# Module 18 Multivariate Analysis for Genetic data
# Session 13 Multivariate Normal Distribution

Jan Graffelman

jan.graffelman@upc.edu

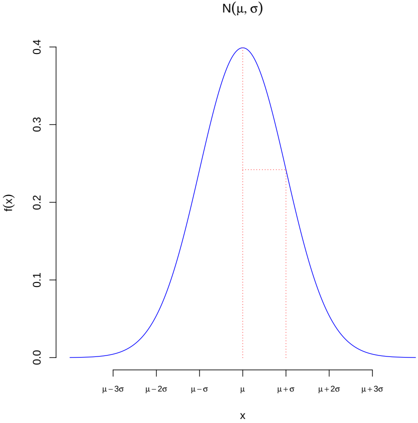
[1]Department of Statistics and Operations Research
Universitat Politècnica de Catalunya
Barcelona, Spain

[2]Department of Biostatistics
University of Washington
Seattle, WA, USA

26th Summer Institute in Statistical Genetics (SISG 2021)

# Contents

# Multivariate Normal Distribution



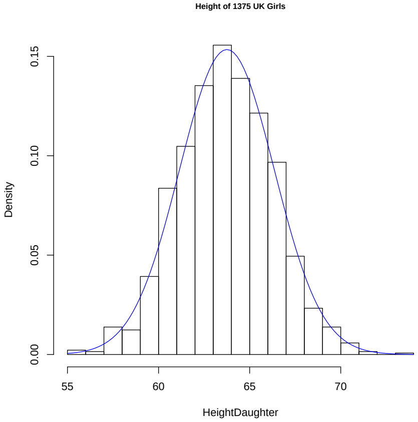

$$X \sim N(\mu, \sigma)$$

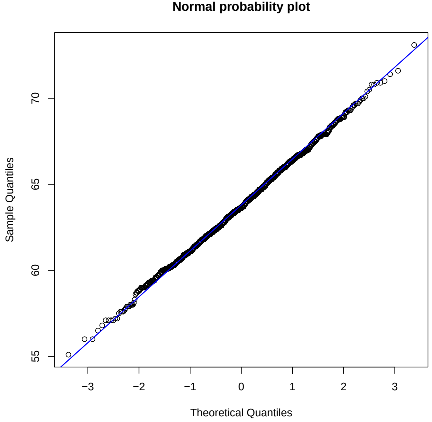
$$f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$E(X) = \mu$$

$$V(X) = \sigma^2$$

# Some normal data (Height UK girls in 1903)

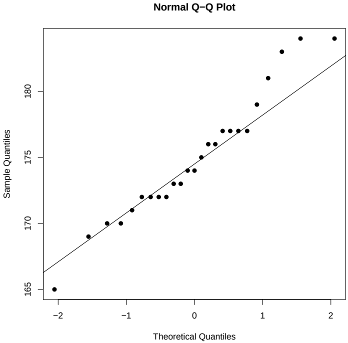


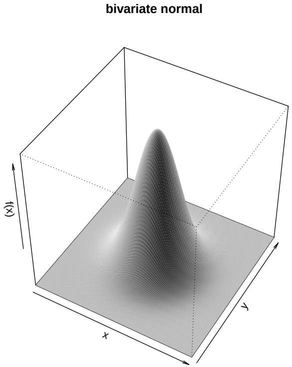

| | N | N* | Mean | Stdev | Med | Q1 | Q3 | Min | Max |
|---|---|---|---|---|---|---|---|---|---|
| Height | 1375 | 0 | 63.751 | 2.6 | 63.6 | 62 | 65.6 | 55.1 | 73.1 |

# Normal probability plot

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | ... | 25 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Height | 172 | 174 | 183 | 175 | 176 | 184 | 177 | 169 | 172 | ... | 172 |
| Sorted | 165 | 169 | 170 | 170 | 171 | 172 | 172 | 172 | 172 | ... | 184 |
| Rank $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | ... | 25 |
| $\frac{i-0.5}{n}$ | 0.02 | 0.06 | 0.10 | 0.14 | 0.18 | 0.22 | 0.26 | 0.30 | 0.34 | ... | 0.98 |
| $z_{(i-0.5)/n}$ | -2.05 | -1.55 | -1.28 | -1.08 | -0.92 | -0.77 | -0.64 | -0.52 | -0.41 | ... | 2.05 |



# Some bivariate normal distributions



# Density multivariate normal

$$f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Exponent univariate normal

$$-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2 = -\frac{1}{2}(x-\mu)(\sigma^2)^{-1}(x-\mu)$$

Exponent multivariate normal

$$-\frac{1}{2}(\mathbf{x}-\mu)'\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)$$

# Multivariate normal distribution

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\mu)'\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)}$$

Parameters:

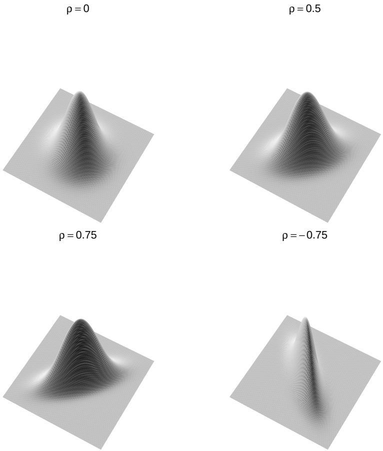
- Population mean vector:

$$\mu = (\mu_1, \mu_2, \ldots, \mu_p)$$

- Population variance-covariance matrix:

$$Cov(\mathbf{X}) = \mathbf{\Sigma}_{p\times p} = E\left((\mathbf{X}-\mu)(\mathbf{X}-\mu)'\right) = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix}$$

# Bivariate normal distribution



# Parameter estimation

Maximum likelihood estimator for $\mu$:

$$\hat{\mu} = \bar{\mathbf{x}} = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p)$$

Maximum likelihood estimator for $\boldsymbol{\Sigma}$:

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

In practice, $\mathbf{S}_{n-1}$ is often used to estimate $\boldsymbol{\Sigma}$:

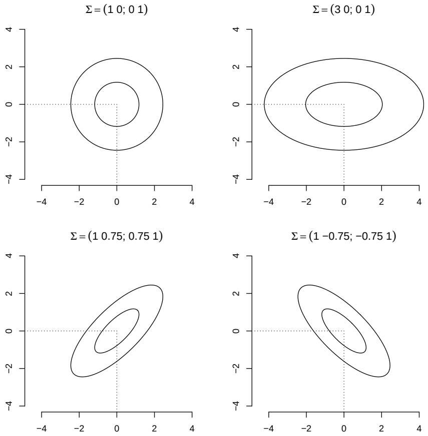
$$\mathbf{S}_{n-1} = \frac{n}{n-1} \mathbf{S}_n$$

# Some Properties of MVN random variates

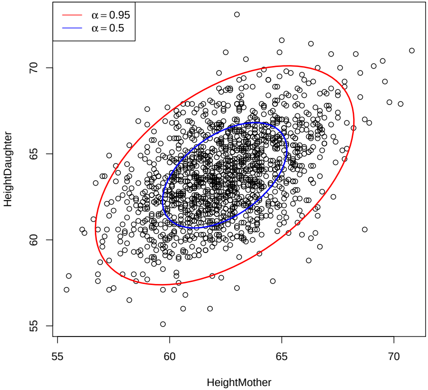
Let $\mathbf{X}$ be a $p \times 1$ random vector, and $\mathbf{X} \sim N_p(\mu, \mathbf{\Sigma})$.

- Linear combinations of the components of $\mathbf{X}$ are normally distributed.
- Basic result: if $\mathbf{X} \sim N_p(\mu, \mathbf{\Sigma})$ and $\mathbf{A} q \times p$, then $\mathbf{AX} \sim N_q(\mathbf{A}\mu, \mathbf{A\Sigma A}')$
- Subsets of components have a (multivariate) normal distribution.
- Components with covariance zero $\Leftrightarrow$ components are independent.
- Conditional distributions of components are (multivariate) normal.

# Contours of normal densities (0.50 and 0.95)



# Contours for empirical data



# Assessing multivariate normality

Some basic ideas:

- Individual variables (marginal distributions) should have bell-shaped (normal) histograms
- Bivariate scatterplots should have clouds of points with an elliptic shape
- Some outliers can be expected, in particular in larger samples

# $\chi^2$ plot for multivariate normality

$$(\mathbf{x} - \mu)' \mathbf{\Sigma}^{-1} (\mathbf{x} - \mu) \sim \chi^2_p$$

The ellipsoid traced by $\mathbf{x}$ described by

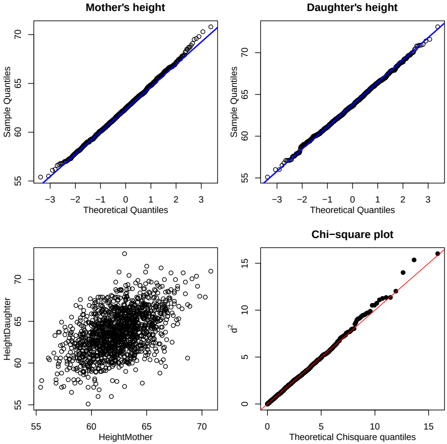
$$(\mathbf{x} - \mu)' \mathbf{\Sigma}^{-1} (\mathbf{x} - \mu) \leq \chi^2_p (1 - \alpha)$$

should contain $100 \cdot (1 - \alpha)\%$ of the observations.

For sample data:

1. Calculate $d_i^2 = (\mathbf{x}_i - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}})$
2. Order the distances from small to large
3. Calculate the rank $(i - \frac{1}{2})/n$
4. Calculate corresponding quantiles $q_i$ according to a $\chi^2_p$ distribution.
5. Plot $(d_i^2, q_i)$
6. Compare with a reference line with intercept 0 and slope 1

# Example $\chi^2$ plot for multivariate normality



# Normality of the NIST STR data

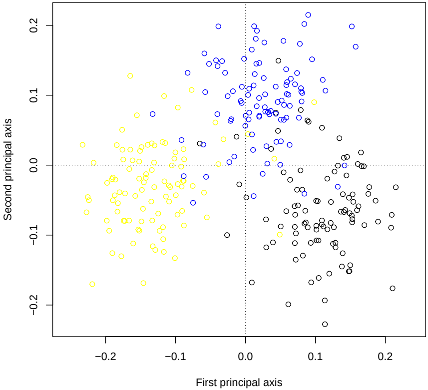
The data:

- 29 autosomal STRs
- Consider individuals with African-American, Asian and Caucasian ancestry
- Sample sizes balanced by subsampling
- STRs coded as binary variables
- Quantification of the data by MDS based on Jaccard metric

The multivariate normality of the MDS principal axis was seen to be important for:

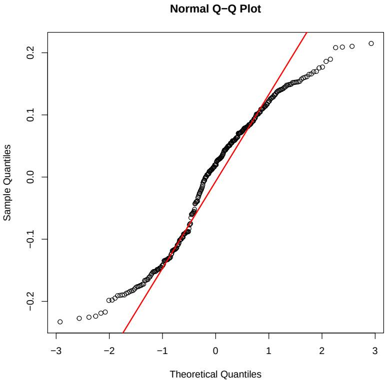
- Model-based clustering
- Linear and quadratic discriminant analysis
- ....

# Normality of the NIST STR data

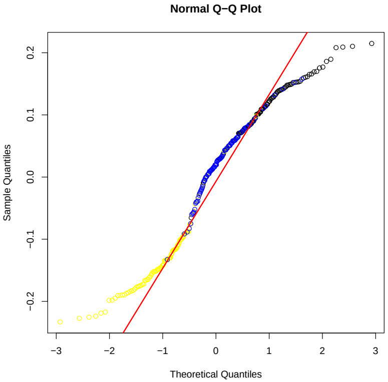



Multivariate normal?

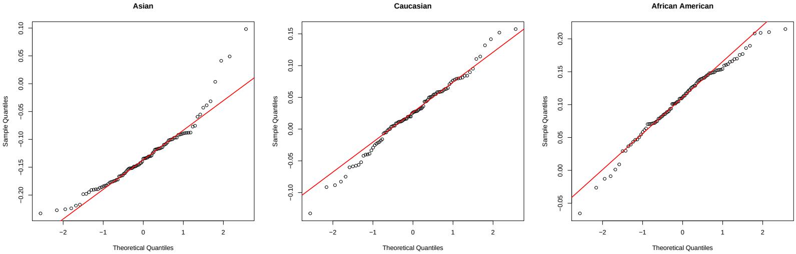
# Normality of the NIST STR data



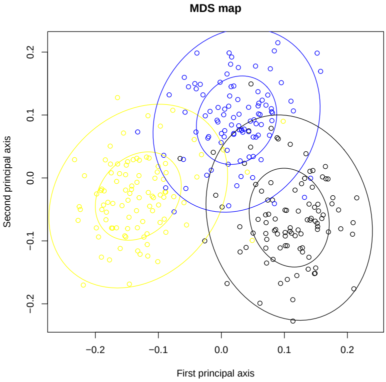
# Normality of the NIST STR data



# Stratifying

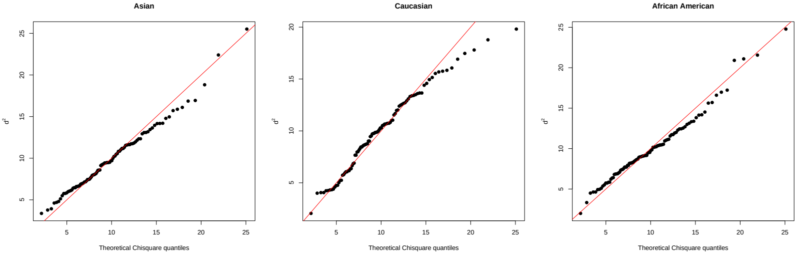


# Bivariate exploration




with 0.95 and 0.50 ellipses

# Multivariate exploration ($\chi^2$ plots for $p = 10$)



# Bibliography


- Johnson & Wichern, (2002) Applied Multivariate Statistical Analysis, Chapters 4 and 5, 5th edition, Prentice Hall.