國立交通大學

統計學研究所

碩士論文

覆蓋區間之平均連串長度

& 基因分析之 p 值

Concept of Average Run Length for Coverage Interval

& p values for Gene Expression Analysis

研 究 生：曾鈺婷

指導教授：陳鄰安　博士

中 華 民 國 九 十 七 年 六 月

覆蓋區間之平均連串長度

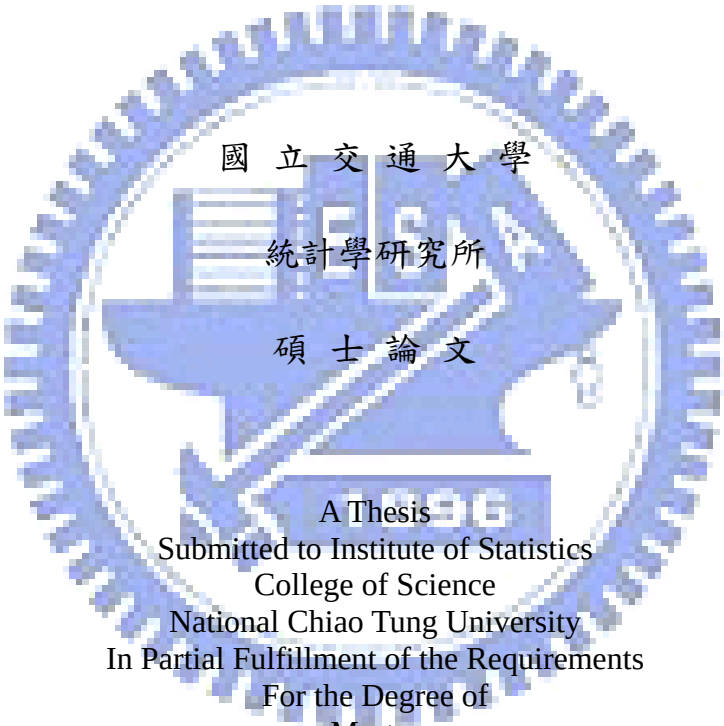
& 基因分析之 p 值

Concept of Average Run Length for Coverage Interval

& p values for Gene Expression Analysis

研 究 生：曾鈺婷　　Student：Yu-Ting Tseng

指導教授：陳鄰安 博士　　Advisor：Dr. Lin-An Chen



國 立 交 通 大 學

統計學研究所

碩 士 論 文

A Thesis
Submitted to Institute of Statistics
College of Science
National Chiao Tung University
In Partial Fulfillment of the Requirements
For the Degree of
Master
In

Statistics

June 2008

Hsinchu, Taiwan, Republic of China

中 華 民 國 九 十 七 年 六 月

覆蓋區間之平均連串長度 & 基因分析之 p 值

研究生：曾鈺婷　　指導教授：陳鄰安 教授

國立交通大學統計學研究所

摘　　要

主題一：

覆蓋區間的使用是來檢測一個人是否是健康的。如果未來的觀察值是被判斷正確且要經過多久這個觀察值會被判斷錯誤，我們就希望去計算這個覆蓋區間的檢定力。爲此，我們研究檢定力和平均連串長度，以評估的覆蓋區間。最後將這兩項工作運用在幾個分佈的研究。

關鍵字：平均連串長度；覆蓋區間；假設檢定；檢定力；參考區間。

主題二：

離群總和的概念已在 Tibshirani 和 Hastie（ 2007 年）和 Wu（ 2007 年）等論文中提出，是在癌症研究中用來檢測許多不同基因，而一個或數個疾病團體指出顯示異常高的基因表達的一個子樣本。我們這裡建議一個新的離群總和的定義，使我們能夠發展其漸近分佈理論，並訂定出它的 P 值。這個 P 值的計算可以用在參數或非參數的分佈。我們進一步地在常態的假設下導出 p 值的公式。爲了研究這個 P 值，我們執行了一些模擬及進行實際的數據分析。這個離群總和，不僅讓我們來計算基因的 P 值，而且是有彈性的處理各種結構的分佈基因的變數。

關鍵字：基因分析；離群總和；p 值。

Concept of Average Run Length for Coverage Interval & *p* values for Gene Expression Analysis

Student：Yu-Ting Tseng　　Advisor：Dr. Lin-An Chen

Institute of Statistics
National Chiao Tung University

Abstract

Topic 1：

One use of coverage interval is monitor if an individual should be classified as healthy one. It is then desired to evaluate the coverage interval for its power if a future observation is classified correctly and how often that this observation could be mis-classified. For this, we study the power and implement the concept of average run length to evaluate the coverage interval. Some distributions are examined for these two tasks.

Key words: Average run length; coverage interval; hypothesis testing; power; reference interval.

Topic 2：

Outlier sum has been proposed in Tibshirani and Hastie(2007) and Wu(2007) for detection of differential genes in cancer studies where one or several disease groups show unusually high gene expression in a subset of their samples. A new outlier sum is proposed that allows us to develop its asymptotic distribution theory for formulating p value. Since it is a function of some distributional parameters, this p value may be computed parametrically or nonparametrically. We further formulate parametrically this p value when normal distribution for gene variables is assumed. To investigate this p value, we perform a simulation and conduct a real data analysis which indicates that this outlier sum not only allows us to compute p values for genes but is also flexible for treatment of various structures of distribution for gene variables.

Key words: Gene expression analysis; outlier sum; p value.

致　　謝

從大學到研究所，轉眼間在交大已經過了這麼多個年頭，又要畢業了，回想研究所兩年的時光，雖然時間過得很快，但也過得很充實，一方面在課業及論文研究上，另一方面則是結識了更多厲害的朋友，不論在學業或者玩樂的功力，總是能拿捏得當，的確都是值得學習的對象。

先要感謝的當然是我的指導教授 陳鄰安老師，他總是能很有耐性的將一個觀念解說的非常清楚，即使在自己忙碌的情況下，依然不厭其煩的與我討論論文內容，非常願意花時間一起研究一些小細節，他同時也是生活上的好老師，告訴我很多人生哲學，並且也是一同討論棒球賽事好伙伴，和老師一起做研究的這一年絕對是一段愉快又難忘的回憶；也要感謝江永進老師、彭南夫老師以及賴怡璇老師對我這篇論文的指導與寶貴的建議。

其次就是感謝我的家人，一直在背後支持我，因為我暴躁的脾氣常常會因為一點不順心就爆發出來，但還是能感受到你們對我的關心與體諒，真的要說聲對不起以及謝謝。還有一些多年來的朋友，經常要聽我抱怨東抱怨西，除了安撫我的心情外，同時也給我很多鼓勵。

在碩士班兩年又認識了許多朋友，先是大一屆的學長姐，經常給予很多課業上的指導，以及日常上的照顧，還有就是同班同學們一直以來的幫助及陪伴，平常時功課上的討論、每個人的驚喜慶生還有難忘的畢旅等，都將成為我珍藏的回憶。

在此，將本篇論文獻給我的師長、家人、好朋友以及同學，並致上我最誠摯的謝意。

曾鈺婷　謹致于

國立交通大學統計研究所

中華民國九十七年六月

Contents

Concept of Average Run Length for Coverage Interval and p values for Gene Expression Analysis

Topic 1: Concept of Average Run Length for Coverage Interval

Abstract

One use of coverage interval is monitor if an individual should be classified as healthy one. It is then desired to evaluate the coverage interval for its power if a future observation is classified correctly and how often that this observation could be mis-classified. For this, we study the power and implement the concept of average run length to evaluate the coverage interval. Some distributions are examined for these two tasks.

Key words: Average run length; coverage interval; hypothesis testing; power; reference interval.

1. Introduction

The coverage interval, in accordance with the recommendation of the *Guide to the Expression of Uncertainty in Measurement* for measuring the uncertainty, refers to population-based measurement values obtained from a well-defined group of reference individuals. This is an interval with two confidence limits which covers the measurement values in the population in some probabilistic sense. Laboratory test results are commonly compared to a coverage interval, called a reference interval in clinical chemistry, before caregivers make physiological assessments, medical diagnoses, or management decisions. An individual who is being screened for some disorder according their relevant measurement from that invidual is suspected to be abnormal if their measurement value lies outside the coverage interval.

The coverage interval can be estimated either parametrically or non-parametrically. The parametric method classically assumes that the underlying distribution of the measurement variable is normal whereas, recently, Chen, Huang and Chen (2007) has proposed a technique for constructing coverage intervals for asymmetric distributions. On the other hand, the non-parametric approach

estimates the quantiles (percentile) directly; the most popular technique for estimating the unknown quantiles is through the empirical quantile.

Basically the coverage interval is to assay the measurement units if they meet defined criteria. In radiation protection, it provides a range of maximum acceptable uncertainty in a dose measured under workplace conditions. In its application to clinical chemistry, it serves as reference standards for measurement units such as head circumference, length and mid-arm sircumference/head circumference ratio for the evaluation of exclusively breastfed infants and it provides some guidance in the interpretation of patient results. When the measurement values do not meet the defined criteria (falling in the coverage interval), these units may be suspected as unsafe or unhealthy and are required for further investigation. These concerns are all statistical hypothesis problems.

However used as an acceptance region for some hypothetical assumption, little has been known the statistical properties of the test based on coverage intervals.

We say that a manufacturing process is in statistical control if the process distribution for the quality characteristic is constant over time and if there is change over time, the process is said to be statistically out of control. A control chart provides the most popular technique for monitoring the process. For a control chart, the most popularly used technique to evaluate its risk is the average run length (ARL) which is the average number of sample points that must be plotted before a point indicates an out-of-control condition. For a control chart, the ARL is

$$ARL = \frac{1}{\alpha} \tag{1.1}$$

where α is the probability that a single sample point exceeds the control limits.

Coverage intervals in clinical chemistry are used for mass screening, to confirm a diagnosis and to monitor a patient's disease status. Diagnosis is test or procedure that helps detect, confirm, document or exclude a disease. An individual is normal if his or her test result falls within a pre-specified coverage interval. Once a disease is suspected, testing result falling outside the coverage interval, further intensive tests may be performed aiming to increase

or decrease the diagnostic certainty of one diagnosis.

How can we measure a coverage interval in terms of effectiveness for its role in diagosis in clinical practice? This is important in reducing the risk of classifying a patient with diseased as non-diseseased person and the risk of classifying an healthy people as diseseased person. One way for this measurement is to transfer the concept of ARL in quality control to measurement science. Suppose that there is a sequence of individuals physically healthy. How many individuals, on the average, in this class that will be examined before a decision of disorder will be claimed is tolerant for the laboratory? Can we design a coverage interval that is more effective in detecting a disorder individual?

2. Specifications for Evaluating the Coverage Interval

The International Federation of Clinical Chemists (IFCC) standard coverage interval for a measurement variable with distribution function F_θ is an estimate of the central interfractile interval

$$C(1-\alpha) = [F_\theta^{-1}(\frac{\alpha}{2}), F_\theta^{-1}(1-\frac{\alpha}{2})] \tag{2.1}$$

(usually with $\alpha = 0.05$) where $F_\theta^{-1}(\delta)$ is the δth fractile for measurement variable. The parametric method generally assumes that the underlying distribution of the measurement variable is normal. If it is not normal, the classical technique to deal with this case is applying a known transformation to normality, setting the normal limits and then transforming to obtain the required interval.

We consider the parametric coverage interval where the underlying distribution is known that we need not to make transformation for approximate normality. Suppose that the parameter value for healthy people is θ_0. Then the true coverage interval is

$$[F_{\theta_0}^{-1}(\frac{\alpha}{2}), F_{\theta_0}^{-1}(1-\frac{\alpha}{2})]. \tag{2.2}$$

However, parameter value θ_0 for distribution of healthy people is usually unknown so that an estimate is required. All approaches to establishing coverage intervals require large groups of individuals (e.g., a minimum of 120 individuals

in the IFCC recommendation). When an appropriate estimate $\hat{\theta}$ for θ is computed from the measurement values is available, the coverage interval based on the central interfractile interval is

$$\hat{C}(1-\alpha) = [F_{\hat{\theta}}^{-1}(\frac{\alpha}{2}), F_{\hat{\theta}}^{-1}(1-\frac{\alpha}{2})]. \tag{2.3}$$

Our interest is, as long as we have an established coverage interval $\hat{C}(1-\alpha)$, how is it performed for diagnosis of disease? The use of coverage interval in diagnosis is, in fact, testing the follwoing hypotheses:

$$H_0: \text{ The individual is healthy vs. } H_1: \text{ The individual is unhealthy.} \tag{2.4}$$

The test is then set as the following:

$$\begin{array}{c}\text{Accepting } H_0 \text{ when the measurement value falls in } \hat{C}(1-\alpha) \text{ and} \\ \text{not rejecting } H_0 \text{ when the measurement value falls outside } \hat{C}(1-\alpha).\end{array} \tag{2.5}$$

An individual will be suspected to be abnormal when H_0 is rejected. There are two errors may happen in the diagnosis based on coverage interval:

Type I error: The individual is healthy but he/she is claimed to be unhealthy

Type II error: The individual is unhealthy but he/she is claimed to be healthy

Our interest in diagnosis of disease through the estimated coverage interval includes the followings:

(a) A $100(1-\alpha)\%$ coverage interval is expected to have probability $1-\alpha$ to claim a healthy people to be healthy. How is it performed in sample coverage interval?

(b) On the other hand, a coverage interval is expected to have large probability to claim a diseased people to be diseased. How is it performed in sample coverage interval for this case? The test procedure is based on coverage interval.

3. A Study for Normal Distribution

Let $X_1, ..., X_n$ be a random sample drawn from the normal distribution $N(\mu_0, \sigma_0^2)$. However, μ_0 and σ_0^2 are assumed to be unkown. The true $100(1-\alpha)\%$ coverage interval is

$$(\mu_0 - z_{1-\frac{\alpha}{2}}\sigma_0, \mu_0 + z_{1-\frac{\alpha}{2}}\sigma_0) \tag{3.1}$$

which is also unknown. Hence, it is estimated by the $100(1-\alpha)\%$ normal coverage interval as

$$(\bar{X} - z_{1-\frac{\alpha}{2}}S, \bar{X} + z_{1-\frac{\alpha}{2}}S). \tag{3.2}$$

Now, suppose that X_0 is the characteristic variable of interest for diagonosis based on coverage interval estimate of (3.2). If X_0 is in healthy condition, the probability of type I error is derived in the follwoings:

$$\begin{aligned}
&P(\text{Type I error}) = P_{\mu_0,\sigma_0}(X_0 \notin (\bar{X} - z_{1-\frac{\alpha}{2}}S, \bar{X} + z_{1-\frac{\alpha}{2}}S)) \\
&= 1 - P_{\mu_0,\sigma_0}(\bar{X} - z_{1-\frac{\alpha}{2}}S \le X_0 \le \bar{X} + z_{1-\frac{\alpha}{2}}S) \\
&= 1 - P_{\mu_0,\sigma_0}(-z_{1-\frac{\alpha}{2}} \le \frac{X_0 - \bar{X}}{S} \le z_{1-\frac{\alpha}{2}}) \\
&= 1 - P_{\mu_0,\sigma_0}(-\frac{z_{1-\frac{\alpha}{2}}}{\sqrt{1+\frac{1}{n}}} \le \frac{X_0 - \bar{X}}{\sqrt{1+\frac{1}{n}}S} \le \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{1+\frac{1}{n}}}) \\
&= 1 - P(-\frac{z_{1-\frac{\alpha}{2}}}{\sqrt{1+\frac{1}{n}}} \le t(n-1) \le \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{1+\frac{1}{n}}})
\end{aligned}$$

where we use the fact that, under H_0, $\frac{X_0-\bar{X}}{\sqrt{1+\frac{1}{n}}S} \sim t(n-1)$. Next, suppose that X_0 is in unhealthy condition, let μ and σ^2 be the true mean and variance of variable X_0. For deriving the probability of type II error, we first derive the desired test statistic. It is seen that $X_0 - \bar{X}$ has the normal distribution $N(\mu - \mu_0, \sigma^2 + \frac{\sigma_0^2}{n})$ and $\frac{(n-1)S^2}{\sigma_0^2}$ has chi-square distribution $\chi^2(n-1)$ and these two quantities are independent. We then have the following

$$T = \frac{X_0 - \bar{X}}{\sqrt{\left(\frac{\sigma}{\sigma_0}\right)^2 + \frac{1}{n}}\ S} \sim t_{n-1}(\frac{\mu - \mu_0}{\sqrt{\sigma^2 + \frac{\sigma_0^2}{n}}})$$

where $t_k(a)$ represents the noncentral t distribution with degrees of freedom k and noncentrality parameter a. The derivation of type II error is as follows:

$$\begin{aligned}
&P(\text{Type II error}) = P(X_0 \in (\bar{X} - z_{1-\frac{\alpha}{2}}S, \bar{X} + z_{1-\frac{\alpha}{2}}S)) \\
&= P(\frac{-z_{1-\frac{\alpha}{2}}}{\sqrt{\left(\frac{\sigma}{\sigma_0}\right)^2 + \frac{1}{n}}} \le \frac{X_0 - \bar{X}}{\sqrt{\left(\frac{\sigma}{\sigma_0}\right)^2 + \frac{1}{n}}\ S} \le \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{\left(\frac{\sigma}{\sigma_0}\right)^2 + \frac{1}{n}}}) \\
&= P(\frac{-z_{1-\frac{\alpha}{2}}}{\sqrt{\left(\frac{\sigma}{\sigma_0}\right)^2 + \frac{1}{n}}} \le t_{n-1}(\frac{\mu - \mu_0}{\sqrt{\sigma^2 + \frac{\sigma_0^2}{n}}}) \le \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{\left(\frac{\sigma}{\sigma_0}\right)^2 + \frac{1}{n}}}).
\end{aligned}$$

Let $\delta_1 = \frac{\sigma}{\sigma_0}, \delta_2 = \frac{\mu - \mu_0}{\sqrt{\sigma^2 + \frac{\sigma_0^2}{n}}}$. We will evaluate this probability under some values of δ_1 and δ_2. In this design, we have

$$\beta = P(\text{Type II error}) = P(\frac{-z_{1-\frac{\alpha}{2}}}{\sqrt{\delta_1^2 + \frac{1}{n}}} \leq t_{n-1}(\delta_2) \leq \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{\delta_1^2 + \frac{1}{n}}}). \tag{3.3}$$

and the power is $1 - \beta$. When $\delta_1 = 1$ and $\delta_2 = 0$ is true, the power is expected to be the probability of type I error. On the other hand, when this assumption is not true, we expect that the power is large when the deviation is big.

Any sequence of sample points that leads to a disorder signal is called a run. The number of individuals that is taken during a run is called the "run length." Clearly, the run length is of very importance in evaluating how well a coverage interval performs. Because run length can vary run to run, from the statistical point of view, it is more interesting to evaluate the average run length (ARL) that is defined as

$$ARL = \frac{1}{1 - \beta}. \tag{3.4}$$

If the coverage interval is monitoring a sequence of healthy people, a perfect interval would never generate a signal of disorder - thus, the ARL would be infinitely large. If the coverage interval is monitoring a sequence of un-healthy people, a perfect interval would quickly generate a signal of disorder - thus, a coverage interval with an ARL of 1 would be desired. However, statistically this is not possible.

We would like to see a high ARL when the coverage interval is treating a group of healthy people and a low ARL when it is treating a group of un-healthy people. However, from the statistical point, we expect a high ARL when the parameters of the underlying distribution are on target and low ARL when the parameters shift to an unsatisfactory level.

Definition 3.1. The average run length (ARL) represents the length of time the consecuitive diagnoses must run, on the average, before a coverage interval will indicate an disorder.

We display the powers of (3.3) for several alternatives in Table 1.

Table 1. Powers for Normal distribution $N(\mu, \sigma^2)$ (two-sided)

	$(\delta_1, \delta_2) = (1, 0)$	$(1, 1)$	$(2, 0)$	$(1, 2)$	$(2, 1)$	$(2, 2)$
$n = 20$	0.07098	0.20102	0.34234	0.54309	0.54165	0.84932
$n = 30$	0.06369	0.19075	0.33717	0.53437	0.53833	0.84873
$n = 50$	0.05806	0.18250	0.33310	0.52718	0.53571	0.84827
$n = 100$	0.05398	0.17630	0.33008	0.52165	0.53376	0.84792
$n = 500$	0.05079	0.17132	0.32769	0.51714	0.53222	0.84764

Table 2. ARL for Normal distribution $N(\mu, \sigma^2)$ (two-sided)

	$(\delta_1, \delta_2) = (1, 0)$	$(1, 1)$	$(2, 0)$	$(1, 2)$	$(2, 1)$	$(2, 2)$
$n = 20$	14.0885	4.9745	2.9211	1.8413	1.8462	1.1774
$n = 30$	15.7011	5.2423	2.9658	1.8713	1.8576	1.1782
$n = 50$	17.2236	5.4793	3.0021	1.8969	1.8667	1.1789
$n = 100$	18.5254	5.6721	3.0295	1.9170	1.8735	1.1794
$n = 500$	19.6889	5.8370	3.0517	1.9337	1.8789	1.1797

We have several comments drawn from the above two tables:

(a) When H_0 is true, the ARL expected to be 20. This means that, in average, 20 healthy people will have one being classified as an unhealthy individual. However, the results are all not identical to 20 that can be as small as only 14 for sample size $n = 20$. The ARL increases in sample size n and it is seen approached to 20 when n goes to infinity.

(b) When the parameters are moved away from the null one, the power increases and the ARL decreases. This satisfies the expectation for the use of coverage interval in monitoring an individual's health.

There is no other approach that has studied the ARL. So, we can't make comparison for this approach with others.

We may consider a one sided coverage interval as $(-\infty, \mu_0 + z_{1-\alpha}\sigma_0)$ and its estimate is

$$(-\infty, \bar{X} + z_{1-\alpha}S).$$

The probability of type II error of this coverage interval estimate may be shown as

$$\beta = P(\text{Type II error}) = P(-\infty < t_{n-1}(\delta_2) \leq \frac{z_{1-\alpha}}{\sqrt{\delta_1^2 + \frac{1}{n}}})$$

and the power is $1-\beta$. We display the power and ARL results in Tables 3 and 4.

Table 3. Powers for Normal distribution $N(\mu, \sigma^2)$ (one-sided)

	$n=20$	$n=30$	$n=50$	$n=100$	$n=500$
$\delta_1=1, \delta_2=-1$	0.00613	0.00540	0.00485	0.00446	0.00415
$\delta_1=1, \delta_2=1$	0.28592	0.27725	0.27022	06489	0.26059
$\delta_1=1, \delta_2=-2$	0.00025	0.00020	0.00017	0.00015	0.00013
$\delta_1=1, \delta_2=2$	0.65648	0.65076	0.64605	0.64244	0.63950
$\delta_1=2, \delta_2=-1$	0.03662	0.03579	0.03514	0.03466	0.03428
$\delta_1=2, \delta_2=1$	0.57603	0.57415	0.57267	0.57156	0.57068
$\delta_1=2, \delta_2=-2$	0.00269	0.00258	0.00250	0.00244	0.00239
$\delta_1=2, \delta_2=2$	0.81616	0.88124	0.88095	0.88073	0.88056

Table 4. ARL for Normal distribution $N(\mu, \sigma^2)$ (one-sided)

	$n=20$	$n=30$	$n=50$	$n=100$	$n=500$
$\delta_1=1, \delta_2=-1$	162.97	185.09	206.09	224.20	240.40
$\delta_1=1, \delta_2=1$	3.4974	3.6068	3.7006	3.7751	3.8374
$\delta_1=1, \delta_2=-2$	3913.6	4784.8	5676.4	6494.5	7264.3
$\delta_1=1, \delta_2=2$	1.5233	1.5367	1.5479	1.5566	1.5637
$\delta_1=2, \delta_2=-1$	27.307	27.939	28.454	28.843	29.164
$\delta_1=2, \delta_2=1$	1.7360	1.7417	1.7462	1.7496	1.7523
$\delta_1=2, \delta_2=-2$	371.39	386.786	399.583	409.474	417.573
$\delta_1=2, \delta_2=2$	1.2252	1.1348	1.1351	1.1354	1.1356

4. Coverage Intervals for Gamma and Exponential Distributions

Consider the Gamma distribution $\Gamma(k, \beta)$ with pdf of the form

$$f_\beta(x) = \frac{1}{\Gamma(k)\beta^k} x^{k-1} e^{-x/\beta}, x > 0.$$

The αth quantile of this distribution is $F_\beta^{-1}(\alpha) = \frac{\beta}{2}\chi^2_{2k}(\alpha)$. The one sided $1-\alpha$ coverage interval is $C(1-\alpha) = (0, \frac{\beta}{2}\chi^2_{2k}(1-\alpha))$. With mle $\hat{\beta} = \frac{\sum_{i=1}^n x_i}{nk}$, a sample coverage interval is

$$\hat{C}(1-\alpha) = (0, \frac{\sum_{i=1}^n x_i}{2nk}\chi^2_{2k}(1-\alpha)).$$

Suppose that the true coverage interval is $C(1-\alpha) = (0, \frac{\beta_0}{2}\chi^2_{2k}(1-\alpha))$. The

power function is a function of parameter β as

$$\pi(\beta) = P_\beta(X > \frac{\sum_{i=1}^n X_i}{2nk}\chi^2_{2k}(1-\alpha))$$
$$= P_\beta(\frac{2X/2k\beta}{2\sum_{i=1}^n X_i/2nk\beta_0} > \frac{\beta_0}{2k\beta}\chi^2_{2k}(1-\alpha))$$
$$= P(F(2k,2nk) > \frac{\beta_0}{2k\beta}\chi^2_{2k}(1-\alpha))$$

We list the power and ARL results for this Gamma distribution in Tables 5 and 6.

Table 5. Powers for Gamma distribution $\Gamma(k,\beta)$ (one-sided)

	$\beta = 0.5$	$\beta = 1$	$\beta = 5$	$\beta = 20$
$k = 1$	0.00424	0.05753	0.55253	0.86121
$k = 2$	0.00137	0.05613	0.75445	0.97566
$k = 3$	0.00056	0.05550	0.86526	0.99577
$k = 4$	0.00025	0.05513	0.92660	0.99928
$k = 5$	0.00012	0.05487	0.96034	0.99987
$k = 6$	0.00006	0.05468	0.97873	0.99997
$k = 7$	0.00003	0.05454	0.98867	0.99999
$k = 8$	0.00001	0.05442	0.99400	0.99999
$k = 9$	0.00001	0.05433	0.99684	0.99999
$k = 10$	0.00000	0.05425	0.99834	0.99999
$k = 12$	0.00000	0.05412	0.99955	1.00000
$k = 15$	0.00000	0.05397	0.99993	1.00000
$k = 20$	0.00000	0.05381	0.99999	1.00000

Table 6. ARL for Gamma distribution $\Gamma(k,\beta)$ (one-sided)

	$\beta = 0.5$	$\beta = 1$	$\beta = 5$	$\beta = 20$
$k = 1$	235.69	17.381	1.8098	1.1612
$k = 2$	727.76	17.813	1.3255	1.0249
$k = 3$	1780.6	18.015	1.1557	1.0042
$k = 4$	3904.8	18.138	1.0792	1.0007
$k = 5$	8012.0	18.222	1.0413	1.0001
$k = 6$	15702	18.285	1.0217	1.0000
$k = 7$	29748	18.333	1.0115	1.0000
$k = 8$	54888	18.373	1.0060	1.0000
$k = 9$	99132	18.405	1.0032	1.0000
$k = 10$	175905	18.433	1.0017	1.0000
$k = 12$	530669	18.477	1.0004	1.0000
$k = 15$	2562981	18.526	1.0001	1.0000
$k = 20$	30419787	18.581	1.0000	1.0000

For two sided coverage interval $\frac{\beta}{2}(\chi^2_{2k}(\frac{\alpha}{2}), \chi^2_{2k}(1-\frac{\alpha}{2}))$, its estimate is

$$\hat{C}(1-\alpha) = \frac{\sum_{i=1}^{n} X_i}{2nk}(\chi^2_{2k}(\frac{\alpha}{2}), \chi^2_{2k}(1-\frac{\alpha}{2})).$$

We then see that the power of this coverage interval estimate is

$$\pi(\beta) = 1 - P(\frac{\beta_0}{2k\beta}\chi^2_{2k}(\frac{\alpha}{2}) \leq F(2k, 2nk) \leq \frac{\beta_0}{2k\beta}\chi^2_{2k}(1-\frac{\alpha}{2})).$$

Some of the power and ARL results for this two sided consideration are listed in Tables 7 and 8.

Table 7. Powers for Gamma distribution $\Gamma(k, \beta)$ (two-sided)

	$\beta = 0.5$	$\beta = 1$	$\beta = 5$	$\beta = 20$
$k = 1$	0.05069	0.05582	0.48751	0.83330
$k = 2$	0.08650	0.05489	0.69533	0.96742
$k = 3$	0.12999	0.05455	0.82175	0.99384
$k = 4$	0.17816	0.05437	0.89733	0.99887
$k = 5$	0.22926	0.05427	0.94165	0.99979
$k = 6$	0.28194	0.05420	0.96722	0.99996
$k = 7$	0.33506	0.05415	0.98176	0.99999
$k = 8$	0.38771	0.05411	0.98994	0.99999
$k = 9$	0.43913	0.05408	0.99449	0.99999
$k = 10$	0.48874	0.05406	0.99701	0.99999
$k = 12$	0.58083	0.05402	0.99913	1.00000
$k = 15$	0.69790	0.05398	0.99986	1.00000
$k = 20$	0.83608	0.05395	0.99999	1.00000

Table 8. ARL for Gamma distribution $\Gamma(k, \beta)$ (two-sided)

	$\beta = 0.5$	$\beta = 1$	$\beta = 5$	$\beta = 20$
$k = 1$	19.724	17.913	3.2461	1.4376
$k = 2$	11.559	18.218	2.2077	1.1215
$k = 3$	7.6925	18.330	1.2169	1.0062
$k = 4$	5.6128	18.389	1.1144	1.0011
$k = 5$	4.3618	18.425	1.0620	1.0002
$k = 6$	3.5468	18.449	1.0339	1.0000
$k = 7$	2.9845	18.466	1.0186	1.0000
$k = 8$	2.5792	18.479	1.0102	1.0000
$k = 9$	2.2772	18.489	1.0055	1.0000
$k = 10$	2.0461	18.497	1.0030	1.0000
$k = 12$	1.7217	18.510	1.0009	1.0000
$k = 15$	1.4329	18.522	1.0001	1.0000
$k = 20$	1.1960	18.534	1.0000	1.0000

Let $X_1, ..., X_n$ be a random sample drawn from the exponential distribution with probability density function

$$f(x, \theta) = \frac{1}{\theta} e^{-x/\theta}, x > 0.$$

The distribution function is $F(x) = 1 - e^{-x/\theta}$. Hence, the population quantile function is $F^{-1}(\alpha) = -\theta ln(1-\alpha)$ indicating that a $100(1-\alpha)\%$ population coverage interval is

$$(-\theta ln(1 - \frac{\alpha}{2}), -\theta ln(\frac{\alpha}{2})).$$

An appropriate estimate of θ is $\bar{X}$ and then a sample $100(1-\alpha)\%$ coverage interval is

$$(-\bar{X} ln(1 - \frac{\alpha}{2}), -\bar{X} ln(\frac{\alpha}{2})).$$

Suppose that the parameter for healthy people is θ_0. The type I error probability is deriving as follows:

$$\begin{aligned} P(\text{Type I error}) &= P_{\theta_0}(X_0 \notin (-\bar{X} ln(1 - \frac{\alpha}{2}), -\bar{X} ln(\frac{\alpha}{2}))) \\ &= 1 - P_{\theta_0}(\frac{-\sum_{i=1}^n X_i ln(1-\frac{\alpha}{2})}{n} \leq X_0 \leq \frac{-\sum_{i=1}^n X_i ln(\frac{\alpha}{2})}{n}) \\ &= 1 - P_{\theta_0}(\frac{-ln(1-\frac{\alpha}{2})}{n} \leq \frac{X_0}{\sum_{i=1}^n X_i} \leq \frac{-ln(\frac{\alpha}{2})}{n}) \\ &= 1 - P_{\theta_0}(-ln(1 - \frac{\alpha}{2}) \leq F(2, 2n) \leq -ln(\frac{\alpha}{2})) \end{aligned}$$

where we use the fact that $\frac{X_0}{\sum_{i=1}^n X_i} = \frac{2X_0/\theta_0 2}{2\sum_{i=1}^n X_i/\theta_0 2n} \sim F(2, 2n)$. The probbility of type II error when the true parameter is θ is

$$\beta = P(\text{Type II error}) = P(-\frac{1}{\theta^*} ln(1 - \frac{\alpha}{2}) \leq F(2, 2n) \leq -\frac{1}{\theta^*} ln(\frac{\alpha}{2}))$$

where $\theta^* = \frac{\theta}{\theta_0}$. We consider $(1-\alpha) = 0.95$ coverage interval as example and list the results in Tables 9 and 10.

Table 9. Powers for Exponential distribution $Exp(\theta)$ (two-sided) (Assume $\frac{\theta}{\theta_0} = \theta^*$)

	$n=5$	$n=20$	$n=30$	$n=50$
$\theta^*=0.2$	0.11795	0.11855	0.11867	0.11876
$\theta^*=0.5$	0.05988	0.05118	0.05069	0.05037
$\theta^*=0.8$	0.06916	0.04690	0.04484	0.04328
$\theta^*=1$	0.08803	0.05884	0.05582	0.05345
$\theta^*=1.5$	0.15203	0.11506	0.11080	0.10738
$\theta^*=2$	0.22060	0.18388	0.17954	0.17602
$\theta^*=2.5$	0.28451	0.25090	0.24690	0.24366
$\theta^*=3$	0.34147	0.31161	0.30806	0.30518

Table 10. ARL for Exponential distribution $Exp(\theta)$ (two-sided) (Assume $\frac{\theta}{\theta_0}=\theta^*$)

	$n=5$	$n=20$	$n=30$	$n=50$
$\theta^*=0.2$	8.4777	8.4349	8.4267	8.4201
$\theta^*=0.5$	16.697	19.536	19.724	19.850
$\theta^*=0.8$	14.459	21.320	22.296	23.100
$\theta^*=1$	11.358	16.993	17.913	18.706
$\theta^*=1.5$	6.5775	8.6910	9.0245	9.3119
$\theta^*=2$	4.5329	5.4382	5.5697	5.6809
$\theta^*=2.5$	3.5147	3.9856	4.0501	4.1040
$\theta^*=3$	2.9285	3.2091	3.2461	3.2767

Let's now consider the one sided coverage interval $(0, -\theta ln(\alpha))$ that is estimated by $(0, -\bar{X} ln(\alpha))$. The probability of type II error is

$$\beta = P(\text{Type II error}) = P(0 < F(2, 2n) \leq -\frac{1}{\theta^*} ln(\alpha)).$$

Again, $1-\alpha=0.95$, we list the power and ARL in Tables 11 and 12.

Table 11. Powers for Exponential distribution $Exp(\theta)$ (one-sided) (Assume $\frac{\theta}{\theta_0}=\theta^*$)

	$n=5$	$n=20$	$n=30$	$n=50$
$\theta^*=0.2$	0.00098	0.00001	0.00000	0.00000
$\theta^*=0.5$	0.01947	0.00529	0.00424	0.00318
$\theta^*=0.8$	0.06111	0.03230	0.02934	0.02702
$\theta^*=1$	0.09562	0.06172	0.05753	0.05450
$\theta^*=1.5$	0.18631	0.14902	0.14464	0.14109
$\theta^*=2$	0.26977	0.23588	0.21848	0.22858
$\theta^*=2.5$	0.34157	0.31230	0.30882	0.30600
$\theta^*=3$	0.40235	0.37740	0.37444	0.37204

Table 12. ARL for Exponential distribution $Exp(\theta)$ (one-sided) (Assume $\frac{\theta}{\theta_0} = \theta^*$)

	$n = 5$	$n = 20$	$n = 30$	$n = 50$
$\theta^* = 0.2$	1018.5	71428	188679	500000
$\theta^* = 0.5$	51.336	188.80	235.69	286.80
$\theta^* = 0.8$	16.363	30.954	34.081	37.005
$\theta^* = 1$	10.457	16.201	17.381	18.346
$\theta^* = 1.5$	5.3673	6.7101	6.7136	7.0873
$\theta^* = 2$	3.7068	4.2394	4.5770	4.3748
$\theta^* = 2.5$	2.9276	3.2020	3.2381	3.2679
$\theta^* = 3$	2.4854	2.6497	2.6706	2.6878

Topic 2: p Value of an Outllier Sum in Differential Gene Expression Analysis

Abstract

Outlier sum has been proposed in Tibshirani and Hastie (2007) and Wu (2007) for detection of differential genes in cancer studies where one or several disease groups show unusually high gene expression in a subset of their samples. A new outlier sum is proposed that allows us to develop its asymptotic distribution theory for formulating p value. Since it is a function of some distributional parameters, this p value may be computed parametrically or nonparametrically. We further formulate parametrically this p value when normal distribution for gene variables is assumed. To investigate this p value, we perform a simulation and conduct a real data analysis which indicates that this outlier sum not only allows us to compute p values for genes but is also flexible for treatment of various structures of distribution for gene variables.

Key words: Gene expression analysis; outlier sum; p value.

5. Introduction

Microarray technology by probing thousands of genes simultaneously has been successfully used in medical research to classify different diseases (see this point in, for examples, Agrawal et al. (2002); Alizadeh et al. (200 0) Ohki et al. (2005); Sorlie et al. (2003)). For example, two molecular subtypes of breast cancer (two distinct gene expression patterns), luminal A and basal-like

subtypes, have been reported to have different clinical outcome (see Sorlie et al. (2003)). Another example is diffuse large B-cell lymphoma (DLBCL). Patients with one particular molecular pattern, germinal centre B-like DLBCL, had a significant better overall survival than those with another molecular pattern, activated B-like DLBCL (see Alizadeh et al. (2000)). Furthermore, microarray analysis has been advanced to identify oulier genes which are over-expressed only in a small number of disease samples (see Beer et al. (2002); Tibshirani and Hastie (2007); Tomlins et al. (2005)), such as recurrent chromosomal rearrangements (one type of chromosomal mutation), which is common in lymphoma and leukemia, but rare in other cancers. Standard statistical methods for two-group comparisons (e.g., t-tests) have a limitation to identify these genes to distinguish tumor versus normal samples.

Several statistical approaches have been proposed to address this issue of finding those genes where only a subset of the samples has high expression. Among the proposals, Tomlins et al. (2005) introduced a method called cancer outlier profile analysis (COPA). Latter, Tibshirani and Hastie (2007) introduced a sum of the values in the cancer group, called the outlier sums, and showed that the technique of outlier sums is noticeably better in simulation of p values than the technique of COPA. There is an alternative outlier sums - like statistic proposed by Wu (2007). Basically, these methods of outlier sums pool outlier score which is a standardized score centered at median and scales by median absolute deviation in various ways. A larger outlier score indicates an outlier gene. The outlier sum statistics are very promising in detecting genes where only a subset of their samples have high expression. Unfortunately, without development of distribution theory for the outlier sum statistic, its power (see the simulations in Tibshirani and Hastie (2007)) in gene expression analysis relies on that the number of genes with samples having high expression is known. However, this is usually not true in practice and then there is no natural cut off point to decide the number of influential genes.

We propose the non-standardized outlier sum statistics and develop a technique for computing p values for genes. One interesting result is that this technique will generally produce a cut off point to classify the genes into class

of outlier genes and non-outlier genes. So, this would not require that there is only one outlier gene. The studies of gene expression detection such as the t test, Tibshirani and Hastie (2007) and Wu (2007) all assume that the underlying distributions for all genes are normal distributions. Hence, under this distribution, we further derive a simpler formula for p values and perform simulations evaluate its ability in detection of outlier genes. A formula developed in this paper makes the study of p values in parameteric of other distributions and nonparametric techniques is straight forward, however, we would not go further for this.

6. General Formulation for Outlier Means

Suppose that there are m genes to be cocerned and for each gene there are two groups of subjects, one normal or healthy group and one cancer (disease) group. We assume that there are available n_1 and n_2 expression variables respectively for two groups forming as follows:

$$\begin{array}{lcc} & \text{Normal group} & \text{Cancer group} \\ \text{Gene 1} & X_{11},...,X_{1n_1} & Y_{11},...,Y_{1n_2} \\ \text{Gene 2} & X_{21},...,X_{2n_1} & Y_{21},...,Y_{2n_2} \\ \vdots & \vdots & \vdots \\ \text{Gene m} & X_{m1},...,X_{mn_1} & Y_{m1},...,Y_{mn_2} \end{array} \tag{6.1}$$

The outlier sums for gene expression in literature actually implicitly defined three parameters:

H_1 : Centering parameter for measuring distance of observations in Y group

H_2 : Threshold for identifying observations from Y group as outliers

H_3 : Scale parameter for standardizing an outlier sum

Let H_{1j}, H_{2j}, H_{3j} represent, respectively the above three parameters for gene j and we assume that there are appropriate estimators $\hat{H}_{1j}, \hat{H}_{2j}, \hat{H}_{3j}$, based on variables in gene j, available for estimating these parameters.

The outlier sum statistic for gene j defined by Tibshirani and Hastie (2007) and Wu (2007) may be represented in a general form as

$$W_j = \sum_{i=1}^{n_2} \frac{Y_{ji} - \hat{H}_{1j}}{\hat{H}_{3j}} I(Y_{ji} > \hat{H}_{2j}), \tag{6.2}$$

where $\hat{H}_{1j}, \hat{H}_{2j}$ and $\hat{H}_{3j}$ are estimates of H_{1j}, H_{2j} and H_{3j} respectively.

Let F_{xj} and F_{yj}, respectively, be the distribution functions that $\{X_{ji}, i = 1, ..., n_1\}$ and $\{Y_{ji}, i = 1, ..., n_2\}$ are drawn. Let's denote

$$\hat{F}_{xj}^{-1}(\alpha) : \alpha\text{th percentile of the set } \{X_{ji}, i = 1, ..., n_1\}$$
$$\hat{L}_{j}^{-1}(\alpha) : \alpha\text{th percentile of the set } \{X_{ji}, i = 1, ..., n_1, Y_{ji}, i = 1, ..., n_2\}$$
$$\text{med}_{xj} = \hat{F}_{xj}^{-1}(0.5), \text{med}_{yj} = \hat{F}_{yj}^{-1}(0.5), \text{med}_{j} = \hat{L}_{j}^{-1}(0.5)$$
$$IQR_{xj} = \hat{F}_{xj}^{-1}(0.75) - \hat{F}_{xj}^{-1}(0.25), IQR_{j} = \hat{L}_{j}^{-1}(0.75) - \hat{L}_{j}^{-1}(0.25),$$
$$\text{mad}_{xj} = 1.4826 \times \text{median}\{|Y_{ji} - \text{med}_{xj}|, i = 1, ..., n_2\}$$

where the constant 1.4826 is chosen such that mad_{xj} is approximately equal to the normal standard error.

For comparison of the two approaches on outlier sums by Tibshirani and Hastie (2007) and Wu (2007), we use a table to express their formulations of outlier sums. This expression allows us to generate alternative outlier sums when thresholds $\hat{H}_{1j}, \hat{H}_{2j}$ and $\hat{H}_{3j}$ are chosen in different ways that could be in consideration of robustness or efficiency.

Table 14. Comparison of parameter estimates for outlier sums method and outlier robust t method

Parameter estimate	Tibshirani and Hastie	Wu
$\hat{H}_{1j}$	med_j	med_{xj}
$\hat{H}_{2j}$	$\hat{L}_j^{-1}(0.75) + IQR_j$	$\hat{F}_{xj}^{-1}(0.75) + IQR_{xj}$
$\hat{H}_{3j}$	mad_{xj}	$\text{median}\{\lvert X_{ji} - \text{med}_{xj}\rvert_{i=1}^{n_1}, \lvert Y_{ji} - \text{med}_{yj}\rvert_{i=1}^{n_2}\}$

When gene expression values $x_{ji}, i = 1, ..., n_1, y_{ji}, i = 1, ..., n_2$ are available, we can evaluate statistic values w_j for the outlier sum statistics W_j of (6.2). The technique applied in Tibshirani and Hastie (2007) of gene expression analysis computes the p values as

$$p_{jw} = \frac{1}{m} \sum_{j' \neq j} I(w_{j'} \geq w_j), j = 1, ..., m. \tag{6.3}$$

The genes with smaller p values are suspected to be significant genes.

It is desired to evaluate p values with probability sense. Suppose that we have a statistic $t(Z)$ where Z is a random sample from a distribution involving parameter θ and we consider the null hypothesis $H_0 : \theta = \theta_0$. The classical significance test defines the p value as

$$p_t = P_{\theta_0}\{t(Z) \text{ at least as extreme as the observed } t(z)\}, \tag{6.4}$$

where z is the realization of the random sample Z. Extending this concept, the proposal of p value for gene expression based on outlier sums is appropriate in the form as

$$p_j^* = P_{F_{xj}}\{W_j \geq w_j\}, j = 1, ..., m, \tag{6.5}$$

where statistic W_j involves distributions F_{xj} and F_{yj} since it is function of $\{X_{ji}\}$ and $\{Y_{ji}\}$ but we consider that $F_{xj} = F_{yj}$ in (6.5).

We consider a non-centered and non-scaled outlier sum statistic in the following and use it to introduce a test statistic that does involve centering and scaling estimates.

Definition 6.1. The outlier sum statistic for jth gene is

$$\tilde{\Pi}_j = \sum_{i=1}^{n_2} Y_{ji} I(Y_{ji} > \hat{H}_j). \tag{6.6}$$

The aim in this paper is to develop p values for outlier sum statistics $\tilde{\Pi}_j, j = 1, ..., m$.

7. Formulation of p Value with Normal Samples

From now on, for simplicity, we drove the index j. The threshold suggested by Wu (2007) is

$$\hat{H}_a = \hat{F}_x^{-1}(0.75) + IQR_x = 2\hat{F}_x^{-1}(0.75) - \hat{F}_x^{-1}(0.25).$$

For latter comparison, we suggested a flexible type of threshold as

$$\hat{H}_b = \hat{F}_x^{-1}(0.5) + 1.5kIQR_x.$$

We now further denote the outlier mean Π by Π_a when its threshold is $\hat{H} = \hat{H}_a$ and it by Π_b when $\hat{H} = \hat{H}_b$.

We have notes on the design of threshold $\hat{H}_b$:

(a) Consider that the underlying distributions F_x is normal. We then see that $\hat{H}_a$ and $\hat{H}_b$ when $k = 1$ are both estimates of $\mu_x + 3\sigma_x z_{0.75}$. Hence, $\hat{H}_b$ when $k = 1$ is asymptotically equlivalent to $\hat{H}_a$.

(b) Small k will make the outlier sum able to detect any positive outliers in second group. The larger the outliers the more the efficiency will be. However, it could happen that there are many genes to be identified as outlier genes since their p values all indicate significant different.

(c) Larger k can only detect larger shift in distribution and it will probably not be able to detect smaller shift in distribution.

(d) We latter will see that when $k = 1$ the p values p_a and p_b are identical.

We now assume that $\{X_i\}$ and $\{Y_i\}$ are two random sample, respectively, from normal distributions $N(\mu_x, \sigma_x^2)$ and $N(\mu_y, \sigma_y^2)$. With denoted ϕ as the probability density function of the standard normal distribution $N(0, 1)$, we further let $\phi_{\mu,\sigma}$ be the probability density function of the normal distribution $N(\mu, \sigma^2)$.

With the normality assumptions, $F_x^{-1}(\alpha) = \mu_x + z_\alpha \sigma_x$ indicates that $F_x^{-1}(0.5) + 1.5k(F_x^{-1}(0.75) - F_x^{-1}(0.25)) = \mu_x + 3kz_{0.75}\sigma_x$. Hence, the outlier sum may be reformulated as

$$\tilde{\Pi} = \sum_{i=1}^{n_2} Y_i I(Y_i > \hat{\mu}_x + 3kz_{0.75}\hat{\sigma}_x)$$

that requires only estimators $\hat{\mu}_x$ and $\hat{\sigma}_x$. Furthermore, the p value is evaluated under that H_0 is assumed to be true. Hence, we may let $\mu_x = \mu_y, \sigma_x = \sigma_y$ and, with careful checking, we may see that some elements for evaluating p value in

Section 4 are as follows:

$$\tilde{\pi} = \sum_{i=1}^{n_2} y_i I(y_i > \hat{\mu}_x + 3kz_{0.75}\hat{\sigma}_x),$$

$$\beta = \int_{3kz_{0.75}}^{\infty} \phi(z)dz, \text{ a known constant},$$

$$\mu_\pi = \mu_x + \frac{\sigma_x}{\beta}\int_{3kz_{0.75}}^{\infty} z\phi(z)dz,$$

$$b_1 = \frac{1}{\beta}3kz_{0.75}\sigma_x\phi(3kz_{0.75})\sqrt{h}\phi^{-1}(0),$$

$$b_2 = b_3 = 1.5k\frac{b_1}{\phi^{-1}(0)}\phi^{-1}(z_{0.75}),$$

$$v = \frac{\sigma_x^2}{\beta^2}[\int_{3kz_{0.75}}^{\infty} z^2\phi(z)dz - (\int_{3kz_{0.75}}^{\infty} z\phi(z)dz)^2],$$

where b_1, b_2, b_3 and v are to formulate $\sigma_\pi^2 = \sigma_\pi^2(b_1, b_2, b_3, v)$ where

$$\begin{aligned}\sigma_\pi^2 &= \sigma_\pi^2(b_1, b_2, b_3, v)\\ &= 0.25 \times 0.75[(0.5b_1 + 0.25b_2 - 0.75b_3)^2 + (0.5b_1 + 0.25b_2 + 0.25b_3)^2\\ &+ (-0.5b_1 + 0.25b_2 + 0.25b_3)^2 + (-0.5b_1 - 0.75b_2 + 0.25b_3)^2] + v\end{aligned}$$

From the formulations stated earlier, we need only to specify estimators of h, μ_x and σ_x.

Theorem 7.1 Suppose that $\{X_i\}$ and $\{Y_i\}$ are, respectively, random samples from distributions $N(\mu_x, \sigma_x^2)$ and $N(\mu_y, \sigma_y^2)$. Then, under $H_0 : \mu_x = \mu_y, \sigma_x^2 = \sigma_y$,

$$W = W(X_i, Y_i) = \sqrt{n_2}(\frac{\tilde{\Pi} - \beta n_2\mu_\pi}{\sqrt{\beta n_2}\sigma_\pi}) \tag{7.1}$$

converges asymptotically to the standard normal distribution.

We then apply an estimator of W of (7.1) as the test statistic

Definition 7.2. Suppose that we have appropriate estimators of β, μ_π and σ_π. Then we define the test statistic as

$$\tilde{W} = \tilde{W}(X_i, Y_i) = \sqrt{n_2}(\frac{\tilde{\Pi} - \hat{\beta} n_2\hat{\mu}_\pi}{\sqrt{\hat{\beta} n_2}\hat{\sigma}_\pi}). \tag{7.2}$$

Definition 7.3. Suppose that the outlier mean Π_j has the asymptotic property of (6.2) and there are $\hat{\beta}_j, \hat{\mu}_{j\pi}$ and $\hat{\sigma}_{j\pi}$, estimates, respectively, of $\beta_j, \mu_{j\pi}$ and $\sigma_{j\pi}$ based on observations x_{ji}'s. We define the p value for gene j as

$$p_j = \int_{\sqrt{n_2}(\frac{\tilde{\pi}_j - \hat{\beta}_j n_2 \hat{\mu}_{j\pi}}{\sqrt{\hat{\beta}_j n_2}\hat{\sigma}_{j\pi}})}^{\infty} \phi(z)dz, j = 1, ..., m. \tag{7.3}$$

We have two notes for the specified p values:
(a) The estimates $\hat{\beta}_j, \hat{\mu}_{j\pi}$ and $\hat{\sigma}_{j\pi}$ are designed to be computed from the data x_{ji}'s since p values try to see how significant the observation $\tilde{\pi}_j$'s it is when y_{ji} are drawn from the same distribution of x_{ji}'s.
(b) Suppose that p_j's for all j are available. The genes with indexes j's such that their p's are relatively smaller are then suspected to be influential and those with relatively larger p_j's are not influential. This resolve the difficulty of ordinal p values proposed in the literature for outlier sums statistics for not been able to determine a finite set of influential genes when it is not known the true number of influential genes.

Let $\hat{h} = \frac{n_2}{n_1}, \hat{\mu}_x = \bar{x} = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i, \hat{\sigma}_x^2 = s_x^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(x_i - \bar{x})^2$. Some elements for computing the observation of the following test statistic

$$\tilde{W}(X_i, Y_i) = \sqrt{n_2}(\frac{\tilde{\Pi} - \beta n_2 \hat{\mu}_\pi}{\sqrt{\beta n_2}\hat{\sigma}_\pi})$$

are the followings:

$$\begin{aligned}
&\tilde{\pi} = \sum_{i=1}^{n_2} y_i I(y_i > \bar{x} + 3kz_{0.75}s_x) \\
&\beta = \int_{3kz_{0.75}}^{\infty} \phi(z)dz, \text{ a known constant} \\
&\hat{\mu}_\pi = \bar{x} + \frac{s_x}{\beta}\int_{3kz_{0.75}}^{\infty} z\phi(z)dz \\
&\hat{b}_1 = \frac{1}{\beta}3kz_{0.75}s_x\phi(3kz_{0.75})\sqrt{\hat{h}}\phi^{-1}(0) \\
&\hat{b}_2 = \hat{b}_3 = 1.5k\frac{\hat{b}_1}{\phi^{-1}(0)}\phi^{-1}(z_{0.75}) \\
&\hat{v} = \frac{s_x^2}{\beta^2}[\int_{3kz_{0.75}}^{\infty} z^2\phi(z)dz - (\int_{3kz_{0.75}}^{\infty} z\phi(z)dz)^2].
\end{aligned} \tag{7.4}$$

Then the asymptotic variance σ_π^2 is estimated as

$$\hat{\sigma}_\pi^2 = \sigma_\pi^2(\hat{b}_1, \hat{b}_2, \hat{b}_3, \hat{v}) \tag{7.5}$$

and then the p value of (6.4) is

$$p = \int_{\sqrt{n_2}(\frac{\tilde{\pi} - \beta n_2 \hat{\mu}_\pi}{\sqrt{\beta n_2}\hat{\sigma}_\pi})}^{\infty} \phi(z)dz. \tag{7.6}$$

The p value of (7.6) uses only $\bar{x}$ and s_x to estimate μ_x and σ_x for formulating $\hat{\mu}_\pi$ and $\hat{\sigma}_\pi$. The computation of p value under normality assumption is very simple. If it is the situation that G_x and G_y are known but not normal, this procedure of establishing p value may be analogously derived.

8. Simulation and Data Analysis

It is desired to evaluate the ability of outlier sum in detecting significant genes through the p values of genes. We restrict this evaluation for that the underlying distributions are normal that are generally assumed in the approaches of Tibshirani and Hastie (2007) and Wu (2007). Under the normal assumption, the outlier sum statistic may be formulated as

$$\tilde{\pi}_b = \sum_{i=1}^{n_2} Y_i I(Y_i > \bar{X} + 3kz_{0.75}S_x) \tag{8.1}$$

where $\bar{X}$ and S_x are, respectively, sample mean and sample standard deviation based on sample of normal group people. This outlier sum is equivalent to the proposals of Wu (2007) when $k = 1$. It is then interesting to study the choice of constant k for detecting significant genes through simulation and data analysis.

We conduct two simulations. First, the classical t test has been criticized that when there are occassionally hundreds of influential genes if 10 thounsands genes are investigated. Hence, we generate $n_1 = 20$ and $n_2 = 20$ observations from $N(0, 1)$ and conduct 1 million replications of this data generation to compute p values of (7.6). Setting significance level $\alpha = 0.001, 0.01, 0.05$ and constant $k = 1, 2, 3$, we compute the numbers of p values smaller than the

corresponding specified significance level α. The results are displayed in Table 15.

Table 15. Numbers in 1 millions replications with p values smaller than α

α	$k=1$	$k=2$	$k=3$
0.05	57808	460	5
0.01	25231	86	2
0.001	9632	23	1

We have two conclusions drawn from the results in Table 1:
(a) Consider that $k=1$. If $\alpha=0.05$, there are more than 50 thousands genes to be claimed influential. So, if there are totally 10 thounsands genes, then there are about 500 or more genes to be identified as influential. Similarly, $\alpha=0.01$ and $\alpha=0.001$ indicate to have, respectively, 200 and 90 or more genes to be identified as influential. This shows that outlier sum of $k=1$ which is equivalent to Wu (2007) is still struggled in having too many influential genes. (b) Consider that $k=2$. The results show that when the gene number is about 10 thousands, there will be very small numbers of influential genes to be identified. On the other hand, $k=3$ will be almost none to be identified as influential gene. Hence, based on this simulation, $k=2$ or 3 is an appropriate constant to contruct the outlier sum.

We first consider a simulation to evaluate the efficiency of the approach of p value for differential sum in detcting outlier genes. Let (s,h) be a fixed index for gene data generation. We generate $n_1=20$ and $n_2=20$ observations from $N(0,1)$. However, we add h units for s of the samples in the second group of n_2 observations. We then compute the p value of (7.6).

For the next simulation, we consider that there are influential genes and see the efficiency of the approach of p value for detection of influential genes. Again, we generate $n_1=20$ and $n_2=20$ observations from $N(0,1)$. However, we add h units for s of the samples in the second group of n_2 observations. This process is repeated 10 thousands times and we compute the averaging p value. For several values of s and h, we perform this simulation and display the simulation results of averaged p values in Tables 16 and 17.

Table 16. Average p values of outlier sum

(s,h)	$k=1$	$k=2$	$k=3$
$(0,0)$	0.4726	0.4972	0.5
$(2,2)$	0.2441	0.4642	0.4985
$(2,4)$	0.0198	0.2018	0.4475
$(2,6)$	0.00075	0.0162	0.2103
$(2,8)$	$1.75E-05$	0.00056	0.0313
$(4,2)$	0.1271	0.4354	0.4973
$(4,4)$	0.00038	0.1052	0.4160
$(4,6)$	$2.88E-08$	0.0013	0.1293
$(4,8)$	$2.89E-13$	$5.76E-07$	0.0070
$(4,10)$	$6.19E-18$	$3.72E-12$	$2.77E-05$
$(6,2)$	0.0694	0.4145	0.4960
$(6,4)$	$1.74E-05$	0.0672	0.3891
$(6,6)$	$5.37E-13$	0.00027	0.0948
$(6,8)$	$5.34E-24$	$1.95E-11$	0.0029
$(6,10)$	$2.03E-36$	$3.25E-21$	$3.74E-06$

Table 17. Average p values of outlier sum

(s,h)	$k=4$	$k=5$	$k=6$
$(0,0)$	0.5	0.5	0.5
$(2,2)$	0.4999	0.5	0.5
$(2,4)$	0.4958	0.4997	0.4998
$(2,6)$	0.4280	0.4890	0.4986
$(2,8)$	0.2170	0.4084	0.4798
$(4,2)$	0.4999	0.5	0.5
$(4,4)$	0.4911	0.4992	0.4998
$(4,6)$	0.3901	0.4821	0.4978
$(4,8)$	0.1463	0.3712	0.4695
$(4,10)$	0.0179	0.1657	0.3518
$(6,2)$	0.4999	0.5	0.5
$(6,4)$	0.4886	0.4990	0.4999
$(6,6)$	0.3633	0.4766	0.4970
$(6,8)$	0.1152	0.3482	0.4614
$(6,10)$	0.0106	0.1322	0.3291

We have several conclusions drawn from Tables 2 and 3:

(a) Consider the case that $(s,h)=(0,0)$. It is nice that the outlier sums in all cases of k all have average p values more than 0.4 that indicates not statistical significant for practically non-influential genes.

(b) Consider that $k=1$ and $(s,h)\neq(0,0)$. Besides few cases, the average p values are small enough that would efficiently classify these genes as influential

genes. Is $k = 1$ appropriate for constructing outlier sum? We should remind that $k = 1$ may occassionally generate too many influential genes as we have seen in Table 15. So, it is good in detecting influential genes but would produce non negligible type I error.

(c) Consider that $k = 2$. The simulation results for $(s, h) = (0, 0)$ in Table 16 shows that it would produce only negligible type I error. For $(s, h) \neq (0, 0)$, when h is far enough away from 0, the outlier sum performs very well. From consideration of balanced two errors, $k = 2$ seems to be an appropriate choice of outlier sum.

(d) From the table results that $k > 2$, it seems to be not efficient to detect influential genes in all situations of $(s, h) \neq (0, 0)$.

We now consider an application of p value of outlier sum on a real gene data. The breast cancer microarray data reported by Huang et al. (2003) contained the expression levels of 12625 genes from 37 (or 52) breast tumor samples. Each sample had a binary outcome describing the status of lymph node involvement in breast cancer (breast cancer recurrence). Among them, 19 samples had no positive nodes. (Or 34 samples had no cancer recurrence and 18 samples had breast cancer recurrence). The gene expressions, obtained from the Affymetrix human U95a chip. We pre-processed the data using RMA (Irizarry et al. (2003)).

We first compute the p values of (7.6) for various values of k and we display the numbers $no_{<0.001}$ of genes that are classified to be significant for that theirs p values are less than 0.001 in the following table.

Table 18. Numbers of genes with p values smaller than 0.001

	$no_{<0.001}$		$no_{<0.001}$
$k = 1$	5583	$k = 4$	35
$k = 1.5$	2407	$k = 5$	8
$k = 2$	922	$k = 6$	5
$k = 3$	158		

We have several comments drawn from the results in Table 18:

(a) We have seen that $\hat{H}_a$ is the proposal of Wu (2007) and $\hat{H}_b$ with $k = 1$ is asymptotically equivalent to $\hat{H}_a$ when the underlying distribution is assumed

to be normal. The number of siginificant genes when $k = 1$ for $\hat{H}_b$ is 5583. This huge number shows that this gene data is definitely not appropriate to be analyzed by the outlier sum proposals been introduced. The other cases with $k \leq 3$ the numbers of genes claimed to be significant are still too big for further investigation.

(b) When k is as large as 4 the number of siginificant genes is down to 35 and it further goes down to 8 when $k = 5$. This shows that gene data may need outlier sum of more extreme threshold to simplify the pothetial group of genes for further study.

In the following table, we select the cases $k = 5$ and 6 and list their corresponding gene numbers that are with significant p values and the outlier sum values for reference.

Table 19. Gene numbers with their outlier sums associated with p value

Gene number	OS	Gene number	OS
$k = 5$		$k = 6$	
4029	27.88125	4029	27.88125
4028	31.40937	4028	31.40937
10210	16.62765	10210	16.62765
3758	7.615114	3758	7.615114
8972	6.014273	8972	6.014273
10987	5.93685		
10019	10.82669		
198	10.14491		

Detection of significant genes through the p values of outlier sum solves the difficulty of classical outlier sum technique that is not not able to detect significant genes when the number of them is not known. But how to decide constant k for the outlier sum of (8.1)? We propose to list the numbers of significant genes for various values of k and select k for that has a moderate small group of significant genes.

References

2004 *Guide to the Expression of Uncertainty in Measurement* Supplement 1 *Numerical Methods for the Propagation of Distributions* Draft of JCGM document. p. 38.

Chen, L.-A., Huang, J.-Y. and Chen, H.-C. (2007). Parametric coverage interval. *Metrologia*. 44, L7-L9.

Agrawal, D., Chen, T., Irby, R., et al. (2002). Osteopontin identified as lead marker of colon cancer progression, using pooled sample expression profiling. *J. Natl. Cancer Inst.* 94, 513-521.

Alizadeh, A. A., Eisen, M. B., Davis, R. E., et al. (2000). Distint types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403, 503-511.

Beer, D. G., Kardia, S. L., Huang, C. C., et al. (2002). Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat. Med.*, 8, 816-824.

Chen, L.-A. and Chiang, Y.-C. (1996). Symmetric quantiles and trimmed means for location and linear regression model. *Journal of Nonparametric Statistics*. 7, 171-185.

Huang, E., Cheng, S. H., Dressman, H., et al. (2003). Gene expression predictors of breast cancer outcomes. *Lancet*, 361, 1590-1596.

Irizarry, R., Hobbs, B., Collin, F., Beazer-Barclay, Y., Antonellis, K., Scherf, U. and Speed, T. (2003). Exploration, normalization, and summarizes of high density oligonucleotide array probe level data. *Biostatistics*, 2, 249-64.

Ohki, R., Yamamoto, K., Ueno, S., et al. (2005). Gene expression profiling of human atrial myocardium with atrial fibrillation by DNA microarray analysis. *Int. J. Cardiol.* 102, 233-238.

Sorlie, T., Tibshirani, R., Parker, J., eta l. (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl. Acad. Sci. U.S.A.*, 100, 8418-8423.

Ruppert, D. and Carroll, R. J. (1980). Trimmed least squares estimation in the linear model. *Journal of the American Statistical Association*. 75, 828-838.

Tibshirani, R. and Hastie, T. (2007). Outlier sums differential gene expression analysis. *Biostatistics*, 8, 2-8.

Tomlins, S. A., Rhodes, D. R., Perner, S., eta l. (2005). Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science*, 310, 644-648.

Wu, B. (2007). Cancer outlier differential gene expression detection. *Biostatistics*,