavior. For example, in this study we saw that multivitamin usage had the largest weight and thus the most influence on the score.

There are at least two potential explanations for this finding. First, the promotion of multivitamin usage may require less participant burden when compared to the other health behaviors (e.g., physical activity). Thus, it may be easier for participants to modify their multivitamin use; this supposition appears to be supported by the finding of an almost thirty percent increase from baseline to follow-up of the number of subjects taking a multivitamin daily. However, it is important not to undermine the significance of a change in multivitamin usage which is strongly related to the prevention of disease outcomes. Sustained use of multivitamins containing folic acid have been associated with the reduction in risks for numerous conditions including colorectal cancer and cardiovascular disease (Ggiovannucci *et al.*, (2002); Fairfield and Fletcher, 2002). Physical activity on the other hand, is among the most challenging health behaviors addressed in the study to intervene upon. In this population, 66 percent of the subjects were getting the recommended level of physical activity at baseline, and 69 percent at follow-up. Of those subjects that were not at or above the target level of physical activity at baseline, almost 9.5 percent were at or above the target level at follow-up. Another factor to consider is that multivitamin usage was treated as a binary variable in our models. That is, many potential changes are captured in the categorization of either taking a multivitamin 6 or more times a week or not doing so. Relative to increasing one serving of fruits and vegetables a week, decreasing one serving of red meat in a week, or increasing an hour of physical activity a week, this is a substantial change.

Although the purpose of our method was to develop a health behavior score (composite variable), there are some limitations to using this type of variable. The purpose of such a variable is to allow for easy comparisons of the four factors with one number. When there are changes in the score, however, a composite variable does not provide any insight into which individual risk factor(s) have contributed to the change.

Another potential limitation of applying this method to the HCPPP data is the merging of the two cohorts, small businesses and health centers. Our method develops a score function that is independent of the population but not independent of the intervention. By combining the two data sets, we have made the assumption that the interventions given to these two populations are the same. In reality, although the two interventions were quite similar, they were not identical. We decided, however, to combine the two cohorts in order to increase power, and to create a universal score that could be applied to both cohorts. This not only allowed us to make comparisons within

a cohort, but between cohorts. Taking these limitations into account, our methodology remains preferable compared to existing techniques that do not accord weights to the risk factors or adjust for cluster effects.

In summary, we have developed a score that effectively integrates multiple behavioral cancer risk factors into one measure, irrespective of individual demographic factors. We believe that the methods are generalizable to other working class multi-ethnic populations, and future research should be done to evaluate the effectiveness of these methods in other groups. The primary strength of the methodology used to develop the score is that it can be easily implemented to develop scores for other populations, for other combinations of behavioral risk factors, or for other disease outcomes (e.g., cardiovascular disease). Given the increasing attention being paid to the development of MRF interventions, we believe the described method to be the preferred means of analysis in comparison to previously used strategies. Ultimately, we believe that analytic focus on examining clusters of behavioral risk factors will enhance the design of multiple risk factor intervention approaches.

Figure 1 about here

Figure 2 about here

## Acknowledgements

The research of Melody Goodman was supported by National Institute of Child Health and Human Development grant 5 F31 HD043695. The research of Dr. Li was supported by National Institute of Health grant R01CA95747. The research of Dr. Bennett, Dr. Stoddard, and Dr. Emmons was supported by grant 5 P01 CA75308 from the National Institutes of Health, and support to the Dana-Farber Cancer Institute by Liberty Mutual, National Grid, and the Patterson Fellowship Fund.

The statistical output for this paper was generated using SAS/STAT software, Version 8 of the SAS System for Windows. Copyright© 1999-2001 SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA.

## References

- American Cancer Society (2004). Cancer Facts and Figures. Technical Report, American Cancer Society.
- Anonymous (1996). Harvard Report on Cancer prevention Volume: 1 Causes of Human Cancer, *Cancer Causes and Control* **7**, S3-S9.
- Barbeau, E. M., Wallace, L., Lederman, R., Lightman, N., Stoddard, A. and Sorensen, G. (2004). Recruiting small manufacturing worksites that employ multi-ethnic, low-wage workers to a cancer prevention research trial. *Preventing Chronic Disease* **1** 1-9.
- Butterfield, R. M., Park, E. R., Puleo, E., Mertens, A., Gritz, E. R., Li, F. P., and Emmons, K. (2004). Multiple risk behaviors among smokers in the childhood cancer survivors cohort. *Psychooncology* **13**, 619-629.

- Campbell, M. K., James, A., Hudson, M. A., Carr, C., Jackson, E., Oakes, V., Demissie, S., Farrell, D. and Tessaro, I. (2004). Improving multiple behaviors for colorectal cancer prevention among african american church members, *Health Psychol* **23**, 492-502.
- Colditz, G. A., Cannuscio, C., and Frazier, A. (1997). Physical activity and reduced risk of colon cancer: Implications for prevention. *Cancer Causes and Control* **8**, 649-667.
- Doll, R. and R. Peto (1981). The causes of cancer : quantitative estimates of avoidable risks of cancer in the United States today, *J. Natl. Cancer Inst.* **66**, 1191-1308.
- Emmons, K. M., Marcus, B. H., Linnan, L. A., Rossi, J. S. and Abrams, D. B. (1999) Mechanisms in multiple risk factor interventions: Smoking, physical activity, and dietary fat intake among manufacturing workers. *Preventive Medicine* **23**, 481-489.
- Emmons, K. M., Stoddard, A. M., Gutheil, C., Suarez, E. C., Lobb, R. and Fletcher, R. (2003). Cancer prevention for working class, multi-ethnic populations through health centers: The healthy directions study. *Cancer Causes and Control* **14**, 727-737.
- Fairfield, K. M and Fletcher, R. H. (2002). Vitamins for chronic disease prevention in adults: scientific review. *Journal of the American Medical Association* **287**, 3116-3126.
- Friedenreich, C. M., and Rohan, T. E. (1995). Physical activity and risk of breast cancer. *European Journal of Cancer Prevention* **4**, 145-151.
- Giovannucci, E., Stampfer, M. J., Colditz, G., Hunter, D., Fuchs ,C., Rosner, B., Speizer, F. and Willett, W. (1998). Multivitamin use, folate, and colon cancer in women in the Nurse's Health Study. *Annals of Internal Medicine* **129**, 517-524.
- Hosmer, D. W., and Lemeshow, S. (1998). *Encyclopeida of Biostatistics*, 2327-2333. John Wiley.
- Hunt, M. K., Stoddard, A., Barbeau, E. M., Wallace, L. and Sorensen, G. (2003). Cancer prevention for working class, multiethnic populations through small businesses: The Healthy Directions Study. *Cancer Causes & Control* **14**, 749-760.
- Institute of Medicine (2000). *Promoting Health: Intervention Strategies from Social and Behavioral Research*. National Academy Press.
- Michaud, D. S., Augustsson, K., Rimm, E. B., Stampfer, M. J., Willett, W. C. and Giovannucci, E. (2001). A prospective study on intake of animal products and risk of prostate cancer. *Cancer Causes & Control* **12**, 557-567.
- National Cancer Policy Board and Institute of Medicine (2003). Fulfilling the potential of cancer prevention and early detection, Technical Report, Washington D. C.
- Prochaska, J. J. and Sallis, J.F. (2004). A randomized controlled trial of single versus multiple health behavior change: promoting physical activity and nutrition among adolescents. *Health Psychology* **23**, 314-318.
- Sandhu, M. S. and White, I. R. and McPherson, K. (2001). Systematic review of the prospective cohort studies on meat consumption and colorectal cancer risk: A meta-analytical approach. *Cancer Epidemiology, Biomarkers & Prevention* **10**, 439-446.
- Vainio, H. and Bianchini, F. (2002). *IARC Handbooks of Cancer Prevention. Volume 6: Weight Control and Physical Activity*. IARC Press.

**第三篇** (此文的摘要和緒論曾用作習題)**Reducing Subjectivity in the Likelihood**

*Abstract:* Some scientists prefer to exercise substantial judgment in formulating a likelihood function for their data. Others prefer to try to get the data to tell them which likelihood is most appropriate. We suggest here that one way to reduce the judgment component of the likelihood function is to adopt a mixture of potential likelihoods and let the data determine the weights on each likelihood. We distinguish several different types of subjectivity in the likelihood function and show with examples how these subjective elements may be given more equitable treatment.

*Key words:* Mixture likelihood, model averaging, subjectivity.

**1. Introduction**

We propose methods for modeling the likelihood function that will require fewer subjective judgments. We first discuss the nature of the problem of subjectivity in the likelihood function; then we review some related research; and finally, we define a mixture likelihood function and suggest estimation procedures that reduce the effects of subjective views imposed on the observed data.

**1.1 Statement of the problem**

It is sometimes desirable that beliefs of experimenters should be brought into a scientific analysis in ways that minimally distort the measured data (see, for example, Hogarth, 1980; Kyberg and Smokler, 1980; Lad, 1996). But that having been said, scientists observing data sometimes interpret the data points subjectively, according to what they want the data to show, and according to how precisely they believe the data points were measured. The latter procedure is of course quite common. This subjective interpretation of observed data may be totally at the unconscious level, or it may be purposeful (with the purposeful interpretation, the analysis may become fraudulent; see for example, Grayson, 1995, 1997; Howson and Urbach, 1990; and Press and Tanur, 2001).

The subjective interpretation of empirical data in medicine was discussed by Kaptchuk (2003). He stated (page 1, *op. cit.*):

Doctors are being encouraged to improve their critical appraisal skills to make better use of medical research. But when using these skills, it is important to remember that interpretation of data is inevitably subjective and can itself result in bias. Facts do not accumulate on the blank slates of researchers' minds, and data simply do not speak for themselves. Good science inevitably embodies a tension between the empiricism of concrete data and the rationalism of deeply held convictions. Unbiased interpretation of data is as important as performing rigorous experiments. This evaluative process is never totally objective or completely independent of scientists' convictions or theoretical apparatus.

Statistical analysis of a data set most often proceeds by summarizing the distribution of the data in terms of its likelihood function. In order to specify the form of the likelihood function, various assumptions are made about the data, such as mutual independence, identical distributions, unimodality, etc. After the likelihood function has been specified, additional assumptions are sometimes made (significance levels thought to be appropriate are specified, a prior distribution about the underlying unobservable quantities may be brought in, etc.). Analysis of the data generally proceeds by trying to keep the likelihood function treatment of the data as simple as possible, so that the scientist or analyst will introduce minimal distortion of the data. The analyst tries not to discard data, and tries to maximize the chance of understanding what nature is trying to tell us through the revealed data about the underlying phenomenon. In this way, when the analysis of the data has been completed, the claim can reasonably be made that the conclusions drawn from the analysis approximate, if not precisely reflect, the laws of nature, rather than the possible misinterpretations and misunderstandings of the laws of nature by human beings. It will be useful to first briefly define what we mean by objectivity and subjectivity, in this context.

According to Mandik (2001)<sup>1</sup>,

The word *objectivity* refers to the view that the truth of a thing is independent from the observing subject. The notion of objectivity entails that certain things exist independently from the mind, or that they are at least in an external sphere. Objective truths are independent of human wishes and beliefs. The notion of objectivity is especially relevant to the status of our various ideas, and the question is to what extent objectivity is possible for thought, and to what extent it is necessary.

This is but one of many definitions that have been suggested. The elusive quest for objectivity in science has been, and remains, an important topic of discussion among historians and philosophers of science (for extensive additional discussions of the meaning of “objectivity”, see for example, Bower, 1998; Porter, 1995, 1996; and Daston and Galison 1992). For some, scientific objectivity involves the search for *certainty* in knowledge about one of nature’s well-kept secrets, independent of what human beings believe; but in many cases, we find that what we earlier thought to be true about nature, turns out later to be questionable.

In an interesting example from physics, Folger, 2003, pointed out that:

Pioneer 10, launched in 1972, is now some 8 billion miles from home. But it has been slowing down, as if the gravitational pull on it from the sun is growing progressively stronger the farther away it gets. Milgrom proposed (see the MOND pages—MODified Newtonian Dynamics)<sup>2</sup> that Newton’s laws might change at these accelerations. If Milgrom is right, Newton’s and Einstein’s laws will be in for some major tweaking.

Sometimes the scientist has such deep understanding and insight into the phenomenon he/she is studying that the scientist’s own predictions of what should be found from the analysis are far superior to what the data analysis seems to indicate. In some cases the beliefs of the scientist or analyst are so strong, even before actually taking any data that bear on the phenomenon, that the data are interpreted or manipulated so that they will reflect these preconceived views of the scientist. Any preconceived personal views (views held before taking any data), weak or strong, are what we refer to in this context as *subjectivity*.

---

<sup>1</sup>Mandik, P. (2001). *The Internet Encyclopedia of Philosophy*, <http://www.utm.edu/research/iep/o/objectiv.htm>.

<sup>2</sup>MOND pages — <http://www.astro.umd.edu/ssm/mond/>

## 1.2 Related Research

One approach to reducing the effects of differing assumptions about likelihoods may be found in a line of research that involves use of the *empirical likelihood function*. In this approach, most useful in large samples, a discretized, binned, version of the empirical cdf, instead of a specific likelihood function, is used. Inference is then made from a multinomial distribution. An unfortunate feature of this approach is the additional unknown parameters that are concomitantly introduced into the model. See: Owen, 1988, 2001. For typically small and moderate size samples this could be a problem, but for the massive data sets typical of data mining applications (see, for example: Berry and Linoff (1997); and Hastie, Tibshirani, and Friedman, 2001) such an approach could be a helpful alternative.

We show in the next section how we might understand and account for some types of subjectivity that sometimes enters the *likelihood function*, and might not be desired. We will use the definition and form of the likelihood function in which for absolutely continuous random variables, up to a proportionality constant, it is the joint probability density function of the observables given the unobservables.

## 2. Types of Subjectivity in the Likelihood Function

We distinguish three of the types of likelihood subjectivity problems that may occur:

- (a) how to determine the distributional form of the likelihood function in a way that is largely objective, but permits the data themselves to guide the modeling as to whether the data are Normally-distributed, or Gamma-distributed, or possibly follow some other convenient distribution. We call this problem, “*distributional subjectivity*”;
- (b) how to treat observed data that have possibly been weighted subjectively so that some data points are valued more heavily than others, and some are even ignored; we call this problem, “*weighted-data subjectivity*”;
- (c) how to account for the nature of the experiment used to obtain the data that may have favored one type of response over another; we call this problem, “*experiment subjectivity*”.

We treat each of these types of subjectivity in Section 3.

## 3. Reducing Likelihood Subjectivity

### 3.1 The mixture likelihood

We use a convex mixture of various likelihoods for the data; the usual likelihood function results as a special case.

Suppose an experiment is repeated  $n$  times with the resulting one-dimensional data outcomes:  $x_1, x_2, \dots, x_n$ . We suppose that there are  $J$  models for the data that potentially we might reasonably entertain. For simplicity, merely to suggest a general type of approach, we consider problems

involving only one unknown parameter, namely, the means of the  $J$  distributions,  $\omega$ .

In some situations, the parameters may be quite different from one another but they can generally be related functionally. For example, the case of distinguishing between the means of normal and log-normal distributions, where the mean parameter has different meanings in the two cases is sometimes particularly interesting. In such cases, functional relationships among the parameters are required.

Suppose, in the one-parameter problem, we can assume these data to be mutually independent and identically distributed, and we agree to adopt the likelihood function for Model  $m_j$ :

$$\ell_j(x_1, \dots, x_n | m_j, \omega) \equiv \ell_j(\underline{x} | m_j, \omega).$$

Define a “mixture likelihood function”,

$$L_M(x_1, \dots, x_n | \omega) \equiv L_M(\underline{x} | \omega),$$

such that:

$$L_M(\underline{x} | \omega) = E\{\text{likelihood}\} = E_{Model}[\ell(\underline{x} | \omega)] = \sum_{j=1}^J \ell_j(\underline{x} | \omega) P(m_j | \omega) \quad (3.1)$$

where  $\ell_j(\underline{x} | m_j, \omega)$  denotes the usual likelihood function of the data under model  $m_j$ ,  $\ell(\underline{x} | \omega)$  denotes a model-independent likelihood function, and  $P(m_j | \omega)$  denotes the prior probability of model  $m_j$ . The mixture likelihood function is of course a likelihood function itself. If there were only one model ( $J = 1$ ),  $L_M$  reduces to the ordinary likelihood. The mixture likelihood function explicitly assumes that we should combine different models in a linear way. Other possibilities exist of course, and perhaps in certain cases, they are even more desirable. But because for a wide variety of cases, the linear assumption seems appropriate, we will retain this assumption throughout. We next address the issue of how to reduce model subjectivity (how to choose the weights).

### 3.2 Reducing “model subjectivity”

In some instances, the scientist has very strong, theory-based, beliefs about how the data were generated, and how the corresponding likelihood function should behave. In such instances, especially in small samples, the analyst should surely use that information to permit the desired likelihood function to emerge. In other situations where the scientist/analyst wants the data to speak as loudly as possible relative to the scientist’s pre-conceived beliefs, there is no unique way to accomplish this objective. The approach suggested here is to take equal weights in the mixture. Accordingly, take all  $P(m_j | \omega)$  in eqn. (3.1) to be equal (discrete uniform distribution). This interpretation of equal treatment for the different models is:

- (1) in keeping with the approach frequently used for weighting in mixture models to express indifference or ignorance among the various components in the mixture;
- (2) it is the procedure suggested by Laplace when he adopted his Principle of Insufficient Reason (Laplace, 1812, 1814);
- (3) it is consistent with a basic result of information theory that the distribution that corresponds to maximum entropy, or minimum information, is the uniform distribution.

This gives the mixture likelihood function (equally-weighted average likelihood):

$$L_M(\underline{x}|\omega) = \frac{1}{J} \sum_{j=1}^J \ell_j(\underline{x}|m_j, \omega). \quad (3.2)$$

For example, suppose there are just two potential models ( $J = 2$ ) that might reasonably represent the data:  $N(\omega, 1)$  and a Student  $t$ -distribution centered at  $\omega$ , with 3 degrees of freedom (a fat-tail distribution that has a population mean). Then, the mixture likelihood function becomes:

$$L_M(\underline{x}|\omega) = \frac{1}{2} \left\{ \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(x_i - \omega)^2\right\} + \prod_{i=1}^n \frac{m^{m/2}/B(1/2, m/2)}{[m + (x_i - \omega)^2]^{(m+1)/2}} \right\}. \quad (3.3)$$

Clearly each term in equation (3.3) is non-negative and integrates to one (with respect to  $\underline{x}$ ), so  $L_M\{\underline{x}|\omega\}$  is a bone fide likelihood function for the data (as would be the case whichever models we choose). In some situations, one scientist might favor the normal distribution for representing the distribution of the data, while another might favor the Student  $t$ -distribution. By using  $L_M\{\underline{x}|\omega\}$  to represent the likelihood function for all inferences, the analyst reduces the model subjectivity in the description of the data distribution. Maximum likelihood estimation of  $\omega$  is now more complicated numerically than it would be with use of either the normal or the Student  $t$  distributions separately, but the numerical problem is straightforward (see numerical example below) and easily generalizes to more than two possible ordinary likelihoods.

We next numerically illustrate the example suggested in this section of how to reduce model subjectivity when the models under consideration are the  $N(\omega, 1)$  and the Student  $t_3$  centered at  $\omega$ . We randomly generated a total of 20 observations, 10 observations from  $t_3$ , a Student  $t$ -distribution with 3 degrees of freedom centered at  $x = 10$ , and 10 observations from  $N(10, 1)$ . The resulting data are shown in columns 2 and 3 of Table 1a. Then, using the Newton-Raphson method, we calculated the mixed MLE. It is given at the bottom of Column 2 as:  $\hat{\omega} = 9.9168$ . To illustrate variability, there are four replications of this entire process shown in Table 1a; the four resulting mixed maximum likelihood estimates (mixed MLE's) are also shown in Table 1a.

Table 1a: Four replications of model subjectivity

	$t$	normal	$t$	normal	$t$	normal	$t$	normal
1	11.1861	9.6734	8.1266	11.182	8.9931	8.331	11.476	9.8325
2	9.9749	11.542	11.092	10.175	10.641	10.131	10.922	11.051
3	8.5632	10.259	11.374	11.720	9.8717	7.8108	10.270	10.642
4	5.5471	9.4442	9.3422	10.757	8.8824	8.3177	8.3906	9.0293
5	9.5188	10.779	13.189	9.8871	9.7579	9.4354	11.899	8.8359
6	10.994	9.3448	9.6007	9.715	10.385	10.092	9.5234	10.566
7	9.1875	9.9779	9.9069	9.7106	13.659	9.4326	14.559	9.495
8	11.283	9.2274	8.3785	9.8394	9.5543	10.361	10.177	10.247
9	8.7574	10.724	10.073	11.637	9.2285	9.0399	10.155	9.0938
10	7.9804	11.263	10.379	9.4837	11.274	9.7291	9.2795	10.366
Mixed MLE	9.9168		10.240		9.6237		10.116	

For comparison purposes, we also computed the separate ordinary MLE's assuming all 20 observations were generated from a normal, and then, that all 20 observations were generated from a  $t_3$  distribution. Results are given in Table 1b.

Table 1b: Separate MLE's For normal and Student data

Normal MLE	9.761	10.278	9.7463	10.291
$t_3$ -MLE	9.924	10.187	9.6104	10.111

Thus, it may be seen that in the first instance, while the mixed MLE is 9.9168, the MLE assuming all 20 observations came from a normal is 9.7614, whereas the MLE assuming all 20 observations came from a  $t_3$  is 9.9240. Results for the other 3 cases are shown in Tables 1a and 1b as well. Depending upon the assumptions made for the modeling, results for the mixture MLE obtained from the model averaging may differ substantially from those of the separate models, or not.

### 3.3 Reducing “weighted-data subjectivity”

We examine two distinct cases of weighted data subjectivity and model the two cases separately below.

#### Case 1 — Several Observers (Scientists) Rate the Same Data Points Differently

In this case, different observers (scientists) might interpret the same points differently. Some observers might view certain points as mistakes (outliers that were generated from different distributions from the other points), and therefore delete them from the analysis; and others might, according to their own beliefs, weight certain points more heavily than others (perhaps difficult-to-measure points might be weighted less heavily because the error associated with the measurement might be greater than with most of the other points; perhaps certain points obtained were measured under censored conditions; etc.).

For simplicity, assume the data points are mutually independent. We define the likelihood function for Observer  $O_k$  as:

$$\ell(\underline{x}|\omega) = \prod_{j=1}^n [f(\delta_{jk}x_j | O_k, \omega)]^{p_k(\delta_{jk}x_j | O_k)}, \quad (3.4)$$

where:  $p_k(\delta_{jk}x_j | O_k) = 1$ , if Observer  $O_k$  includes the data point  $x_j$  in the analysis, and  $p_k(\delta_{jk}x_j | O_k) = 0$  if not;  $\delta_{jk}$  denotes the weight that Observer  $k$  places on observation  $x_j$ ,  $f(x_j|\omega)$  denotes the pdf (probability density function) of  $X_j$ , conditional on  $\omega$ . The mixture likelihood function may be defined as:

$$\begin{aligned} L_M(\underline{x}|\omega) &= E\{\text{likelihood}\} = E_{data}\{\ell_k(\underline{x}|\omega)\} \\ &= \sum_{k=1}^K \ell_k(\underline{x}|\omega)P_k(O_k), \end{aligned} \quad (3.5)$$

where  $P_k(O_k)$  denotes the prior probability that the data analyst places on the model that has been developed by Observer  $k$ . To be objective (or indifferent among the choices), in the sense we have been discussing, we take  $P_k(O_k) = 1/K$ , for all  $k$ . Then,

$$L_M(\underline{x}|\omega) = \frac{1}{K}\ell_k(\underline{x}|\omega) \quad (3.6)$$

As a simple example, suppose that all  $K$  observers adopt the same distribution for the data, say,  $N(\omega, 1)$  (in Section 3.2 the analyst adopted two different possible distributions for the data),

and assume that they weight the points in the same way, so that  $\delta_{jk} = 1$  for all  $k$ , for all points they include in their analyses, but they may include different points. Then, since the  $n$  observations are independent,

$$\ell_k(\underline{x}|\omega) = \prod_{j=1}^n \left[ \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(x_j - \omega)^2\right\} \right]^{p_k(x_j|O_k)}. \quad (3.7)$$

To be specific, suppose that  $n = 102$ , and that there are two observers,  $O_1$  and  $O_2$ . Suppose further that  $O_1$  believes  $x_{102}$  is an outlier, and  $O_2$  believes that both  $x_{101}$  and  $x_{102}$  are outliers, but they agree that the first 100 points  $(x_1, \dots, x_{100})$  should be included in their analyses. Then,

$$\begin{aligned} p_1(x_j|O_1) &= 1 \quad \text{for } j = 1, 2, \dots, 101, \\ &= 0, \quad \text{for } j = 102, \end{aligned}$$

also,

$$\begin{aligned} p_1(x_j|O_2) &= 1 \quad \text{for } j = 1, 2, \dots, 100, \\ &= 0, \quad \text{for } j = 101, 102, \end{aligned}$$

Then,

$$\ell_1(\underline{x}|\omega) \equiv \ell_1 = \left( \frac{1}{\sqrt{2\pi}} \right)^{101} \exp\left\{-\frac{1}{2} \sum_{j=1}^{101} (x_j - \omega)^2\right\}, \quad (3.8)$$

and

$$\ell_2(\underline{x}|\omega) \equiv \ell_2 = \left( \frac{1}{\sqrt{2\pi}} \right)^{100} \exp\left\{-\frac{1}{2} \sum_{j=1}^{100} (x_j - \omega)^2\right\}, \quad (3.9)$$

Then,

$$\begin{aligned} L_M(\underline{x}|\omega) &= \frac{1}{2} \left\{ \left( \frac{1}{\sqrt{2\pi}} \right)^{101} \exp\left\{-\frac{1}{2} \sum_{j=1}^{101} (x_j - \omega)^2\right\} \right. \\ &\quad \left. + \left( \frac{1}{\sqrt{2\pi}} \right)^{100} \exp\left\{-\frac{1}{2} \sum_{j=1}^{100} (x_j - \omega)^2\right\} \right\} \end{aligned} \quad (3.10)$$

We may now estimate  $\omega$  by maximizing  $L_M(\underline{x}|\omega)$  with respect to  $\omega$ . Note first that if we let  $n_1, n_2$  be the numbers of data points used in the respective analyses of Observers  $O_1$  and  $O_2$ , they are also the numbers of terms in the two summations, and in this example,  $n_1 = 101$  and  $n_2 = 100$ . We may readily find by ordinary differentiation, the mixture maximum likelihood estimator (mixture MLE) to be:

$$\omega = \alpha(\omega)\bar{x}_1 + [1 - \alpha(\omega)]\bar{x}_2, \quad 0 \leq \alpha(\omega) \leq 1, \quad (3.11)$$

where:

$$\alpha(\omega) \equiv \frac{n_1 \ell_1(\omega)}{n_1 \ell_1(\omega) + n_2 \ell_2(\omega)} \quad (3.12)$$

$$\bar{x}_1 \equiv \frac{1}{101} \sum_{j=1}^{101} x_j, \quad \bar{x}_2 \equiv \frac{1}{100} \sum_{j=1}^{100} x_j. \quad (3.13)$$

That is, we find the interesting result that  $(\hat{\omega}|\ell_1, \ell_2)$  is a weighted average (actually a convex combination) of the separate MLE's that the two observers might adopt separately, and the weights

are their respective proportions of their ordinary likelihoods, an intuitively sensible result. But note that because  $\alpha(\omega)$  depends upon  $\omega$ , equations (3.11) and (3.12) must be jointly solved numerically for  $\hat{\omega}$ .

While in large samples, the (continuous) data will generally ultimately swamp any prior distribution weights placed on the data points (see Le Cam, 1956), in small or moderate size samples, certain very influential points that may have been deleted from an analysis can have substantial effects on the interpretation of the experiment outcomes.

We next illustrate this example numerically. We randomly generated 18 points from  $N(0, 1)$ . We then ordered the points, and added 2 larger outliers. We assumed the first observer dropped the largest point as an outlier, and the second observer dropped the two largest points as outliers. We then calculated the mixture MLE numerically from equation (3.11) using the Newton-Raphson method. We replicated the procedure four times to examine variation. Data are shown in Table 2a.

Table 2a: Four replications of weighted-data subjectivity

Observation	$N(10, 1)$	$N(10, 1)$	$N(10, 1)$	$N(10, 1)$
1	8.3959	7.6748	8.1260	7.7977
2	8.4063	7.8796	8.5249	8.8122
3	8.559	7.9954	9.6225	8.9922
4	8.7975	8.7684	9.6490	9.0079
5	9.3082	8.9002	9.7041	9.0501
6	9.6001	8.9819	9.7444	9.1783
7	9.8433	9.2957	9.7660	9.2580
8	9.9802	9.3553	10.0400	9.3645
9	10.2570	9.3687	10.1180	9.4404
10	10.5710	9.5069	10.3150	9.7344
11	10.6690	9.6790	10.4280	9.8685
12	10.6900	9.8179	10.5690	10.0880
13	10.7120	9.8868	10.5780	10.2120
14	10.7140	10.0860	10.6230	10.2380
15	10.8160	10.3790	10.6770	10.3900
16	10.8580	10.4620	10.7310	10.4440
17	11.1910	10.5510	10.7990	10.5690
18	11.2540	10.9440	10.8960	10.7810
19	12.0000	12.0000	12.0000	12.0000
20	13.0000	13.0000	13.0000	13.0000

Calculations of MLE's for the data in Table 2a are given in Table 2b:

Table 2b: MLE's for data with outliers

Mixture MLE	10.0406	9.4205	10.0568	9.6267
$\bar{x}_{18}$	10.0346	9.4185	10.0506	9.6237
$\bar{x}_{19}$	10.1380	9.5544	10.1532	9.7487

We see that for the data in column 2 of Table 2b, for example, the mixture MLE was 10.0406. Had the observers carried out separate MLE's, with Observer 1 dropping only the last observation, he would have found his MLE to be 10.1380, while Observer 2 who dropped both of the last 2 observations would have found her MLE to be 10.0346. While the differences are not large they are intended to be illustrative.

### Case 2 — One Observer (Scientist) Rates Each Data Point Differently

The second case of weighted-data subjectivity involves a single scientist weighting the importance of the data points differently from one another. Here we envision a single scientist who has carried out an experiment many times, but sometimes, for one reason or another, the scientist carried out the experiment with extremely small error, whereas on some other occasions, the scientist associated the experimental outcomes with considerably more error. Thus, which observed results had small associated error, and which had large associated error might differ from one replication of the experiment to the next.

In this context there is just one scientist who rates his/her experimental data differentially, according to how "well" the data point was measured, or what he/she thought should have occurred, or whatever. This is the more typical situation, compared with the first case. The mixture likelihood function is obtained from equations (3.4) and (3.5), for  $K = 1$ , as:

$$L_M(\underline{x}|\omega) = \ell_1(\underline{x}|\omega)P_1\{O_1\} = \prod_{j=1}^n [f(\delta_{j1}x_j | O_1, \omega)]^{p_1(\delta_{j1}|O_1)} . \quad (3.14)$$

To follow the paradigm suggested here we should take  $\delta_{j1} = 1$  for every  $j$ . Of course the individual scientist would often argue that he/she knows better than anyone else that certain points were really not as good as others, and should therefore be down-weighted.

A now-classical example of this type of subjectivity of special historical interest has been documented with real data. It involves the data collected by R. A. Millikan (1868-1953). Dr. Millikan was an American physicist who successfully measured the charge on a single electron, winning a Nobel Prize in 1923 for this famous oil-drop experiment (as well as other prizes). Holton (1978) scrutinized Millikan's laboratory notebooks and found that Millikan had repeated his oil-drop experiment 39 times, obtaining outcomes:  $x_1, \dots, x_{39}$  for the charge on the electron. Holton reported that Millikan had given each of his original sets of observations a personal quality-of-measurement rating: "best", "very good", "good", "fair", and no rating at all for discarded measurements (we interpret his weights to represent his prior probabilities for these measurements). The distribution of his rating results is summarized in the Table 3.

Table 3: Millikan's measurements

rating descriptions	effective raating	$\delta_{j1} = \text{Weight}$	number of measurements
best	4	4/10	2
very good	3	3/10	7
good	2	2/10	10
fair	1	1/10	13
discard	no rating	—	7

For Millikan,  $p_1(\cdot) = 1$ , for 32 data points and  $p_1(\cdot) = 0$  for the discarded 7 points. We order the measurements according to their effective ratings, from “best” to “fair”, and form the weighted average. The estimated value of the charge on the electron is then given by the weighted average:

$$\hat{e} = \frac{4}{10} \sum_{j=1}^2 x_j + \frac{3}{10} \sum_{j=3}^9 x_j + \frac{2}{10} \sum_{j=10}^{19} x_j + \frac{1}{10} \sum_{j=20}^{32} x_j.$$

Millikan formed the weighted average of his measurements and accordingly estimated the charge on the electron as  $4.85 \times 10^{-10}$  esu (electrostatic units). The ordinary equally weighted average would have been  $4.70 \times 10^{-10}$  esu. In his reported value he also averaged in the values obtained by other researchers. By contrast, the accepted value for “ $e$ ”, the charge on the electron, today, is  $4.77 \times 10^{-10}$  esu. But the impressive closeness of Millikan's values with today's accepted value is deceptive; it occurred only because his values were based upon, “a faulty value for the viscosity of air, which when corrected, increases the discrepancy with the modern value by over 40%” (Mathews, 1998).

### 3.4 Reducing “experiment subjectivity”

Suppose there are two experiments that might be performed:  $E_g$  (“ $g$ ” for “good”), and  $E_{\bar{g}}$  (“ $\bar{g}$ ” for “not good”). In  $E_g$  the scientist knows that the experiment will contain one or more variables that might produce effects that will be confounded with the effect of fundamental interest. In  $E_{\bar{g}}$ , there are likely to be fewer such confounding variables, so the scientist believes that he/she is more likely to be able to distinguish the effect he/she is seeking. Concomitantly, it may be that by carrying out  $E_g$ , the scientist is missing the important variables that suggest that the effect sought is really artifactual, and the seeming effect is explainable in other ways. Because the scientist is so convinced that the effect sought is real and not artifactual, he/she reasons that  $E_g$  is a “cleaner” and more promising experiment. The scientist might even argue, in a moment of enthusiastic zeal, that  $E_g$  is cheaper and/or less subject to error.

In both experiments, for simplicity of interpretation, we assume the data are normally distributed with variance equal to 1. Suppose that the scientist referred to above, call him/her Scientist A, would like to show that the population mean for the underlying phenomenon of interest is positive. If Scientist A carries out  $E_g$ , it is more likely that the sample mean  $\bar{x}$  will be positive than if Scientist A carries out  $E_{\bar{g}}$  wherein the sample mean  $\bar{y}$  will imply the alternative hypothesis  $H_{\bar{g}}$ : that the population mean is not positive. If  $E_{\bar{g}}$  is performed the scientist believes results are either unlikely to be supportive of the theory, or they are likely to be sufficiently marginal so that the theory will be in doubt. A priori, the experimenter adjudges the chances for concluding  $H_g$ : the population mean is positive, when performing  $E_g$  as greater than the chances for concluding that the population mean is positive when performing  $E_{\bar{g}}$ . Consequently, Scientist A decides to perform  $E_g$ .

Suppose some other scientist, say Scientist B, performs  $E_{\bar{g}}$ , and subsequently observes  $\bar{y}$  (using the same sample size,  $n$ ). Let  $\theta$  denote an indexing parameter such that  $\theta = 1$  if the hypothesis  $H_g$  is true, and  $\theta = 0$  if the hypothesis  $H_{\bar{g}}$  is false.

$$\begin{aligned} L_M\{\text{data} \mid \theta\} &= E\{\text{likelihood}\} = E_{\text{experiment}}[\ell(\text{data} \mid \theta)] \\ &= \ell(\bar{x} \mid E_g, \theta)P\{E_g\} + \ell(\bar{y} \mid E_{\bar{g}}, \theta)P\{E_{\bar{g}}\} \end{aligned}$$

The mixture likelihood function becomes:

$$L_M\{\text{data} \mid \theta\} = P\{E_g\} \frac{\sqrt{n}}{\sqrt{2\pi}} \exp\left\{-\frac{n}{2}(\bar{x} - \theta)^2\right\} + P\{E_{\bar{g}}\} \frac{\sqrt{n}}{\sqrt{2\pi}} \exp\left\{-\frac{n}{2}(\bar{y} - \theta)^2\right\}.$$

An investigator cognizant of both experiments has both available. In the same spirit of a desire for equity of treatment in the likelihood function, the investigator takes  $P\{E_g\} = P\{E_{\bar{g}}\} = 0.5$ . Then,

$$L_M\{\bar{x}, \bar{y} \mid \theta\} = \frac{1}{2} \left[ \frac{\sqrt{n}}{\sqrt{2\pi}} \exp\left\{-\frac{n}{2}(\bar{x} - \theta)^2\right\} + \frac{\sqrt{n}}{\sqrt{2\pi}} \exp\left\{-\frac{n}{2}(\bar{y} - \theta)^2\right\} \right].$$

Define  $z = (\bar{x} + \bar{y})/2$ . Then, combining terms shows that:

$$L_M\{z \mid \theta\} = \frac{1}{2} \frac{\sqrt{n}}{\sqrt{2\pi}} \exp\{-n/4\} \exp\{-n(z - \theta)^2\}.$$

Thus, the MLE for  $\theta$  is clearly:  $\hat{\theta} = z = (\bar{x} + \bar{y})/2$ . If Scientist A were correct in his/her a priori assessments of what was likely to happen in the experiment,  $\hat{\theta}$  is likely to be closer to zero than  $\bar{x}$  (or even negative), a result that would tend to vitiate Scientist A's conclusions.

For example, for Scientist A's experiment,  $E_g$ , we generated 100 observations from  $N(1, 1)$  and found  $\bar{x} = 1.0598$ . Then, for Scientist B's experiment,  $E_{\bar{g}}$ , we generated 100 observations from  $N(-1, 1)$  and found  $\bar{y} = -.9531$ . So the generalized MLE,  $\hat{\theta}$ , is 0.053, a sample value just barely positive, which might not be convincing in many contexts for asserting that the population mean is really positive.

#### 4. Conclusions

We have been concerned with how to reduce the effects of a scientist's pre-conceived beliefs in the analysis of his/her supposedly objectively-observed data. We have found that we can reduce the effect of some of those subjective interpretations by using a mixture likelihood function, and then choosing the mixture weights that weigh the various interpretations of the data equally.

#### References

- Berry, M. J., and Linoff, G. (1997). *Data Mining Techniques*. John Wiley.
- Bower, B. (1998). Objective visions: Historians track the rise and times of scientific objectivity. *Science News* **154**, 360-362.
- Daston, L. J. and Galison, P. (1992). The image of objectivity. *Representations* **40**, 81-128.

- Folger, T. (2003). Nailing down gravity: New ideas about the most mysterious power in the universe. *Discover Magazine*, Oct., 2003, 34-41.
- Grayson, L. (1995). *Scientific Deception*. The British Library.
- Grayson, L. (1997). *Scientific Deception – An Update.*: The British Library.
- Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer-Verlag.
- Hogarth, R. (1980). *Judgment and Choice*. John Wiley.
- Holton, G. (1978). Sub-electrons, presuppositions, and the Millikan-Ehrenhaft dispute. In *The Scientific Imagination: Case Studies* (Edited by Gerald Holton), 25-83. The Cambridge University Press.
- Howson, C. and Urbach, P. (1990). *Scientific Reasoning: The Bayesian Approach*. Open Court Publishing Co.
- Kaptchuk, T. J. (2003). Effect of interpretive bias on research evidence. *British Medical Journal* **326**, 1453-1455.
- Kyberg, H. E. Jr., and Smokler, H. E., Editors (1980). *Studies in Subjective Probability*. Robert E. Krieger Publishing Co.
- Lad, F. (1996). *Operational Subjective Statistical Methods: A Mathematical, Philosophical, and Historical Introduction*. John Wiley.
- Laplace, P. S. (1812). *Theorie Analytique des Probabilités*.<sup>3</sup> Paris: Courcier.
- Laplace, P. S. (1814). *Essai Philosophique sur les Probabilités*.<sup>4</sup> Paris.
- Le Cam, L. (1956). On the asymptotic theory of estimation and testing hypotheses. *Proc. Third Berkeley Symposium on Math. Statist. And Prob.* **1**, 128-156. University of California Press. p. 308. \*\*\*\*\* Please check where the p.308 come from? \*\*\*
- Mathews, Robert A. J. (1998). Facts versus factions: The use and abuse of subjectivity in scientific research. Cambridge, England: The European Science and Environment Forum, Working Paper 2/98, September, 1998.<sup>5</sup>
- Owen, A. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75**, 237-249.
- Owen, A. (2001). *Empirical Likelihood*. Chapman and Hall.
- Porter, T. (1995). *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton University Press.
- Porter, T. (1996). Statistics, social science, and the culture of objectivity. *Oesterreichische Zeitschrift für Geschichtswissenschaften* **7**, 177-191.
- Press, S. J., and Tanur, J. M. (2001). *The Subjectivity of Scientists and the Bayesian Approach*. John Wiley and Sons.

---

<sup>3</sup>The second, third, and fourth editions appeared in 1814, 1818, and 1820, respectively. It is reprinted in *Oeuvres Completes de Laplace*, Vol. VII, 1847, Paris: Gauthier Villars.

<sup>4</sup>This book went through five editions (the fifth was in 1825) revised by Laplace. The sixth edition appeared in English translation by Dover Publications, New York, in 1951. While this philosophical essay appeared separately in 1814, it also appeared as a preface to his earlier work, *Theorie Analytique des Probabilités*.

<sup>5</sup>See <http://www.esef.org>, p.5.

## 第四篇 (此文的摘要和緒論曾用作習題)

## Application of One Sided $t$ -tests and a Generalized Experiment Wise Error Rate to High-Density Oligonucleotide Microarray Experiments: An Example Using Arabidopsis

**Abstract: Motivation:** A formidable challenge in the analysis of microarray data is the identification of those genes that exhibit differential expression. The objectives of this research were to examine the utility of simple ANOVA, one sided  $t$  tests, natural log transformation, and a generalized experiment wise error rate methodology for analysis of such experiments. As a test case, we analyzed a Affymetrix GeneChip microarray experiment designed to test for the effect of a CHD3 chromatin remodeling factor, PICKLE, and an inhibitor of the plant hormone gibberellin (GA), on the expression of 8256 *Arabidopsis thaliana* genes.

**Results:** The GFWER( $k$ ) is defined as the probability of rejecting  $k$  or more true null hypothesis at a given  $p$  level. Computing probabilities by GFWER( $k$ ) was shown to be simple to apply and, depending on the value of  $k$ , can greatly increase power. A  $k$  value as small as 2 or 3 was concluded to be adequate for large or small experiments respectively. A one sided  $t$ -test along with GFWER(2)=.05 identified 43 genes as exhibiting PICKLE-dependent expression. Expression of all 43 genes was re-examined by qRT-PCR, of which 36 (83.7%) were confirmed to exhibit PICKLE-dependent expression.

*Key words:* \*\*\*\*\* Please add keywords \*\*\*

### 1. Introduction

The advent of inexpensive microarray technology has enabled individual laboratories to easily obtain a global perspective on the expression pattern of thousands of genes. This powerful technology has allowed investigators to diagnose early cancers (Kim, J. W. and Wang, X. W., 2003; Zhang *et al.*, 2003), discover genes that contribute to quantitative traits (Gu *et al.*, 2002), and detect coordinated gene regulation during pivotal developmental events such as embryogenesis and sexual maturation (Girke *et al.*, 2000; Lo *et al.*, 2003; Ruuska *et al.*, 2002).

The first generation microarrays were generally based on two dye methodologies. These cDNA microarray experiments involve hybridizing two mRNA samples, each of which has been converted into cDNA and labelled with its own fluorophore, on a single glass slide that has been spotted with 10,000-20,000 cDNA probes. In contrast, more recent high-density oligonucleotide microarrays, such as those offered by Affymetrix, provide direct information about the expression levels in an mRNA sample and can have a much higher density (Yang and Speed, 2002).

The majority of methodologies for microarray analysis have been developed for two dye spotted arrays (Kerr *et al.*, 2000; Kerr and Churchill, 2001; Lee *et al.*, 2003, Nguyan *et al.*, 2004, for review see Quackenbush, 2001 and Yang and Speed, 2002). Unfortunately these two-dye spotted arrays also pose other statistical issues, such as normalization to correct for dye bias. Furthermore if more than 2 treatments are used, it is not possible to compare all treatments on the same chip thus necessitating an Incomplete Block Design (IBD) type design (Kerr and Churchill, 2001). As such, special experimental designs, such as the reference and rotational design are needed for correct analysis (Kerr and Churchill, 2001; Quackenbush, 2001 and Yang and Speed, 2002).

In contrast, oligonucleotide microarrays use a single dye technology and pose some advantages, including a greatly increased density of genes and simplified experimental design because treatment effects are tested independently on each chip, eliminating the need for IBD designs. Nevertheless, statistical issues remain, such as normality of residuals, homogeneity of residual variance, correlation of errors within an array, and correlation of biological samples across arrays.

Mixed model methods for analysis of microarray experiment, proposed by Wolfinger *et al.* (2001), solves most of these issues (see Craig *et al.*, 2003 for review). However, the complexity of analysis dramatically increases with these advanced methods. Unfortunately, many of the current practitioners of microarray technology do not possess the mathematical expertise necessary to meaningfully employ these methods. On the other hand ANOVA is a tool that is easy to implement with methods common to most researchers. Kerr and Churchill (2001) conclude that “The analysis of variance (ANOVA) is a natural tool for studying data from experiments with multiple categorical factors”.

The first objective of this research was to examine the utility of simple ANOVA for analysis of replicated oligonucleotide microarrays experiments. The motivation was given eloquently by Kerr and Churchill (2001) who stated “An advantage of model based data analysis such as ANOVA is that a model helps the analyst explore the data. If one finds a model inadequate, discovering *why* it is inadequate can help the analyst identify sources of variation and bias.” A secondary objective of this study was to show how using a one sided *t*-test can be used to increase power. The final objective was to introduce an alternative method to increase power by accepting a base number of false positive with high probability.

The ANOVA is particularly suited to analyzing data from microarray experiments that employ a replicated factorial arrangement of treatments. An example of such an experimental design is one in which the investigator looks at gene expression in wild-type and mutant plants in the presence or absence of an added chemical. Many microarray studies incorporate this type of experimental design, e.g. the response of genes in nontumorigenic and tumorigenic tissues to different concentrations of toxic or therapeutic drugs (Lundquist *et al.*, 2002; Martinez *et al.*, 2002) or the response of genes from different tissues to estrogen or other hormones (Abe *et al.*, 2003; Faccioli *et al.*, 2002; Fujita *et al.*, 2003; Goda *et al.*, 2002). This design easily extends into any number of genotypes (or tissues) by any number of developmental time points (or biochemical exposures).

The primary biological objective of this research was to understand how a CHD3-chromatin remodeling factor, PICKLE, and a plant growth regulator, gibberellin (GA), regulate gene expression during germination of Arabidopsis seeds (Rider *et al.*, 2003). PICKLE is necessary for repression of embryonic traits in Arabidopsis (Ogas *et al.*, 1997). Expression of the embryonic state in pickle seedlings is inhibited by the plant growth regulator gibberellin (GA) and is enhanced by application of uniconazole-P, an inhibitor of GA biosynthesis (Izumi *et al.*, 1985; Ogas *et al.*, 1997). Specifically, gene expression was examined in wild-type and *pickle* seeds grown in the absence and presence of  $10^{-8}$  M uniconazole-P. Thus the genotypes were ‘wild type’ vs. the *pickle* mutant, and biochemical exposure was to either  $10^{-8}$  M uniconazole-P or no uniconazole-P during seed germination.

Our working hypothesis was that PICKLE functions during germination to repress genes that promote embryonic identity. In support of such a hypothesis, the transcript levels of two positive regulators of embryogenesis, LEAFY COTYLEDON1 (LEC1) and LEAFY COTYLEDON2 (LEC2) (Lotan *et al.*, 1998; Stone *et al.*, 2001), are elevated during germination of pickle seedlings (Ogas *et al.*, 1999; Rider *et al.*, 2003). Our interest was to find new genes that exhibited PICKLE-dependent expression, i.e. were up regulated. As such, we had a natural one sided test.

## 2. 21 Biological Methods

Seeds and tissues from the *Arabidopsis pickle-1* mutant (in a Columbia ecotype background) and wild-type Columbia were used for all investigations. Plants were grown as described previously (Ogas *et al.*, 1997; Rider *et al.*, 2003).

The Affymetrix GeneChip *Arabidopsis* Genome Array<sup>6</sup> contained 8256 sets of oligos representing approximately 30% of the *Arabidopsis thaliana* transcriptome. A  $2 \times 2$  factorial arrangement of treatments was examined. The first treatment was genotype (*pickle* mutant vs. wild type), the second treatment was uniconazole (applied vs. control), the treatment combinations were designated pkl, Upkl, wt and Uwt (*pickle* mutant untreated, *pickle* mutant treated with uniconazole-P, wild type untreated, and wild type treated with uniconazole-P) were each represented by four biological replicates ( $n = 4$ ) for a total of 16 chips (Rider *et al.*, 2003).

### 3. Statistical Methods

#### 3.1 The ANOVA, partitions, and transformations

The model for the completely randomized design (CRD) associated with the  $k$ -th spot (or gene) is  $Y_{ij}^k = \mu + \tau_i^k + \epsilon_{(i)j}^k$ , where  $Y_{ij}^k$  is the expression (or log transform) for the  $k$ -th gene, in the  $j$ -th replicate of the  $i$ -th treatment;  $\mu^k$  is the overall mean;  $\tau_i^k$  is the effect of the  $i$ -th treatment on that gene, and  $\epsilon_{(i)j}^k$  is random residual. For maximum information treatment effects are further partitioned into main effects and interactions. The partitioning should be reduced to single degree of freedom tests by use of orthogonal contrasts. Because the ANOVA must be completed for each spot on the array, methods to automate the test are needed. To accomplish this goal, we use the well-known result that any single degree of F tests can equivalently be constructed as a  $t$ -test (Gill, 1978). A simple  $t$ -test for any contrast can be computed with the means procedure in SAS or in any standard spreadsheet, such as Excel. The  $t$ -test also offers the advantage of being able to test for a one sided alternative. In some experiments, as in this one, the researchers may only be interested in genes that are either up or down regulated, as a result, the power to detect those genes will be greatly increased.

For expression type data, the variance is usually correlated with the mean, violating a critical assumption for the ANOVA. For such data, transforming to logs will usually correct this problem. Interpretation of log transformed data also better meets the interest of the biologist as significant differences are interpreted as being significant ratios on a non-transformed basis, i.e. the difference between logs of numbers is the same as the log of a ratio. A log base 2 is interpreted as fold change, while base 10 is interpreted as orders of magnitude difference. Natural logs have not been widely used for array data but perhaps represents the most valid biological interpretation due to kinetics. A common rate equation in chemistry is where the rate of change in product ( $\partial Y$ ) per unit of time ( $\partial t$ ) is proportional ( $c$ ) to the product ( $Y$ ), thus  $\partial Y = cY \partial t$ . The solution to this differential equation is  $Y = ce^t$ . Therefore by taking natural logs, the expression is linearized into a rate equation,  $\ln(Y) = \ln c + t$ . If  $t$  is constant across biological replications, then variation in expression is due to linear differences in the rate constant  $c$ , the gene regulatory factor. Differences due to treatments are then interpreted as linear differences in gene regulatory factors (rate constants).

The vast majority of array data will require such a transformation, however, curiously these data better met the assumption when non-transformed. To check this assumption for any data, compute the within gene variance for each gene (the residual error variance in the ANOVA), then plot that against the average expression level for that gene. Any slope significantly different from

<sup>6</sup>part no. 510429, Affymetrix, Santa Clara, CA

zero (a zero slope is parallel to the  $x$  axis) indicates that the data require a transformation before the analysis proceeds.

For a given gene, because each treatment combination was randomized onto each of 4 biological replicates, the experiment as detailed above is a  $2 \times 2$  factorial arrangement of treatments in a completely randomized design (CRD). The ANOVA for this design with treatment effects partitioned is given in Table 1.

Table 1: ANOVA table with partitions.

Source of Variation	Degrees of Freedom	Mean Square
Treatments	$t - 1$	MS(T)
Genotypes ( $C_1$ )	1	MS( $C_1$ )
Inhibitor ( $C_2$ )	1	MS( $C_2$ )
Interaction of Genotype x Inhibitor ( $C_3$ )	1	MS( $C_3$ )
Within Error	$t(r - 1)$	MS(E)

The mean squares for the partitions can be found using the following formula along with the contrast coefficients given in Table 2

Table 2: Coefficients for partitions of treatment effects.

Treatment	Treatment Combination		Contrast Coefficients ( $C_{mj}$ )		
	Genotype	Inhibitor	$C_{1j}$	$C_{2j}$	$C_{3j}$
1- pkl	<i>pickle</i>	None	1	1	1
2- Upkl	<i>pickle</i>	Uniconazole	1	-1	-1
3- wt	wild type	None	-1	1	-1
4- Uwt	wild type	Uniconazole	-1	-1	1

$$MS(C_m) = r \left( \sum_j C_{mj} \bar{Y}_{ij} \right)^2 I \left( \sum_j C_{mj}^2 \right). \tag{3.1}$$

The  $F$  test, which is distributed as  $F$  with 1 and  $t(r - 1)$  degrees of freedom, is then computed as the ratio of  $F = MS(C_m)/MS(E)$ . This test is equivalently computed as  $t$

$$T_m = \frac{\sum_j C_{mj} \bar{Y}_{ij}}{\sqrt{\frac{1}{r} MS(E) \sum_j C_{mj}^2}} \tag{3.2}$$

which has  $t(r - 1)$  degrees of freedom. From these formula it is easy to verify that the calculated value  $F = t^2$  and from tables one can verify corresponding critical values, i.e.  $F_{1,t(r-1)} = (t_{t(r-1)})^2$ .

However, when calculated as a  $t$ -test the sign of the contrast is preserved, thus allowing a one tailed test. This approach will extend to any contrast for any number of treatments, provided the sum of the coefficients for that contrast is zero. To be orthogonal with other contrasts the sum of the cross products must also sum to zero.

For this analysis, our hypothesis was that one or more genes existed for which the expression level was elevated in pickle mutants, regardless of uniconazole treatment. This hypothesis was based on an expression pattern similar to that of LEC1 and LEC2. Thus the primary contrast of interest was the main effect of genotype ( $C_1$ ). Because we were only looking for a similar pattern (up regulation), the power to detect up regulated genes increased. Use of prior information to increase power is more cost effective than increasing the number of biological replicates. In other experiments additional contrasts may be of equal or greater importance, this may be particularly true of the interaction of genotypes with uniconazole treatment ( $C_3$ ), which test the hypothesis that application of uniconazole has a different effect on one genotype than the other.

The critical value of  $t$  depends on a number of factors, including one- vs. two-sided alternatives, degrees of freedom (df) for estimation of error variance, and acceptable type I error rates. Choosing an acceptable Type I error rate is discussed in the next section.

### 3.2 Generalized experiment wise error rate (GFWER(k))

Experimenters have long recognized that if a comparison wise type I error rate (CWER) is used across a great number of tests, a large proportion of declared significant differences would be false. For example analysis of array data involves thousands of comparisons, consequently, if a per comparison error rate of 0.05 were used for our analysis, more than 413 of the 8256 tests would be expected to be declared significant by chance alone. The most widely used approaches to control Type I errors in multiple tests is based on controlling the family wise Type I error rate (FWER) (Fernando *et al.*, 2004). The FWER is the probability of rejecting one or more true null hypotheses, i.e. the probability of accepting one or more false positives. A common method for controlling the FWER is the Bonferroni or Sidak (1967) adjustments.

However, the FWER with those adjustments is too conservative if the cost of false negatives is high relative to the cost of false positives, i.e. they sacrifice power to avoid accepting false positives. Methods have been developed to address this issue by allowing for some false positives among those declared significant, such as the false discovery rate (FDR, Benjamini and Hochberg, 1995; Reiner. *et al.*, 2003; see Nguyen, 2005 for general discussion on this issue). Alternatives to the FDR have since been proposed that take into account the expected number of false null hypothesis and other modifications (see Fernando *et al.*, 2004 for review). However, all methods used to estimate an FDR make assumptions about the distribution of truly expressed genes. As a result, the FDR will either be too liberal or conservative.

Here we present an alternative that does not attempt to establish an FDR. Rather the method is an extension of the FWER methodology to allow for a higher family wise error rate. The development is as follows: Assume a strictly null distribution from which  $N$  independent test statistics are computed, from which  $N$  independent decisions are made at the same critical threshold level. The probability that any one decision is incorrect is  $p$ . An incorrect decision is defined as rejecting a true null hypothesis. With multiple tests, the probability of exactly  $m$  incorrect and  $N - m$  correct decisions is

$$P(m = \text{incorrect} | N = \text{decisions}) = \binom{N}{m} p^m (1 - p)^{N-m} \quad (3.3)$$

The usual FWER =  $\xi(1)$  is the probability of rejecting 1 or more true null hypotheses found as:

$$\xi(1) = \sum_{m=1}^N p^m (1-p)^{N-m} \quad (3.4)$$

or equivalently, 1 minus the probability of no incorrect decisions,

$$\xi(1) = 1 - (1-p)^N \quad (3.5)$$

which is Sidak's (1967) equation. The value of  $p$  per comparison (CWER) is found such that the  $\xi(1)$  is achieved, i.e.

$$p = 1 - e^{\{\ln\{1-\xi(1)\}/N\}} \quad (3.6)$$

Stated in the reverse, there is a 1-FWER probability of *no* incorrect decisions among the  $N$  decisions made, i.e.

$$\omega(k) = \sum_{m=1}^{k-1} \binom{N}{m} p^m (1-p)^{N-m} \quad (3.7)$$

A generalization of this procedure is to divide the total probability of making Type I errors into parts associated with how many errors are likely to be made at a given probability. Among the  $N$  decisions made, define  $\xi(k)$  as the probability of rejecting  $k$  or more true null hypotheses and  $\omega(k)$  as the probability of rejecting fewer than  $k$  true null hypotheses,  $\xi(k) + \omega(k) = 1$ , where

$$\xi(k) = \sum_{m=k}^N p^m (1-p)^{N-m} \quad (3.8)$$

$$\omega(k) = \sum_{m=0}^{k-1} p^m (1-p)^{N-m} \quad (3.9)$$

If for a given  $k$ , the value for  $\xi(k)$  is set to a small value, then among those tests declared significant, one accepts that there will be a high probability of  $k-1$  false positives plus a low probability of  $k$  or more false positives. Therefore, a new type of error rate is defined as GFWER( $k$ ), which is strictly the probability of making  $k$  or more incorrect decisions at a given level of  $p$ , and ignores the probability of less than  $k$  Type I errors. The latter type of errors are considered acceptable in order to gain power and decrease the Type II error rate. For a more general development of the generalized family wise error rate see van der Laan (2004). The GFWER( $k$ ) cannot be solved for directly, but solutions can be found by iteration. SAS source code used to compute adjusted  $p$  values for any  $\xi(k)$  and  $N$  is given at our web site. However, Equations (3.8) and (3.9) can also be approximated by the normal as follows: If  $X$  is binomial with  $n$  trials and probability of success  $p$ , then

$$P[X > r] \approx \Phi \left( \frac{r - np}{\sqrt{np(1-p)}} \right),$$

where  $\Phi$  is the cumulative distribution of standard normal distribution.

Tables 3 and 4 give  $p$  values for, respectively, a one- and two-tailed alternative, and  $\xi(k) = .05$ . Associated critical values of  $t$  are given in Tables 5 and 6 for experiments with 6 and 60 df for estimating error variance. Note that for all  $k$  values, the critical value of  $t$  for a one-sided test

is between 12 and 13% smaller, with corresponding increases in power. Table 3 shows that by allowing for 2 or more false positives in a 1 sided t-test increases the adjusted  $p$  value by 6.5 times, and thereby also increasing the power of the test. Results presented in Tables 3 show that for the range of  $N$  examined (i.e.  $N > 1000$ ) the ratios of  $p$  values for GFWER( $k$ ) to that of GFWER(1) are independent of  $N$ . Thus, for such chips, once the Sidak  $p$  values are found, GFWER( $k$ ) can be found by multiplication using the constants given in the table.

Table 3: Adjusted  $p$  values for  $k = 1$  to 5, chips of size 1,000 and 50,000 and a one-tailed GFWER( $k$ )=5%.

$k$	Number of Tests			
	1,000		50,000	
	$p\text{-value} \times 10^6$	Ratio*	$p\text{-value} \times 10^6$	Ratio*
1	51.29		1.025	
2	335.02	6.5	6.70	6.5
3	783.41	15.3	15.66	15.3
4	1320.01	25.7	26.38	25.7
5	1913.31	37.3	38.23	37.3

\* Ratio of  $p$ -values to that of GFWER(1)

Table 4: Adjusted  $p$ -values for  $k = 1$  to 5, chips of size 1,000 and 50,000 and a two-tailed GFWER( $k$ )=5%.

$k$	Number of Tests	
	1,000	50,000
	$p\text{-value} \times 10^6$	$p\text{-value} \times 10^6$
1	25.63	.521
2	167.5	3.35
3	391.7	7.83
4	660.0	13.19
5	956.65	19.15

An important issue is what value of  $k$  should one use. The value of  $k$  should be set as small as possible without sacrificing too much power. For an experiment of a given size, the rate at which power increases is dependent on the critical value of  $t$ . Examination of Tables 5 and 6 shows that the greatest decrease in the critical value of  $t$ , with either large or small experiment, comes from increasing  $k$  from 1 to 2. For large experiments increasing  $k$  beyond 2, or for small increasing  $k$  beyond 3, brings about much smaller incremental decreases in  $t$ . From these results, some general guidelines can be deduced for choice of  $k$ . Regardless of the number of spots on a chip, a  $k$  value of 2 or 3 should be adequate for large and small experiments respectively.

Table 5: The six genes for which the qRT-PCR assay detected no expression in untreated wild type seed. Transcripts were detected in untreated *pickle* seeds. Transcripts were also detected for both wild type and *pickle* seeds (Uwt and Upkl) germinated in the presence of uniconazole-*p*, thus permitting calculation of Upickle fold change relative to Uwt. The mean values from the arrays are included for illustration. #Pr is the number of times Affymetrix Microarray Suite software (v. 5.0) labeled a gene 'present' for the 16 gene chips used for this investigation.

AGI Code	#Pr	Mean values (4 chips)				qRT-PCR fold change				Putative ID/function
		wt	pickle	Uwt	Upickle	wt	pickle	Uwt	Upickle	
At3g16410	16	4535	14225	3630	18074	-	+	1	204.3	Jacalin type lectin
At4g27140	15	1000	2365	417	3214	-	+	1	8.95	2S1 seed storage protein
At1g67330	4	170	701	97	1255	-	+	1	3.75	uncharacterized
At5g13930	16	18459	37623	18471	46304	-	+	1	1.75	TT4/chalcone synthase
At3g23220	16	1611	2606	1364	2560	-	+	1	1.43	ERF1/transcription factor
At1g09750	16	2005	3200	2178	5193	-	+	1	0.58	nucleoid-like protein

Table 6: Presence of uniconazole-*p* increases derepression of *PICKLE*-dependent genes in *pickle* seedlings.

AGI Code	qPCR Ratios			Putative Function
	pkl/Wt	Upk/Wt	Upk/pkl	
At5g01600	2.90	8.82	3.04	maturation
At3g16420	4.75	11.41	2.40	defense
At1g73190	1.79	2.96	1.65	maturation
At2g28790	1.77	2.90	1.64	
At1g20620	2.33	3.51	1.50	maturation
At5g54740	2.97	4.10	1.43	
At4g19810	2.43	3.33	1.37	
At3g52500	5.55	7.55	1.36	
At3g16430	1.55	2.08	1.34	defense
At1g05510	1.88	2.53	1.34	
At3g16460	4.64	5.22	1.13	defense
At4g08685	2.61	2.78	1.06	
At2g35810	4.24	4.24	1.00	
At2g19590	2.50	2.24	0.90	
At4g37410	2.38	1.95	0.82	
At5g12030	5.36	2.68	0.50	desiccation

#### 4. Biological Verification: QRT-PCR analysis

Those genes found significant with ANOVA were re-analyzed using qRT-PCR to compare results. The qRT-PCR method, while more precise than the chip analysis, is still subject to error. The method is based on PCR amplification of mRNA in the sample until a pre-determined threshold is obtained. Because the amplification is a doubling with each cycle, the accuracy of the method is questionable if there exists less than a 2 fold difference in mRNA between the two treatments. qRT-PCR is also subject to biological variability between samples and should therefore also be replicated and treated to statistical analysis. However, replicated qRT-PCR analysis for

each gene would be extremely expensive and time consuming. Therefore within the limitations of this experiment, and recognizing those limitations, we defined confirmation of *PICKLE*-dependent expression as a two-fold or greater increase in expression level of a given gene in *pickle* versus wild-type seed when grown in either the absence or presence of uniconazole-P. qRT-PCR was used to compare transcript levels in *pickle* versus wild-type seed grown in the absence of uniconazole-P as well as transcript levels in *pickle* versus wild-type seed grown in the presence of uniconazole-P.

Quantitative RT-PCR was performed on an ABI sequence detection system using RNA from one of the biological replicates previously generated (Rider *et al.*, 2003). Oligonucleotide primer sequences and primer concentrations used are listed in supplementary Table 2S available at the web site.

## 5. Results and Discussion

### 5.1 Statistical issues

For this experiment, we used  $\xi(2) = .05$ . Allowing for one false positive raised the adjusted  $p$  value from  $6.21 \times 10^{-6}$  to  $4.1 \times 10^{-5}$  and correspondingly increased the power of the test. The ANOVA method selected 43 genes, less than one of which was expected to be a false positive based on the experimentwise selection criteria that we employed ( $8256 \times 4.1 \times 10^{-5} = .33$ ). Our qRT-PCR analysis supported 36 of the 43 genes (Figure 1). A surprising result of this study was that qRT-PCR did not detect transcripts in wild-type seeds for 6 of the 43 genes identified as having expression differences based on analysis of the array data (Table 5). Although this observation is consistent with the hypothesis that *PICKLE* represses expression of these genes in wild-type seeds to facilitate the developmental transition from embryo to seedling, the array expression values did not suggest absence of transcripts in wild-type seeds.

figure 1 about here

There are at least two possible explanations for the elevated number of observed false positives. Affymetrix constructed this GeneChip when the sequence of the Arabidopsis genome was only partially completed. Inflated expression values for some oligos may have arisen from cross hybridization to unintended targets. In fact, two of the false positives were false because qRT-PCR detected no expression in germinating seeds under any condition. Alternatively, as previously discussed, the discrepancy may be due to different criteria used to determine success for each method. The qRT-PCR data should only be viewed as supporting evidence, not confirmatory.

### 5.2 Biological Inferences

*PICKLE* is necessary to repress expression of embryonic traits in Arabidopsis seedlings. Previous analysis of genes that exhibit *PICKLE*-dependent repression identified genes associated with various stages of seed development. ANOVA identified genes associated with seed development, including 2S albumin genes, HSP17.6, and several lectin-like genes (Guerche *et al.*, 1990; Lenman *et al.*, 1993; Ruuska *et al.*, 2002; Sun *et al.*, 2001). In all, 10 of the genes (28%) identified and confirmed by qRT-PCR analysis were associated with embryo development or exhibit sequence similarity to genes involved in embryo development (Table 1S, available at the web site). Additional studies will be necessary to determine if the other 26 genes that showed *PICKLE*-dependent expression in the germinating seed are also involved in some aspect of embryo development. Previous expression analysis did not suggest a specific role for uniconazole-P in increasing penetrance

of the pickle root phenotype in *pickle* seedlings (Rider *et al.*, 2003). This analysis revealed that the extent of derepression of many of the genes that exhibit *PICKLE*-dependent repression is enhanced by the presence of uniconazole-P (Figure 1, black bars versus white bars). The magnitude of this enhancement, however, was often due in large part to the fact that the presence of uniconazole-P resulted in decreased expression of the gene relative to wild-type seed imbibed in the absence of uniconazole-P (data not shown).

In order to examine the effect of combining the *pickle* mutation with exposure to uniconazole-P, we compared the fold change values of genes in *pickle* versus wt seedlings (pkl/wt) and the fold change values of genes in *pickle* treated with uniconazole-P versus wt seedlings (Upkl/wt) as determined by qRT-PCR (Table 6). In order to make this analysis comparable to previous analysis of the dataset, either the ratio pkl/wt or the ratio Upkl/wt or both had to be  $\geq 2$  for a gene to be included in this analysis. Genes for which a transcript was not detected in wild-type seedlings were excluded from this analysis. Sixteen genes identified with ANOVA met these expression criteria. We found that the presence of uniconazole-P did increase expression of many of these genes in *pickle* seedlings; the transcript level of 10 genes increased 33% or more when *pickle* seeds were imbibed in the presence of uniconazole-P. In contrast, a previous analysis of the same array data identified no genes for which the corresponding transcript was increased by treatment of *pickle* seeds with uniconazole-P (Rider *et al.*, 2003).

Uniconazole-P increases the probability that primary roots of the *pickle* mutant will express embryonic differentiation traits (Ogas *et al.*, 1997). Genes associated with seed development exhibit elevated expression in *pickle* seedlings, suggesting that the expression of these genes contributes to the ability of *pickle* seedling to express embryonic traits after germination (Rider *et al.*, 2003). The discovery that the presence of uniconazole-P enhances the expression of 10 genes in *pickle* seedlings, 5 of which (50%) are involved in seed development or exhibit sequence similarity to genes involved in seed development, suggests for the first time that the increased penetrance of embryonic traits in *pickle* seedlings treated with uniconazole-P may be mediated in part through changes in gene expression. Specifically, our results are consistent with the hypothesis that GA acts in concert with *PICKLE* during germination to repress expression of genes that promote embryonic traits. Further characterization of the genes identified here may facilitate subsequent genetic and biochemical analysis of the GA signal transduction pathway that mediates this response.

### 5.3 Utility

We have shown that a simple ANOVA method can identify a manageable number of candidate genes for differential expression from a gene expression array, most of which were real. Although we only applied the approach to an experiment that incorporated a simple class-by-treatment design, it is applicable to any full factorial design and is computationally straightforward. Previous analysis of the array data employed a modified fold change (MFC) approach (Rider *et al.*, 2003) and failed to detect many of the genes identified by ANOVA. In addition, our current analysis demonstrates that treatment of *pickle* seedlings with uniconazole-P enhances the derepression of *PICKLE*-dependent genes during germination. These results reinforce the power of ANOVA versus a method that emphasizes fold-change.

The practical utility of the  $\text{GFWER}(k)$  method is derived from allowing the user to influence the number of genes identified by selecting the appropriate value for  $k$ , the number of false positives allowed above the threshold significance level. A critical question is what value of  $k$  will result in the greatest increase in power with the lowest number of Type I errors. A simple power analysis showed that regardless of the number of spots on a chip, a  $k$  value of 2 or 3 should be adequate for large and small experiments respectively. Although the  $\text{GFWER}(k)$  and FDR are closely related and greatly increase the power of the experiment by relaxing the Type 1 error rate, the application of the  $\text{GFWER}(k)$  does not attempt to project an FDR, rather, we only set the maximum number of false

positives under the null hypothesis. Calculations for an exact FDR would require knowledge of 1) the number of truly expressed genes, 2) the signal to noise ratio, and 3) their distribution. Without knowing these factors, the FDR as calculated by any of the current methods is an approximation. As a result, the GFWER( $k$ ) may be more or less conservative than FDR methods, depending on the particular experiment. However, the GFWER( $k$ ) is constant and independent of the experiment, which in itself is appealing. This gives rise to another interesting difference between the methods. The expected number of false positives can be determined *a priori* with the GFWER( $k$ ) because the rate is independent of the data, whereas with the FDR (and newer methods as reviewed by Fernando *et al.*, 2004) calculations are dependent on the data and one has to wait until the list is generated to determine what the expected number of false positives will be. This difference could be critical in the planning stage of an experiment.

## Acknowledgements

We thank Guilherme Rosa for helpful comments and discussion during preparation of the manuscript, Jim Henderson for helping to generate the RNA used for this study and Howard Edenberg for chip hybridization and calculation of expression values. **Funding:** Partial Funding for this research was provided by: JO National Institutes of Health (R01GM059770-01A1 and 5R01GM59770-02); JRS The Indiana 21-st Century Research and Development Fund and Purdue Research Foundation; SDR BASF ; WM USDA/NRI 9803430 Animal Genetic Mechanisms. This is journal paper number 17605 of the Purdue University Agricultural Experiment Station. **Supplemental tables** used in the analysis are available at: [http://www.biochem.purdue.edu/research/ogas\\_lab/arrays/JPmethod/](http://www.biochem.purdue.edu/research/ogas_lab/arrays/JPmethod/).

## References

- Abe, H., Urao, T., Ito, T., Seki, M., Shinozaki, K., and Yamaguchi-Shinozaki, K. (2003). Arabidopsis AtMYC2 (bHLH) and AtMYB2 (MYB) function as transcriptional activators in abscisic acid signaling. *Plant Cell* **15**, 63-78.
- Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate — A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B* **57**, 289-300.
- Black M. A., and Doerge R. W. (2002) Calculation of the minimum number of replicate spots required for detection of significant gene expression fold change in microarray experiments. *Bioinformatics* **18**, 1609-1616.
- Craig B. A., Black M. A., and Doerge R. W. (2003). Gene expression data: The technology and statistical analysis. *J. Agricultural Biological and Environmental Statistics* **8**, 1-28.
- Faccioli, P., Lagonigro, M., Cecco, L., de Stanca, A., Alberici, R., Terzi, V., and de Cecco, L. (2002). Analysis of differential expression of barley ESTs during cold acclimatization using microarray technology. *Plant Biol.* **4**, 630-639.
- Fernando, R. L., Nettleton, D., Southey, B. R., Dekkers, J. C. M., Rothschild, M. F., and Soller M. (2004). Controlling the proportion of false positives in multiple dependent tests. *Genetics* **166**, 611-619.
- Fujita, N., Jaye, D. L., Kajita, M., Geigerman, C., Moreno, C. S., and Wade, P. A. (2003). MTA3, a Mi-2/NuRD complex subunit, regulates an invasive growth pathway in breast cancer. *Cell* **113**, 207-219.
- Gill, J. L. (1978). *Design and Analysis of Experiments in the Animal and Medical Sciences. Volume 1.* The Iowa State University Press.

- Girke, T., Todd, J., Ruuska, S., White, J., Benning, C., and Ohlrogge, J. (2000). Microarray analysis of developing Arabidopsis seeds. *Plant Physiol* **124**, 1570-1581.
- Goda, H., Shimada, Y., Asami, T., Fujioka, S., and Yoshida, S. (2002). Microarray analysis of brassinosteroid-regulated genes in arabidopsis. *Plant Physiol* **130**, 1319-1334.
- Gu, C. C., Rao, D. C., Stormo, G., Hicks, C., T., and Province, M. A. (2002). Role of gene expression microarray analysis in finding complex disease. *Genes. Genet Epidemiol* **23**, 37-56.
- Guerche, P., Tire, C., De Sa, F. G., De Clercq, A., Van Montagu, M., and Krebbers, E. (1990). Differential expression of the arabidopsis 2S albumin genes and the effect of increasing gene family Size. *Plant Cell* **2**, 469-478.
- Izumi, K., Kamiya, Y., Sakurai, A., Oshio, H., and Takahashi, N. (1985). Studies of sites of action of a new plant growth retardant (E)-1-(4-chlorophenyl)-4,4-dimethyl-2-(1,2,4-triazol-1-yl)-1-penten-3-ol (S-3307) and comparative effects of its stereoisomers in a cell-free system from Cucurbita maxima. *Plant Cell Physiol* **26**, 821-827.
- Kerr M. K., and Churchill G. A. (2001). Statistical design and the analysis of gene expression microarray data. *Genetical Research* **77**, 123-128.
- Kerr M. K., Martin M., and Churchill G. A. (2000). Analysis of variance for gene expression microarray data. *J. Computational Biology* **7**, 819-837.
- Kim, J. W., and Wang, X. W. (2003). Gene expression profiling of preneoplastic liver disease and liver cancer: A new era for improved early detection and treatment of these deadly diseases? *Carcinogenesis* **24**, 363-369.
- Lee, M. T., Whitmore, G. A., Yukhananov, R. Y. (2003). Analysis of unbalanced microarray data. *Journal of Data Science* **1**, 103-121.
- Lenman, M., Falk, A., Rodin, J., Hoglund, A. S., Ek, B., and Rask, L. (1993). Differential expression of myrosinase gene families. *Plant Physiol* **103**, 703-711.
- Lo, J., Lee, S., Xu, M., Liu, F., Ruan, H., Eun, A., He, Y., Ma, W., Wang, W., Wen, Z., and Peng, J. (2003). 15,000 unique zebrafish EST clusters and their future use in microarray for profiling gene expression patterns during embryogenesis. *Genome Research* **13**, 455-466.
- Lotan, T., Ohto, M., Yee, K. M., West, M. A., Lo, R., Kwong, R. W., Yamagishi, K., Fischer, R. L., Goldberg, R. B., and Harada, J. J. (1998). Arabidopsis LEAFY COTYLEDON1 is sufficient to induce embryo development in vegetative cells. *Cell* **93**, 1195-1205.
- Lundquist, H., Oredsson, S., and Akesson, B. (2002). Effect of quercetin on gene expression in human cells as measured by microarrays — a pilot study. Paper presented at: Health promoting compounds in vegetables and fruit. Proceedings of a workshop in Karrebaeksminde, Denmark, 6 8 November, 2002. DIAS Report, Horticulture. 2002, No.29, 77 80; 10 ref. (Danmarks JordbrugsForskning; Tjele; Denmark).
- Martinez, J. M., Afshari, C. A., Bushel, P. R., Masuda, A., Takahashi, T., and Walker, N. J. (2002). Differential toxicogenomic responses to 2,3,7,8-tetrachlorodibenzo-p-dioxin in malignant and nonmalignant human airway epithelial cells. *Toxicol Sci.* **69**, 409-423.
- Nguyen, D. V., Wang, N. and Carroll, R. J. (2004). Evaluation of missing value estimation for microarray data. *Journal of Data Science* **2**, 347-370.
- Nguyen, D. V. (2005). A unified computational framework to compare direct and sequential false discovery rate algorithms for exploratory DNA microarray studies. *J. Data Science* (In Press).
- Nitin, J., Thatte, j. Braciale, T., Ley, K., O'Connell, M. and Lee. J. K. (2003). Local-pooled-error test for identifying differentially expressed genes with a small number of replicated microarrays. *Bioinformatics* **19**, 1945-1951.
- Ogas, J., Cheng, J. C., Sung, Z. R., and Somerville, C. (1997). Cellular differentiation regulated by gibberellin in the arabidopsis thaliana pickle mutant. *Science* **277**, 91-94.

- Ogas, J., Kaufmann, S., Henderson, J., and Somerville, C. (1999). PICKLE is a CHD3 chromatin-remodeling factor that regulates the transition from embryonic to vegetative development in arabidopsis. *Proc. Natl. Acad. Sci. USA* **96**, 13839-13844.
- Quackenbush, J. (2001). Computational analysis of microarray data. *Nature Reviews Genetics* **2**, 418-429.
- Reiner, A., Yekutieli, D., and Benjamini, Y. (2003). Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics* **19**, 368-375.
- Rider, S. D., Henderson, J. T., Jerome, R. E., Edenberg, H. J., Romero-Severson, J., and Ogas, J. (2003). Coordinate repression of regulators of embryonic identity by PICKLE during germination in arabidopsis. *Plant J.* **35**, 33-43.
- Ruuska, S. A., Girke, T., Benning, C., and Ohlrogge, J. B. (2002). Contrapuntal networks of gene expression during Arabidopsis seed filling. *Plant Cell* **14**, 1191-1206.
- Sidak, Z. (1967). Rectangular confidence regions for means of multivariate normal distributions. *J. Amer. Statist. Asso.* **62**, 626-\*\*\*\*\*
- Sokal, R. R., and Rohlf, F. J. (1995). *Biometry, Third edition*. W. H. Freeman and Company.
- Stone, S. L., Kwong, L. W., Yee, K. M., Pelletier, J., Lepiniec, L., Fischer, R. L., Goldberg, R. B., and Harada, J. J. (2001). LEAFY COTYLEDON2 encodes a B3 domain transcription factor that induces embryo development. *Proc. Natl. Acad. Sci. USA* **98**, 11806-11811.
- Storey, J. D. and Tibshiran, R. (2003a). Statistical significance for genome wide studies. *Proc. Natl. Acad. Sci, USA* **100**, 9440-9445.
- Storey, J. D. and R. Tibshirani, R. (2003b). SAM thresholding and false discovery rates for detecting differential gene expression in DNA microarrays. In *The Analysis of Gene Expression Data: Methods and Software* (Edited by G. Parmigiani, E. S. Garrett, R. A. Irizarry, S.L . Zeger). Springer.
- Sun, W., Bernard, C., van de Cotte, B., Van Montagu, M., and Verbruggen, N. (2001). At-HSP17.6A, encoding a small heat-shock protein in Arabidopsis, can enhance osmotolerance upon overexpression. *Plant J.* **27**, 407-415.
- van der Laan, M. J., Dudoit, S. and Pollard, K. S. (2004). Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. *Statistical Applications in Genetics and Molecular Biology*<sup>7</sup>
- Wolfinger R. D., Gibson G., Wolfinger E. D., Bennett L., Hamadeh H., Bushel P., Afshari C., Paules, R. S. (2001). Assessing gene significance from cDNA microarray expression data via mixed models. *J. Computational Biology* **8**, 625-637.
- Yang, Y. H. and Speed, T. (2002). Design issues for cDNA microarray experiments. *Nature Reviews Genetics* **3**, 579-588.
- Zhang, H., Yu, C. Y., and Singer, B. (2003). Cell and tumor classification using gene expression data: Construction of forests. *Proc. Natl. Acad. Sci. USA* **100**, 4168-4172.

September 4, 2006

---

<sup>7</sup>See <http://www.bepress.com/sagmb/vol3/iss1/art14/>