Biometrical Letters
Vol. 45(2008), No. 2, 57-64

Influence of the averaging observations on ellipsoidal confidence regions in bivariate normal samples

Joanna Tarasińska

Department of Applied Mathematics and Computer Science, University of Life Sciences in Lublin, Akademicka 13, 20-934 Lublin, Poland, e-mail:joanna.tarasinska@up.lublin.pl

SUMMARY

It is assumed that a large number of observations from a bivariate normal population are given. These can be used for classical statistical inference about the mean. Sometimes the investigator averages data and makes the inference based on this “sample of means”. Such an averaging procedure, when not justified by the non-normality of the data, causes loss of information.
The aim of this paper is to establish by how much the quality of an ellipsoidal confidence region based on the “sample of means” is inferior compared with the “raw sample”.

Key words: bivariate normal distribution, confidence region for mean

1. Introduction

Most classic statistical methods of multivariate analysis are based on the assumption that data have multivariate normal distribution. There are many tests investigating this assumption, in fact more than fifty; see for example Mecklin and Mundfrom (2004). When there is lack of normality, according to the central limit theorem, a fixed number of data can be averaged and the inference made on this “sample of means”. On the other hand when the data have normal distribution such an averaging procedure is superfluous, and can even degrade statistical inference, as it causes loss of information. In this paper it is shown by how much the quality of an ellipsoidal confidence region based on the “sample of means” can be inferior compared with the “raw sample”.

Let us assume we have a random sample of size $n = m \cdot k$ from a bivariate normal distribution:

$$\mathbf{X}_{ij} \sim N_2(\boldsymbol{\mu},\boldsymbol{\Sigma}) \quad i.i.d. \quad i=1\ldots m, \quad j=1\ldots k\,. \tag{1}$$

Now, let us assume that instead of the "raw" sample $\mathbf{X}_{ij}$ we are given only m arithmetic means

$$\overline{\mathbf{X}}_{i\bullet} = \frac{1}{k}\sum_{j=1}^{k} \mathbf{X}_{ij} \quad i=1\ldots m,$$

$$\mathbf{Y}_i = \overline{\mathbf{X}}_{i\bullet} \sim N_2\left(\boldsymbol{\mu},\frac{1}{k}\boldsymbol{\Sigma}\right) \quad i=1\ldots m. \tag{2}$$

The aim of this paper is to establish how much the "loss of information" caused by averaging in model (2) influences the quality of estimation of $\boldsymbol{\mu}$. The same problem in the univariate case was discussed in Tarasińska (2003).

2. Results

The point estimates of $\boldsymbol{\mu}$ in both models (1) and (2) have the same values, as

$$\hat{\boldsymbol{\mu}} = \overline{\mathbf{X}} = \frac{1}{mk}\sum_{i=1}^{m}\sum_{j=1}^{k} \mathbf{X}_{ij} = \frac{1}{m}\sum_{i=1}^{m}\overline{\mathbf{X}}_{i\bullet} = \overline{\mathbf{Y}}\,.$$

However the situation is quite different with the ellipsoidal confidence regions for $\boldsymbol{\mu}$, or more precisely with their areas. The 100(1-α)% confidence region for $\boldsymbol{\mu}$ under model (1) based on Hotteling's T² is the ellipsoid

$$\left\{\boldsymbol{\mu} : n(\overline{\mathbf{X}}-\boldsymbol{\mu})'\mathbf{S}^{-1}(\overline{\mathbf{X}}-\boldsymbol{\mu}) \le \frac{2(n-1)}{n-2}\cdot F_{2,n-2,\alpha}\right\} \tag{3}$$

where

$$\mathbf{S} = \hat{\boldsymbol{\Sigma}} = \frac{1}{n-1}\sum_{i=1}^{m}\sum_{j=1}^{k}\left(\mathbf{X}_{ij}-\overline{\mathbf{X}}\right)\left(\mathbf{X}_{ij}-\overline{\mathbf{X}}\right)'\,.$$

$F_{n_1,n_2,\alpha}$ is the upper $\alpha \cdot 100\%$ point of the F distribution with (n_1,n_2) degrees of freedom.

The 100(1-α)% confidence region for μ under model (2) is

$$\left\{\boldsymbol{\mu} : m(\overline{\mathbf{X}}-\boldsymbol{\mu})'\overline{\mathbf{S}}^{-1}(\overline{\mathbf{X}}-\boldsymbol{\mu}) \le \frac{2(m-1)}{m-2}\cdot F_{2,m-2,\alpha}\right\} \tag{4}$$

where

$$\overline{\mathbf{S}} = \frac{1}{m-1}\sum_{i=1}^{m}\left(\mathbf{Y}_i - \overline{\mathbf{Y}}\right)\left(\mathbf{Y}_i - \overline{\mathbf{Y}}\right)' .$$

Let us compare the areas of ellipsoids (3) and (4). The area of ellipsoid (3) is given by

$$A = 2\pi\frac{n-1}{n(n-2)}\cdot F_{2,n-2,\alpha}\cdot\sqrt{|\mathbf{S}|}$$

while the area of ellipsoid (4) is

$$\overline{A} = 2\pi\frac{m-1}{m(m-2)}\cdot F_{2,m-2,\alpha}\cdot\sqrt{|\overline{\mathbf{S}}|}$$

Because (Johnson and Kotz, 1972)

$$\frac{|(n-1)\mathbf{S}|}{|\mathbf{\Sigma}|} \sim \chi^2_{n-1}\cdot\chi^2_{n-2} \ ,$$

where χ^2_{n-1} and χ^2_{n-2} are independent χ^2 variables with *n*-1 and *n*-2 degrees of freedom we can write

$$A \sim \frac{2\pi}{n(n-2)}\cdot F_{2,n-2,\alpha}\cdot\sqrt{|\mathbf{\Sigma}|}\cdot\chi_{n-1}\cdot\chi_{n-2}$$

and analogously

$$\overline{A} \sim \frac{2\pi}{n(m-2)}\cdot F_{2,m-2,\alpha}\cdot\sqrt{|\mathbf{\Sigma}|}\cdot\chi_{m-1}\cdot\chi_{m-2}$$

Using the equalities

$$E(\chi_p) = \sqrt{2}\frac{\Gamma\left(\frac{p+1}{2}\right)}{\Gamma\left(\frac{p}{2}\right)} \quad \text{and} \quad \Gamma\left(\frac{p}{2}\right) = \left(\frac{p}{2}-1\right)\Gamma\left(\frac{p}{2}-1\right)$$

we have

$$E(\chi_{n-1})E(\chi_{n-2}) = \frac{\sqrt{2}\,\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \cdot \frac{\sqrt{2}\,\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n}{2}-1\right)} = 2\left(\frac{n}{2}-1\right) = n-2$$

and in the same manner $E(\chi_{m-1})E(\chi_{m-2}) = m-2$.
Thus finally we get

$$\frac{E(\overline{A})}{E(A)} = \frac{n-2}{m-2} \cdot \frac{F_{2,m-2,\alpha}}{F_{2,n-2,\alpha}} \cdot \frac{E(\chi_{m-1})E(\chi_{m-2})}{E(\chi_{n-1})E(\chi_{n-2})} = \frac{F_{2,m-2,\alpha}}{F_{2,n-2,\alpha}}\,.$$

In the case of variance we have

$$\begin{aligned} Var(A) &= \frac{4\pi^2}{n^2(n-2)^2} F^2_{2,n-2,\alpha}|\Sigma| \cdot [E(\chi^2_{n-1})E(\chi^2_{n-2}) - (E(\chi_{n-1}))^2 (E(\chi_{n-2}))^2] = \\ &= \frac{4\pi^2}{n^2(n-2)^2} F^2_{2,n-2,\alpha}|\Sigma| \cdot [(n-1)(n-2) - (n-2)^2] = \\ &= \frac{4\pi^2}{n^2(n-2)^2} F^2_{2,n-2,\alpha}|\Sigma|(n-2) = \frac{4\pi^2}{n^2(n-2)} F^2_{2,n-2,\alpha}|\Sigma| \end{aligned}$$

and

$$Var(\overline{A}) = \frac{4\pi^2}{n^2(m-2)} F^2_{2,m-2,\alpha}|\Sigma|\,.$$

Thus we obtain

$$\frac{Var(\overline{A})}{Var(A)} = \frac{n-2}{m-2} \cdot \frac{F^2_{2,m-2,\alpha}}{F^2_{2,n-2,\alpha}}\,.$$

Table 1 contains the quotients

$$\frac{E(\overline{A})}{E(A)} \text{ and } \frac{Var(\overline{A})}{Var(A)}$$

in the case n=100 and different numbers of averaged observations for model (2) i.e. k=2, 5, 10, 20. The upper values in the cells are for a 90% confidence ellipsoid, the middle ones for a 95% confidence ellipsoid, and the lower ones for a 99% confidence ellipsoid.

Table 1. $\frac{E(\overline{A})}{E(A)}$ and $\frac{Var(\overline{A})}{Var(A)}$ for *n*=100, *k*=2, 5, 10, 20

k	2	5	10	20
$\frac{E(\overline{A})}{E(A)}$	1.025	1.113	1.320	2.317
	1.033	1.151	1.443	3.092
	1.051	1.245	1.791	6.382
$\frac{Var(\overline{A})}{Var(A)}$	2.145	6.744	21.360	175.368
	2.178	7.208	25.522	312.327
	2.257	8.443	39.305	1330.591

Now let us consider the distribution of $Q = A/\overline{A}$. We have

$$(n-1)\mathbf{S} = k(m-1)\overline{\mathbf{S}} + \sum_{i=1}^{m}\sum_{j=1}^{k}\left(\mathbf{X}_{ij} - \overline{\mathbf{X}}_{i\cdot}\right)\left(\mathbf{X}_{ij} - \overline{\mathbf{X}}_{i\cdot}\right)'$$

and

$$Q = \frac{A}{\overline{A}} = \frac{m-2}{n-2} \cdot \frac{F_{2,n-2,\alpha}}{F_{2,m-2,\alpha}} \cdot \frac{1}{\sqrt{L}} ,$$

where $\sqrt{L} = \frac{k(m-1)}{n-1}\sqrt{\frac{|\overline{\mathbf{S}}|}{|\mathbf{S}|}}$ and $\frac{1-\sqrt{L}}{\sqrt{L}} \cdot \frac{m-2}{m(k-1)} \sim F_{2m(k-1),2(m-2)}$ (Srivastava, 2002; Johnson and Kotz, 1972, p.202). Hence the probability density function (p.d.f.) of Q is $f(x) = \frac{1}{a} g\left(\frac{x-b}{a}\right)$, where $g(\cdot)$ is the p.d.f. of $F_{2m(k-1),2(m-2)}$, $a = \frac{m(k-1)}{n-2} \cdot \frac{F_{2,n-2,\alpha}}{F_{2,m-2,\alpha}}$, $b = \frac{m-2}{n-2} \cdot \frac{F_{2,\, n-2,\alpha}}{F_{2,\, m-2,\alpha}}$.

Figure 1 shows the p.d.f. of Q for a 95% confidence ellipsoid, *n*=100 and different *k*. Figure 2 shows the p.d.f. of Q for *n*=100, *k*=10 and three confidence levels (α=0.01, 0.05, 0.1).

If we are interested in $Pr(\overline{A} > A)$ we have

$$Pr(\overline{A} > A) = G\left[\left(\frac{n-2}{m-2} \cdot \frac{F_{2,m-2,\alpha}}{F_{2,n-2,\alpha}} - 1\right) \cdot \frac{m-2}{m(k-1)}\right],$$

where G is the c.d.f. of $F_{2m(k-1),2(m-2)}$. Table 2 gives the probabilities of $\overline{A} > A$ for *n*=100; *k*=2, 5, 10, 20; α=0.1, 0.05, 0.01.

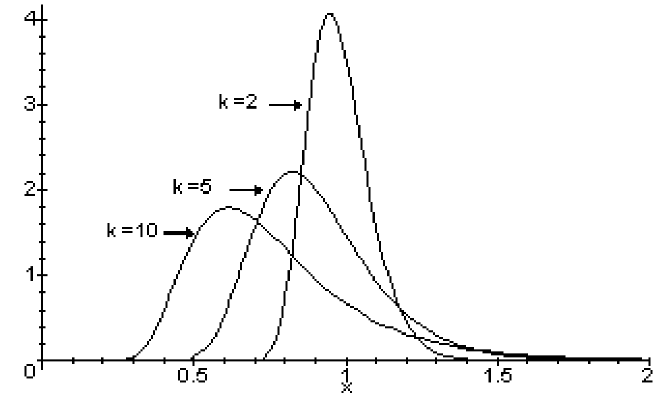

Figure 1. The p.d.f.s of Q for n=100; α=0.05; k=2, 5, 10

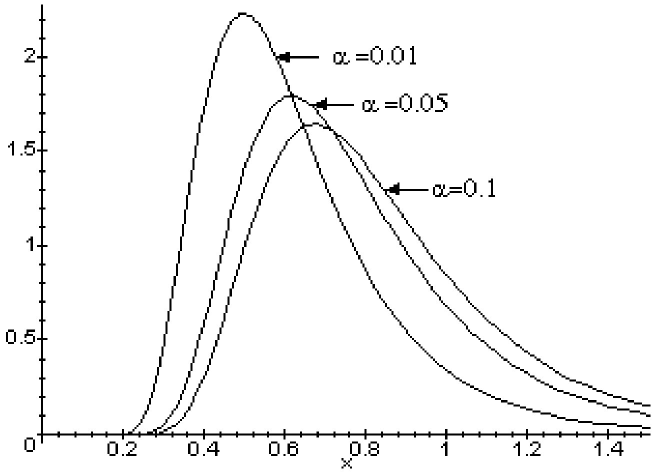

Figure 2. The p.d.f.s of Q for n=100; k=10; α=0.1, 0.05, 0.01

Table 2. $Pr\left(\overline{A} > A\right)$ for n=100

k α	2	5	10	20
0.1	0.593	0.668	0.749	0.862
0.05	0.620	0.719	0.817	0.928
0.01	0.682	0.821	0.925	0.988

3. Conclusions

1. Averaging a small number of observations causes hardly any increase in the expected area of the ellipsoidal confidence region for the vector of means (see *n*=100, *k*=2 in Table 1). However, it does have a great influence on the variance of the area. The probability of enlargement of the area is also significant (see Table 2).
2. The greater the confidence level, the greater is the influence of averaging on the area (see Figure 2 and Table 2).

4. Example

As an example let us consider a part of Fisher's famous data on Iris Setosa (Fisher, 1936). The part being considered contains 50 observations for sepal length and sepal width. The hypothesis of bivariate normal distribution of data is not rejected, as the *p*-values for tests based on Mardia's measures of skewness and kurtosis are, respectively, ≈ 1 and 0.849.

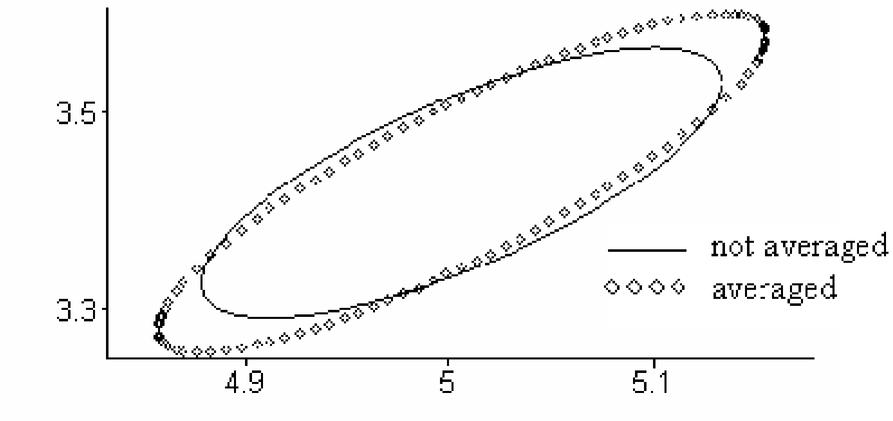

Figure 3. 95% confidence regions for μ

Figure 3 presents 95% confidence regions for $\boldsymbol{\mu}$, one based on the original 50 pairs of observations and the other on 25 pairs of averaged data (hence $k = 2$). The areas of the regions are $A \approx 0.0366$ and $\overline{A} \approx 0.0402$. Therefore averaging (unjustified, because the data are normal) enlarged the confidence region. With $n = 50$ and $k = 2$ the probability of enlargement is 0.672.

REFERENCES

Fisher R.A. (1936): The use of multiple measurements in taxonomic problems. Annals of Egenics 7: 179–188.

Johnson N.L., Kotz S. (1972): Distributions in statistics: continuous multivariate distributions. John Wiley & Sons, New York.

Mecklin C.J., Mundfrom D.J. (2004): An appraisal and bibliography of tests for multivariate normality. International Statistical Review 72/1: 123–138.

Srivastava M.S. (2002): Methods of multivariate statistics. John Wiley & Sons.

Tarasinska J. (2003): On certain consequences of averaging observations derived from normal population. Tatra Mt. Math. Publ. 26: 391–397.