STAT 530 – Applied Multivariate Statistics

- We will study methods of analyzing multivariate data.
  1. What is a Multivariate Data Set?
  2. Displaying Multivariate Data
  3. Summarizing Multivariate Data
  4. Numerous Exploratory Multivariate Data Analyses
  5. Some Inferential Methods
- This course will be primarily applied, with a bit of theory to explain/justify the methods.

What is a Multivariate Data Set?

- A data set in which several variables are measured on each sampled unit is *multivariate*.
- *Notation*: We have n units ("individuals") and q variables in a multivariate data set, where $q > 1$.
- The observations can be represented in matrix form:

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1q} \\ \vdots & & \ddots & \\ x_{n1} & x_{n2} & \cdots & x_{nq} \end{bmatrix}$$

Types of Measurement

- The variables in a multivariate data set may be measured on different *levels*.
- Four levels of measurement are possible:
  1. *Nominal:* Categorical variables with no meaningful order (Examples: Gender, Hair color)
  2. *Ordinal:* Categorical variables where a meaningful order exists (Examples: Social class, Rating of instructor)
  3. *Interval:* Numerical variables where taking differences is meaningful, but there is no fixed zero position (Examples: Temperature using Celsius/Fahrenheit)
  4. *Ratio:* Numerical variables where taking ratios is meaningful since there is a fixed zero (Examples: Age, Height, Weight)
- We should take care to use statistical analyses that are appropriate for the measurements we have.

Missing Data Issues

- With many variables being measured, we may have observations with missing values for some variables.
- One option: **Complete-case analysis** – Deleting any observations that have any missing values.
- Problem: With many variables being measured, this can lead to **a lot** of observations being deleted and highly reduced sample size.
- Another problem: Can lead to biased estimates unless the missing data are *missing completely at random*.
- A better solution: Multiple imputation, which "fills in" missing values in a sound way (accounts for the extra uncertainty this induces)

Summary Statistics for Multivariate Data: The Mean Vector

- *Mean vector*: For a multivariate data set with q variables $X_1, \ldots, X_q$ and n units, the *population* mean vector $\boldsymbol{\mu}$ is:

$$\boldsymbol{\mu} = \left[\mu_1, \ldots, \mu_q\right]'$$

  where $\mu_j = E(X_j), j = 1, \ldots, q$.

- An estimate of $\boldsymbol{\mu}$ is the *sample* mean vector $\bar{\boldsymbol{x}}$:

$$\bar{\boldsymbol{x}} = \left[\bar{x}_1, \ldots, \bar{x}_q\right]'$$

  where $\bar{x}_j = n^{-1} \sum_{i=1}^{n} x_{ij}$, simply the sample mean of the j-th variable.

- This can be calculated in `R` with the `colMeans` function.

Summary Statistics for Multivariate Data: The Covariance Matrix

- For a multivariate data set with q variables $X_1, \ldots, X_q$ and n units, the *population* covariance matrix $\mathbf{\Sigma}$ is:

$$\begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1q} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{q1} & \sigma_{q2} & \cdots & \sigma_{qq} \end{bmatrix}$$

- The diagonal elements of $\mathbf{\Sigma}$ are the variances of the q variables: $\sigma_{jj} = \sigma_j^2 = E[(X_j - \mu_j)^2]$.

- The off-diagonal elements of $\mathbf{\Sigma}$ are the covariances between two of the variables: $\sigma_{ij} = cov(X_i, X_j) = E[(X_i - \mu_i)(X_j - \mu_j)]$.

Summary Statistics: The Sample Covariance Matrix

- The *population* covariance matrix $\boldsymbol{\Sigma}$ is estimated by *sample* covariance matrix $\mathbf{S}$:

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\boldsymbol{x}})(\mathbf{x}_i - \bar{\boldsymbol{x}})'.$$

- The diagonal elements of $\mathbf{S}$ are the sample variances of the q variables.
- The off-diagonal elements of $\mathbf{S}$ are the sample covariances between two of the variables.
- The sample covariance matrix can be found in `R` using the `var` function or the `cov` function.

Summary Statistics for Multivariate Data: The Correlation Matrix

- Covariances can be difficult to interpret, so often it is useful to work with the *correlation*, which is always between -1 and +1.
- Any covariance σ_{ij} can be standardized into a correlation ρ_{ij} by:

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$$

- The *correlation matrix* $\boldsymbol{\rho}$ for a multivariate data set is one whose diagonal elements are 1's and whose off-diagonals are the respective correlation values ρ_{ij}.
- The *sample correlation matrix* replaces ρ_{ij} with the sample correlation coefficient and can be found by $\mathbf{R} = \mathbf{D}^{-1/2}\mathbf{S}\mathbf{D}^{-1/2}$ where $\mathbf{D}^{-1/2}$ is the diagonal matrix with diagonal elements $1/s_j, j = 1, \ldots, q$ (note s_j^2 is the j-th diagonal element of $\mathbf{S}$).
- The sample correlation matrix can be found in `R` using the `cor` function.
- A covariance matrix can be converted to a correlation matrix using the `cov2cor` function.

Distances

- The distance between two multivariate observations is important in several common analyses.
- The *Euclidean distance* between the i-th and j-th multivariate observations is

$$d_{ij} = \Big[\sum_{k=1}^{q} (x_{ik} - x_{jk})^2\Big]^{1/2}.$$

- If the q variables are measured on quite different scales, it makes sense to *standardize* each variable before calculating distances.
- Sometimes we also consider the distance between two *variables*.

The Multivariate Normal Distribution

- Some multivariate analyses assume the data follow a *multivariate normal distribution*.
- Such assumptions should be checked when dealing with sample data.
- With univariate data, normality is often checked via a quantile-quantile (Q-Q) plot.
- This plots ordered sample values against the corresponding quantiles of a normal distribution.
- For multivariate data, we could do separate Q-Q plots for each variable, but . . .
- the fact that each variable is marginally normally distributed *does not imply* that the multivariate data has a multivariate normal distribution.

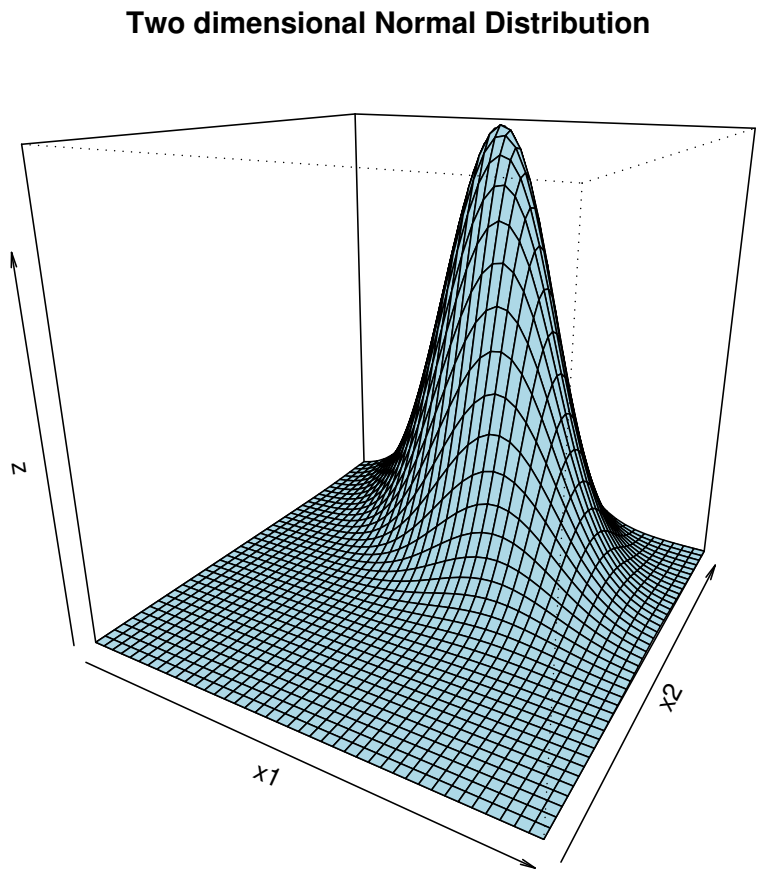

Figure 1: A 3-D plot of a bivariate normal density.

Chi-square Plots

- A better plot to check multivariate normality is a *chi-square plot*
- Consider the generalized (Mahalanobis) distance of each multivariate observation from the mean vector $\bar{\boldsymbol{x}}$:

$$d_i^2 = (\mathbf{x}_i - \bar{\boldsymbol{x}})' \mathbf{S}^{-1} (\mathbf{x}_i - \bar{\boldsymbol{x}}), i = 1, \ldots, n.$$

- If the data are truly multivariate normal, these distances have a χ^2 distribution with q degrees of freedom.
- So we can plot these d_i^2 values against the corresponding χ_q^2 quantiles.
- If the resulting plot is roughly linear, it is reasonable that the data follow a multivariate normal distribution.
- This plot can be done with the `chisplot` function from the text's website.

Transformations to Attain Normality

- Certain analyses require multivariate normality for their conclusions to be valid.
- If our data do not appear to be multivariate normal, we can try to transform some (or all) of the variables.
- We hope that the transformed data will be closer to normal.
- A common class of transformations is the Box-Cox Transformation. For each variable ($j = 1, \ldots, q$):

$$x_j^* = \begin{cases} \frac{(x_j)^{\lambda_j} - 1}{\lambda_j}, & \lambda_j \neq 0 \\ \ln(x_j), & \lambda_j = 0. \end{cases}$$

- Note that this incorporates a wide range of power-type transformations, as well as a log-transformation.
- The powers $\lambda_1, \lambda_2, \ldots, \lambda_q$ could be chosen subjectively, or preferably objectively by choosing these powers to maximize a normal likelihood criterion.

Goals of Multivariate Analysis

- Some multivariate data analyses are *exploratory*, in which the investigator is merely seeking patterns in the data and looking for compelling behavior.
- Many exploratory methods involve data summary, data reduction, and graphics.
- Other multivariate data analyses are *confirmatory*, in which the investigator has a well-defined hypothesis he or she wishes to test.
- Here, inference such as a significance test can be useful.
- Many confirmatory analyses have specific conditions required for the conclusions to be valid.