J4 - The Multivariate Normal Distribution

Large Sample behaviour of $\overline{X}$ and S

It turns out that certain multivariate statistics, like $\overline{X}$ and S, have large-sample properties analogous to their univariate counterparts.

Theorem (Law of large numbers). *Let $Y_1, Y_2, ..., Y_n$ be independent observations from a population with mean $\mathbb{E}(Y_i) = \mu_i$. Then*

$$\overline{Y} = \frac{Y_1 + Y_2 + ... + Y_n}{n} \xrightarrow{p} \mu \qquad \text{for } n \to \infty,$$

that is $\overline{Y}$ converges in probability to μ.

Remark. *Convergence in probability means in our case that for every prescribed accuracy $\epsilon > 0$*

$$P(|\overline{Y} - \mu| < \epsilon) \longrightarrow 1 \qquad \text{as } n \to \infty.$$

Theorem (The central limit theorem). *Let $X_1, X_2, ..., X_n$ be independent observations from any population with mean $\mu \in \mathbb{R}^p$ and finite covariance $\Sigma \in \mathbb{R}^{p \times p}$. Then*

$$\sqrt{n}(\overline{X} - \mu) \overset{a}{\sim} \mathcal{N}_p(0, \Sigma)$$

for large sample sizes n.

A corollary of the central limit theorem is that $n(\overline{X} - \mu)^T \Sigma^{-1} (\overline{X} - \mu)$ is approximately χ^2_p (replacing Σ by S does not seriously affect this approximation).

Assessing the Assumption of Normality

Based on the central limit theorem, in situations where the sample size is large and the techniques depend on the behavoir of $\overline{X}$ or $n(\overline{X} - \mu)^T \Sigma^{-1} (\overline{X} - \mu)$ the assumption of normality is made.

There may be cases where the sample size isn't large enough and our assumption of normaility is violated. We should address these questions in order to check if our assumption was appropriate:

(a) Do the marginal distributions of the elements of X appear to be normal?
You can use a Q-Q-plot to assess the assumption of normality and you could do a correlation coefficient test for normality.

(b) Do the scatter plots of pairs of observations on different characteristics give the elliptical appearance expected from normal populations?

(c) Are there any "wild" observations that should be checked for accuracy?

Detecting Outliers and Cleaning Data

Outliers are best detected visually whenever this is possible.

Steps for Detecting Outliers:

(a) Make a dot plot for each variable.

(b) Make a scatter plot for each pair of variables.

(c) Calculate the standardized values $z_{jk} = (x_{jk} - \bar{x}_k)/\sqrt{s_{kk}}$ for $j = 1, 2, ..., n$ and each column $k = 1, 2, ..., p$. Examine these standardized values for large or small values.

(d) Calculate the generalized squared distances $(x_j - \bar{x})^T S^{-1}(x_j - \bar{x})$. In a chi-square plot, these would be the points farthest from the origin.

Transformations to Near Normality

If normality is not a viable assumption, there are some transformations to make the data more "normal looking".

Definition (Box-Cox-Transformation)**.** *Let $x > 0$ be an observation. The Box-Cox-Transformation of x is given by*

$$x^{(\lambda)} = \begin{cases} \dfrac{x^\lambda - 1}{\lambda} & \textit{für } \lambda \neq 0 \\ \ln(x) & \textit{für } \lambda = 0 \end{cases}$$

The Box-Cox solution for the choice of an appropriate power λ is the solution that maximizes the expression

$$l(\lambda) = -\frac{n}{2} \ln\left[\frac{1}{n}\sum_{j=1}^{n}(x_j^{(\lambda)} - \overline{x_j^{(\lambda)}})^2\right] + (\lambda - 1)\sum_{j=1}^{n} \ln x_j.$$

or in the multivariate case

$$l(\lambda_1, \lambda_2, ..., \lambda_p) = -\frac{n}{2} \ln|S(\lambda)| + (\lambda - 1)\sum_{j=1}^{n} \ln x_{j1} + (\lambda_2 - 1)\sum_{j=1}^{n} \ln x_{j2} + ... + (\lambda_p - 1)\sum_{j=1}^{n} \ln x_{jp},$$

where $S(\lambda)$ is the sample covariance matrix computed from the transformed observations $x_j^{(\lambda)}$.

As the second term is very difficult to maximize, it is a good practical approach to maximize the first equation for each variable separately. In other words, you make each marginal distribution approximately normal. Although normal marginals are not sufficient to ensure the joint distribution is normal, in practice this may be good enough.