THE UNIVERSITY OF MINNESOTA

Statistics 5401 September 16, 2005

Statistical Distances

Figures 1 – 4 on the following pages contain scatter plots of artificially generated data. Each plot displays random samples from two bivariate normal populations A and B with the same covariance matrices ($\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$), but differing mean vectors $\boldsymbol{\mu}_A = [\mu_{XA}, \mu_{YA}]'$ and $\boldsymbol{\mu}_B = [\mu_{XB}, \mu_{YB}]'$. The figures differ in mean vectors and correlation coefficient ρ_{XY}. Here are the characteristics of the populations in each figure:

Figure	μ_{XA}	μ_{YA}	μ_{XB}	μ_{YB}	ρ_{XY}	δ
1	95	95	110	110	0	2.828
2	95	110	110	95	0	2.828
3	95	95	110	110	0.9	2.052
4	95	110	110	95	0.9	8.944

The quantity δ is the *multistandardized* or *covariance standardized* (Mahalanobis) distance between the population means defined by

$$\delta = \{(\boldsymbol{\mu}_A - \boldsymbol{\mu}_B)'\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_A - \boldsymbol{\mu}_B)\}^{1/2}$$

In every case the population marginal standard deviations are $\sigma_X = \sigma_Y = 7.5$. Note that in Figures 1 and 2, X and Y are independent ($\rho_{XY} = 0$), while in Figures 3 and 4 they are fairly strongly positively correlated ($\rho_{XY} = .9$).

In all four cases the individual means satisfy $|\mu_{XA} - \mu_{XB}|/\sigma_X = |\mu_{YA} - \mu_{YB}|/\sigma_y = 2$, and hence, standardized distances between the *univariate* population means are the same for both X and Y in every plot.

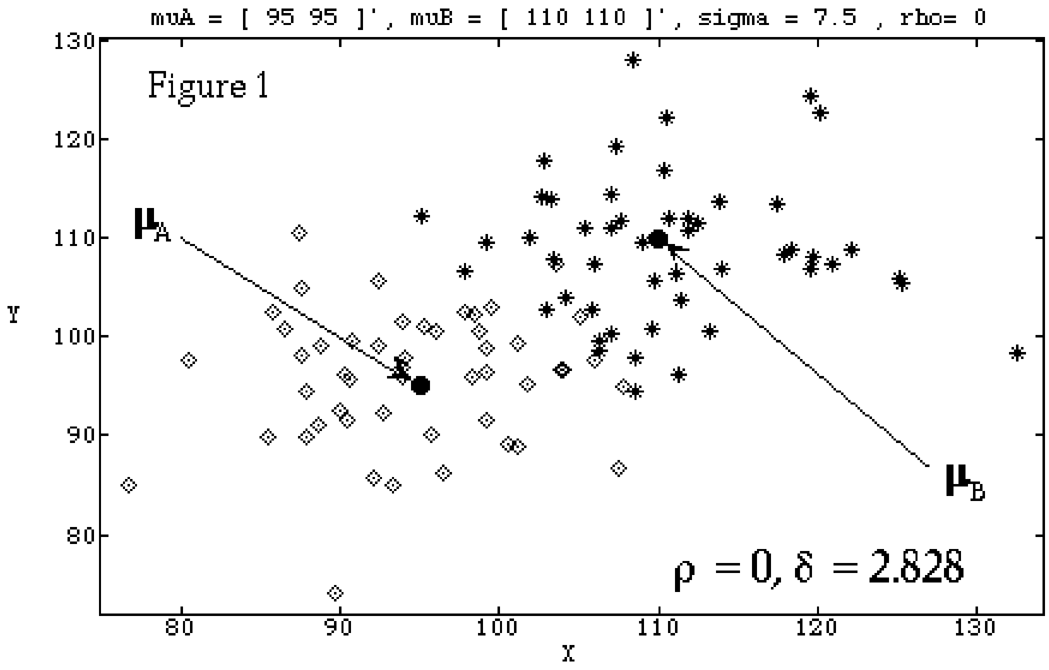
Also, in all four cases, the *Euclidean* distances (straight line) between the two population means vectors are the same, that is,

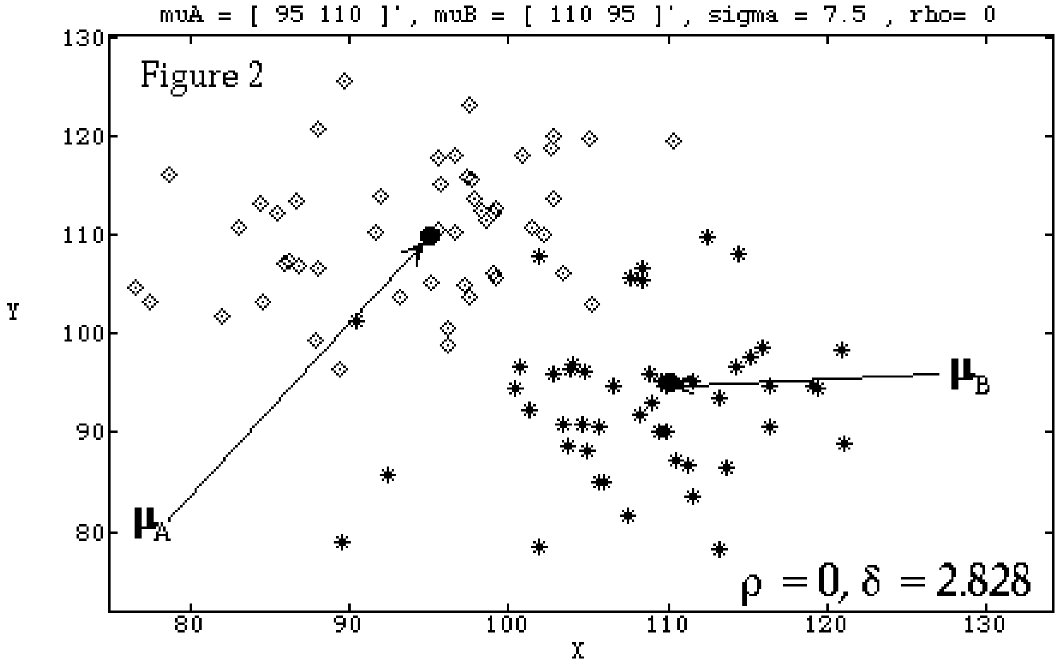
$$||\boldsymbol{\mu}_A - \boldsymbol{\mu}_B|| \equiv \{(\mu_{XA} - \mu_{XB})^2 + (\mu_{YA} - \mu_{YB})^2\}^{1/2} = 15\sqrt{2} = 21.213.$$

If X and Y were each *standardized* by dividing by their standard deviations, σ_X and σ_Y, the Euclidean distances between the means of the standardized variables are the same in all four groups, namely

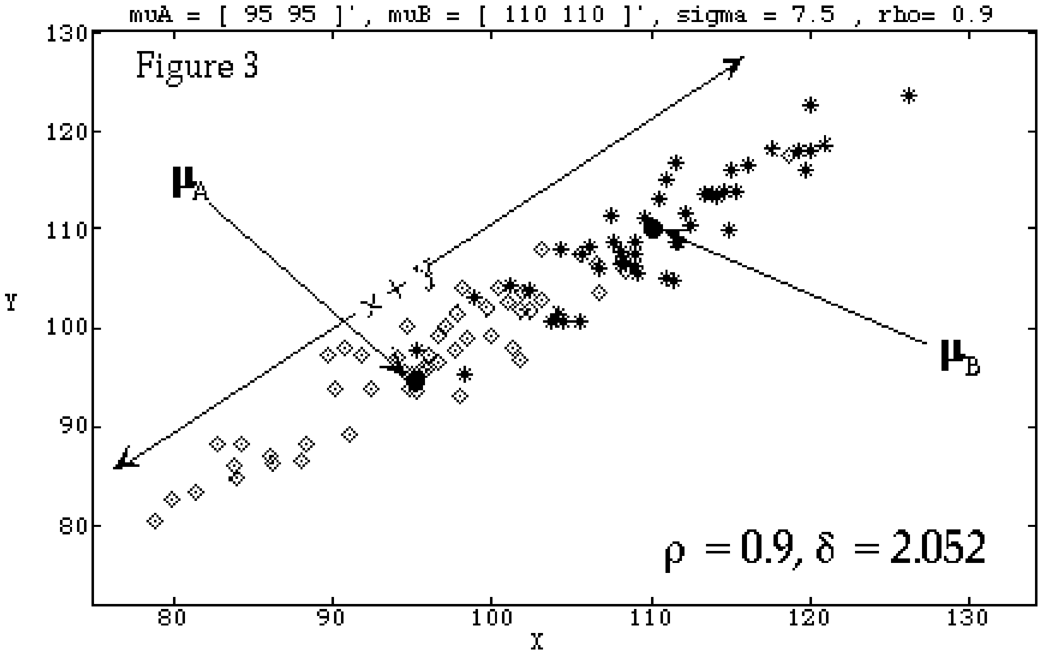
$$\{[(\mu_{XA} - \mu_{XB})/7.5]^2 + ([(\mu_{YA} - \mu_{YB})/7.5]^2\}^{1/2} = 2\sqrt{2} = 2.828.$$

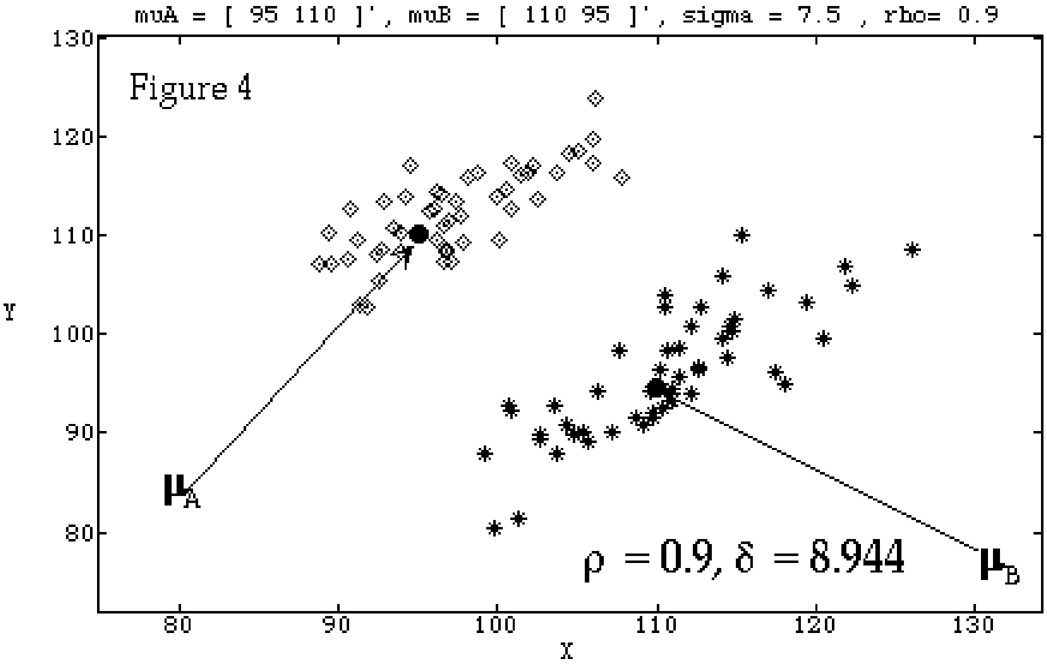
Note that there are substantial variation among the figures in the Mahalanobis distance δ between the two population means. δ ranges from 2.052 in Figure 3 to 8.944 in Figure 4.





There is a a fair amunt of overlap between the observations from the two populations in both these figres. If the same plotting symbol had been used for both samples you might not guess that there were samples from two distinct populations.





There is more overlap of the sample in Figure 3 for which the Mahalanobis distance δ is smaller than for Figures 1 and 2.

However, in Figure 4, the two samples do not overlap at all. Even with just one plotting symbol, you would almost certainly guess tha there were samples from two quite distinct populations. This reflects the fact that the population Mahalanobis distance δ is much larger for Figure 4 than for the other figures.

I'll try to clarify the differences among these four cases.

Suppose $W_c = c_1X+c_2Y = \mathbf{c}'\mathbf{y}$ is an arbitrary linear combination of X and Y, where $\mathbf{y} = [X, Y]'$ with means μ_{WA} and μ_{WB} and $\mathbf{c} = [c_1 , c_2]'$. Then let γ_c be the standardized distance between the means of W_c for populations A and B. That is, because $\mathbf{c}'\mathbf{\Sigma}\mathbf{c} = \sigma_W{}^2 = \mathrm{Var}[\mathbf{c}'\mathbf{y}\ ]$,

$$\begin{aligned}\gamma_c &\equiv |\mu_{WA} - \mu_{WB}|/\sigma_{W_c} = |\mathbf{c}'(\boldsymbol{\mu}_A - \boldsymbol{\mu}_B)|/(\mathbf{c}'\mathbf{\Sigma}\mathbf{c})^{1/2} \\ &= |c_1(\mu_{XA} - \mu_{XB})+c_2(\mu_{YA} - \mu_{YB})|/\{c_1{}^2\sigma_X{}^2 + 2c_1c_2\rho_{XY}\sigma_X\sigma_Y + c_2{}^2\sigma_Y{}^2\}^{1/2}.\end{aligned}$$

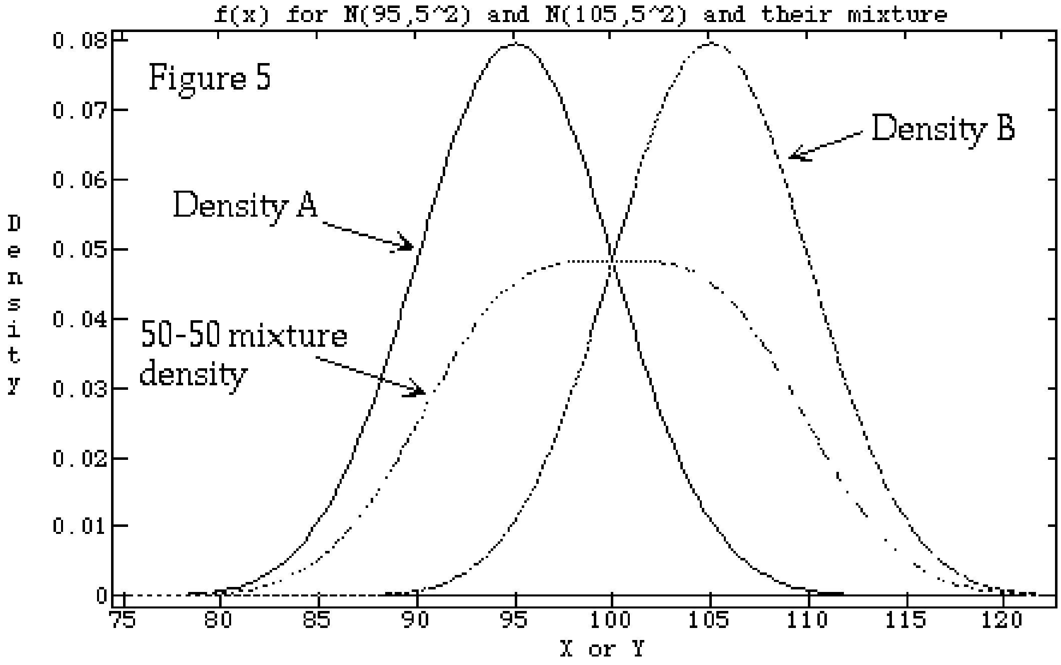
Mathematics shows that the Mahalanobis distance $\delta = \{(\boldsymbol{\mu}_A - \boldsymbol{\mu}_B)'\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_A - \boldsymbol{\mu}_B)\}^{1/2}$ is the largest value that γ_c can have for any possible choice of c_1 and c_2.

In Figures 1 and 2, for every choice of c_1 and c_2, $\gamma_c = \delta = 2\sqrt{2} = 2.828$. No linear combinations has a difference of means as large as 3 standard deviations.

In Figure 3, $c_1 = c_2 = 1$, $W_c = X+Y$ gives the maximum $\gamma_c = \delta = 2.052$. This says that the means of the most separated linear combination differ between the two populations by only a little more than 2 standard deviations. Other linear combinations differ by less. In particular when $c_1 = 1$ and $c_2 = -1$, $\delta = 0$ and the two samples of $X - Y$ completely overlap.

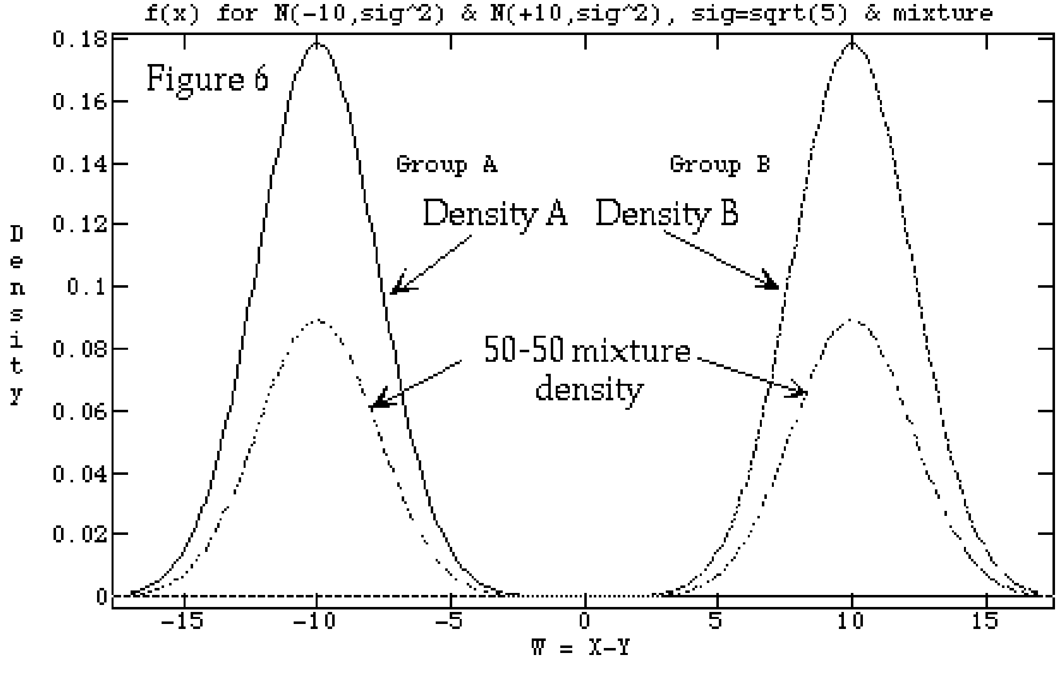
In Figure 4, the choice of $c_1 = 1$ and $c_2 = -1$, gives $\gamma_c = 20/\sqrt{5} = 8.94 = \delta$ and the visible separation is in the "direction" defined by $W = X - Y$. In a sense, the Mahalanobis distance automatically finds the direction of greatest separation. The means of $W_c = X - Y$ differ by almost 9 times the standard deviation of W_c, and thus there is a large separation between the populations in the direction determined by the vector $\mathbf{c} = [1, -1]'$.

Figure 5 shows why it is so hard to separate the two groups based solely on univariate means. Plotted are the separate densities $N(95, 10^2)$ and $N(105, 10^2)$ as well as the density for a population which is 50-50 mixture of these two populations. From the separate densities we see there is a lot of overlap of the separate populations; from the flat-topped density for the mixture it would be hard to guess even that it represents a combination of two distinct populations, let alone to characterize the populations.



The flat top curve labeled "50-50 mixture density" is a weighted combination of the two normal curves above it, each curve getting weight 1/2.

Figure 6 is a similar plot for W = X – Y for the situation in Figure 4. The density of the mixture has two widely separated modes, each of which well represents the density for A or B.



The lower curves labeled "Mixture density" are parts of a single curve that is a weighted combination of the two normal curves above it, each curve getting weight 1/2.