Volume 53 • Number 4 • 2024
communications in statistics
SIMULATION AND COMPUTATION
Taylor & Francis

Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Simulation and Computation

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/lssp20

Combining independent tests for a common parameter of several continuous distributions: a new test and power comparisons

K. Krishnamoorthy, Shanshan Lv & Md Monzur Murshed

To cite this article: K. Krishnamoorthy, Shanshan Lv & Md Monzur Murshed (2024) Combining independent tests for a common parameter of several continuous distributions: a new test and power comparisons, Communications in Statistics - Simulation and Computation, 53:4, 1837-1856, DOI: 10.1080/03610918.2022.2058546

To link to this article: https://doi.org/10.1080/03610918.2022.2058546

Published online: 06 Apr 2022.

COMMUNICATIONS IN STATISTICS - SIMULATION AND COMPUTATION®
2024, VOL. 53, NO. 4, 1837–1856
https://doi.org/10.1080/03610918.2022.2058546

Combining independent tests for a common parameter of several continuous distributions: a new test and power comparisons

K. Krishnamoorthy[a], Shanshan Lv[b], and Md Monzur Murshed[a]

[a]Department of Mathematics, University of Louisiana at Lafayette, Lafayette, LA, USA; [b]Department of Statistics, Truman State University, Kirksville, MO, USA

ABSTRACT
The problem of testing a common parameter of several independent continuous populations is considered. Among all tests, Fisher's combined test is the most popular one and is routinely used in applications. In this article, we propose an alternative method of combining the *p*-values of independent tests using chi-square scores, referred to as the inverse chi-square test. The proposed test is as simple as other existing tests. We compare the powers of the combined tests for (i) testing a common mean of several normal populations, (ii) testing the common coefficient of variation of several normal populations, (iii) testing the common correlation coefficient of several bivariate normal populations, (iv) testing the common mean of several lognormal populations and (v) testing the common mean of several gamma distributions. Our comparison studies indicate that the inverse chi-square test is a better alternative combined test with good power properties. An illustrative example with real-world data is given for each problem.

ARTICLE HISTORY
Received 2 August 2021
Accepted 22 March 2022

KEYWORDS
Common mean; Exact test; Inverse normal; χ^2 score; Meta analysis; Power; *z* score

MATHEMATICS SUBJECT CLASSIFICATION
62F03; 62P10

1. Introduction

In many applications, it is desired to combine the results of several independent studies to seek evidence to support some common hypotheses of interest. Such problems arise, for example, when two or more independent agencies are involved in measuring the effect of a new drug or when different measuring instruments/laboratories are used to measure the same variable to assess the overall average quality. There are many ways to combine the results for estimating/testing the common parameter of interest. In this article, we address hypothesis tests that are developed by combining *p*-values of several independent tests for a common parameter or parametric function. The combined test that we propose and other existing tests are applicable to any continuous distribution under some assumptions.

To describe the problem formally, let us suppose that there are k independent populations with the same parameter or function of parameters such as the coefficient of variation and correlation coefficient. Let us denote the common parametric function by $\xi = \xi(\theta_{11}, ..., \theta_{l1}) = \cdots = \xi(\theta_{1k}, ..., \theta_{lk})$. Assume that independent samples, each of size n_i, $i = 1, ..., k$, are available from these populations. Consider testing

$$H_0 : \xi = \xi_0 \quad \text{vs.} \quad H_a : \xi \neq \xi_0, \tag{1}$$

CONTACT K. Krishnamoorthy krishna@louisiana.edu Department of Mathematics, University of Louisiana at Lafayette, Lafayette, LA 70504-1010, USA.

where ξ_0 is a specified value. Let P_i denote the p-value of a test for the above hypotheses based on the ith sample, $i = 1, ..., k$. The p-values are based on independent samples, so they are independent uniform(0, 1) random variables. To arrive at a combined test based on all samples, some transformed p-values are combined to find a single test statistic for the hypotheses in (1). Fisher (1932) has proposed the combined statistic $-2\sum_{i=1}^{k} \ln P_i$, which has the χ^2_{2k} distribution under the null hypothesis. Fisher's method is one of the popular methods in meta-analysis, and comparison of several combined tests for the normal case by Kifle and Sinha (2021) indicated that Fisher's test is better than other combined tests.

As noted by Whitlock (2005), Fisher's method treats large and small p-values asymmetrically. Rice (1990) provided an example to describe this asymmetry problem and Whitlock (2005) provided an extreme case as follows. Suppose there were p-values from two studies on a topic that would be combined to arrive at a single combined test. One of these studies rejected the null hypothesis with p-value of 0.001, while the other did not reject with p-value of 0.999. Fisher's combined test produces a p-value of 0.008. The combined p-value is statistically significant despite the fact that the individual p-values clearly indicate that the data are equivocal. In this sense, Fisher's method is asymmetrically sensitive to small p-values compared to large p-values. Stouffer et al. (1949) have proposed another combined test based on the z-scores, where the z-score based on P_i is defined as $Z_i = \Phi^{-1}(P_i)$, $i = 1, ..., k$, where Φ is the standard normal distribution function. This test is also easy to use, because the null distribution of $\sum_{i=1}^{k} Z_i$ is normal with men zero and variance k. For the extreme case described earlier, the p-value of the combined z-score test is 0.5, which indicates that the z-score test does not have asymmetrical problem. Whitlock (2005) also suggested the weighted z-score test based on the combined z-scores $\frac{\sum_{i=1}^{k} w_i Z_i}{\sqrt{\sum_{i=1}^{k} w_i^2}}$. For testing a common mean of several normal populations, Whitlock recommended $w_i = n_i - 1$, the degrees of freedom (df) associated with the t-test based on the ith sample of size n_i. Combining independent studies with weights seems to be reasonable as one would want to weight studies with more information more strongly than those with less information. For testing a common mean of several normal populations, Whitlock has noted, on the basis of simulation results, that the weighted score test is better than the Fisher test in terms of power.

In general, to arrive at a combined test, the p-values P_i's of independent studies are transformed so that the distributions of transformed p-values do not depend on any unknown quantity and have the additive property. On the basis of this idea, we propose yet another combined test where the p-values P_i's are converted to $\chi^2_{n_i;P_i}$ variates, where $\chi^2_{m;\alpha}$ denotes the 100α percentile of the chi-square distribution with degrees freedom (df) m. Note that, since P_i is a uniform random variable, by inverse transformation method, $\chi^2_{n_i;P_i}$ has the chi-square distribution with df $=$ n_i. As the chi-square distribution is stochastically increasing in the degrees of freedom, this χ^2 score test gives more weights to the individual tests based on larger sample sizes.

In the following section, we describe the Fisher test, z-score test, the weighted z-score test by Whitlock (2005) and the new inverse chi-square test. These tests are applicable as along as the individual tests are based on some continuous distributions. In Sec. 3, we address the problem of combining independent tests for a common mean of several normal populations and compare the combined tests in terms of power. In the subsequent sections, we consider the problems of testing a common coefficient of variation of normal populations, testing a common correlation coefficient of bivariate normal populations and testing a common mean of several lognormal populations. For each problem, we compare the combined tests with respect to power and illustrate them using examples. In Sec. 7, we describe a combined test for a common mean of several gamma distributions and compare them in terms of power. An illustrative example with real-world data is given for each problem. Some concluding remarks along with discussion are given in Sec. 8.

2. Combined tests

We shall describe some combined tests that are applicable to any continuous distributions. Other combined tests that are applicable to only normal family is described in the next section.

2.1. Fisher's test

Fisher's test is based on the combination of the p-values of the individual tests given by $F = -2\sum_{i=1}^{k} \ln P_i$. Under H_0, P_i's are independent uniform(0, 1) random variables and $-2\ln P_i$ are independent χ_2^2 random variables. As a result, F has the χ_{2k}^2 distribution under the null hypothesis in (1). Let p_i be an observed value of P_i, $i = 1,\ldots,k$. For a given level α, Fisher's test rejects H_0 when $-2\sum_{i=1}^{k} \ln p_i > \chi_{2k;1-\alpha}^2$, where $\chi_{m;q}^2$ denote the $100q$ percentile of the χ_m^2 distribution. Equivalently, the Fisher test rejects H_0 whenever the p-value

$$P\left(\chi_{2k}^2 > -2\sum_{i=1}^{k} \ln p_i\right) < \alpha.$$

2.2. Inverse normal test

Let Φ denote the standard normal cumulative distribution function (CDF), and Φ^{-1} denotes the inverse function. Since P_i's are independent uniform(0, 1) random variables, the corresponding normal scores $Z_i = \Phi^{-1}(P_i)$ are independent standard normal random variables. The inverse normal test rejects H_0 when

$$Z = \frac{1}{\sqrt{k}}\sum_{i=1}^{k} \Phi^{-1}(P_i) < z_\alpha,$$

where z_α denotes the 100α percentile of the standard normal distribution. This test was proposed in Stouffer et al. (1949).

2.3. Weighted inverse normal test

Note that the weights for the individual p-values in the inverse normal test are the same. Instead, Whitlock (2005) has proposed the weight $\nu_i = n_i - 1$ for the p-value of the t-test for a normal mean based on the ith sample. The resulting combined test rejects the null hypothesis when

$$Z_w = \frac{1}{\sqrt{\sum_{j=1}^{k} \nu_j^2}}\sum_{i=1}^{k} \nu_i \Phi^{-1}(P_i) < z_\alpha.$$

2.4. Inverse χ^2 test

Let $\chi_{m;q}^2$ denote the $100q$ percentile of the chi-square distribution with df m. Notice that the Fisher test combines the independent p-values by converting them to $\chi_{2;P_i}^2$ variates, and then uses the single combined statistic $\sum_{i=1}^{k} \chi_{2;P_i}^2 = -2\sum_{i=1}^{k} \ln(P_i)$ which has the χ_{2k}^2 distribution. Instead, we convert the p-value P_i to $\chi_{n_i;1-P_i}^2$ variate, where n_i is the sample size from the ith population. Since $\sum_{i=1}^{k} \chi_{n_i;1-P_i}^2$ has the χ_N^2 distribution, where $N = \sum_{i=1}^{k} n_i$, the inverse χ^2 test rejects the null hypothesis in (1) when

$$\sum_{i=1}^{k} \chi^2_{n_i;1-P_i} > \chi^2_{N;1-\alpha}.$$

Since the χ^2_m distribution is stochastically increasing in m, for any $0 < u < 1$, $\chi^2_{n_i;u} > \chi^2_{n_j;u}$, provided $n_i > n_j$. Thus, we see that, unlike the Fisher test and the inverse normal test, the inverse χ^2 test gives more "weights" to individual tests with large sample sizes. As a result, this test is expected to be better than the Fisher test when the sample sizes are unequal.

In the following sections, we apply these tests for several examples and compare them with respect to power. For each example, the tests are illustrated using real life data.

3. Tests for a common mean of several normal distributions

Let $(\bar{X}_i, S_i^2)$ denote the (mean, variance) based on a sample of size n_i from a $N(\mu, \sigma_i^2)$ distribution, $i = 1, ..., k$. The variance S_i^2 is defined with the divisor $m_i = n_i - 1$, $i = 1, ..., k$. For testing hypotheses

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_a : \mu \neq \mu_0 \tag{2}$$

consider the t-statistic $\sqrt{n_i}(\bar{X}_i - \mu_0)/S_i$, based on the ith sample. Let t_m denote the t random variable with df $= m$. The p-values for testing above hypotheses is given by

$$P_i = P\left(|t_{m_i}| > \left|\frac{\sqrt{n_i}(\bar{X}_i - \mu_0)}{S_i}\right|\right), \quad i = 1, ..., k. \tag{3}$$

As the p-values are independent, we can combine them to arrive at a single combined test using the methods described in the preceding section. Furthermore, the following combined test due to Zhou and Mathew (1993), which is valid only for testing a common mean of several normal distributions, will also be considered for power comparison.

3.1. The Zhou-Mathew test

This test is valid only for testing two-sided hypothesis in (1). Let $t_i^2 = n_i(\bar{X}_i - \mu_0)^2/S_i^2$, $i = 1, ..., k$. Further, let $Q_i = -2\ln(1 - F_i(t_i^2))$, where F_i is the cumulative distribution function of an F_{1,m_i} random variable with dfs 1 and $m_i = n_i - 1$, $i = 1, ..., k$. Let $W = \sum_{i=1}^k Q_i$. Note that the Fisher test rejects H_0 if $W > \chi^2_{2k;\alpha}$, and it gives equal weight to all the individual tests even though the sample sizes and variances could be unequal. Instead, Zhou and Mathew (1993) (also see Jordan and Krishnamoorthy 1995) have proposed the test based on $Q = \sum_{i=1}^k \gamma_i Q_i$, where

$$\gamma_i = \frac{n_i/T_i}{\sum_{j=1}^k (n_j/T_j)} \text{ and } T_i = \frac{m_i S_i^2 + n_i(\bar{X}_i - \mu_0)^2}{n_i}$$

is an unbiased estimate of σ_i^2 when H_0 in (2) is true. Zhou and Mathew (1993) have shown that the test that rejects H_0 whenever

$$\sum_{i=1}^{k} \frac{\gamma_i^{k-1} e^{-Q/\gamma_i}}{\prod_{j=1;j\neq i}(\gamma_i - \gamma_j)} \leq \alpha(1 + \eta), \tag{4}$$

where

$$\eta = \frac{2}{k(k-1)} \sum_{i<j} \frac{(\bar{X}_i - \mu_0)(\bar{X}_j - \mu_0)}{|\bar{X}_i - \mu_0||\bar{X}_j - \mu_0|}$$

is an exact level α test.

3.2. Power studies

All the combined tests in Sec. 2 and the Zhou-Mathew test are exact in the sense that their null distributions do not depend on any unknown parameter. Therefore, the type I error rates of these tests should be the same as the nominal level. To compare the tests in terms of power, we estimated the powers using simulation with 100,000 runs. The estimated powers for testing $H_0 : \mu = \mu_0$ vs. $H_a : \mu \neq \mu_0$ at the level 0.05 are reported in Table 1 for $k = 2, 3, 4, 5$ and 8. The sample sizes and the standard deviations (SDs) are chosen according to following combinations:

i. sample sizes are equal, but SDs are different,
ii. sample sizes are different, but the SDs are equal,
iii. sample sizes and SDs are negatively associated; that is, smaller SDs are associated with larger sample sizes,
iv. sample sizes and SDs are positively associated,
v. no apparent association between sample sizes and SDs, and
vi. Sample sizes are approximately equal and the population SDs are approximately equal.

For convenience, we write F to denote the Fisher test, χ^2 to denote the inverse χ^2 test, ZM to denote the Zhou-Mathew test, Z to denote the inverse normal test and WZ denote the weighted inverse normal test. Furthermore, we write $X > Y$ to indicate the comparison that the test X is more powerful than the test Y and $X \simeq Y$ to indicate that the powers of the tests X and Y are approximately equal.

Case 1: For this case of equal sample size, but different SDs, we find the following power comparisons from Table 1. Note that, for this case, the inverse normal test Z and the weighted inverse normal test WZ are identical, and so only the test Z is included in the power comparison.

n	σ	Power comparison
(5)	(1, 2)	$ZM > \chi^2 = Z = F$
(12)	(1, 5)	$F > \chi^2 > ZM > Z$
(15)	(1, 4)	$F > \chi^2 > ZM > Z$
(8)	(1, 2)	$ZM > F > \chi^2 > Z$
(10)	(1, 4, 5)	$ZM > \chi^2 \simeq F > Z$
(4)	(1–4)	$\chi^2 \simeq ZM > F > Z$
(5)	(2, 3)	$\chi^2 \simeq Z > F > ZM$
(5)	(1–6)	$ZM > \chi^2 \simeq F > Z$

Power comparisons in the above table clearly indicate that the Zhou-Mathew test followed by the Fisher test are preferable when sample sizes are equal or close by, and SDs are different. In some cases, the inverse χ^2 test is better than the other tests. The inverse normal test is less powerful than all other tests.

Case 2: Power comparisons are given in the following table when the sample sizes are unequal and the population SDs are the same.

n	σ	Power comparison
(5, 10)	(2)	$ZM > WZ > \chi^2 > Z > F$
(5, 10)	(4)	$ZM > WZ > \chi^2 > Z \simeq F$
(5, 10, 15, 30)	(4)	$\chi^2 \simeq WZ > F > ZM > Z$
(3–5, 8–10, 15)	(3)	$\chi^2 \simeq WZ > F > Z > ZM$

Table 1. Powers of the combined tests for a common mean of several normal populations.

Tests for $H_0 : \mu = 0$ vs. $H_a : \mu \neq 0$

μ	$\mathbf{n} = (5,5)$, $\boldsymbol{\sigma} = (1,2)$					$\mathbf{n} = (5,5)$, $\boldsymbol{\sigma} = (1,4)$					$\mathbf{n} = (8,8)$, $\boldsymbol{\sigma} = (1,2)$					$\mathbf{n} = (10,5)$, $\boldsymbol{\sigma} = (4,4)$				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
0.0	.051	.051	.049	.050	.050	.050	.050	.050	.050	.050	.049	.049	.049	.049	.049	.050	.050	.049	.051	.050
0.5	.128	.127	.170	.129	.129	.112	.112	.147	.113	.113	.208	.208	.275	.205	.205	.065	.063	.069	.064	.065
1.0	.391	.384	.500	.392	.392	.300	.304	.408	.284	.284	.672	.676	.764	.651	.651	.111	.106	.127	.106	.113
1.25	.564	.558	.676	.563	.563	.425	.436	.556	.393	.393	.857	.863	.907	.834	.834	.148	.140	.171	.141	.152
1.5	.726	.721	.811	.718	.718	.557	.574	.681	.502	.502	.955	.958	.968	.937	.937	.197	.185	.225	.188	.202
1.75	.850	.846	.898	.837	.837	.677	.700	.774	.606	.606	.988	.991	.989	.979	.979	.256	.241	.289	.244	.262
2.0	.928	.927	.947	.916	.916	.780	.804	.835	.696	.696	.998	.998	.996	.994	.994	.325	.305	.360	.311	.333
2.25	.969	.969	.972	.960	.960	.859	.881	.877	.770	.770	.999	.999	.998	.998	.998	.402	.380	.436	.384	.412
2.5	.988	.988	.985	.983	.983	.914	.934	.906	.830	.830	1.00	1.00	.999	.999	.999	.483	.458	.513	.464	.492
μ	$\mathbf{n} = (10,10)$, $\boldsymbol{\sigma} = (3,3)$					$\mathbf{n} = (12,12)$, $\boldsymbol{\sigma} = (1,5)$					$\mathbf{n} = (15,15)$, $\boldsymbol{\sigma} = (1,4)$					$\mathbf{n} = (15,14)$, $\boldsymbol{\sigma} = (1.5,2)$				
0.0	.049	.049	.049	.049	.049	.049	.050	.050	.049	.049	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050
0.5	.084	.085	.100	.084	.084	.252	.267	.312	.234	.234	.327	.349	.399	.304	.304	.247	.246	.308	.242	.245
1.0	.210	.207	.259	.210	.210	.736	.790	.717	.649	.649	.861	.902	.815	.788	.788	.776	.769	.828	.767	.773
1.25	.314	.307	.374	.314	.314	.898	.936	.802	.806	.806	.967	.984	.883	.914	.914	.933	.929	.951	.926	.930
1.5	.438	.428	.503	.440	.440	.970	.987	.851	.901	.901	.994	.999	.925	.969	.969	.987	.986	.990	.984	.985
1.75	.570	.557	.631	.573	.573	.993	.998	.888	.952	.952	.999	1.00	.954	.989	.989	.998	.998	.998	.997	.998
2.0	.695	.681	.745	.699	.699	.998	1.00	.917	.976	.976	1.00	1.00	.974	.996	.996	.999	.999	.999	.999	.999
2.25	.802	.789	.837	.806	.806	.999	1.00	.940	.989	.989	1.00	1.00	.985	.999	.999	1.00	1.00	.000	1.00	1.00
2.5	.884	.873	.904	.885	.885	1.00	1.00	.958	.995	.995	1.00	1.00	.992	.999	.999	1.00	1.00	1.00	1.00	1.00
μ	$\mathbf{n} = (10,5)$, $\boldsymbol{\sigma} = (2,2)$					$\mathbf{n} = (10,5)$, $\boldsymbol{\sigma} = (2,2.5)$					$\mathbf{n} = (10,5)$, $\boldsymbol{\sigma} = (4,1)$					$\mathbf{n} = (10,5)$, $\boldsymbol{\sigma} = (1,4)$				
0.0	.049	.049	.049	.049	.049	.049	.049	.049	.049	.050	.050	.050	.050	.050	.050	.049	.049	.049	.049	.049
0.5	.108	.104	.125	.103	.112	.105	.100	.120	.099	.110	.110	.120	.160	.120	.095	.238	.216	.261	.194	.262
1.0	.326	.306	.362	.311	.334	.294	.274	.331	.271	.308	.301	.345	.463	.334	.231	.707	.675	.637	.543	.738
1.25	.483	.457	.515	.464	.494	.437	.408	.472	.402	.455	.433	.496	.632	.471	.323	.878	.859	.738	.700	.893
1.5	.645	.617	.666	.626	.655	.588	.556	.616	.544	.608	.567	.645	.771	.604	.425	.962	.954	.795	.818	.965
1.75	.785	.760	.792	.768	.793	.730	.699	.743	.682	.749	.691	.775	.864	.721	.526	.991	.989	.833	.891	.991
2.0	.887	.868	.882	.873	.892	.843	.817	.841	.799	.857	.795	.870	.922	.814	.623	.998	.998	.866	.937	.997
2.25	.948	.937	.939	.938	.951	.919	.902	.907	.885	.928	.872	.932	.954	.881	.710	.999	.999	.893	.964	.999
2.5	.980	.974	.970	.975	.981	.963	.953	.947	.939	.968	.925	.967	.972	.926	.787	1.00	1.00	.917	.980	.999
μ	$\mathbf{n} = (5,5,5)$ $\boldsymbol{\sigma} = (3,2,1)$					$\mathbf{n} = (5,5,5)$ $\boldsymbol{\sigma} = (1,3,5)$					$\mathbf{n} = (10,10,10)$ $\boldsymbol{\sigma} = (3,2,1)$					$\mathbf{n} = (10,10,10)$ $\boldsymbol{\sigma} = (5,4,1)$				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
0.0	.051	.051	.051	.051	.051	.050	.050	.050	.050	.050	.051	.050	.050	.051	.051	.049	.049	.049	.049	.049
0.5	.121	.120	.157	.120	.120	.104	.103	.150	.103	.103	.242	.246	.341	.230	.230	.187	.195	.311	.176	.176
1.0	.365	.360	.463	.355	.355	.275	.278	.432	.258	.258	.762	.781	.876	.718	.718	.576	.629	.827	.497	.497
1.25	.533	.528	.637	.514	.514	.398	.405	.594	.364	.364	.924	.937	.970	.887	.887	.764	.825	.951	.659	.659
1.5	.696	.693	.776	.667	.667	.530	.542	.732	.477	.477	.984	.989	.994	.965	.965	.891	.938	.991	.788	.788
1.75	.824	.825	.870	.793	.793	.656	.673	.831	.585	.585	.998	.998	.998	.991	.991	.958	.983	.998	.877	.877
2.0	.912	.913	.924	.882	.882	.763	.783	.892	.683	.683	1.00	.999	.999	.998	.998	.986	.996	.999	.934	.934
2.25	.961	.962	.954	.938	.938	.847	.866	.928	.765	.765	1.00	1.00	.999	1.00	1.00	.996	.999	.999	.966	.966
2.5	.985	.986	.971	.969	.969	.907	.924	.949	.833	.833	1.00	1.00	.999	1.00	1.00	.999	1.00	.999	.984	.984
μ	$\mathbf{n} = (10,7,4)$ $\boldsymbol{\sigma} = (1,3,5)$					$\mathbf{n} = (10,7,4)$ $\boldsymbol{\sigma} = (5,3,1)$					$\mathbf{n} = (20,10,4)$ $\boldsymbol{\sigma} = (6,2,2)$					$\mathbf{n} = (20,10,4)$ $\boldsymbol{\sigma} = (2,2,6)$				
0.0	.050	.050	.050	.049	.050	.050	.050	.050	.049	.050	.050	.049	.049	.049	.050	.050	.050	.051	.050	.051
0.5	.214	.190	.304	.170	.238	.089	.096	.128	.098	.079	.095	.099	.109	.097	.085	.188	.161	.194	.153	.200
1.0	.657	.619	.808	.478	.685	.228	.257	.367	.262	.178	.265	.282	.305	.277	.210	.626	.559	.621	.502	.651
1.25	.840	.817	.934	.639	.849	.335	.381	.522	.380	.253	.397	.425	.450	.415	.308	.828	.771	.812	.699	.846
1.5	.942	.934	.978	.767	.940	.458	.516	.673	.509	.342	.543	.585	.601	.570	.421	.941	.910	.925	.848	.950
1.75	.983	.982	.989	.860	.978	.583	.649	.794	.632	.441	.684	.733	.736	.713	.537	.985	.974	.974	.936	.988
2.0	.996	.996	.993	.920	.992	.698	.764	.878	.740	.541	.803	.849	.841	.826	.649	.997	.994	.991	.976	.998
2.25	.999	.999	.993	.957	.997	.795	.854	.932	.827	.638	.888	.926	.912	.906	.745	.999	.999	.996	.993	1.00
2.5	.999	.999	.994	.978	.999	.870	.918	.962	.891	.723	.943	.969	.954	.954	.824	1.00	.999	.997	.998	1.00
μ	$\mathbf{n} = (10,10,10,10)$ $\boldsymbol{\sigma} = (3,3,3,3)$					$\mathbf{n} = (15,13,10,5)$ $\boldsymbol{\sigma} = (4,3.5,3,2)$					$\mathbf{n} = (15,13,10,5)$ $\boldsymbol{\sigma} = (2,3,3.5,4)$					$\mathbf{n} = (15,13,10,5)$ $\boldsymbol{\sigma} = (4,4,4,4)$				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
0.0	.049	.049	.049	.050	.050	.049	.050	.050	.049	.050	.049	.049	.049	.050	.049	.050	.050	.050	.049	.050
0.5	.098	.097	.103	.097	.097	.096	.096	.096	.094	.092	.133	.124	.156	.118	.138	.082	.080	.081	.077	.081
1.0	.305	.299	.308	.297	.297	.283	.283	.274	.284	.263	.451	.419	.503	.383	.467	.200	.190	.194	.184	.199

(*continued*)

Table 1. Continued.

Tests for $H_0 : \mu = 0$ vs. $H_a : \mu \neq 0$

μ	$\mathbf{n} = (5,5)$, $\boldsymbol{\sigma} = (1,2)$					$\mathbf{n} = (5,5)$, $\boldsymbol{\sigma} = (1,4)$					$\mathbf{n} = (8,8)$, $\boldsymbol{\sigma} = (1,2)$					$\mathbf{n} = (10,5)$, $\boldsymbol{\sigma} = (4,4)$				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
1.25	.470	.460	.465	.463	.463	.441	.441	.420	.445	.407	.663	.627	.704	.576	.678	.306	.289	.289	.279	.306
1.5	.653	.639	.637	.645	.645	.615	.613	.577	.622	.572	.830	.803	.849	.744	.838	.433	.408	.401	.401	.432
1.75	.802	.789	.778	.795	.795	.772	.770	.728	.778	.726	.933	.919	.936	.873	.937	.578	.548	.530	.539	.576
2.0	.910	.900	.885	.904	.904	.883	.882	.842	.888	.846	.979	.973	.973	.944	.980	.709	.679	.648	.671	.705
2.5	.995	.993	.964	.995	.995	.969	.963	.916	.968	.968	.969	.974	.961	.935	.935	.999	.999	1.00	.985	.994
2.5	.989	.986	.978	.987	.987	.982	.982	.963	.983	.967	.999	.998	.994	.993	.999	.902	.882	.842	.875	.899
	$\mathbf{n} = (15,10,5,3)$ $\boldsymbol{\sigma} = (2,3,5,6)$					$\mathbf{n} = (15,10,5,3)$ $\boldsymbol{\sigma} = (6,5,3,2)$					$\mathbf{n} = (10,8,5,3)$ $\boldsymbol{\sigma} = (1,2,4,6)$					$\mathbf{n} = (10,8,5,3)$ $\boldsymbol{\sigma} = (6,4,2,1)$				
0.0	.049	.049	.049	.049	.049	.049	.049	.049	.049	.049	.050	.050	.049	.050	.050	.050	.051	.051	.050	.050
0.5	.125	.107	.144	.101	.139	.066	.067	.066	.068	.064	.222	.191	.302	.171	.249	.081	.086	.097	.088	.072
1.0	.397	.330	.441	.282	.442	.129	.131	.125	.135	.117	.697	.638	.813	.511	.736	.192	.216	.258	.227	.150
1.25	.581	.498	.621	.416	.633	.184	.187	.174	.194	.163	.877	.839	.935	.685	.896	.287	.324	.385	.340	.214
1.5	.751	.672	.774	.559	.796	.258	.263	.238	.274	.224	.964	.948	.976	.819	.968	.401	.452	.524	.468	.295
1.75	.876	.816	.875	.694	.906	.346	.353	.319	.373	.299	.992	.988	.987	.905	.992	.524	.586	.661	.600	.386
2.0	.949	.914	.934	.803	.964	.450	.457	.412	.484	.385	.998	.998	.991	.954	.998	.644	.710	.779	.717	.483
2.25	.982	.966	.962	.882	.988	.558	.568	.511	.599	.478	.999	.999	.992	.980	.999	.750	.813	.867	.814	.581
2.5	.995	.989	.974	.934	.997	.663	.674	.612	.705	.573	1.00	1.00	.992	.991	1.00	.836	.889	.926	.885	.671
	$\mathbf{n} = (6,5,4,4,5)$ $\boldsymbol{\sigma} = (2,2,2.2,1.8,2)$					$\mathbf{n} = (5,5,5,5,5)$ $\boldsymbol{\sigma} = (2,3,2,3,2)$					$\mathbf{n} = (12,8,7,4,3)$ $\boldsymbol{\sigma} = (6,4,2,1,1)$					$\mathbf{n} = (12,8,7,4,3)$ $\boldsymbol{\sigma} = (1,1,2,4,6)$				
0.0	.050	.050	.050	.050	.050	.050	.050	.049	.049	.049	.048	.049	.050	.049	.048	.050	.050	.049	.050	.049
0.5	.099	.096	.092	.099	.100	.087	.086	.085	.087	.087	.108	.117	.124	.122	.088	.368	.310	.406	.276	.411
1.0	.312	.296	.257	.313	.319	.245	.236	.223	.245	.245	.342	.397	.419	.408	.237	.938	.901	.943	.796	.957
1.25	.488	.465	.394	.492	.500	.390	.375	.347	.391	.391	.518	.593	.618	.600	.351	.994	.986	.987	.928	.996
1.5	.672	.645	.551	.679	.686	.548	.530	.482	.554	.554	.686	.767	.791	.766	.477	.999	.999	.994	.980	.999
1.75	.826	.801	.700	.830	.836	.712	.693	.628	.715	.715	.823	.891	.905	.880	.601	1.00	1.00	.996	.995	1.00
2.0	.924	.908	.819	.927	.931	.838	.821	.753	.840	.840	.913	.957	.964	.946	.711	1.00	1.00	.996	.998	1.00
2.25	.973	.965	.900	.975	.976	.923	.911	.853	.923	.923	.963	.986	.988	.978	.802	1.00	1.00	.996	.999	1.00
2.50	.992	.989	.950	.993	.993	.950	.951	.922	.954	.926	.986	.996	.996	.991	.871	1.00	1.00	.996	1.00	1.00
	$\mathbf{n} = (3,4,5,5,4,5)$ $\boldsymbol{\sigma} = (4.1,3.9,4.5,3.9,4.2,3.8)$					$\mathbf{n} = (4,4,4,4,4,4)$ $\boldsymbol{\sigma} = (4,3,2,2,1,1)$					$\mathbf{n} = (9,7,7,6,5,5)$ $\boldsymbol{\sigma} = (1,9,3,6,4,2)$					$\mathbf{n} = (9,7,7,6,5,5)$ $\boldsymbol{\sigma} = (1,1,2,2,3,4)$				
0.0	.048	.048	.048	.048	.049	.050	.049	.049	.050	.050	.049	.049	.049	.050	.050	.049	.049	.049	.050	.050
0.5	.057	.058	.056	.058	.057	.121	.118	.130	.121	.121	.158	.147	.254	.135	.168	.270	.249	.315	.232	.285
1.0	.092	.090	.082	.092	.094	.407	.395	.426	.403	.403	.504	.494	.726	.386	.500	.853	.830	.879	.756	.858
1.25	.122	.119	.104	.123	.125	.604	.592	.608	.590	.590	.703	.702	.876	.546	.680	.970	.964	.969	.915	.968
1.5	.165	.157	.133	.164	.169	.778	.769	.760	.755	.755	.852	.859	.946	.694	.816	.997	.996	.991	.978	.995
1.75	.220	.208	.170	.221	.228	.899	.893	.860	.872	.872	.940	.949	.972	.810	.906	.999	.999	.995	.995	.999
2.0	.289	.272	.218	.292	.300	.962	.960	.918	.942	.942	.980	.985	.984	.894	.957	1.00	1.00	.997	.999	.999
2.25	.370	.349	.273	.374	.384	.988	.988	.949	.976	.976	.994	.996	.990	.946	.982	1.00	1.00	.998	.999	1.00
2.50	.460	.435	.338	.465	.477	.997	.996	.965	.991	.991	.998	.999	.993	.974	.993	1.00	1.00	.998	1.00	1.00
	$\mathbf{n} = (5,4,6,5,4,5,6,4)$ $\boldsymbol{\sigma} = (3.2,3,2.7,3.1,3.3,3,3,2.7)$					$\mathbf{n} = (5,5,5,5,5,5,5,5)$ $\boldsymbol{\sigma} = (2,1,3,4,2,6,5,5)$					$\mathbf{n} = (4,3,8,9,4,10,15,5)$ $\boldsymbol{\sigma} = (4,4,5,6,5,3,5,6)$					$\mathbf{n} = (4,3,8,9,4,10,15,5)$ $\boldsymbol{\sigma} = (3,3,3,3,3,3,3,3)$				
0.0	.050	.050	.051	.050	.050	.050	.049	.050	.050	.050	.049	.049	.049	.049	.049	.049	.049	.050	.049	.049
0.5	.076	.076	.069	.076	.077	.103	.102	.135	.101	.101	.071	.068	.066	.066	.072	.103	.095	.082	.093	.105
1.0	.189	.182	.141	.185	.191	.312	.309	.414	.292	.292	.160	.143	.131	.138	.164	.340	.300	.217	.288	.347
1.25	.296	.282	.208	.290	.299	.477	.477	.589	.440	.440	.241	.214	.187	.203	.246	.531	.474	.338	.458	.536
1.5	.434	.412	.299	.426	.439	.648	.652	.739	.596	.596	.347	.308	.261	.291	.350	.725	.668	.484	.648	.726
1.75	.590	.563	.412	.582	.597	.796	.801	.846	.739	.739	.472	.424	.352	.399	.472	.870	.826	.632	.810	.868
2.0	.738	.712	.535	.732	.744	.899	.904	.912	.849	.849	.605	.553	.450	.520	.599	.953	.928	.756	.916	.950
2.25	.854	.833	.654	.849	.859	.956	.961	.950	.920	.920	.730	.679	.552	.641	.720	.987	.977	.847	.970	.985
2.50	.929	.915	.761	.926	.932	.984	.986	.970	.962	.962	.831	.789	.647	.750	.819	.997	.994	.905	.992	.996

1 = Inverse chi-square; 2 = Fisher's test; 3 = Zhou-Mathew's test; 4 = Inverse normal; 5 = Weighted inverse normal.

The χ^2 test and the weighted inverse normal test are more powerful than the other tests. For smaller number of groups, the Zhou-Mathew test also performs better. The Fisher test and the inverse normal test are less powerful than the other tests.

Case 3: Power comparisons, when the sample sizes and the SDs are negatively associated, are given in the following table.

n	σ	Power comparison
(5, 10)	(1, 4)	$ZM > WZ > \chi^2 > F > Z$
(4, 7, 10)	(1, 3, 5)	$WZ > ZM > \chi^2 > F > Z$
(4, 10, 20)	(2, 6)	$ZM > WZ \simeq \chi^2 > F > Z$
(5, 10, 13, 15)	(2,3,3.5,4)	$WZ > ZM > \chi^2 > F > Z$
(3, 5, 10, 15)	(2, 3, 5, 6)	$WZ > ZM > \chi^2 > F > Z$
(3, 5, 8, 10)	(1, 2, 4, 6)	$WZ > ZM > \chi^2 > F > Z$
(3, 4, 7, 8, 12)	(1, 2, 4, 6)	$WZ > ZM > \chi^2 > F > Z$

For Case 3, the inverse normal test is the least powerful among all five tests. The weighted inverse normal test and the Zhou-Mathew test are preferable to other tests. The inverse χ^2 test is better than the Fisher test for all cases in the above table.

Case 4: Power comparisons on the basis of reported powers in Table 1 are given in the following table for the case where sample sizes and the SDs are positively associated.

n	σ	Power comparison
(5, 10)	(1, 4)	$ZM > F > Z > \chi^2 > WZ$
(4, 7, 10)	(1, 3, 5)	$ZM > F > Z > \chi^2 > WZ$
(4, 10, 20)	(2, 6)	$ZM > F > Z > \chi^2 > WZ$
(5, 10, 13, 15)	(4,3.5,3,2)	$Z > \chi^2 > F > ZM > WZ$
(3, 5, 10, 15)	(2, 3, 5, 6)	$Z > F > \chi^2 > ZM > WZ$
(3, 5, 8, 10)	(1, 2, 4, 6)	$ZM > Z > F > \chi^2 > WZ$

For Case 4, the Zhou-Mathew test and the inverse normal test are preferable to others. The weighted inverse normal test is inferior to all other tests.

Case 5: When there is no apparent relation between sample sizes and SDs, the inverse χ^2 test and the weighted inverse normal test are similar and they are better than others; see the powers for $\mathbf{n} = (4, 3, 8, 9, 4, 10, 15, 5), \sigma = (4, 4, 5, 6, 5, 3, 5, 6)$. The Zhou-Mathew test is the worst. For the case $\mathbf{n} = (9, 7, 7, 6, 5, 5), \sigma = (1, 9, 3, 6, 4, 2)$, the Zhou-Mathew test is better than all other tests. In general, the inverse χ^2 test is stable and performs satisfactorily for Case 5.

Case 6: When the sample sizes are approximately the same and the population SDs are approximately equal, the Fisher test seems to be less powerful than all other tests. If the sample sizes are small, the Zhou-Mathew test is less powerful than all other tests; see the powers for $\mathbf{n} = (6, 5, 4, 4, 5), \sigma = (2, 2, 2.2, 1.8, 2)$; $\mathbf{n} = (3, 4, 5, 5, 4, 5), \sigma = (4.1, 3.9, 4.5, 3.9, 4.2, 3.8)$; $\mathbf{n} = (5, 4, 6, 5, 4, 5, 6, 4), \sigma = (3.2, 3, 2.7, 3.1, 3.3, 3, 3, 2.7)$. For these cases, the inverse χ^2 test and the inverse normal test perform very similar and they are more powerful than the other tests.

On overall basis, no test dominates others uniformly. Among the five tests, the inverse χ^2, the Zhou-Mathew test and the weighted inverse normal test are competing tests for a common mean of several normal populations.

3.3. Example

This example is concerned with the estimation of Selenium in nonfat milk powder by combining the results of four different analytical methods. The data, taken from Eberhardt, Reeve, and

Table 2. Selenium content in nonfat milk powder using four methods.

Methods	Sample size	Mean	Variance
1. Atomic absorption Spectrometry	8	105.00	85.711
2. Neutron activation: Instrumental	12	109.75	20.748
3. Neutron activation: Radiochemical	14	109.50	2.729
4. Isotope dilution mass spectrometry	8	113.25	33.640

Spiegelman (1989), are reproduced here in Table 2. Application of Bartlett's test by Jordan and Krishnamoorthy (1996) have shown that the variances are significantly different, and so the t-test for the mean based on the pooled data is not appropriate. To illustrate the testing methods, consider testing $H_0 : \mu = 108$ vs. $H_a : \mu \neq 108,$ where μ is the mean selenium content in nonfat milk powder. The p-values of the t test based on the measurements by methods 1, 2, 3 and 4 are $P_1 = 0.3899$, $P_2 = 0.2102$, $P_3 = 0.0048$ and $P_4 = 0.0375,$ respectively. The results based on the five combined tests are given in the following table.

Tests	Statistics	p-value
Inv χ^2	$\sum \chi^2_{n_i;1-P_i} = 71.887$	$P(\chi^2_{42} > 71.887) = .0028$
Fisher	$-2\sum \ln(P_i) = 22.261$	$P(\chi^2_8 > 22.261) = .0045$
Zhou-Mathew	—	LHS of (4) $=.0036;\ \eta = 0$
Inv Normal	$\frac{1}{2}\sum \Phi^{-1}(P_i) = -2.7290$	$\Phi(-2.7290) = .0032$
Weighted Inv Norm	$\sum w_i\Phi^{-1}(P_i)/\sqrt{\sum w_i^2} = -2.8929$	$\Phi(-2.8929) = .0019$

P-values of all the combined tests are much smaller than the nominal level of 0.05, and so all the tests reject the null hypothesis. The p-values of the inverse χ^2 test is smaller than that of the Fisher test. Among the five tests, the weighted normal test produced the smallest p-value.

Note that the p-values of the two individual tests based on samples 3 and 4 are $P_3 = 0.0048$ and $P_4 = 0.0375,$ respectively, which are smaller than 0.05. All the combined tests produced p-values that are smaller than the smallest p-value $P_3 = 0.0048$.

4. Tests for a common coefficient of variation

The coefficient of variation is commonly used as a measure of precision and repeatability of data. For example, Plesch and Klimpel (2002) have noted that the coefficient of variation is often used to assess the meter-to-meter variability when comparing different types of equipment that perform the same task. In practical situations where the CV is an appropriate measure of variability, the variable is usually positive. For a normal population, the ratio of the mean to the standard deviation has to be on the order of three or more, for the probability of a negative value is negligible. As a consequence, the CV must be at most.33 in practical situations where the CV is a suitable measure of variability (Johnson and Welch 1940).

To describe the problem of testing a common coefficient of variation for the normal case, let $X_{i1}, \ldots, X_{in_i}$ be a sample from a normal distribution with mean μ_i and the variance $\sigma_i^2 = \mu_i^2\tau^2$, $i = 1, \ldots, k$. Let $m_i = n_i - 1$, $i = 1, \ldots, k$. Note that τ is the common coefficient of variation for all populations.

4.1. An exact test for τ

To derive a test for τ based on the ith sample, let us find the distribution of the quantity $\bar{X}_i/S_i$ as follows. Let Z_i be a standard normal random variable and U_i^2 be a $\chi^2_{m_i}/m_i$ random variable. Noting that $(\bar{X}_i - \mu_i)/\sigma_i \sim Z_i/\sqrt{n_i}$ independently of $S_i/\sigma_i \sim U_i,$ we can write

Table 3. Powers of the combined tests for testing a common coefficient of variation.

Tests for $H_0 : \tau = \tau_0$ vs. $H_a : \tau > \tau_0$; $\tau_0 = .05$																				
n = (5,5)					**n** = (10,5)				**n** = (15,5)				**n** = (20,10) f				**n** = (30,5)			
τ	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
.05	.049	.049	.049	.049	.049	.049	.049	.049	.051	.051	.051	.051	.049	.049	.049	.050	.051	.051	.051	.051
.06	.216	.216	.209	.209	.275	.269	.263	.267	.332	.318	.308	.319	.428	.415	.415	.418	.469	.435	.413	.455
.07	.443	.441	.429	.429	.575	.566	.553	.559	.677	.660	.639	.658	.819	.810	.804	.807	.867	.844	.808	.852
.08	.638	.637	.623	.623	.791	.783	.770	.775	.881	.869	.851	.865	.964	.960	.957	.959	.980	.974	.958	.975
.09	.778	.777	.764	.764	.905	.899	.890	.894	.961	.957	.946	.954	.994	.993	.992	.993	.997	.996	.992	.996
.10	.866	.865	.856	.856	.959	.957	.950	.951	.988	.986	.981	.985	.999	.999	.999	.999	1.00	1.00	.999	1.00
.11	.920	.919	.911	.911	.982	.981	.977	.978	.996	.995	.993	.995	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
n = (5,5,5,5)					**n** = (10,5,5,5)				**n** = (20,10,4,5)				**n** = (15,15,5,5)				**n** = (30,5,7,4)			
.05	.050	.050	.050	.050	.050	.050	.050	.051	.049	.050	.049	.050	.050	.051	.051	.051	.050	.050	.051	.049
.06	.311	.309	.296	.296	.359	.350	.338	.339	.487	.456	.438	.459	.493	.469	.456	.470	.540	.490	.460	.501
.07	.640	.637	.616	.616	.729	.719	.699	.701	.883	.861	.835	.856	.889	.874	.856	.871	.921	.896	.858	.895
.08	.851	.849	.830	.830	.915	.909	.895	.895	.984	.979	.969	.976	.986	.982	.977	.981	.993	.989	.978	.988
.09	.944	.943	.933	.933	.978	.975	.970	.969	.998	.998	.995	.997	.999	.998	.997	.998	.999	.999	.997	.999
.10	.981	.981	.976	.976	.994	.994	.991	.991	1.00	1.00	.999	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
.11	.993	.993	.990	.990	.998	.998	.998	.998	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
n = (4,4,4,4,4,4)					**n** = (14,4,4,4,4,14)				**n** = (16,3,4,5,4,7)				**n** = (10,10,4,5,4,3)				**n** = (9,7,3,5,4,3)			
.05	.051	.051	.053	.053	.051	.049	.052	.049	.052	.051	.054	.049	.049	.048	.047	.048	.049	.049	.051	.048
.06	.332	.335	.308	.308	.526	.495	.467	.487	.477	.449	.417	.442	.439	.419	.397	.410	.387	.376	.353	.362
.07	.678	.678	.643	.643	.905	.887	.857	.877	.864	.837	.801	.820	.843	.825	.794	.810	.782	.764	.735	.746
.08	.883	.884	.858	.858	.989	.985	.977	.982	.981	.974	.958	.965	.969	.963	.949	.958	.946	.940	.922	.926
.09	.958	.959	.945	.945	.999	.999	.998	.998	.998	.997	.993	.994	.996	.995	.992	.994	.991	.989	.982	.983
.10	.988	.989	.982	.982	1.00	1.00	1.00	1.00	1.00	.999	.999	.999	1.00	.999	.998	.999	.998	.998	.995	.996
.11	.997	.997	.996	.996	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.999	.999

1 = Inverse chi-square; 2 = Fisher's test; 3 = Inverse normal; 4 = Weighted inverse normal.

$$\frac{\bar{X}_i}{S_i} = \frac{(\bar{X}_i - \mu_i + \mu_i)/\sigma_i}{S_i/\sigma_i} \sim \frac{Z_i + \sqrt{n_i}/\tau}{\sqrt{n_i}U_i}.$$

Thus, $\sqrt{n_i}\bar{X}_i/S_i \sim t_{m_i}\left(\sqrt{n_i}/\tau\right)$, where $t_m(\delta)$ denotes the noncentral t random variable with df $= m$ and the noncentrality parameter δ. For an observed value $\sqrt{n_i}\bar{x}_i/s_i$, Johnson and Welch (1940) have obtained an exact $1 - 2\alpha$ confidence interval $(\hat{\tau}_L, \hat{\tau}_U)$ as the solutions of equations

$$t_{m_i;\alpha}\left(\sqrt{n_i}/\hat{\tau}_L\right) = \sqrt{n_i}\frac{\bar{x}_i}{s_i} \quad \text{and} \quad t_{m_i;1-\alpha}\left(\sqrt{n_i}/\hat{\tau}_U\right) = \sqrt{n_i}\frac{\bar{x}_i}{s_i}. \tag{5}$$

Now consider testing

$$H_0 : \tau = \tau_0 \quad \text{vs.} \quad H_a : \tau > \tau_0, \tag{6}$$

where τ_0 is a specified value, based on the ith sample. The test that rejects the null hypothesis whenever $\hat{\tau}_L > \tau_0$, or equivalently, the p-value

$$P_i = P\left(t_{m_i}\left(\sqrt{n_i}/\tau_0\right) \le \sqrt{n_i}\frac{\bar{x}_i}{s_i}\right) < \alpha, \tag{7}$$

is an exact level α test. A combined test for testing the hypotheses in (6) can be obtained using the p-values $P_1, \ldots, P_k$, where P_i is based on the ith sample.

4.2. Power comparisons

We evaluated the powers of the inverse χ^2 test, Fisher's test, inverse normal test and weighted inverse normal test. The powers of the tests depend only on τ, and so the powers were estimated as a function τ and reported in Table 3 for $k = 2$, 4 and 6 and various sample sizes. The reported powers are Monte Carlo estimates based on 100,000 simulation runs. We observe from the estimated powers in Table 3 that the inverse χ^2 test performs slightly better than all other tests when sample sizes are equal, and appreciably better than the other tests for unequal sample sizes. See the powers for $\mathbf{n} = (5,5)$, $\mathbf{n} = (5,5,5,5)$ and $\mathbf{n} = (4,4,4,4,4,4)$; for the cases of unequal sample sizes, see the powers when $\mathbf{n} = (20,10)$, $\mathbf{n} = (30,5)$, $\mathbf{n} = (30,5,7,4)$ and $\mathbf{n} = (10,10,4,5,4,3)$. For all the cases, the inverse χ^2 test performs better than other tests, and so this inverse χ^2 test can be recommended for applications. The Fisher test also performs better than the inverse normal test and the weighted inverse normal test.

4.3. Example

Hong Kong Medical Technology Association conducted a quality assurance program for medical laboratories in Hong Kong in 1989. In the specialty of hematology and serology, one normal and one abnormal hematology and serology blood samples were sent to participants for measurements of variables Hb, RBC, MCV, Hct, WBC and Platelet in each survey. Tests by Fung and Tsang (1998) on the equality of coefficients of variation of the measurements on the six variables showed that the coefficient of variation for MCV in 1995 is not significantly different from that of 1996. Therefore, we are interested in making inference about the common coefficient of variation based on the 1995 and 1996 data on MCV. For 1995 survey, the sample size $n_1 = 63$, sample mean $\bar{x}_1 = 84.13$, sample variance $s_1^2 = 3.390$ and the coefficient of variation $\hat{\tau}_1 = 0.0219$. For the 1996 survey, $n_2 = 72$, $\bar{x}_2 = 85.68$, $s_2^2 = 2.946$ and $\hat{\tau}_2 = 0.0200$.[1] Tian (2005) has also used the data to find a confidence interval for the common coefficient of variation.

Suppose we take 0.023 as a hypothetical value for the criteria of good precision for the measurements of MCV in the survey. That is, we want to test $H_0 : \tau = .023$ vs. $H_a : \tau < .023$. The p-value (7) for the 1995 survey is $P_1 = 0.3104$ and the p-value for 1996 survey is $P_2 = 0.0675$. The combined statistic for the Fisher test is $-2\sum \ln(P_i) = 7.732$ and the p-value is $P(\chi_4^2 > 7.732) = 0.1019$. The statistic for the inverse χ^2 test is $\sum \chi^2_{n_i;1-P_i} = 158.71$ and the p-value is $P(\chi^2_{135} > 158.71) = 0.0798$. The statistic for the inverse normal test is $\sum \Phi^{-1}(P_i)/\sqrt{2} = -1.407$ and the p-value is $\Phi(-1.407) = 0.0797$. The weighted inverse normal test statistic is $\sum w_i \Phi^{-1}(P_i)/\sqrt{\sum w_i^2} = -1.451$ with p-value $\Phi(-1.451) = .0733$. Thus, all the tests, except the Fisher test, reject the null hypothesis at the level of 0.10, and supports the hypothesis that the common coefficient of variation is less than.023. We also note that all the combined tests produced p-values closer to the smaller of the individual p-values which is $P_2 = 0.0675$.

5. Combined tests for a common correlation coefficient

Let $\boldsymbol{X}_{i1}, \ldots, \boldsymbol{X}_{in_i}$ be a sample from a bivariate normal distribution with mean vector $\boldsymbol{\mu}_i$ and variance-covariance matrix

$$\begin{pmatrix} \sigma_{i1}^2 & \rho\sigma_{i1}\sigma_{i2} \\ \rho\sigma_{i1}\sigma_{i2} & \sigma_{i2}^2 \end{pmatrix}, \; i = 1, \ldots, k.$$

Let $\boldsymbol{S}_i = (s_{i,lj})$ be the sample variance-covariance matrix based on the ith sample. Then the sample correlation coefficient is given by $R_i = s_{i,12}/\sqrt{s_{i,11}s_{i,22}}$, $i = 1, \ldots, k$. The problem of estimating a common correlation coefficient of several bivariate normal populations has been addressed by Donner and Rosner (1980), Paul (1988) and Tian and Wilding (2008). These authors have proposed some approximate tests and confidence intervals for a common

[1] Our values of $\hat{\tau}_1$ and $\hat{\tau}_2$ are based on the means and variances reported in Fung and Tsang (1998). It is not clear how the authors Fung and Tsang (1998) and Tian (2005) obtained the value of 0.0406 for $\hat{\tau}_1$ and the value of 0.0346 for $\hat{\tau}_2$.

correlation coefficient. In the following, we propose combined tests by combining the Fisher Z tests based on individual samples.

5.1. Fisher's Z test for ρ

Even though there are other improved tests for correlation coefficient (e.g., Krishnamoorthy and Xia 2007), Fisher's test for the correlation coefficient is simple with practical accuracy. It is based on the distributional result that

$$Z_i = \tanh^{-1}(R_i) = \frac{1}{2}\ln\frac{1+R_i}{1-R_i} \sim N(\mu_\rho, (n_i-3)^{-1}) \text{ asymptotically,} \tag{8}$$

where $\mu_\rho = \tanh^{-1}(\rho) = \frac{1}{2}\ln\frac{1+\rho}{1-\rho}$. As μ_ρ is an increasing function of ρ, inferential procedures about ρ can be obtained from the above asymptotic distribution. Specifically, let z_i be an observed value of Z_i. That is, $z_i = \frac{1}{2}\ln\left(\frac{1+r_i}{1-r_i}\right)$, where r_i is an observed value of R_i, $i = 1, ..., k$. Furthermore, the p-value for testing

$$H_0 : \rho = \rho_0 \quad \text{vs.} \quad H_a : \rho > \rho_0$$

on the basis of the ith sample is given by

$$P_i = P(Z_i > z_i | n_i, \rho_0) = 1 - \Phi\left(\sqrt{n_i - 3}(z_i - \mu_{\rho_0})\right), \tag{9}$$

where Φ is the standard normal distribution function.

The above p-values can be combined using one of the methods in Sec. 2 to arrive at a single test. Apart from these combined tests, a test based on a linear combination of Fisher's Z statistics can be obtained (Donner and Rosner 1980). This combined test, described below, is also simple and easy to use.

5.2. A combined test based on fisher's Z statistics

Using the distributional result in (8), we see that, under $H_0 : \rho = \rho_0$,

$$Z_c = \sum_{i=1}^{k} Q_i Z_i \sim N\left(\mu_{\rho_0}, 1/(N-3k)\right), \tag{10}$$

where $Q_i = \frac{(n_i-3)}{\sum_{j=1}^{k}(n_j-3)}$ and $N = \sum_{i=1}^{k} n_i$. Thus, this combined test rejects H_0 in favor of $H_a : \rho > \rho_0$ if $\sqrt{N-3k}(Z_c - \mu_{\rho_0}) > z_{1-\alpha}$, or equivalently the p-value $1 - \Phi(\sqrt{N-3k}(Z_c - \mu_{\rho_0})) < \alpha$.

5.3. Power comparisons

As the powers of all combined tests depend only on the population correlation coefficient ρ, we estimated the powers of all the tests as a function of ρ and reported them in Table 4 for $k = 2$, 4 and 6 and some small to moderate sample sizes. Examination of the estimated powers clearly indicates that the test based on the combined statistic Z_c in (10), say, Z_c tests, outperforms all other tests. Even though, this Z_c test offers only little improvements over the weighted inverse normal test and the inverse χ^2 test, still it is better than other tests for all the cases. The inverse χ^2 test and the weighted inverse normal test are more powerful than the Fisher test. Indeed, the Fisher test is less powerful than all other tests. In general, the Z_c test and the weighted inverse normal test perform very similar, and they are preferable to other tests for a common correlation coefficient of several bivariate normal populations. An advantage of the Z_c test over the weighted

Table 4. Powers of the combined tests for testing common correlation coefficient.

	n = (5,5)					n = (10,5)					n = (10,10)					n = (20,5)				
Tests for $H_0 : \rho = \rho_0$ vs. $H_a : \rho > \rho_0;\ \rho_0 = .1$																				
ρ	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
.1	.048	.048	.047	.047	.046	.051	.051	.048	.051	.049	.052	.053	.051	.051	.052	.050	.049	.048	.050	.050
.2	.074	.073	.074	.074	.074	.093	.088	.089	.094	.094	.107	.104	.107	.107	.106	.120	.108	.111	.119	.121
.3	.111	.107	.114	.114	.117	.162	.151	.158	.166	.168	.205	.197	.209	.209	.211	.240	.212	.216	.242	.248
.4	.168	.161	.175	.175	.177	.266	.247	.263	.275	.279	.355	.340	.362	.362	.363	.425	.380	.385	.429	.433
.5	.251	.239	.266	.266	.264	.418	.390	.413	.431	.432	.551	.531	.560	.560	.564	.654	.603	.601	.656	.660
.6	.372	.356	.395	.395	.395	.606	.575	.599	.618	.620	.761	.742	.769	.769	.770	.856	.821	.812	.856	.859
.7	.536	.517	.562	.562	.564	.803	.778	.795	.813	.814	.919	.909	.922	.922	.921	.966	.954	.945	.965	.968
	n = (5,5,5,5)					n = (10,5,10,5)					n = (15,5,5,5)					n = (30,10,5,5)				
.1	.051	.052	.049	.049	.049	.053	.053	.051	.052	.052	.052	.052	.050	.051	.051	.052	.052	.050	.052	.051
.2	.093	.090	.095	.095	.095	.121	.112	.120	.124	.124	.119	.108	.115	.115	.124	.163	.138	.147	.162	.164
.3	.158	.149	.170	.170	.169	.241	.219	.243	.253	.254	.238	.208	.230	.227	.252	.381	.318	.337	.378	.391
.4	.263	.245	.286	.286	.284	.429	.389	.431	.447	.450	.424	.371	.407	.398	.446	.664	.581	.596	.659	.679
.5	.414	.385	.450	.450	.449	.657	.610	.658	.677	.681	.654	.590	.629	.615	.677	.892	.837	.836	.884	.898
.6	.601	.567	.643	.643	.642	.859	.825	.857	.871	.870	.855	.809	.834	.816	.870	.984	.970	.965	.981	.985
.7	.804	.776	.833	.833	.832	.969	.958	.967	.972	.973	.968	.953	.958	.948	.972	.999	.998	.997	.999	.999
	n = (8,8,8,8,8,8)					n = (20,4,4,14,4,4)					n = (10,10,10,20,20,20)					n = (30,10,10,10,5,5)				
.1	.055	.056	.053	.053	.054	.045	.043	.043	.049	.051	.055	.056	.054	.053	.054	.054	.055	.053	.051	.054
.2	.156	.147	.160	.160	.158	.141	.111	.127	.153	.158	.231	.209	.230	.233	.236	.194	.166	.182	.193	.197
.3	.341	.315	.356	.356	.359	.331	.255	.293	.359	.359	.582	.530	.579	.586	.589	.472	.403	.439	.467	.488
.4	.603	.562	.624	.624	.628	.597	.490	.540	.627	.632	.887	.851	.883	.887	.893	.787	.713	.747	.776	.799
.5	.843	.811	.857	.857	.860	.844	.757	.788	.859	.866	.989	.982	.988	.989	.990	.959	.930	.939	.954	.963
.6	.969	.958	.972	.972	.972	.970	.940	.945	.973	.975	.999	.999	.999	.999	1.00	.997	.994	.994	.996	.998
.7	.998	.996	.998	.998	.998	.998	.995	.994	.998	.998	1.00	1.00	.999	.999	1.00	.999	.999	.999	.999	1.00

1 = Inverse chi-square; 2 = Fisher's test; 3 = Inverse normal; 4 = Weighted inverse normal; 5 = test based on Z_c in (10).

inverse normal test is that the former test yields simple closed-form confidence intervals whereas the confidence intervals based on the weighted inverse normal test can be obtained only by numerically.

5.4. An example

The data for this example are from Tian and Wilding (2008) who has used these data to compute confidence intervals for a common correlation coefficient in three age groups of proposita girls. These data have also been analyzed by many researchers in the context of making inference on interclass and intraclass correlations. The data are measurements on diastolic and systolic blood pressures of proposita girls for age groups 6–8, 9–11, and 12–14 with sample sizes $n_1 = 7$, $n_2 = 6$ and $n_3 = 7$, respectively. The sample correlation coefficients corresponding to the three age groups are $r_1 = 0.7454$, $r_2 = 0.6391$ and $r_3 = 0.7379$. On the basis of Fisher's Z test on the equality of correlation coefficients, Tian has noted that the homogeneity of correlation coefficients among these three groups is tenable.

Let us consider testing the common correlation coefficient ρ using the hypotheses $H_0 : \rho = .30$ vs. $H_a : \rho > .30$. The p-values based on the individual samples are $P_1 = .07865$, $P_2 = .19265$ and $P_3 = .08366$. Fisher's combined statistic is $-2\sum \ln(P_i) = 12.300$ and the p-value is $P(\chi^2_6 > 12.300) = .0556$. The inverse χ^2 statistic is $\sum \chi^2_{n_i;1-P_i} = 32.383$ and the p-value is $P(\chi^2_{20} > 32.383) = .0394$. The inverse normal statistic is $\sum \Phi^{-1}(P_i)/\sqrt{3} = -1.936$ and the p-value is $\Phi(-1.936) = .0264$. The weighted inverse normal statistic is $\sum w_i \Phi^{-1}(P_i)/\sqrt{\sum w_i^2} = -1.961$ and the p-value is $\Phi(-1.961) = .0250$. The test statistic Z_c in (10) is 1.959 with p-value $1 - \Phi(1.959) = .0250$. All the tests, except the Fisher test, indicate that there is enough evidence to support H_a at the level.05. Furthermore, the weighted normal test and the test based on Z_c produced the same smallest p-value.

Finally, we note in this example that the tests based on individual samples do not reject the null hypothesis at the level 0.05, but some combined tests reject it at the same level of significance.

6. Tests for a common mean of several lognormal populations

In order to find a combined test for a common mean of several lognormal populations, we first need to identify a one-sample test for the mean of a lognormal distribution. Let $X_1, ..., X_n$ be a log-transformed sample from a lognormal distribution with parameters μ and σ^2, say, LN(μ, σ). Since the mean of a lognormal distribution is given by $\exp(\mu + \sigma^2/2)$, it is enough to consider a test for $\psi = \mu + \sigma^2/2$. Even though, there are several tests available in the literature (e.g., Krishnamoorthy and Mathew (2003) and Zou, Huo, and Taleban (2009)) we shall describe one of the most accurate tests by Wu, Wong, and Jiang (2003).

6.1. The MLR test

To derive a test for a common mean of lognormal populations, let us first describe the modified likelihood ratio test (MLRT) for a single sample by Wu, Wong, and Jiang (2003). Let $(w_1, w_2) = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right)$. The maximum likelihood estimators (MLEs) are

$$\hat{\sigma}^2 = \bar{w}_2 - \bar{w}_1^2 \quad \text{and} \quad \hat{\psi} = \bar{w}_1 + \frac{1}{2}\hat{\sigma}^2. \tag{11}$$

where $\bar{w}_1 = w_1/n$ and $\bar{w}_2 = w_2/n$. For fixed ψ, the constrained MLE of σ^2 is

$$\hat{\sigma}_\psi^2 = 2\left\{(\psi + 1)^2 + \bar{w}_2 - 2\psi\bar{w}_1 - 2\psi\right\}^{1/2} - 2.$$

The minimum (over ψ) value of the expression within the curly brackets is $1 + \hat{\sigma}^2$ and hence $\hat{\sigma}_\psi^2$ is nonnegative. Define

$$r(\psi) = \text{sgn}(\hat{\psi} - \psi)\left\{n \ln\frac{\hat{\sigma}_\psi^2}{\hat{\sigma}^2} + n\left(\bar{w}_1 - \psi + \frac{1}{2}\hat{\sigma}_\psi^2\right)\right\}^{\frac{1}{2}} \tag{12}$$

and

$$u(\psi) = \sqrt{n}(\hat{\psi} - \psi)\left(\frac{\hat{\sigma}}{\hat{\sigma}_\psi^3}\right) \Big/ \sqrt{\frac{1}{2} + \frac{1}{\hat{\sigma}_\psi^2}}. \tag{13}$$

For testing $H_0 : \psi = \psi_0$ vs. $H_a : \psi > \psi_0$, the MLRT statistic is given by

$$r^*(\psi_0) = r(\psi_0) + \frac{1}{r(\psi_0)}\ln\frac{u(\psi_0)}{r(\psi_0)}, \tag{14}$$

which follows a standard normal distribution asymptotically. This asymptotic result has third-order accuracy, and is valid even for small samples. For testing $H_0 : \psi = \psi_0$ vs. $H_a : \psi > \psi_0$, the p-value of the MLRT is given by

$$P(Z > r^*(\psi_0)) = 1 - \Phi(r^*(\psi_0)). \tag{15}$$

To develop a combined test for a common mean of several lognormal distributions, let us suppose that the means of all the k populations are the same. That is, $\exp\left(\mu_1 + \frac{1}{2}\sigma_1^2\right) = \cdots = \exp\left(\mu_k + \frac{1}{2}\sigma_k^2\right)$, or equivalently, $\psi = \mu_1 + \frac{1}{2}\sigma_1^2 = \cdots = \mu_k + \frac{1}{2}\sigma_k^2$. Let P_i denote the p-value for testing

Table 5. Powers of the combined tests for testing a common mean of lognormal populations.

Tests for $H_0 : \psi = \psi_0$ vs. $H_a : \psi > \psi_0$; $\psi_0 = 1$; $\boldsymbol{\sigma} = (.4, .8)$																				
n = (5,5)					**n** = (10,5)				**n** = (5,10)				**n** = (15,5)				**n** = (5,15)			
ψ	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1.0	.049	.049	.049	.049	.050	.050	.049	.050	.050	.050	.050	.050	.048	.048	.048	.048	.050	.050	.050	.050
1.1	.141	.139	.143	.143	.196	.182	.182	.204	.151	.153	.157	.139	.247	.222	.216	.257	.162	.166	.173	.146
1.2	.303	.298	.304	.304	.475	.446	.425	.496	.339	.349	.361	.303	.614	.567	.524	.634	.373	.386	.404	.319
1.3	.520	.516	.514	.514	.770	.741	.697	.787	.578	.595	.608	.510	.898	.870	.811	.910	.635	.656	.677	.545
1.4	.729	.727	.713	.713	.937	.924	.882	.945	.791	.810	.817	.711	.987	.981	.950	.990	.847	.867	.877	.757
1.5	.880	.880	.859	.859	.990	.987	.965	.992	.923	.935	.935	.858	.999	.999	.991	.999	.953	.965	.967	.893
$\boldsymbol{\sigma} = (.4, .8, 1.2, 1.6)$																				
n = (6,6,6,6)					**n** = (8,9,11,12)				**n** = (12,11,9,8)				**n** = (4,8,12,15)				**n** = (15,12,8,4)			
1.0	.048	.048	.048	.048	.050	.051	.050	.050	.049	.050	.049	.049	.048	.049	.048	.047	.049	.049	.048	.049
1.1	.139	.138	.137	.137	.167	.169	.173	.153	.204	.194	.189	.212	.133	.145	.147	.115	.238	.208	.198	.255
1.2	.314	.312	.303	.303	.394	.411	.401	.344	.514	.497	.459	.521	.288	.319	.328	.228	.601	.548	.483	.633
1.3	.546	.549	.514	.514	.673	.706	.671	.585	.816	.810	.741	.812	.503	.554	.561	.389	.890	.861	.768	.904
1.4	.764	.772	.717	.717	.879	.904	.868	.794	.962	.962	.915	.956	.712	.769	.768	.563	.986	.981	.930	.988
1.5	.907	.916	.862	.862	.970	.982	.961	.917	.996	.996	.981	.993	.870	.913	.905	.729	.999	.999	.985	.999
$\boldsymbol{\sigma} = (.4, .8, 1.2, 1.6, 2.0, 2.4)$																				
n = (4,4,4,4,4,4)					**n** = (8,4,8,4,4,4)				**n** = (4,8,4,8,4,4)				**n** = (10,10,4,4,4,4)				**n** = (4,4,10,10,4,4)			
1.0	.045	.045	.043	.043	.045	.045	.044	.045	.045	.046	.044	.045	.046	.047	.044	.046	.047	.047	.045	.046
1.1	.104	.104	.100	.100	.139	.126	.121	.147	.115	.112	.111	.109	.170	.146	.136	.192	.112	.112	.110	.101
1.2	.208	.207	.197	.197	.324	.293	.260	.341	.242	.237	.227	.222	.421	.358	.302	.467	.229	.234	.227	.190
1.3	.350	.351	.319	.319	.569	.533	.442	.578	.425	.421	.387	.380	.713	.644	.522	.754	.398	.413	.388	.320
1.4	.527	.533	.474	.474	.792	.765	.635	.785	.625	.627	.566	.556	.907	.870	.725	.924	.589	.612	.565	.470
1.5	.692	.702	.615	.615	.928	.917	.791	.914	.797	.803	.727	.717	.981	.970	.867	.984	.760	.788	.728	.622

1 = Inverse chi-square; 2 = Fisher's test; 3 = Inverse normal; 4 = Weighted inverse normal.

$$H_0 : \psi = \psi_0 \quad \text{vs.} \quad H_a : \psi > \psi_0,$$

based on the ith sample. These p-values can be combined using any of the methods described in Sec. 2 to arrive at a single test for the common mean of several lognormal populations.

6.2. Power comparisons

The modified likelihood ratio tests based on individual samples are very accurate in terms of type I error rates even for small samples. So all the combined tests control the type I error rates very close to the nominal level (see Table 5) and so it is fair to compare them in terms power. The powers of the combined tests were estimated using Monte Carlo simulation for $k = 2$, 4 and 6, and presented in Table 5. For each case of k, the values σ and ψ are given in the table. Note that $\mu = \psi - \frac{1}{2}\sigma^2$.

We first observe from Table 5 that the powers of the inverse χ^2 test and the Fisher test are approximately the same and are larger than those of the inverse normal test and the weighted inverse normal test when the sample sizes are equal. See the results for $\mathbf{n} = (5, 5)$, $\mathbf{n} = (6, 6, 6, 6)$ and $\mathbf{n} = (4, 4, 4, 4, 4, 4)$. The weighted inverse normal test dominates all other tests when the sample sizes and the variances are negatively associated; that is, larger sample sizes are associated with smaller variances. See the cases $\mathbf{n} = (10, 5), (15, 5)$; $\mathbf{n} = (12, 11, 9, 8), (15, 12, 8, 4)$; $\mathbf{n} = (10, 10, 4, 4, 4, 4)$. However, all tests seem to dominate the inverse weighted normal test when the smaller variances are associated with smaller sample sizes; see the powers for the cases $\mathbf{n} = (5, 10), (5, 15)$; $\mathbf{n} = (8, 9, 11, 12), (4, 8, 12, 15)$. In general, the powers of the weighted inverse normal test heavily depend on the relations between the variances and the sample sizes. Between the inverse χ^2 test and the weighted inverse normal test, the former is better than the latter when σ_i's are positively associated with sample sizes or when the sample sizes are not very much different, and worse than the latter when the variances are negatively associated with the sample sizes. The gain in power of using the inverse χ^2 test over the weighted inverse normal test or inverse normal test outweighs the loss.

There is no clear-cut winner between the inverse χ^2 tests and the Fisher test. For the cases where sample sizes and the variances are positively associated, the inverse χ^2 test is better than the Fisher test; see the powers when $\mathbf{n} = (10, 5), (15, 5)$ and $\boldsymbol{\sigma} = (.4, .8)$; $\mathbf{n} = (12, 11, 9, 8), (15, 12, 8, 4)$ and $\boldsymbol{\sigma} = (.4, .8, 1.2, 1.6)$. On the other hand, the Fisher test outperforms the inverse χ^2 test when the sample sizes and the variances are negatively associated; see the powers when $\mathbf{n} = (5, 10), (5, 15)$ and $\boldsymbol{\sigma} = (.4, .8)$; $\mathbf{n} = (8, 9, 11, 12), (4, 8, 12, 15)$ and $\boldsymbol{\sigma} = (.4, .8, 1.2, 1.6)$. Even though, no test is uniformly better than others, on an overall basis, the Fisher test and the inverse χ^2 test can be recommended in practical applications.

6.3. An example

We shall illustrate the combined tests for a common mean of several lognormal populations using the example given in Tian and Wu (2007). The data were obtained from an alcohol interaction study in men (Bradstreet and Liss 1995). For illustrative purpose, we shall consider only the measurements on maximum concentration (Cmax) and estimate the common mean of active treatment groups. The group sizes are equal with $n_1 = n_2 = n_3 = 22$. The sample mean $\bar{x}_i$ and the sample variance s^2 of the log-transformed data are $(\bar{x}_1, s_1^2) = (2.601, 0.24)$, $(\bar{x}_2, s_2^2) = (2.596, 0.20)$ and $(\bar{x}_3, s_3^2) = (2.599, 0.17)$ for the three groups.

For illustration purpose, let us test if the common mean is larger than 12.5; that is, $H_0 : \exp(\psi) = 12.5$ vs. $H_a : \exp(\psi) > 12.5$. In terms of ψ, we test $H_0 : \psi = 2.5257$ vs. $H_a : \psi > 2.5257$. The p-values based on the samples 1, 2, and 3 are $P_1 = 0.03245$, $P_2 = 0.03881$, and $P_3 = 0.03751$, respectively. The Fisher test statistic is $-2\sum \ln(P_i) = 19.921$ and the p-value is $P(\chi^2_6 > 19.921) = .0029$. The inverse χ^2 statistic is $\sum \chi^2_{n_i;1-P_i} = 105.858$ and the p-value is $P(\chi^2_{66} > 105.858) = .0013$. Since the sample sizes are the same, the inverse normal statistic and the weighted inverse normal statistic are the same and is $\sum \Phi^{-1}(P_i)/\sqrt{3} = -3.1125$ with p-value is $\Phi(-3.1125) = .0009$. Note that the inverse χ^2 test produced smaller p-value than that of the Fisher test. Among all three tests, the inverse normal test produced the smallest p-value.

7. Tests for a common mean of several gamma distributions

Let $\bar{X}$ and $\tilde{G}$ denote respectively the arithmetic mean and geometric based on a sample of size n from a gamma distribution with the shape parameter a and the scale parameter b, say, gamma(a, b). The mean of the gamma(a, b) distribution is given by μ. In the following, we shall describe the modified likelihood ratio test (MLRT) for the mean $\mu = ab$ by Fraser, Reid, and Wong (1997) as given in Krishnamoorthy and Len-Novelo (2014).

7.1. The MLR test

The log-likelihood function is expressed as

$$l(a, b|\bar{X}, \tilde{G}) = -n \ln \Gamma(a) - na \ln b - n\bar{X}/b + (a-1)n \ln \tilde{G}. \tag{16}$$

The MLE $\hat{a}$ is the solution of the equation $\ln(a) - \psi(a) = \ln(\bar{X}/\tilde{G})$, where ψ is the digamma function. The MLE can be evaluated by the Newton-Raphson iterative scheme

$$a_{\text{new}} = a_{\text{old}} - \frac{\ln a_{\text{old}} - \psi(a_{\text{old}}) - s}{1/a_{\text{old}} - \psi'(a_{\text{old}})} \quad \text{with} \quad a_{\text{old}} = \frac{3 - s + \sqrt{(s-3)^2 + 24s}}{12s},$$

where $s = \ln(\bar{X}/\tilde{G})$ and $\psi'(x) = \partial\psi(x)/\partial x$ is the trigamma function. The MLE of b is $\hat{b} = \bar{X}/\hat{a}$. The signed likelihood ratio test (SLRT) statistic is given by

$$r(\mu_0) = \text{sign}(\hat{\mu} - \mu_0)\left\{2\left[l(\hat{a}, \hat{b}|\bar{X}, \tilde{G}) - l(\hat{a}_{\mu_0}, \mu_0|\bar{X}, \tilde{G})\right]\right\}^{1/2}, \tag{17}$$

where $l(a, b|\bar{X}, \tilde{G})$ is the log-likelihood function in (16),

$$l(a, \mu|\bar{X}, \tilde{G}) = -n\ln\Gamma(a) - na\ln(a/\mu) - na\bar{X}/\mu + (a-1)n\ln\tilde{G},$$

and $\hat{a}_{\mu_0}$ is the MLE of a at $\mu = \mu_0$. This constrained MLE $\hat{a}_{\mu_0}$ is obtained as the root of the equation $\ln a - \psi(a) = \ln(\mu_0/\tilde{G}) + \bar{X}/\mu_0 - 1$.

The modified LRT by Fraser, Reid, and Wong (1997) is given by

$$r^*(\mu_0) = r(\mu_0) - \frac{1}{r(\mu_0)}\ln\left(\frac{r(\mu_0)}{Q(\mu_0)}\right), \tag{18}$$

where $R(\mu_0)$ is defined in (17), and $Q(\mu_0) = \sqrt{n\hat{a}}(\hat{\mu}/\mu_0 - 1)(\psi'(\hat{a}) - 1/\hat{a})^{\frac{1}{2}}/(\psi'(\hat{a}_{\mu_0}) - 1/\hat{a}_{\mu_0})^{\frac{1}{2}}$. This MLRT has third-order accuracy in the sense that the standard normal approximation to the distribution of MLRT(μ_0) is accurate up to $O(n^{-3/2})$. The MLRT for $H_0 : \mu = \mu_0$vs. $H_a : \mu > \mu_0$, rejects the null hypothesis if the p-value $P(Z > r^*(\mu_0)) < \alpha$, where Z is the standard normal random variable.

7.2. Power comparisons

To judge the accuracy and powers of the combined tests, which are based on the p-values of the individual MLRT in the preceding section, for testing a common mean of several gamma populations, we estimated powers using Monte Carlo simulation. The powers of the tests for $H_0 : \mu = 0.5$ vs. $H_a : \mu > 0.5$ were estimated at the level 0.05 for values of $k = 2$, 4 and 6 and various values of sample sizes. The values of the parameters are chosen as $a = 1, 2, 3, 4, 5$ and 6 and $b_i = \mu/a_i$, $i = 1, ..., 6$. The estimated powers are reported in Table 6.

Comparison of the powers of the inverse χ^2 test, inverse normal test and the weighted inverse normal test, we see that the powers of the inverse χ^2 test are quite similar to those of other tests for some cases (see $\mathbf{n} = (5, 5), (5, 10)$ and (5, 20); $\mathbf{n} = (5, 10, 15, 20)$; $\mathbf{n} =$ (5,4,7,10,4,14)). For all other cases, the inverse χ^2 test is better than the other two tests. Between the inverse χ^2 test and the Fisher test, the former is better than the latter in most cases. These two tests have similar powers for some cases; see $\mathbf{n} = (20, 15, 10, 15)$ and $(15, 5, 5, 5)$. Overall, the inverse χ^2 test is preferable to other tests.

7.3. An example

To illustrate the combined tests for a common mean of several gamma distributions, we consider the data given in Table 1 of Bhaumik and Gibbons (2006). The data represent vinyl chloride concentrations (in g/L) collected from clean upgradient monitoring wells. A quantile-quantile plot by Bhaumik and Gibbons showed an excellent fit of these data to a gamma distribution. Krishnamoorthy and Len-Novelo (2014) have used the data to outline some inferential methods for gamma distributions. The data set includes 34 measurements, and we divided them randomly into two samples so the first sample includes $n_1 = 20$ measurements and the second one includes $n_2 = 14$ measurements as shown below (Table 7). Let us test if the common mean concentrations is less than 2.5 g/L; that is $H_0 : \mu = 2.5$ vs. $H_a : \mu < 2.5$. The p-values of the MLRT based on samples 1 and 2 are $P_1 = 0.07101$ and $P_2 = 0.31297$, respectively. The Fisher test statistic is $-2\sum\ln(P_i) = 7.613$ and the p-value is $P(\chi^2_4 > 7.613) = .1068$. The inverse χ^2 statistic is

Table 6. Powers of the combined tests for testing a common mean of several gamma populations.

$H_0: \mu = \mu_0$ vs. $H_a: \mu > \mu_0$; $\mu_0 = 0.5$; $a = (1,2,3,4,5,6)$; $b_i = \mu/a_i$																				
n = (5,5)					**n** = (10,5)				**n** = (5,10)				**n** = (20,5)				**n** = (5,20)			
Tests for $H_0: \mu = \mu_0$ vs. $H_a: \mu > \mu_0$; $\mu_0 = 0.5$																				
μ	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4				
0.5	.052	.052	.052	.052	.052	.052	.052	.052	.052	.051	.052	.052	.049	.050	.049	.049	.050	.051	.052	.050
0.6	.100	.099	.100	.100	.114	.113	.112	.110	.133	.128	.124	.136	.153	.147	.142	.150	.217	.193	.178	.224
0.7	.202	.201	.199	.199	.263	.262	.260	.254	.331	.315	.299	.338	.394	.380	.366	.381	.571	.524	.469	.585
0.8	.347	.344	.341	.341	.457	.457	.452	.439	.566	.543	.509	.571	.654	.641	.624	.634	.840	.806	.733	.848
0.9	.500	.497	.490	.490	.638	.641	.635	.613	.755	.737	.691	.758	.835	.828	.811	.812	.954	.939	.885	.957
1.0	.635	.634	.622	.622	.776	.778	.772	.751	.873	.861	.816	.874	.929	.925	.913	.912	.988	.984	.954	.989
n = (5,5,5,5)					**n** = (15,5,5,5)				**n** = (5,5,5,15)				**n** = (5,10,15,20)				**n** = (20,15,10,5)			
0.5	.053	.054	.053	.053	.052	.053	.053	.051	.052	.051	.053	.051	.050	.050	.050	.050	.050	.051	.050	.050
0.6	.153	.151	.149	.149	.178	.179	.175	.163	.286	.250	.229	.302	.441	.407	.370	.449	.278	.283	.269	.252
0.7	.418	.412	.406	.406	.501	.509	.497	.445	.752	.699	.622	.772	.937	.921	.877	.940	.761	.771	.752	.702
0.8	.711	.707	.689	.689	.801	.812	.799	.727	.961	.944	.885	.966	.998	.997	.992	.999	.968	.973	.966	.941
0.9	.890	.887	.871	.871	.943	.951	.943	.891	.996	.993	.974	.996	.999	.999	.999	.999	.998	.998	.997	.992
1.0	.965	.964	.953	.953	.986	.989	.986	.959	.999	.999	.995	.999	1.00	1.00	1.00	1.00	.999	1.00	.999	.999
n = (4,4,4,4,4,4)					**n** = (8,4,5,11,7,6)				**n** = (10,4,4,14,4,4)				**n** = (5,4,7,10,4,14)				**n** = (5,4,15,4,8,4)			
0.5	.057	.058	.057	.057	.051	.052	.051	.051	.055	.056	.056	.053	.053	.053	.052	.052	.054	.056	.055	.053
0.6	.190	.186	.188	.188	.330	.318	.302	.321	.311	.286	.270	.305	.435	.389	.353	.447	.335	.307	.289	.331
0.7	.544	.534	.532	.532	.858	.846	.813	.837	.817	.793	.749	.795	.941	.919	.871	.943	.852	.823	.786	.836
0.8	.850	.844	.834	.834	.991	.991	.982	.986	.984	.981	.964	.975	.999	.998	.992	.998	.990	.987	.975	.985
0.9	.970	.968	.960	.960	.999	.999	.999	.999	.999	.999	.997	.998	1.00	1.00	.999	1.00	.999	1.00	.998	.999
1.0	.995	.995	.992	.992	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.999	1.00	1.00	1.00	1.00	1.00	1.00	.999	1.00

1 = Inverse Chi-square; 2 = Fisher's test; 3 = Inverse normal; 4 = Weighted inverse normal.

Table 7. Vinyl chloride concentrations in monitoring wells.

Sample 1	1.3	0.4	0.1	2.5	0.6	1.2	0.9	0.2	0.8	5.3
	1.8	0.6	1.1	2.3	0.2	0.1	3.2	1.0	8.0	2.0
Sample 2	5.1	2.4	0.5	6.8	1.2	0.5	0.9	0.4	0.4	2.7
	0.5	2.0	2.9	4.0						

$\sum \chi^2_{n_i;1-P_i} = 45.94$ and the p-value is $P(\chi^2_{34} > 45.94) = .083$. The inverse normal statistic $\sum \Phi^{-1}(P_i)/\sqrt{2} = -1.383$ with p-value $\Phi(-1.383) = .083$. The weighted inverse normal statistic is -1.482 with the p-value $\Phi(-1.482) = .069$. Once again we see that the inverse χ^2 test and the inverse normal test produced smaller p-value than that of the Fisher test. Among all four tests, the weighted inverse normal test produced the smallest p-value and the Fisher test produced the largest.

8. Conclusion

Our simulation studies clearly indicated that none of the tests dominates others for all sample size and parameter configurations. For unequal sample sizes, the inverse χ^2 test appears to be better than other tests when the variances are not drastically different. We also noted that the weighted z-score test does not improve the corresponding unweighted test uniformly. The weighted z-score test or the Fisher test could be improved by choosing weights that are inversely proportional to population variances and directly proportional to the sample sizes. As the population variances are unknown, one could use the sample variances to weight the individual tests. However, the null distribution of such a combined test with weights depend on sample variances is difficult to obtain except for the normal case. We also note that, even for the normal case, the combined test in Sec. 3.1, does not dominate other tests uniformly. On the basis of our extensive power comparison studies, we see that there is no single combined test that dominates others for all the problems. However, the inverse χ^2 test can be used as an alternative test in place of the Fisher test. The weighted inverse normal test may be preferable to the inverse normal test when the sample sizes are unequal.

We also observe from illustrative examples that if any of the tests based on individual samples rejects the null hypothesis then almost all combined tests also reject the null hypothesis. Furthermore, we noted in Example 5.4 that no individual test rejects the null hypothesis at the nominal level 0.05, but all except the Fisher test, reject the null hypothesis at the same level. The results of these examples clearly indicate that the combined tests, which are based on information from all independent sources or studies, are more powerful than the tests based on individual samples.

Acknowledgment

The authors are grateful to two reviewers for providing valuable comments and suggestions.

ORCID

K. Krishnamoorthy http://orcid.org/0000-0002-3919-918X

References

Bhaumik, D. K., and R. D. Gibbons. 2006. One-sided approximate prediction intervals for at least p of m observations from a gamma population at each of r locations. *Technometrics* 48 (1):112–9. doi:10.1198/004017005000000355.

Bradstreet, T. E., and C. L. Liss. 1995. Favorite data sets from early (and late) phases of drug research – Part 4. In *Proceedings of the section on statistical education of the American Statistical Association*

Donner, A., and B. Rosner. 1980. On inferences concerning a common correlation coefficient. *Applied Statistics* 29 (1):69–76. doi:10.2307/2346412.

Eberhardt, K. R., C. P. Reeve, and C. H. Spiegelman. 1989. A minimax approach to combining means, with practical examples. *Chemometrics and Intelligent Laboratory Systems* 5 (2):129–48. doi:10.1016/0169-7439(89)80009-7.

Fisher, R. A. 1932. *Statistical methods for research workers*. Edinburgh: Oliver and Boyd.

Fraser, D. A. S., N. Reid, and A. Wong. 1997. Simple and accurate inference for the mean of the gamma model. *Canadian Journal of Statistics* 25 (1):91–9. doi:10.2307/3315359.

Fung, W. K, and T. S. Tsang. 1998. A simulation study comparing tests for the equality of coefficients of variation. *Statistics in Medicine* 17 (17):2003–14. doi:10.1002/(SICI)1097-0258(19980915)17:17<2003::AID-SIM889>3.0.CO;2-I.

Johnson, N. L, and B. L. Welch. 1940. Application of the noncentral t-distribution. *Biometrika* 31 (3–4):362–89. doi:10.2307/2332616.

Jordan, S. M, and K. Krishnamoorthy. 1995. On combining independent tests in linear models. *Statistics & Probability Letters* 23 (2):117–22. doi:10.1016/0167-7152(94)00102-E.

Jordan, S. M, and K. Krishnamoorthy. 1996. Exact confidence intervals for the common mean of several normal populations. *Biometrics* 52 (1):77–87. doi:10.2307/2533146.

Kifle, Y. G, and B. K. Sinha. 2021. Comparison of local powers of some exact tests for a common normal mean with unequal variances. In *Strategic management, decision theory, and decision science*, eds. B. K. Sinha and S. B. Bagchi, Singapore: Springer. doi:10.1007/978-981-16-1368-5_6.

Krishnamoorthy, K, and L. Leoń-Novelo. 2014. Small sample inference for gamma parameters: One-sample and two-sample problems. *Environmetrics* 25 (2):107–26. doi:10.1002/env.2261.

Krishnamoorthy, K, and T. Mathew. 2003. Inferences on the means of lognormal distributions using generalized p-values and generalized confidence intervals. *Journal of Statistical Planning and Inference* 115 (1):103–21. doi:10.1016/S0378-3758(02)00153-2.

Krishnamoorthy, K, and Y. Xia. 2007. Inferences on correlation coefficients: One-sample, independent and correlated cases. *Journal of Statistical Planning and Inference* 137 (7):2362–79. doi:10.1016/j.jspi.2006.08.002.

Paul, S. R. 1988. Estimation and testing significance for a common correlation coefficient. *Communications in Statistics – Theory and Methods* 17 (1):39–53. doi:10.1080/03610928808829608.

Plesch, W, and P. Klimpel. 2002. Performance evaluation of the CoaguChek S system. *Haematologica* 87 (5):557–9.

Rice, W. R. 1990. A consensus combined *P*-value test and the family-wide significance of component tests. *Biometrics* 46 (2):303–8. doi:10.2307/2531435.

Stouffer, S. A. E. A. Suchman, L. C. DeVinney, S. A. Star, and R. M. Williams. Jr. 1949. *The American soldier 1 adjustment during army life*. Princeton, NJ: Princeton University Press.

Tian, L. 2005. Inferences on the common coefficient of variation. *Statistics in Medicine* 24 (14):2213–20. doi:10.1002/sim.2088.

Tian, L, and G. E. Wilding. 2008. Confidence interval estimation of a common correlation coefficient. *Computational Statistics & Data Analysis* 52 (10):4872–7. doi:10.1016/j.csda.2008.04.002.

Tian, L, and J. Wu. 2007. Inferences on the common mean of several log-normal populations: the generalized variable approach. *Biometrical Journal. Biometrische Zeitschrift* 49 (6):944–51. doi:10.1002/bimj.200710391.

Whitlock, M. C. 2005. Combining probability from independent tests: The weighted *Z*-method is superior to Fisher's approach. *Journal of Evolutionary Biology* 18 (5):1368–73. doi:10.1111/j.1420-9101.2005.00917.x.

Wu, J., A. C. M. Wong, and G. Jiang. 2003. Likelihood-based confidence intervals for a log-normal mean. *Statistics in Medicine* 22 (11):1849–60. doi:10.1002/sim.1381.

Zhou, L, and T. Mathew. 1993. Combining independent tests in linear models. *Journal of the American Statistical Association* 88 (422):650–5. doi:10.1080/01621459.1993.10476318.

Zou, G. Y., C. Y. Huo, and J. Taleban. 2009. Simple confidence intervals for lognormal means and their differences with environmental applications. *Environmetrics* 20 (2):172–80. doi:10.1002/env.919.