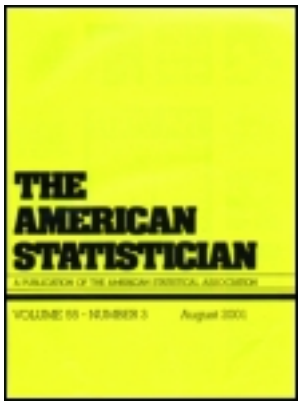

The American Statistician
Publication details, including instructions for authors and subscription information:
http://amstat.tandfonline.com/loi/utas20

The Exact Likelihood Ratio Test for Equality of Two Normal Populations

Lingyun Zhang [a] , Xinzhong Xu [b] & Gemai Chen [c]

[a] Division of Science and Technology, Beijing Normal University and Hong Kong Baptist University United International College (UIC), Zhuhai, P.R. China

[b] Department of Mathematics, Beijing Institute of Technology, Beijing, P. R. China

[c] Department of Mathematics and Statistics, University of Calgary, Calgary, Alberta, T2N 1N4, Canada

Accepted author version posted online: 05 Jul 2012.Published online: 15 Oct 2012.

To cite this article: Lingyun Zhang , Xinzhong Xu & Gemai Chen (2012): The Exact Likelihood Ratio Test for Equality of Two Normal Populations, The American Statistician, 66:3, 180-184

To link to this article: http://dx.doi.org/10.1080/00031305.2012.707083

The Exact Likelihood Ratio Test for Equality of Two Normal Populations

Lingyun ZHANG, Xinzhong XU, and Gemai CHEN

Testing the equality of two independent normal populations is a perfect case of the two-sample problem, yet it is not treated in the main text of any textbook or handbook. In this article, we derive the exact distribution of the likelihood ratio test and implement this test with an R function. This article has supplementary materials online.

KEY WORDS: Dirichlet distribution; Two-sample problem; Testing hypothesis.

1. INTRODUCTION

Testing the equality of two normal populations is of practical importance. For example, a new design/treatment/approach may result in changes in the means, in the variabilities, or in both. Efforts to address this problem were made as early as the 1920s by Pearson and Romanovsky (Plackett 1946). Let $\{x_1, \ldots, x_n\}$ and $\{y_1, \ldots, y_m\}$ be two independent random samples from normal populations $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, respectively. We want to test

$$H_0 : \mu_1 = \mu_2 \text{ and } \sigma_1^2 = \sigma_2^2 \quad \text{versus} \quad H_1 : \mu_1 \neq \mu_2 \text{ or } \sigma_1^2 \neq \sigma_2^2.$$

Pearson and Neyman (1930) considered the likelihood ratio test, with the test statistic given by

$$\lambda_{n,m} = \frac{[\sum_{i=1}^n (x_i - \bar{x})^2/n]^{n/2}[\sum_{j=1}^m (y_j - \bar{y})^2/m]^{m/2}}{\{[\sum_{i=1}^n (x_i - u)^2 + \sum_{j=1}^m (y_j - u)^2]/(n+m)\}^{(n+m)/2}},$$

where $\bar{x}$, $\bar{y}$, and u are the sample means of the x sample, the y sample, and the combined sample, respectively. Pearson and Neyman (1930) showed that under H_0, the limiting distribution of $\lambda_{n,m}$ (as n and m tend to infinity) is the uniform distribution $U(0, 1)$, and they proposed to approximate the exact distribution of $\lambda_{n,m}$ for finite n and m by a beta distribution with the density function

$$f(\lambda) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} \lambda^{p-1}(1-\lambda)^{q-1},$$

where p and q are determined by matching the first two moments

$$\begin{cases} \dfrac{p}{p+q} & = E(\lambda_{n,m}), \\ \dfrac{pq}{(p+q+1)(p+q)^2} & = V(\lambda_{n,m}), \end{cases}$$

which leads to

$$p = -\frac{a}{b}(a^2 - a + b) \quad \text{and} \quad q = \frac{a-1}{b}(a^2 - a + b),$$

where

$$a = E(\lambda_{n,m}) = \left(\frac{(n+m)^{n+m}}{n^n m^m}\right)^{\frac{1}{2}} \frac{\Gamma\left(\frac{2n-1}{2}\right)\Gamma\left(\frac{2m-1}{2}\right)}{\Gamma\left(\frac{2(n+m)-1}{2}\right)} \times \frac{\Gamma\left(\frac{n+m-1}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)\Gamma\left(\frac{m-1}{2}\right)},$$

$$b = V(\lambda_{n,m}) = \frac{(n+m)^{n+m}}{n^n m^m} \frac{\Gamma\left(\frac{3n-1}{2}\right)\Gamma\left(\frac{3m-1}{2}\right)}{\Gamma\left(\frac{3(n+m)-1}{2}\right)} \times \frac{\Gamma\left(\frac{n+m-1}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)\Gamma\left(\frac{m-1}{2}\right)} - a^2.$$

Sukhatme (1935) linked the distribution of $(\lambda_{n,m})^{2/(n+m)}$ to the incomplete beta function. Nagar and Gupta (2004) tabulated the distribution of $\lambda_{n,m}$ for limited n and m as part of the k-sample extension to the current two-sample problem. Although easy to state and naturally a case of the "two-sample" problems in today's terminology, there are hardly any textbooks that treat testing H_0 versus H_1 in the main text, and to the best of our knowledge, there is no easy way to carry out the above test using the exact distribution of $\lambda_{n,m}$. Perng and Littell (1976) considered testing the same H_0 versus H_1, but they used Fisher's method to combine the t-test for equality of the two means and the F-test for equality of the two variances. In this article, we derive the exact distribution of $\lambda_{n,m}$ and offer a simple-to-use R function to make testing H_0 versus H_1 easy to do (R Development Core Team 2011).

Lingyun Zhang is Associate Professor, Division of Science and Technology, Beijing Normal University and Hong Kong Baptist University United International College (UIC), Zhuhai, P.R. China (E-mail: *lyzhang@uic.edu.hk*). Xinzhong Xu is Professor, Department of Mathematics, Beijing Institute of Technology, Beijing, P. R. China (E-mail: *xuxz@bit.edu.cn*). Gemai Chen is Professor, Department of Mathematics and Statistics, University of Calgary, Calgary, Alberta T2N 1N4, Canada (E-mail: *gchen@math.ucalgary.ca*), and School of Statistics and Mathematics, Yunnan University of Finance and Economics, Kunming, Yunnan, P. R. China . The authors thank the editor, the former editor Professor Stufken, and the associate editor for their many helpful comments and suggestions that have helped improve the article. They also thank Dr. Yongchao Ge for sending the dataset that was used in Section 4. The first author's research was supported by the UIC research grant R201206. The second author's research was supported by Chinese National Natural Science Fund 11071015, and the third author's research was supported by the Natural Sciences and Engineering Research Council of Canada.

2. THE EXACT NULL DISTRIBUTION OF $\lambda_{n,m}$

From the general relationship SS Total = SS Error + SS Treatments (SS for sum of squares), we have

$$\sum_{i=1}^{n}(x_i-u)^2+\sum_{j=1}^{m}(y_j-u)^2=\sum_{i=1}^{n}(x_i-\bar{x})^2+\sum_{j=1}^{m}(y_j-\bar{y})^2 + n(\bar{x}-u)^2+m(\bar{y}-u)^2.$$

Under H_0, let $\mu_1=\mu_2=\mu$ and $\sigma_1^2=\sigma_2^2=\sigma^2$. Then, the likelihood ratio test statistic $\lambda_{n,m}$ can be rewritten as

$$\lambda_{n,m}=\frac{(U_1/n)^{n/2}(U_2/m)^{m/2}}{[(U_1+U_2+U_3)/(n+m)]^{(n+m)/2}},$$

where

$$U_1=\sum_{i=1}^{n}(x_i-\bar{x})^2/\sigma^2,\quad U_2=\sum_{j=1}^{m}(y_j-\bar{y})^2/\sigma^2,$$

and

$$U_3=n(\bar{x}-u)^2/\sigma^2+m(\bar{y}-u)^2/\sigma^2.$$

Define

$$W_1=\frac{U_1}{U_1+U_2+U_3},\quad W_2=\frac{U_2}{U_1+U_2+U_3},$$

and

$$W_3=U_1+U_2+U_3.$$

We can further write $\lambda_{n,m}$ as

$$\lambda_{n,m}=\frac{(n+m)^{(n+m)/2}}{n^{n/2}m^{m/2}}W_1^{n/2}W_2^{m/2}.$$

Note that $U_1\sim\chi^2_{n-1}$, $U_2\sim\chi^2_{m-1}$ and

$$\begin{aligned}U_3&=n(\bar{x}-u)^2/\sigma^2+m(\bar{y}-u)^2/\sigma^2\\&=n\left(\bar{x}-\frac{n\bar{x}+m\bar{y}}{n+m}\right)^2\Big/\sigma^2+m\left(\bar{y}-\frac{n\bar{x}+m\bar{y}}{n+m}\right)^2\Big/\sigma^2\\&=\frac{nm(\bar{x}-\bar{y})^2}{(n+m)\sigma^2}\\&=\left(\frac{(\bar{x}-\mu)-(\bar{y}-\mu)}{\sqrt{\sigma^2/n+\sigma^2/m}}\right)^2\sim\chi^2_1,\end{aligned}$$

and U_1, U_2, and U_3 are mutually independent. From

$$(U_1,U_2,U_3)\sim\frac{u_1^{(n-1)/2-1}}{\Gamma\left(\frac{n-1}{2}\right)2^{(n-1)/2}}e^{-u_1/2}\frac{u_2^{(m-1)/2-1}}{\Gamma\left(\frac{m-1}{2}\right)2^{(m-1)/2}}e^{-u_2/2}\times\frac{u_3^{1/2-1}}{\Gamma\left(\frac{1}{2}\right)2^{1/2}}e^{-u_3/2},$$

we have

$$(W_1,W_2,W_3)\sim\frac{\Gamma\left(\frac{n+m-1}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)\Gamma\left(\frac{m-1}{2}\right)\Gamma\left(\frac{1}{2}\right)}w_1^{(n-1)/2-1}w_2^{(m-1)/2-1}\times(1-w_1-w_2)^{-1/2}\times\frac{w_3^{(n+m-1)/2-1}}{\Gamma\left(\frac{n+m-1}{2}\right)2^{(n+m-1)/2}}e^{-w_3/2},\quad w_1>0,\,w_2>0,\,w_3>0,\,w_1+w_2<1.$$

Therefore, (W_1,W_2) and W_3 are independent, and (W_1,W_2) is distributed as the two-dimensional Dirichlet distribution, with parameters $(n-1)/2$, $(m-1)/2$, and $1/2$. Together, for $\lambda\in(0,1)$, the exact distribution of $\lambda_{n,m}$ is given by

$$\begin{aligned}&P(\lambda_{n,m}\le\lambda)\\&=1-P(\lambda_{n,m}>\lambda)\\&=1-C\int\int_D\frac{w_1^{(n-1)/2-1}w_2^{(m-1)/2-1}}{\sqrt{1-w_1-w_2}}dw_1dw_2\\&=1-C\int_{r_1}^{r_2}w_1^{(n-3)/2}\left(\int_{\frac{\lambda^{2/m}n^{n/m}m}{(n+m)^{(n+m)/m}w_1^{n/m}}}^{1-w_1}\frac{w_2^{(m-3)/2}}{\sqrt{1-w_1-w_2}}dw_2\right)\times dw_1,\end{aligned}$$

where

$$C=\frac{\Gamma\left(\frac{n+m-1}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)\Gamma\left(\frac{m-1}{2}\right)\Gamma\left(\frac{1}{2}\right)},$$

$$D=\left\{(w_1,w_2):w_1>0,\,w_2>0,\,w_1+w_2<1,\ \frac{(n+m)^{(n+m)/2}}{n^{n/2}m^{m/2}}w_1^{n/2}w_2^{m/2}>\lambda\right\},$$

and $r_1<r_2$ are the two roots (for variable w_1) of

$$1-w_1-\frac{\lambda^{2/m}n^{n/m}m}{(n+m)^{(n+m)/m}w_1^{n/m}}=0.$$

We compute the above double integral using Gaussian quadrature, implemented in an R function called *plrt*, which is available as the online supplementary materials for this article.

3. EXAMPLES

We illustrate here a testing hypothesis approach to compare the means and the variances of two normal distributions. The x sample and the y sample need to be random samples, independent of each other, and approximately normally distributed.

We first test H_0 versus H_1 using the exact distribution of $\lambda_{n,m}$. If we fail to reject H_0, we conclude that there is no evidence that the two populations have different means or different variances. If we reject H_0, we go on to test $H_0^\sigma:\sigma_1^2=\sigma_2^2$ versus $H_1^\sigma:\sigma_1^2\ne\sigma_2^2$ using the F-test. If H_0^σ holds, we use the pooled t-test to check $H_0^\mu:\mu_1=\mu_2$ versus $H_1^\mu:\mu_1\ne\mu_2$; if H_0^σ is rejected, we use the following Welch t-test to check $H_0^\mu:\mu_1=\mu_2$ versus $H_1^\mu:\mu_1\ne\mu_2$, with the test statistic t and its degrees of freedom df given by

$$t=\frac{\bar{x}-\bar{y}}{\sqrt{\frac{s_x^2}{n}+\frac{s_y^2}{m}}}\quad\text{and}\quad df=\frac{\left(\frac{s_x^2}{n}+\frac{s_y^2}{m}\right)^2}{\frac{1}{n-1}\left(\frac{s_x^2}{n}\right)^2+\frac{1}{m-1}\left(\frac{s_y^2}{m}\right)^2},$$

where s_x^2 and s_y^2 are the sample variances of the x sample and y sample, respectively. We note that three tests are performed using the same data when the likelihood ratio test initially rejects H_0; therefore, one needs to be careful when formally reporting the significance level of a particular test.

In each of the following three examples, the independence between the two samples and within each sample came from the design and execution of the experiment. The normality assumption is checked in Figure 1 near the end of this section.

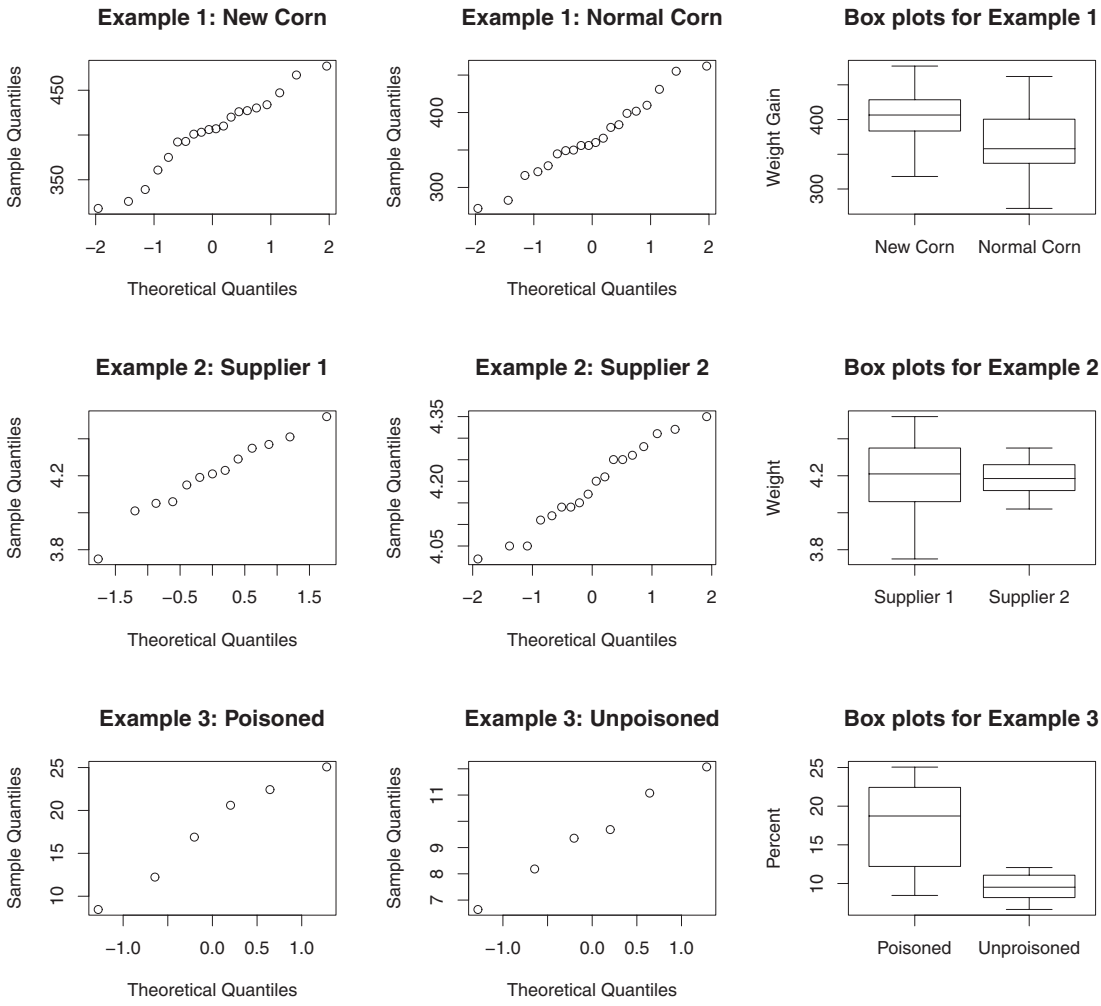

Figure 1. Normal Q-Q plots and boxplots for the data used in Examples 1–3.

Example 1. To test the quality of a new high-lysine corn as animal feed, an experimental group of 20 one-day-old male chicks ate a ration containing the new corn, and a control group of another 20 chicks were given a ration that was identical except that it contained normal corn. After 21 days, the weight gains (in grams) were measured for each group (source: problem 7.35 of Moore 2000, p. 401):

new corn: 361 447 401 375 434 403 393 426 406 318 467 407 427 420 477 392 430 339 410 326;
normal corn: 380 321 366 356 283 349 402 462 356 410 329 399 350 384 316 272 345 455 360 431.

The likelihood ratio test statistic is $\lambda_{20,20} = 0.0379$ with a p-value of 0.0488, suggesting that at 5% level there is some mean and/or variability difference in weight gain between the new corn group and the normal corn group. The F-test statistic is $F = 1.4138$ with a p-value of 0.4575, so there is no evidence that the variabilities in weight gain are different between the two groups. The pooled t-test statistic is $t_{\text{pooled}} = 2.469$ with a p-value of 0.0182, indicating that the means of weight gain are different between the two groups.

Example 2. Weight is a factor that may affect the metabolic rates of white mice, and a researcher wishes to get white mice that are relatively homogeneous with respect to weight. Thirteen mice from Supplier 1 and 18 mice from Supplier 2 are available, and the researcher weighs these mice (in ounces) to get the following data (source: example 9.10 of McClave and Sincich 2009, p. 458):

Supplier 1: 4.23 4.35 4.05 3.75 4.41 4.37 4.01 4.06 4.15 4.19 4.52 4.21 4.29;
Supplier 2: 4.14 4.26 4.05 4.11 4.31 4.12 4.17 4.35 4.25 4.21 4.05 4.28 4.15 4.20 4.32 4.25 4.02 4.14.

The likelihood ratio test statistic is $\lambda_{13,18} = 0.0221$ with a p-value of 0.0328, indicating that at 5% level there is some mean and/or variability difference in the weight of mice from Supplier 1 and Supplier 2. The F-test statistic is $F = 4.2375$ with a p-value of 0.0071, giving strong evidence that the variabilities in the weight of mice are different between the two suppliers. The Welch t-test statistic is $t = 0.1888$ with a p-value of 0.8367, showing no evidence of a difference in the means of the mice weight between the two suppliers.

Example 3. Researchers compared six white rats poisoned with dichlorodiphenyltrichloroethane (DDT) with a control group of six unpoisoned rats in a randomized experiment. A main clue to the nature of DDT poisoning can be obtained

from electrical measurements of nerve activity. When a nerve is stimulated, its electrical response shows a sharp spike followed by a much smaller second spike. The researchers measured the height of the second spike as a percentage of the first spike when a nerve in the rat's leg was stimulated (source: problem 7.37 of Moore 2000, p. 404):

poisoned: 12.207 16.869 25.050 22.429 8.456 20.589;
unpoisoned: 11.074 9.686 12.064 9.351 8.182 6.642.

Running the likelihood ratio test gives $\lambda_{6,6} = 0.0007$ with a p-value of 0.0048; therefore, at 5% level there is strong evidence that the mean and/or variability in the percentage of electrical response is different between the poisoned group and the unpoisoned group. From $F = 10.5707$ with a p-value of 0.0217, the variabilities in the percentage of electrical response are different between the two groups, and from the Welch $t = 2.9912$ with a p-value of 0.0246, the means of the percentage of electrical response are also different between the two groups.

Formal testing of hypothesis does not suit everyone's needs/taste/preference; however, it is always helpful to plot the data in any data analysis. In Figure 1, we display the normal Q-Q plots and boxplots for the data used in the above three examples. From Figure 1, the normality assumption needed to conduct the tests in the three examples is generally met. Moreover, the boxplots in Figure 1 support the conclusions obtained in the three examples.

4. COMPARISON WITH RELATED TESTS

As far as testing H_0 versus H_1 is concerned, a competitor of the likelihood ratio test is the combined test of Perng and Littell (1976). Table 1 shows that the two tests have similar powers.

Normality is one of the assumptions for both the likelihood ratio test and the combined test. Table 2 shows that in general the two tests are not robust to departures from normality.

Table 1. Power comparison between the likelihood ratio test and the combined test of Perng and Littell (1976). Powers for the combined test are given in the parentheses. The significance level is $\alpha = 0.05$, the two samples are generated from $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, respectively, with $\mu_2 = \mu_1 + \delta\sigma_1$ and $\sigma_2 = \rho\sigma_1$, and the powers are the proportions of rejections based on 10,000 simulated pairs of samples in each case

	δ		
ρ	0.0	0.5	1.0
		$n = m = 10$	
0.5	0.40 (0.38)	0.57 (0.57)	0.88 (0.89)
1.0	0.05 (0.05)	0.14 (0.14)	0.45 (0.45)
1.5	0.15 (0.14)	0.22 (0.22)	0.42 (0.43)
		$n = m = 30$	
0.5	0.92 (0.91)	0.99 (0.99)	1.00 (1.00)
1.0	0.05 (0.06)	0.37 (0.37)	0.93 (0.92)
1.5	0.47 (0.46)	0.65 (0.65)	0.93 (0.93)
		$n = 50,\ m = 30$	
0.5	0.97 (0.95)	1.00 (1.00)	1.00 (1.00)
1.0	0.05 (0.05)	0.45 (0.45)	0.97 (0.97)
1.5	0.55 (0.60)	0.75 (0.78)	0.97 (0.98)

Table 2. Robustness of the likelihood ratio test and the combined test of Perng and Littell (1976). The entries are the nominal levels of the two tests (the ones for the combined test are in parentheses) based on 10,000 simulated pairs of samples in each case when H_0 is true for the selected population

	Significance level α		
Population	0.10	0.05	0.01
		$n = m = 10$	
$N(0,\ 1)$	0.098 (0.099)	0.051 (0.052)	0.009 (0.011)
$U(0,\ 1)$	0.057 (0.058)	0.029 (0.031)	0.006 (0.007)
χ_2^2	0.270 (0.275)	0.197 (0.204)	0.091 (0.100)
χ_{10}^2	0.141 (0.143)	0.079 (0.084)	0.018 (0.023)
t_5	0.173 (0.174)	0.103 (0.103)	0.026 (0.029)
		$n = m = 30$	
$N(0,\ 1)$	0.102 (0.103)	0.053 (0.053)	0.009 (0.009)
$U(0,\ 1)$	0.049 (0.047)	0.023 (0.022)	0.004 (0.004)
χ_2^2	0.311 (0.315)	0.239 (0.241)	0.126 (0.136)
χ_{10}^2	0.162 (0.166)	0.098 (0.103)	0.026 (0.033)
t_5	0.224 (0.224)	0.147 (0.148)	0.054 (0.060)
		$n = 50,\ m = 30$	
$N(0,\ 1)$	0.100 (0.009)	0.050 (0.052)	0.008 (0.010)
$U(0,\ 1)$	0.048 (0.047)	0.022 (0.021)	0.003 (0.003)
χ_2^2	0.308 (0.307)	0.232 (0.235)	0.127 (0.138)
χ_{10}^2	0.153 (0.156)	0.092 (0.096)	0.028 (0.034)
t_5	0.222 (0.220)	0.143 (0.143)	0.052 (0.057)

Recently, in the so-called large-scale inference on gene expressions, the pooled t-test or the Welch t-test are commonly used to screen the genes and pinpoint those "nonnull" genes (differently expressed genes) worth possible further study. Suppose that we want to know the top k (a preset number) genes that are most likely "nonnull." One way to screen the genes is to conduct, say, the pooled t-test on all the genes under study with two types of expression levels, order the genes according to the sizes of the p-values from the smallest to the largest, and the genes with the first k smallest p-values are worth more attention. To the best of our knowledge, the likelihood ratio test and the combined test have not been used in such large-scale inference. In the following, we use a dataset from Ge, Dudoit, and Speed (2003) to compare the likelihood ratio test with the pooled t-test, the Welch t-test, and the combined test. This dataset is the result of an apolipoprotein A-1 experiment carried out as part of a study of lipid metabolism and atherosclerosis susceptibility in mice (Callow et al. 2000). The apolipoprotein A-1 is a gene known to play a pivotal role in high-density lipoprotein (HDL) metabolism, such as promoting cholesterol efflux from tissues to the liver for excretion, and mice with the apolipoprotein A-1 gene knocked out have very low HDL cholesterol levels. The goal of this experiment was to identify genes with altered expression in the livers of these knockout mice compared with inbred control mice. The treatment group consisted of eight mice with the apolipoprotein A-1 gene knocked out and the control group consisted of eight wild-type C57Bl/6 mice. Together, 6356 genes were studied, giving a matrix with 6356 rows and 16 columns, where for each row the first $n = 8$ entries form the x sample and the last $m = 8$ entries form the y sample.

We apply the above-mentioned four tests to each row of this data-set and order the 6356 genes using the 6356 p-values of

Table 3. Comparison of the four sets of ordered genes in terms of the pairwise ratio, κ/k, where for any two sets of ordered genes, if the subset of the first k ordered genes from each set is to be taken out for further study, then κ is the number of genes in the intersection of the two related subsets

	The number of the first k ordered genes for further study								
Pair of tests	10	20	30	50	100	200	300	500	1000
Likelihood ratio vs. pooled t	0.5	0.4	0.27	0.18	0.19	0.27	0.28	0.34	0.45
Likelihood ratio vs. Welch t	0.5	0.4	0.27	0.18	0.16	0.23	0.26	0.31	0.44
Likelihood ratio vs. combined	0.7	1.0	0.97	0.96	0.92	0.93	0.94	0.94	0.95
Pooled t vs. Welch t	1.0	1.0	0.93	0.94	0.93	0.94	0.95	0.97	0.99
Pooled t vs. combined	0.8	0.4	0.27	0.18	0.2	0.28	0.3	0.36	0.47
Welch t vs. combined	0.8	0.4	0.27	0.18	0.16	0.24	0.27	0.33	0.46

each test to get four sets of ordered genes. Table 3 compares these four sets of ordered genes in terms of the pairwise ratio, κ/k, where for any two sets of ordered genes, if the subset of the first k ordered genes from each set is to be taken out for further study, then κ is the number of genes in the intersection of the two related subsets.

We observe from Table 3 that the orderings obtained by the pooled t-test and the Welch t-test are similar, and the orderings obtained by the likelihood ratio test and the combined test are similar. However, between these two groups of tests, the orderings are quite different. Although we are not molecular biologists, it is totally inconceivable to us that sophisticated gene manipulations/alterations only lead to possible changes in the means of the gene expression levels. Instead, one perhaps should at least initially look for differences between means and variances, and our likelihood ratio test or the combined test can be used to do so.

5. DISCUSSION

The accuracy of the R function *plrt* was tested on several grounds. Because the integrand involved is unbounded on the boundary $1 - w_1 - w_2 = 0$, more quadrature points (500) are used to achieve the accuracy needed for practical use. For various combinations of n and m, we simulated the percentage points of the distribution of $\lambda_{n,m}$ based on 100,000 samples for probabilities in the set $\{0.010, 0.025, 0.05(0.05)0.95, 0.975, 0.990\}$, and used them to compute $\tilde{p} = P(\lambda_{n,m} \le \tilde{\xi}_p)$ using *plrt*, where $\tilde{\xi}_p$ is a simulated percentage point associated with a value p in the above set. The agreement between p and $\tilde{p}$ was found to be within 0.004 in all the cases tested.

Because nonnormal data can affect the performance of the likelihood ratio test and the combined test, we removed those genes in the apolipoprotein A-1 experiment data, for which the Shapiro test of normality gave a p-value less than 0.01 for either the x sample or the y sample, and redid the comparison in the last section. We also simulated gene expression data where the conditions for the likelihood ratio test and the combined test are satisfied, and redid the comparison. In both cases (results not shown here), the conclusions based on Table 3 remain valid.

Looking for differences between two populations is a commonly encountered problem. When the differences are represented by the means or variances or both, we recommend using the likelihood ratio test to start the investigation. We are aware that looking for differences in means and/or variances to start is not the current practice—many people use a kind of t-test to start. This is so partially because an easy to use test, say, easy to compute and only a few pages of tables were needed, was not available for practitioners to adopt in the past. Even today, the popular statistics textbooks and handbooks do not offer means to do this. With the developments mentioned in the references and in particular the development in this article, we hope that the above situation will change.

SUPPLEMENTARY MATERIALS

A computer program written in R is provided in the supplementary material. This program, called plrt, computes the p-value for the exact likelihood ratio test developed in the article.

[Received June 2011. Revised June 2012.]

REFERENCES

Callow, M. J., Dudoit, S., Gong, E. L., Speed, T. P., and Rubin, E. M. (2000), "Microarray Expression Profiling Identifies Genes With Altered Expression in HDL Deficient Mice," *Genome Research*, 10, 2022–2029. [183]

Ge, Y., Dudoit, S., and Speed, T. P. (2003), "Resampling-Based Multiple Testing for Microarray Data Analysis," *Test*, 12, 1–77. [183]

McClave, J. T., and Sincich, T. (2009), *Statistics* (11th ed.), Upper Saddle River, NJ: Pearson Prentice Hall. [182]

Moore, D. S. (2000), *The Basic Practice of Statistics* (2nd ed.), New York: W. H. Freeman and Company. [182,183]

Nagar, D. K., and Gupta, A. K. (2004), "Percentage Points for Testing Homogeneity of Several Univariate Gaussian Populations," *Applied Mathematics and Computation*, 156, 551–561. [180]

Pearson, E. S., and Neyman, J. (1930), "On the Problem of Two Samples," in *Joint Statistical Papers* (1967), eds. J. Neyman and E. S. Pearson, Cambridge: Cambridge University Press, pp. 99–115. [180]

Perng, S. K., and Littell, R. C. (1976), "A Test of Equality of Two Normal Population Means and Variances," *Journal of the American Statistician Association*, 71, 968–971. [180,183]

Plackett, R. L. (1946), "Literature on Testing the Equality of Variances and Covariances in Normal Populations," *Journal of the Royal Statistical Society*, 109, 457–468. [180]

R Development Core Team (2011), *R: A Language and Environment for Statistical Computing*, Vienna: R Foundation for Statistical Computing. Available at *http://www.R-project.org*. [180]

Sukhatme, P. V. (1935), "A Contribution to the Problem of Two Samples," *Proceedings of the Indian Academy of Sciences,* Series A, 2, 584–604. [180]