Topic Four:
The Multivariate Normal Distribution

Preview

- Motivation:
  - The univariate normal distribution plays a very important role in univariate statistics, both as an assumed population model and in enabling inference with large sample sizes.
  - Similarly, the multivariate normal distribution plays a very important role in multivariate statistics.
- Goals:
  - Define the distribution and its properties.
  - See how sampling distributions of important statistics relate to the distribution.
  - Learn some techniques for assessing the distributional assumption and for transforming data so the assumption holds.

Definition and Properties

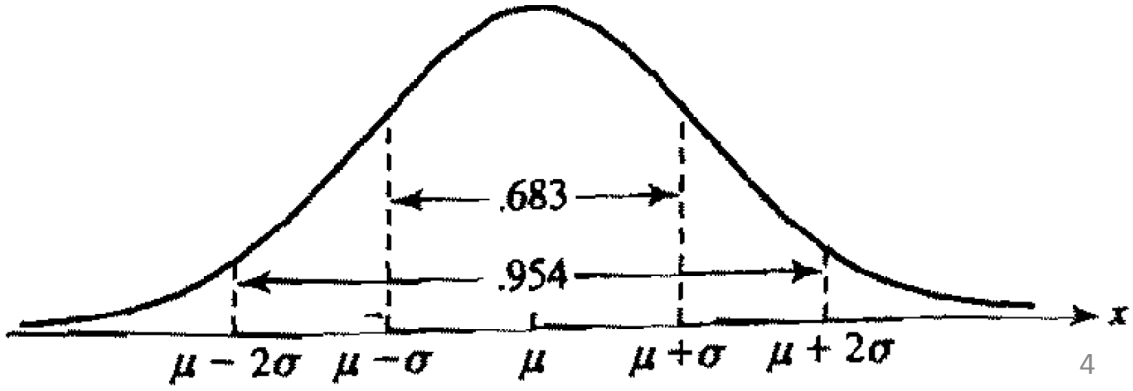
Reminder: When $p = 1$, we have the univariate normal distribution with mean μ, variance σ^2 and probability density function (pdf)

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-[(x-\mu)^2/\sigma^2]/2}$$

for $-\infty < x < \infty$. We will refer to the above distribution as $N(\mu, \sigma^2)$. Probabilities for this distribution equal areas under its pdf. These have to be computed numerically in general, but we have the rules of thumb:

$$P(\mu - \sigma \leq X \leq \mu + \sigma) \approx 0.68$$
$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 0.95$$
$$P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx 0.997$$

Note that the constant in front of the exponent is required to normalize the area (or, total probability) to be one.



When $p \geq 2$, we generalize to the multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and (positive definite) covariance matrix $\boldsymbol{\Sigma}$, with pdf

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

for $-\infty < x_i < \infty$, $i = 1, 2, \ldots, p$. We will refer to the above distribution as $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Note that $(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})$ is a generalization of the $(x - \mu)^2/\sigma^2 = (x - \mu)(\sigma^2)^{-1}(x - \mu)$ in the exponent of the univariate normal pdf. Note also that $(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})$ is the statistical distance from $\mathbf{x}$ to its mean $\boldsymbol{\mu}$. Probabilities now equal *volumes*, and the constant in front of the exponent is required to normalize the total volume (total probability) to be one.

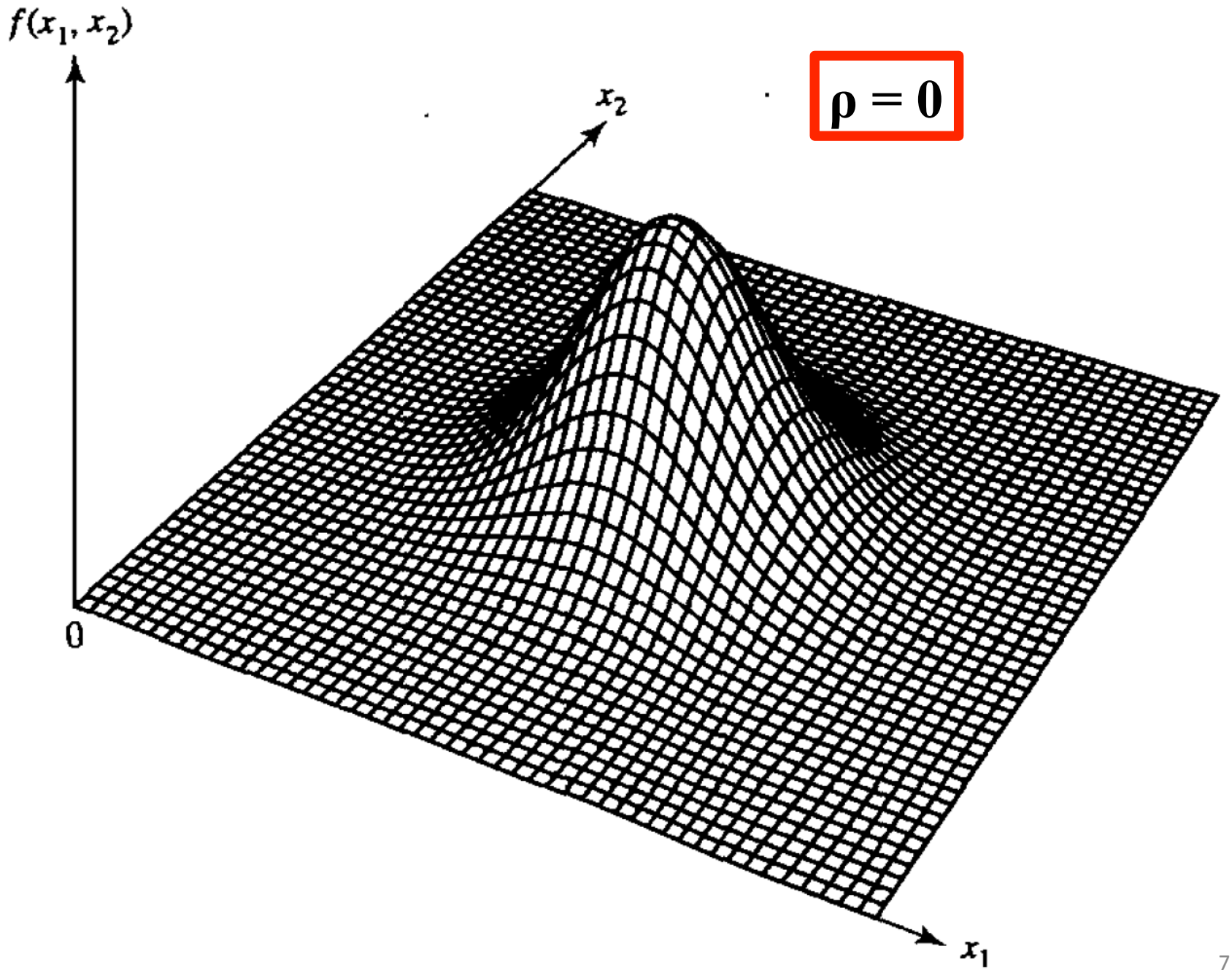
Example: When $p = 2$, we have the bivariate normal: $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. In this simple case, we have $|\boldsymbol{\Sigma}| = \sigma_{11}\sigma_{22} - \sigma_{12}^2 = \sigma_{11}\sigma_{22}(1 - \rho_{12}^2)$ and

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{\sigma_{11}\sigma_{22} - \sigma_{12}^2} \begin{bmatrix} \sigma_{22} & -\sigma_{12} \\ -\sigma_{12} & \sigma_{11} \end{bmatrix} = \frac{1}{\sigma_{11}\sigma_{22}(1 - \rho_{12}^2)} \begin{bmatrix} \sigma_{22} & -\sigma_{12} \\ -\sigma_{12} & \sigma_{11} \end{bmatrix}$$

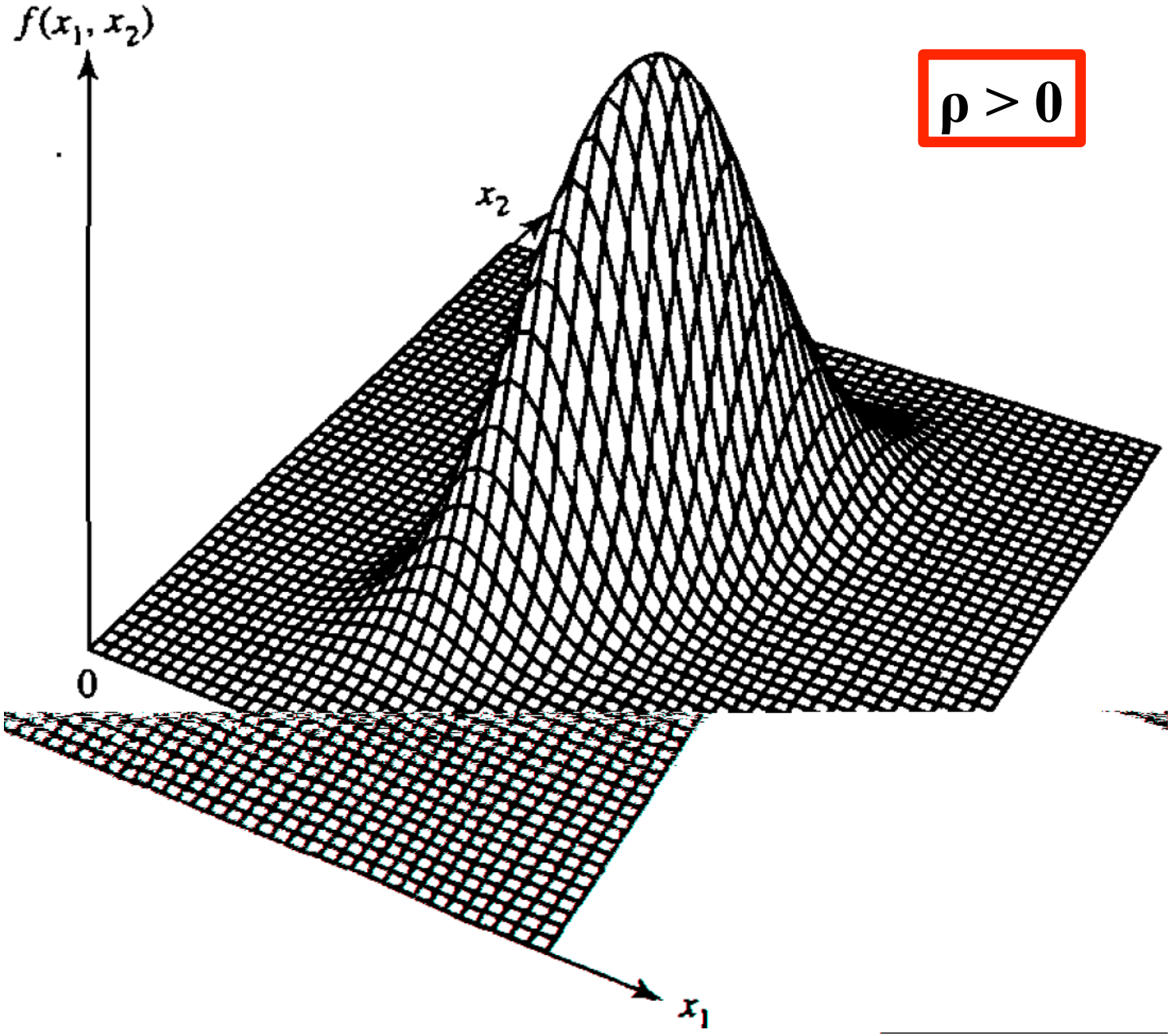
where $\rho_{12} = \text{Corr}(X_1, X_2)$. We can then write the density as

$$\begin{aligned} f(x_1, x_2) = & \frac{1}{2\pi\sqrt{\sigma_{11}\sigma_{22}(1 - \rho_{12}^2)}} \\ & \times \exp\left\{ -\frac{1}{2(1 - \rho_{12}^2)} \left[\left(\frac{x_1 - \mu_1}{\sqrt{\sigma_{11}}} \right)^2 + \left(\frac{x_2 - \mu_2}{\sqrt{\sigma_{22}}} \right)^2 \right.\right. \\ & \left.\left. - 2\rho_{12} \left(\frac{x_1 - \mu_1}{\sqrt{\sigma_{11}}} \right) \left(\frac{x_2 - \mu_2}{\sqrt{\sigma_{22}}} \right) \right] \right\} \end{aligned}$$

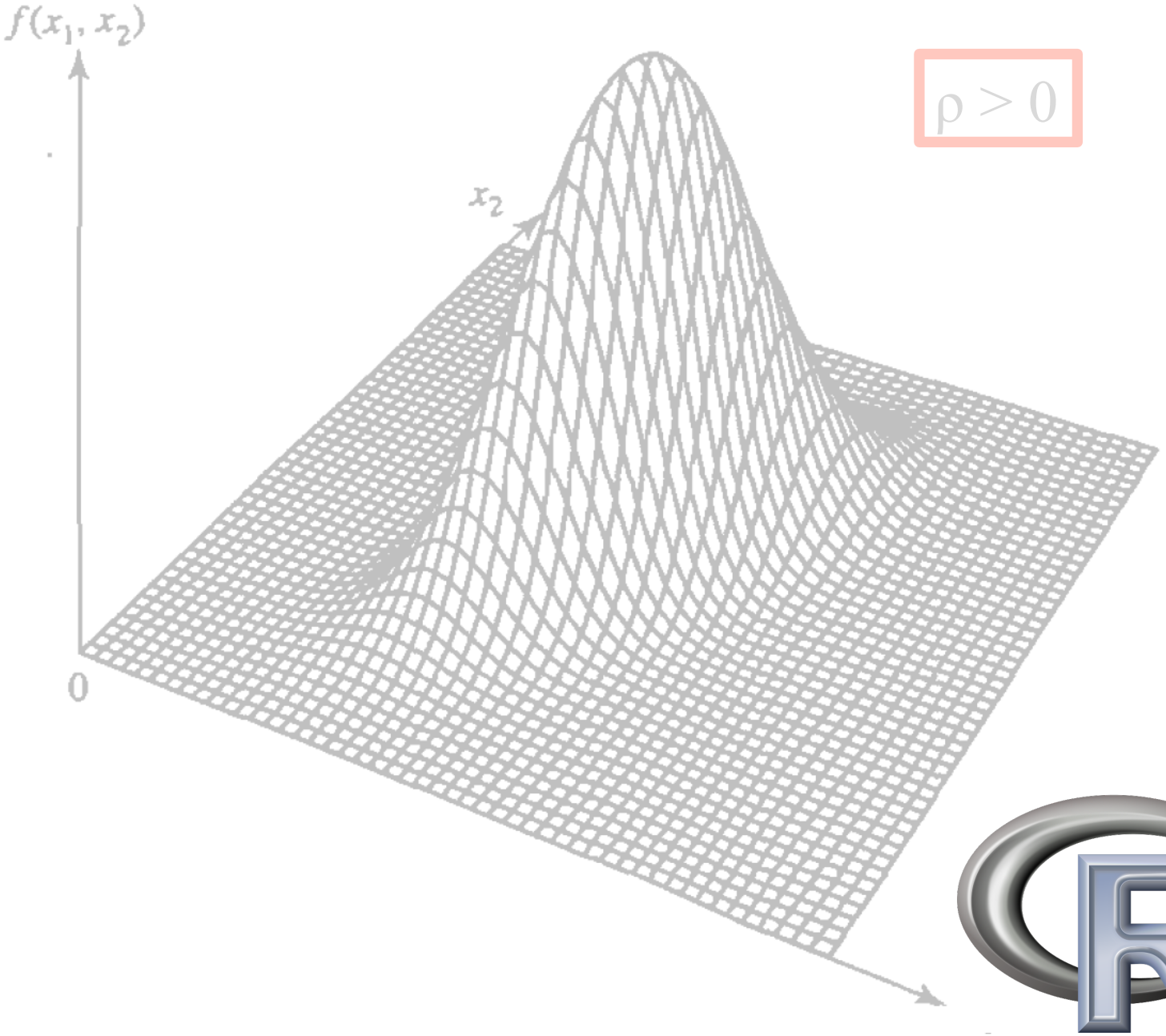
Note that if X_1 and X_2 are uncorrelated ($\rho_{12} = 0$), the joint density factors into the product of two univariate normal densities, which means X_1 and X_2 are independent and marginally normally distributed.



$\rho = 0$



$\rho > 0$


$f(x_1, x_2)$
x_2
0
ρ > 0

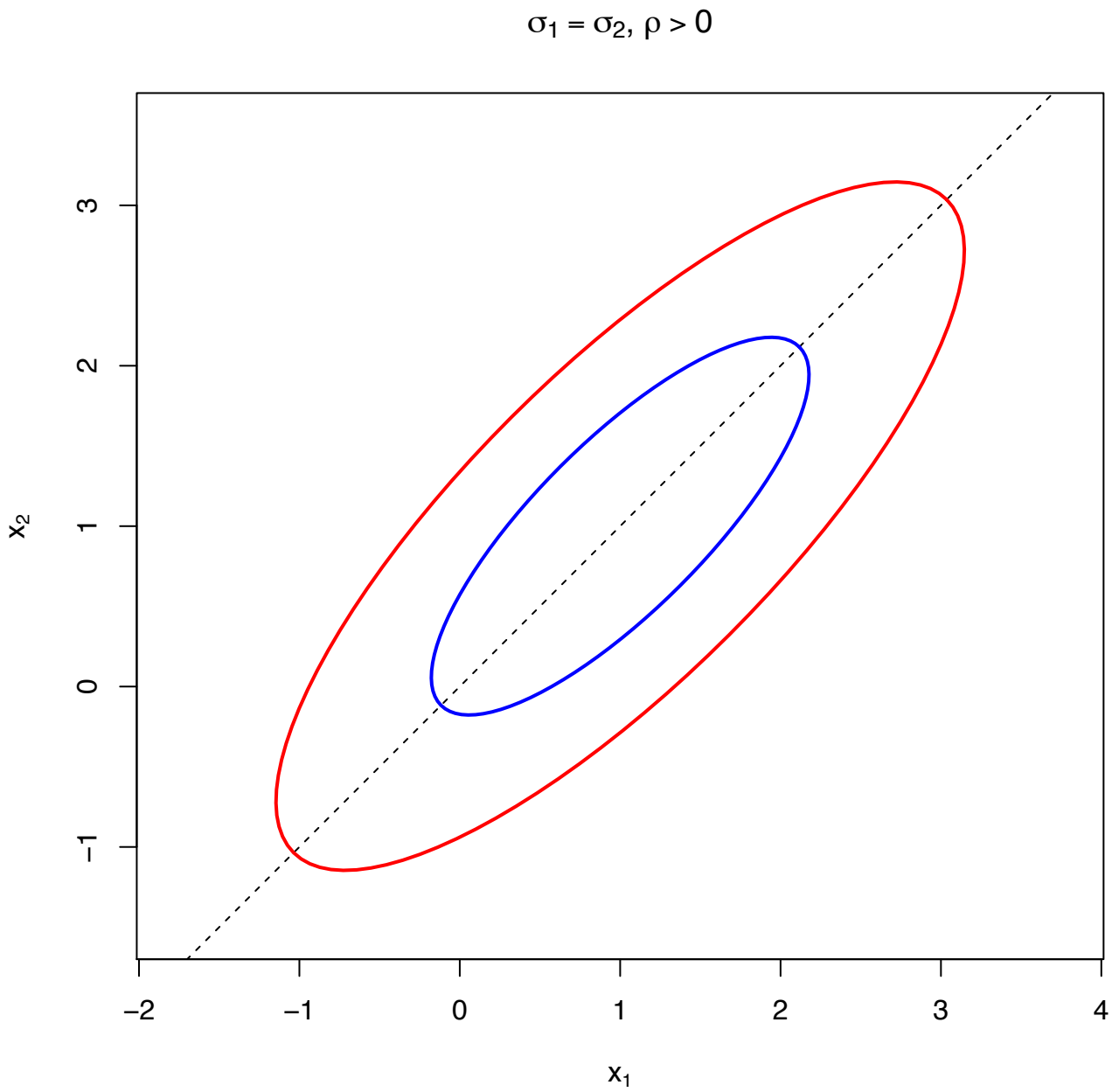
Contours of constant density for the $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution are ellipsoids given by the $\mathbf{x}$ such that

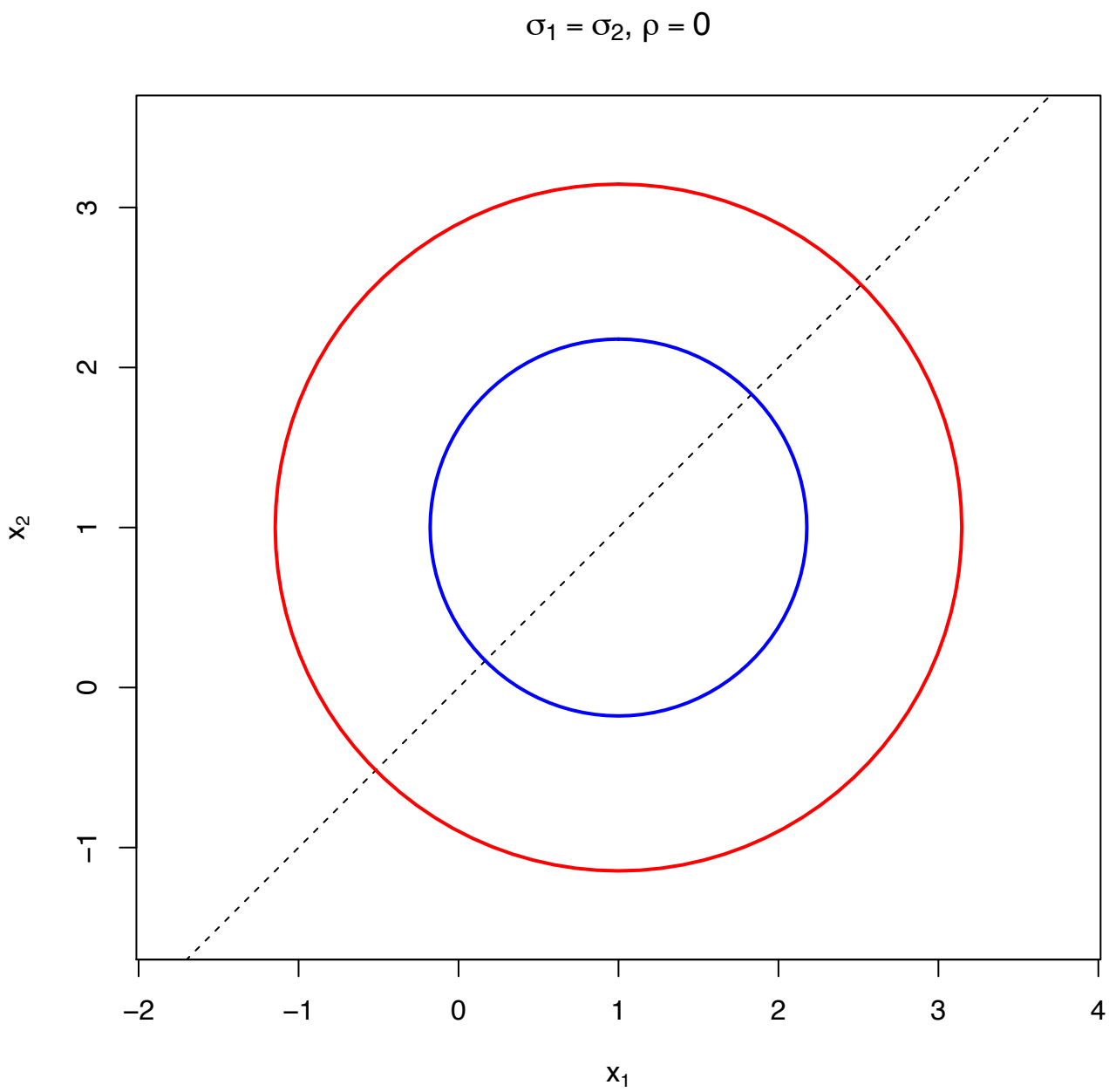
$$(\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) = c^2$$

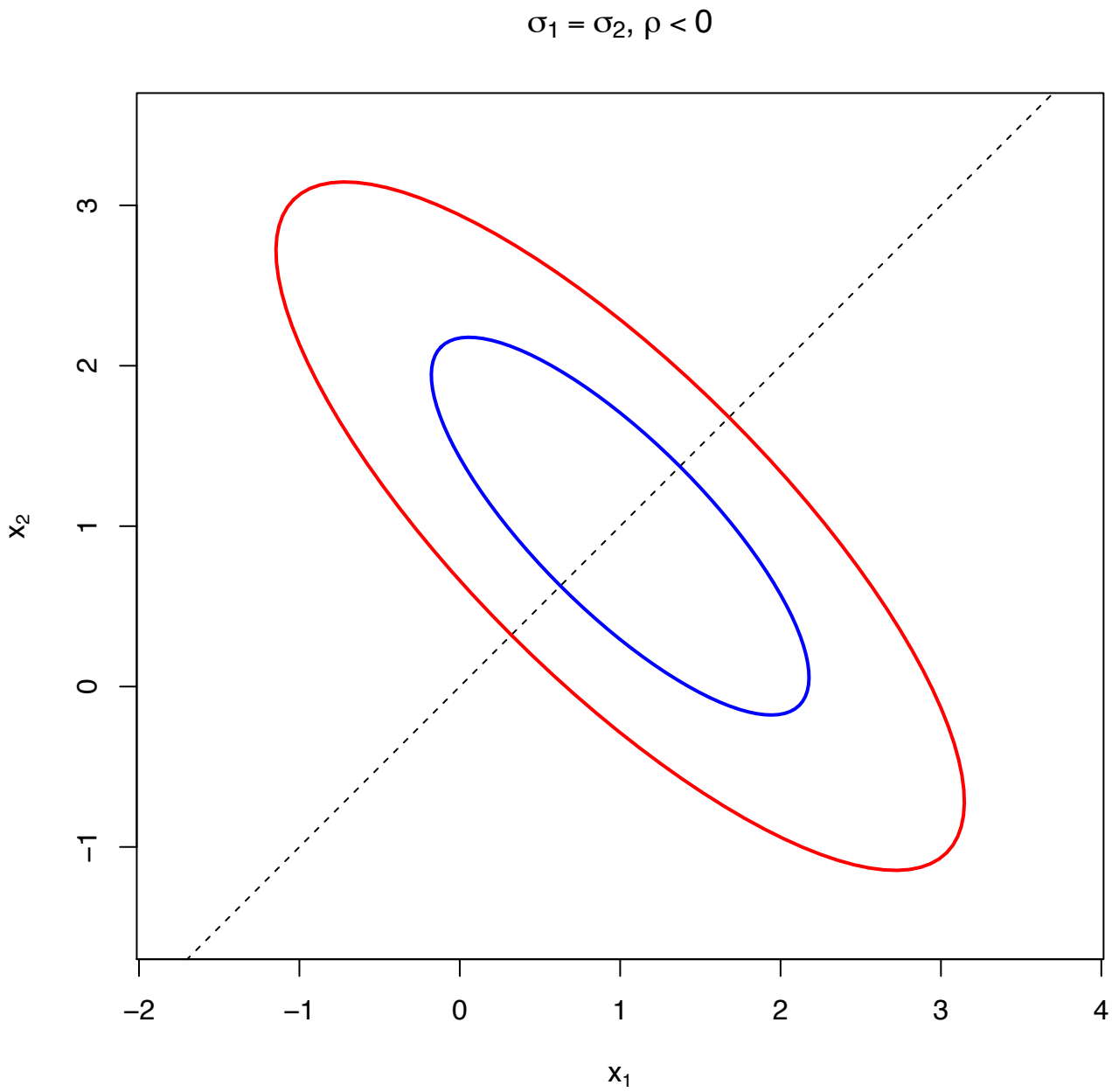
The ellipsoids are centered at $\boldsymbol{\mu}$ and have axes $\pm c\sqrt{\lambda_i}\mathbf{e}_i$, where the $(\lambda_i, \mathbf{e}_i)$, $i = 1, 2, \ldots, p$, are the eigenvalue, eigenvector pairs of $\boldsymbol{\Sigma}$. This uses the fact (which we have already seen) that the eigenvalue, eigenvector pairs of $\boldsymbol{\Sigma}^{-1}$ are $(1/\lambda_i, \mathbf{e}_i)$, $i = 1, 2, \ldots, p$, where $(\lambda_i, \mathbf{e}_i)$ are the eigenvalue, eigenvector pairs of $\boldsymbol{\Sigma}$. Note also that $\boldsymbol{\Sigma}^{-1}$ is positive definite (see textbook for proof).

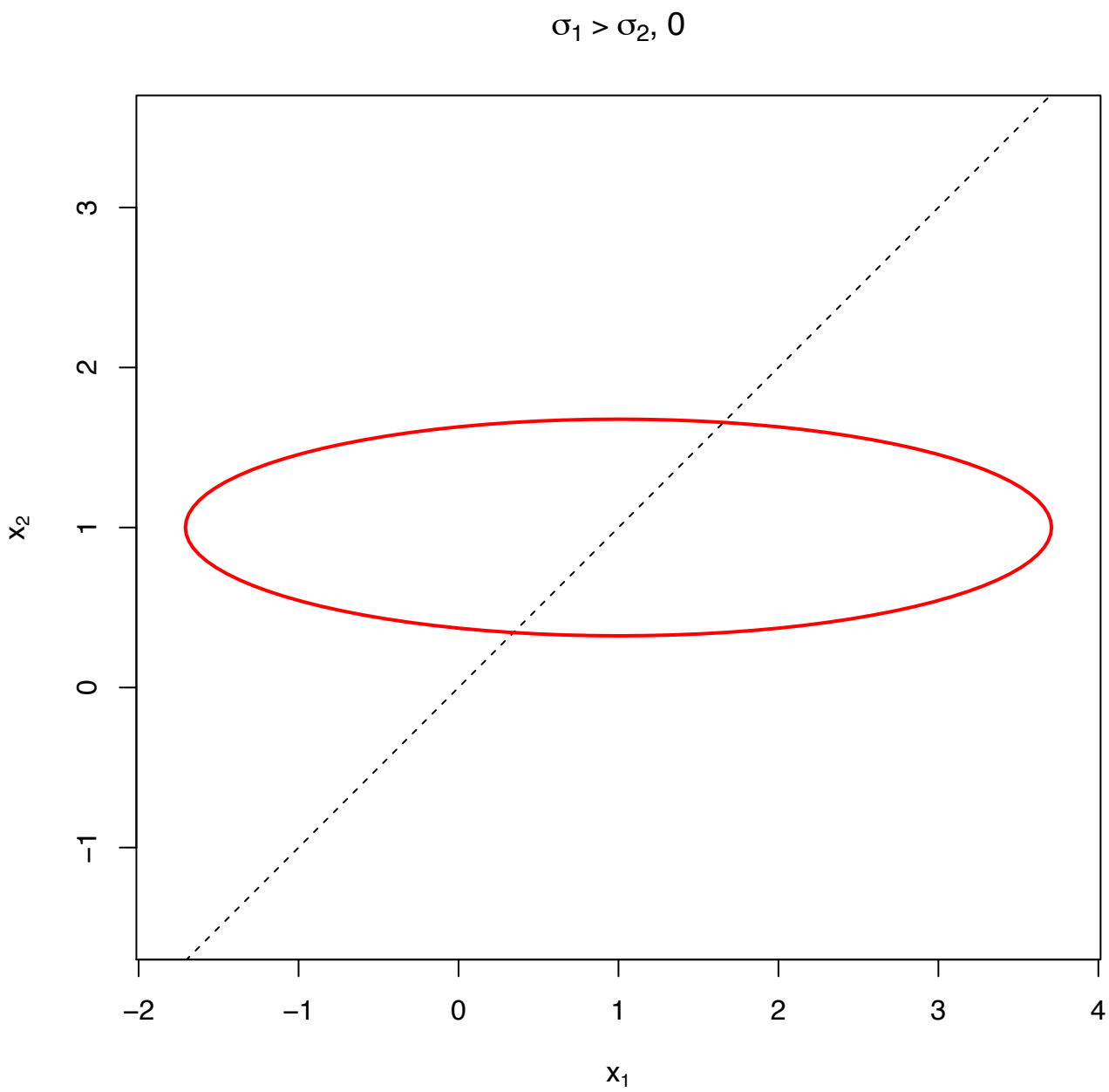
In the bivariate case, when $\sigma_{11} = \sigma_{22}$, the eigenvalues and eigenvectors of $\boldsymbol{\Sigma}$ are $\lambda_1 = \sigma_{11}+\sigma_{12}$, $\lambda_2 = \sigma_{11}-\sigma_{12}$, $\mathbf{e}_1' = [1/\sqrt{2}, 1/\sqrt{2}]$, and $\mathbf{e}_2' = [1/\sqrt{2}, -1/\sqrt{2}]$ (see textbook). As usual, the axes of the ellipses of constant density c^2 are in the direction of $\mathbf{e}_1$ and $\mathbf{e}_2$ with lengths $c\sqrt{\lambda_1} = c\sqrt{\sigma_{11} + \sigma_{12}}$ and $c\sqrt{\lambda_2} = c\sqrt{\sigma_{11} - \sigma_{12}}$, respectively. When $\sigma_{12} > 0$, λ_1 is the largest eigenvalue, and the major axis of the constant-density ellipses will go through $\mathbf{e}_1$ along the 45° line though $\boldsymbol{\mu}$. When $\sigma_{12} < 0$, λ_2 is the largest eigenvalue, and the major axis of the constant-density ellipses will be perpendicular to the 45° line through $\boldsymbol{\mu}$.

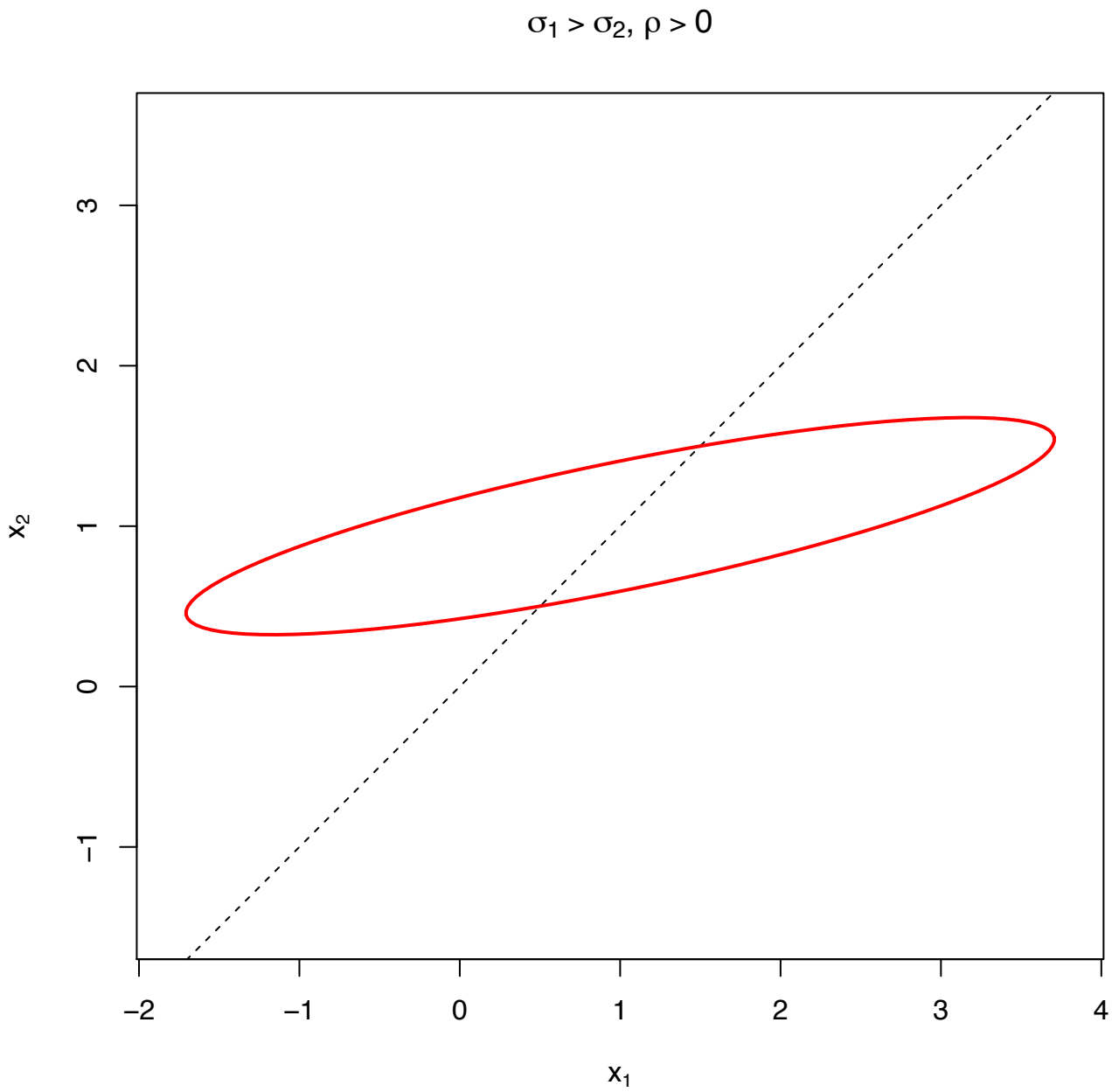
Also, setting $c^2 = \chi^2_p(\alpha)$, where $\chi^2_p(\alpha)$ is the $((1-\alpha)\times 100)$th percentile of the chi-square distribution with p degrees of freedom, specifies contours that contain $(1-\alpha)\times 100\%$ of the probability. We have used this result already, and we will find out why it holds momentarily.

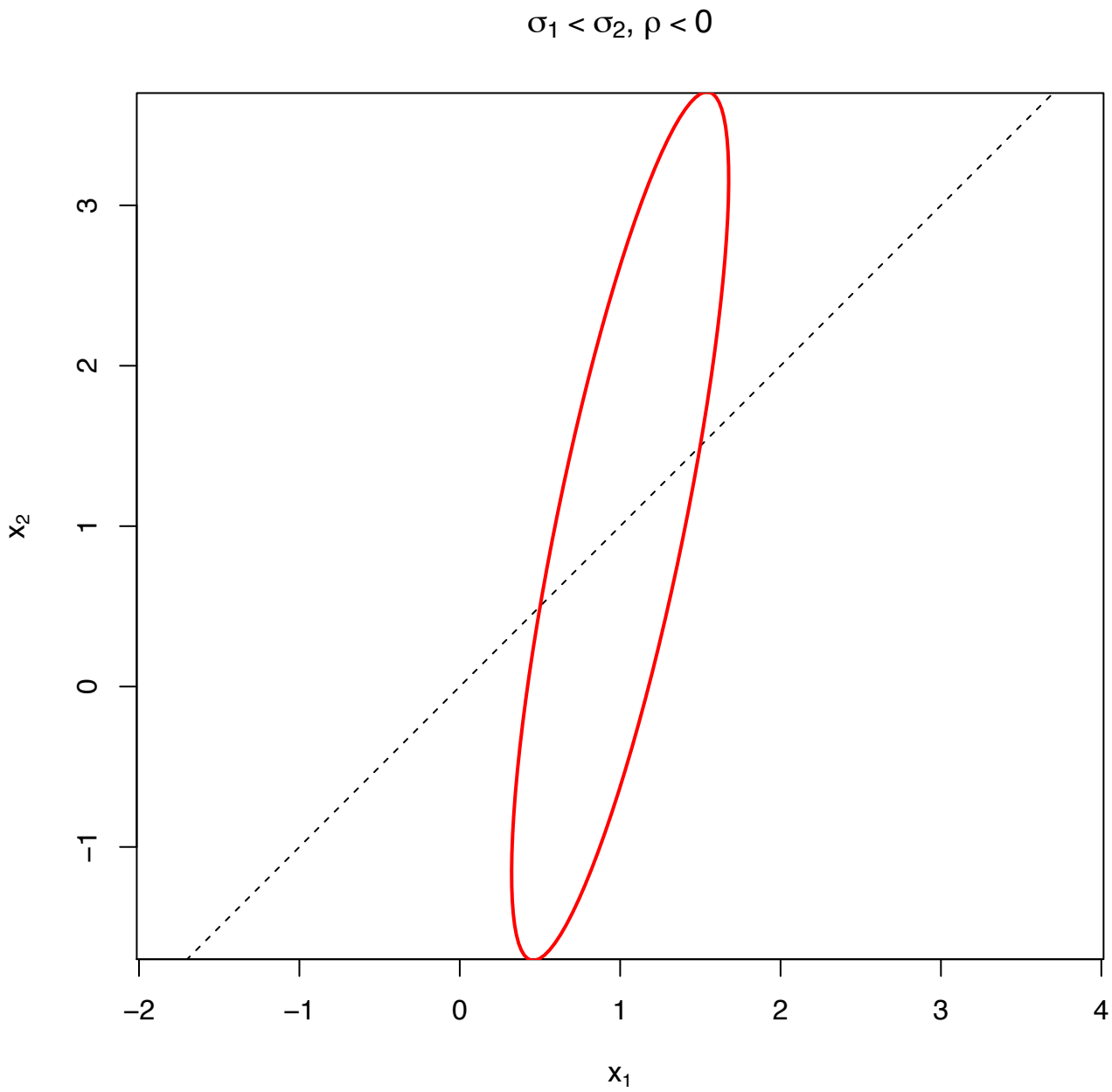

$\sigma_1 = \sigma_2, \rho > 0$
x_2
3
2
1
0
−1
−2
−1
0
1
2
3
4
x_1

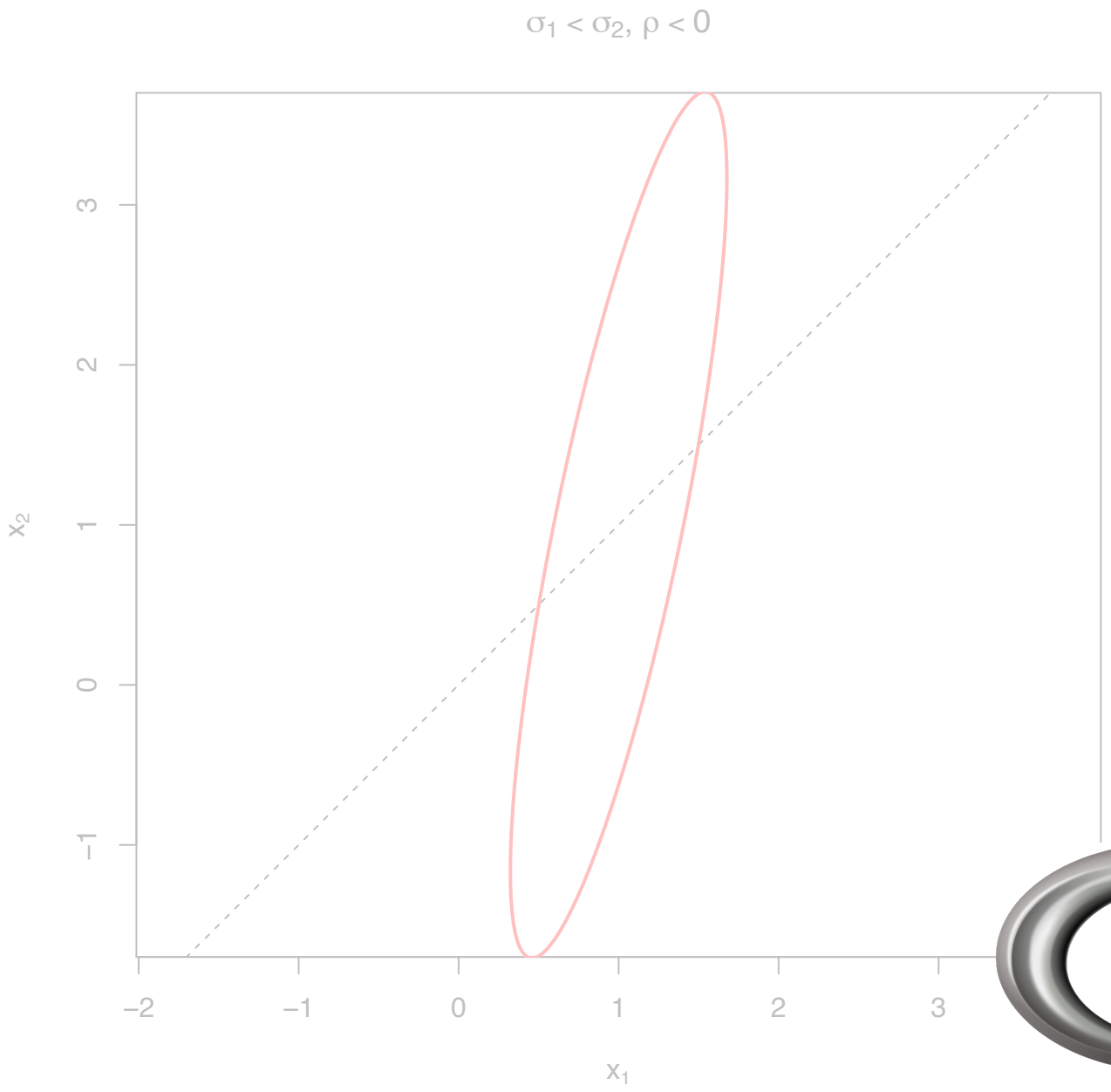

$\sigma_1 = \sigma_2, \rho = 0$
x_2
3
2
1
0
−1
−2
−1
0
1
2
3
4
x_1


$\sigma_1 = \sigma_2, \rho < 0$
x_2
3
2
1
0
−1
−2
−1
0
1
2
3
4
x_1


$\sigma_1 > \sigma_2$, 0
x_2
3
2
1
0
−1
−2
−1
0
1
2
3
4
x_1


$\sigma_1 > \sigma_2, \rho > 0$
x_2
3
2
1
0
−1
−2
−1
0
1
2
3
4
x_1


$\sigma_1 < \sigma_2, \rho < 0$
x_2
3
2
1
0
−1
−2
−1
0
1
2
3
4
x_1


$\sigma_1 < \sigma_2, \rho < 0$
x_2
3
2
1
0
−1
−2
−1
0
1
2
3
x_1

Key Results: With $\mathbf{X}$ multivariate normal...

- Linear combinations of the components of $\mathbf{X}$ are normally distributed.
- All subsets of the components of $\mathbf{X}$ are (potentially multivariate) normal.
- The conditional distributions of the components are (potentially multivariate) normal.
- Zero covariance implies that the corresponding components are independent.

Linear Combinations:

- If $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then any linear combination of the components of $\mathbf{X}$, $\mathbf{a}'\mathbf{X} = a_1X_1 + a_2X_2 + \cdots a_pX_p$, is distributed as $N(\mathbf{a}'\boldsymbol{\mu}, \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a})$.
- If $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then any q linear combinations $\underset{(q\times p)}{\mathbf{A}}\ \underset{(p\times 1)}{\mathbf{X}}$ are distributed as $N_q(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$.
- If $\mathbf{a}'\mathbf{X} \sim N(\mathbf{a}'\boldsymbol{\mu}, \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a})$ for every $\mathbf{a}$, then $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
- If $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\underset{(p\times 1)}{\mathbf{X}} + \underset{(p\times 1)}{\mathbf{d}} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\mathbf{d}$ is a vector of constants.

Example: Let $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathbf{u}_i$ be the p-vector with ith component equal one, all other components equal zero, $i = 1, 2, \ldots, p$.

- With $\mathbf{a} = \mathbf{u}_i$, we have $\mathbf{a}'\mathbf{X} \sim N(\mu_i, \sigma_{ii})$.
- With

$$\underset{(2\times p)}{\mathbf{A}} = \begin{bmatrix} \mathbf{u}_i - \mathbf{u}_j \\ \mathbf{u}_j - \mathbf{u}_k \end{bmatrix}$$

for $i \neq j$, we have $\mathbf{AX}$ q-variate normal with mean

$$\mathbf{A}\boldsymbol{\mu} = \begin{bmatrix} \mu_i - \mu_j \\ \mu_j - \mu_k \end{bmatrix}$$

and covariance

$$\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}' = \begin{bmatrix} \sigma_{ii} - 2\sigma_{ij} + \sigma_{jj} & \sigma_{ij} + \sigma_{jk} - \sigma_{jj} - \sigma_{ik} \\ \sigma_{ij} + \sigma_{jk} - \sigma_{jj} - \sigma_{ik} & \sigma_{jj} - 2\sigma_{jk} + \sigma_{kk} \end{bmatrix}$$

Subsets: Partition $\mathbf{X}$, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$ as

$$\underset{(p\times 1)}{\mathbf{X}} = \begin{bmatrix} \underset{(q\times 1)}{\mathbf{X}_1} \\ \hdashline \underset{((p-q)\times 1)}{\mathbf{X}_2} \end{bmatrix} \qquad \underset{(p\times 1)}{\boldsymbol{\mu}} = \begin{bmatrix} \underset{(q\times 1)}{\boldsymbol{\mu}_1} \\ \hdashline \underset{((p-q)\times 1)}{\boldsymbol{\mu}_1} \end{bmatrix}$$

and

$$\underset{(p\times p)}{\boldsymbol{\Sigma}} = \left[\begin{array}{c:c} \underset{(q\times q)}{\boldsymbol{\Sigma}_{11}} & \underset{(q\times(p-q))}{\boldsymbol{\Sigma}_{12}} \\ \hdashline \underset{((p-q)\times q)}{\boldsymbol{\Sigma}_{21}} & \underset{((p-q)\times(p-q))}{\boldsymbol{\Sigma}_{22}} \end{array}\right]$$

Then $\mathbf{X}_1 \sim N_q(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$.

Example: Let $\mathbf{X} \sim N_4(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Consider the distribution of $\mathbf{X}_1' = [X_1, X_3]$. We can write

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ (2\times 1) \\ \hdashline \mathbf{X}_2 \\ (2\times 1) \end{bmatrix}, \qquad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ (2\times 1) \\ \hdashline \boldsymbol{\mu}_2 \\ (2\times 1) \end{bmatrix}, \qquad \boldsymbol{\Sigma} = \left[\begin{array}{cc} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ (2\times 2) & (2\times 2) \\ \hdashline \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \\ (2\times 2) & (2\times 2) \end{array}\right]$$

where $\boldsymbol{\mu}_1' = [\mu_1, \mu_3]$ and

$$\boldsymbol{\Sigma}_{11} = \begin{bmatrix} \sigma_{11} & \sigma_{13} \\ \sigma_{13} & \sigma_{33} \end{bmatrix}$$

Then

$$\mathbf{X}_1 \sim N_2(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}) = N_2\left(\begin{bmatrix} \mu_1 \\ \mu_3 \end{bmatrix}, \begin{bmatrix} \sigma_{11} & \sigma_{13} \\ \sigma_{13} & \sigma_{33} \end{bmatrix}\right)$$

Conditional Distributions: Let $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}, \qquad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \qquad \boldsymbol{\Sigma} = \left[\begin{array}{c|c} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \hline \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{array}\right]$$

Then the conditional distribution of $\mathbf{X}_1$, given that $\mathbf{X}_2 = \mathbf{x}_2$, is normal with

$$\text{Mean} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2)$$

and

$$\text{Covariance} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$$

Let $\boldsymbol{\mu}_{1:2}(\mathbf{x}_2) = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2)$ and $\boldsymbol{\Sigma}_{11:2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$. You will show in your homework that:

$$|\boldsymbol{\Sigma}| = |\boldsymbol{\Sigma}_{11:2}||\boldsymbol{\Sigma}_{22}|$$

and

$$\begin{aligned}(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = (\mathbf{x}_1 - \boldsymbol{\mu}_{1:2}(\mathbf{x}_2))' \boldsymbol{\Sigma}_{11:2}^{-1} (\mathbf{x}_1 - \boldsymbol{\mu}_{1:2}(\mathbf{x}_2)) \\ \times (\mathbf{x}_2 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}_{22}^{-1} (\mathbf{x}_2 - \boldsymbol{\mu}_2)\end{aligned}$$

So, given that $\mathbf{X}_2 = \mathbf{x}_2$, the $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ density can be factored as

$$f(\mathbf{x}) = f_{1:2}(\mathbf{x}_1) \times f_2(\mathbf{x}_2)$$

where $f_{1:2}$ and f_2 are the densities for the $N_q(\boldsymbol{\mu}_{1:2}(\mathbf{x}_2), \boldsymbol{\Sigma}_{11:2})$ and $N_{p-q}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$ distributions, respectively. Since the conditional distribution of $\mathbf{X}_1$ given $\mathbf{X}_2 = \mathbf{x}_2$ is defined as $f(\mathbf{x})/f_2(\mathbf{x}_2)$, we have the result from the preceding slide.

Example: In the bivariate case, with $\mathbf{X}' = [X_1, X_2]$, we have

$$E(X_1|X_2 = x_2) = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2)$$
$$= \mu_1 + \frac{\sigma_{12}}{\sigma_{22}}(x_2 - \mu_2)$$

and

$$\text{Var}(X_1|X_2 = x_2) = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$$
$$= \sigma_{11} - \frac{\sigma_{12}^2}{\sigma_{22}}$$

Correlation and Independence:

- If $\underset{(q_1\times 1)}{\mathbf{X}_1}$ and $\underset{(q_2\times 1)}{\mathbf{X}_2}$ are independent, then $\mathrm{Cov}(\mathbf{X}_1, \mathbf{X}_2) = \mathbf{0}$, a $q_1 \times q_2$ matrix of zeros.

- If
$$\begin{bmatrix} \mathbf{X}_1 \\ \hline \mathbf{X}_2 \end{bmatrix} \sim N_{q_1+q_2}\left(\begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \hline \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \right)$$
then $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent if and only if $\boldsymbol{\Sigma}_{12} = \mathbf{0}$.

- If $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent with distributions $N_{q_1}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ and $N_{q_2}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$, respectively, then
$$\begin{bmatrix} \mathbf{X}_1 \\ \hline \mathbf{X}_2 \end{bmatrix} \sim N_{q_1+q_2}\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \hline \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \mathbf{0} \\ \hline \mathbf{0} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \right)$$

You can see this from the preceding definition of the distribution for $\mathbf{X}_1$ given $\mathbf{X}_2 = \mathbf{x}_2$.

$\mathbf{X}_1$ and $\mathbf{X}_2$ are independent if and only if $\mathbf{\Sigma}_{12} = \mathbf{0}$.

Ellipsoids of Constant Density: Let $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $|\boldsymbol{\Sigma}| > 0$. Then

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

where χ^2_p denotes the chi-square distribution with p degrees of freedom. As a result,

$$P\left((\mathbf{X} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu}) \leq \chi^2_p(\alpha)\right) = 1 - \alpha$$

where $\chi^2_p(\alpha)$ is the $(1-\alpha) \times 100$th percentile of the χ^2_p distribution.

Recall that the eigenvalues and eigenvector pairs of $\boldsymbol{\Sigma}^{-1}$ are $(1/\lambda_i, \mathbf{e}_i)$, where $(\lambda_i, \mathbf{e}_i)$ are the eigenvalues and eigenvectors of $\boldsymbol{\Sigma}$, $i = 1, 2, \ldots, p$. So, by the spectral decomposition

$$(\mathbf{X} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu}) = \sum_{i=1}^{p}(1/\lambda_k)(\mathbf{X} - \boldsymbol{\mu})'\mathbf{e}_i\mathbf{e}_i'(\mathbf{X} - \boldsymbol{\mu}) = \sum_{i=1}^{p} Z_i^2$$

where $Z_i = (1/\sqrt{\lambda_i})\mathbf{e}_i'(\mathbf{X} - \boldsymbol{\mu})$. Then, since Z_i is a linear combination of the components of $\mathbf{X} - \boldsymbol{\mu}$, and since $(\mathbf{X} - \boldsymbol{\mu}) \sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$, Z_i has the univariate normal distribution with mean 0 and variance

$$\text{Var}(Z_i) = \frac{1}{\lambda_i}\mathbf{e}_i'\boldsymbol{\Sigma}\mathbf{e}_i = \mathbf{e}_i'\left(\sum_{j=1}^{p}\mathbf{e}_j\mathbf{e}_j'\right)\mathbf{e}_i = \mathbf{e}_i'\mathbf{e}_i\mathbf{e}_i'\mathbf{e}_i = 1$$

Similarly, we have

$$\text{Cov}(Z_i, Z_j) = \frac{1}{\sqrt{\lambda_i\lambda_j}}\mathbf{e}_i'\boldsymbol{\Sigma}\mathbf{e}_j = \frac{1}{\sqrt{\lambda_i\lambda_j}}\mathbf{e}_i'\mathbf{e}_i\mathbf{e}_i'\mathbf{e}_j = 0$$

so that the Z_i are mutually independent. Together, this means that $(\mathbf{X} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu})$ is the sum of p independent squared standard normal random variables, which defines the χ_p^2 distribution.

Recall that the eigenvalues and eigenvector pairs of $\boldsymbol{\Sigma}^{-1}$ are $(1/\lambda_i, \mathbf{e}_i)$, where $(\lambda_i, \mathbf{e}_i)$ are the eigenvalues and eigenvectors of $\boldsymbol{\Sigma}$, $i = 1, 2, \ldots, p$. So, by the spectral decomposition

$$(\mathbf{X} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu}) = \sum_{i=1}^{p}(1/\lambda_k)(\mathbf{X} - \boldsymbol{\mu})'\mathbf{e}_i\mathbf{e}_i'(\mathbf{X} - \boldsymbol{\mu}) = \sum_{i=1}^{p} Z_i^2$$

where $Z_i = (1/\sqrt{\lambda_i})\mathbf{e}_i'(\mathbf{X} - \boldsymbol{\mu})$. Then, since Z_i is a linear combination of the components of $\mathbf{X} - \boldsymbol{\mu}$, and since $(\mathbf{X} - \boldsymbol{\mu}) \sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$, Z_i has the univariate normal di

Similarly,

so that the Z_i are mutually independent. Together, this means that $(\mathbf{X} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu})$ is the sum of p independent squared standard normal random variables, which defines the χ_p^2 distribution.

Note that the squared statistical distance can therefore be viewed as the usual squared distance on variables that have been transformed to have unit variance and be mutually independent. We are *adjusting for* unequal variance and correlation.

Another Linear Combination: Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be mutually independent with $\mathbf{X}_j \sim N_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$. Define

$$\mathbf{V}_1 = c_1\mathbf{X}_1 + c_2\mathbf{X}_2 + \cdots + c_n\mathbf{X}_n$$
$$\mathbf{V}_2 = b_1\mathbf{X}_1 + b_2\mathbf{X}_2 + \cdots + b_n\mathbf{X}_n$$

Then $\mathbf{V}' = \underset{(1\times 2p)}{\left[\begin{array}{c|c} \mathbf{V}_1' & \mathbf{V}_2' \end{array}\right]}$ has the ($2p$-dimensional) multivariate normal distribution with

$$\text{Mean} = \begin{bmatrix} \sum_{j=1}^n c_j\boldsymbol{\mu}_j \\ \sum_{j=1}^n b_j\boldsymbol{\mu}_j \end{bmatrix}$$

and

$$\text{Covariance} = \begin{bmatrix} \left(\sum_{j=1}^n c_j^2\right)\boldsymbol{\Sigma} & (\mathbf{b}'\mathbf{c})\,\boldsymbol{\Sigma} \\ (\mathbf{b}'\mathbf{c})\,\boldsymbol{\Sigma} & \left(\sum_{j=1}^n b_j^2\right)\boldsymbol{\Sigma} \end{bmatrix}$$

Thus $\mathbf{V}_1$ and $\mathbf{V}_2$ are independent if $\mathbf{b}'\mathbf{c} = 0$.

Example: Let $\mathbf{X}_1$, $\mathbf{X}_2$, $\mathbf{X}_3$, and $\mathbf{X}_4$ be independent and identically distributed bivariate normal random vectors with

$$\boldsymbol{\mu} = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix}$$

Let $\mathbf{c}' = [1, 0, 1, 0]$, $\mathbf{b}' = [0, 1, 0, 1]$, $\mathbf{V}_1 = c_1\mathbf{X}_1 + c_2\mathbf{X}_2 + c_3\mathbf{X}_3 + c_4\mathbf{X}_4 = \mathbf{X}_1 + \mathbf{X}_3$ and $\mathbf{V}_2 = b_1\mathbf{X}_1 + b_2\mathbf{X}_2 + b_3\mathbf{X}_3 + b_4\mathbf{X}_4 = \mathbf{X}_2 + \mathbf{X}_4$. Both $\mathbf{V}_1$ and $\mathbf{V}_2$ have mean $2\boldsymbol{\mu}$ and covariance $2\boldsymbol{\Sigma}$. Also, since $\mathbf{c}'\mathbf{b} = 0$, $\mathbf{V}_1$ and $\mathbf{V}_2$ are jointly six-variate normal with covariance

$$\begin{bmatrix} 2\boldsymbol{\Sigma} & \mathbf{0} \\ \mathbf{0} & 2\boldsymbol{\Sigma} \end{bmatrix}$$

and $\mathbf{V}_1$ and $\mathbf{V}_2$ are independent.

Sampling and Maximum Likelihood Estimation

Likelihood: Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be a random sample from the $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution. Because the $\mathbf{X}_i$ are mutually independent with the same distribution, their joint pdf is the product of their n marginal densities. For the purposes of estimating the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, we will refer to this joint density as the *likelihood* of observing a particular realization $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ as a function of the unknown parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. We write

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{j=1}^{n} \left\{ \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-(\mathbf{x}_j - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_j - \boldsymbol{\mu})/2} \right\}$$

$$= \frac{1}{(2\pi)^{np/2}|\boldsymbol{\Sigma}|^{n/2}} e^{-\sum_{j=1}^{n} (\mathbf{x}_j - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_j - \boldsymbol{\mu})/2}$$

Maximum likelihood estimation involves finding the values of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ that maximize $L(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

We rewrite the form of the likelihood to simplify what follows. Recall the *trace* of a square matrix is the sum of its diagonal values. Because $\mathbf{x}'\mathbf{A}\mathbf{x} = \text{tr}(\mathbf{x}'\mathbf{A}\mathbf{x}) = \text{tr}(\mathbf{A}\mathbf{x}\mathbf{x}')$ for any $p \times p$ symmetric matrix $\mathbf{A}$*,

$$(\mathbf{x}_j - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_j - \boldsymbol{\mu}) = \text{tr}\left[\boldsymbol{\Sigma}^{-1}(\mathbf{x}_j - \boldsymbol{\mu})(\mathbf{x}_j - \boldsymbol{\mu})'\right]$$

Also, because the trace of a sum of matrices equals the sum of the traces*,

$$\sum_{j=1}^{n}(\mathbf{x}_j - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_j - \boldsymbol{\mu}) = \text{tr}\left[\boldsymbol{\Sigma}^{-1}\left(\sum_{j=1}^{n}(\mathbf{x}_j - \boldsymbol{\mu})(\mathbf{x}_j - \boldsymbol{\mu})'\right)\right]$$

Then, adding and subtracting $\bar{\mathbf{x}}$ in each term $(\mathbf{x}_j - \boldsymbol{\mu})$, we can write*

$$\begin{aligned}\sum_{j=1}^{n}(\mathbf{x}_j - \boldsymbol{\mu})(\mathbf{x}_j - \boldsymbol{\mu})' &= \sum_{j=1}^{n}(\mathbf{x}_j - \bar{\mathbf{x}} + \bar{\mathbf{x}} - \boldsymbol{\mu})(\mathbf{x}_j - \bar{\mathbf{x}} + \bar{\mathbf{x}} - \boldsymbol{\mu})' \\ &= \sum_{j=1}^{n}(\mathbf{x}_j - \bar{\mathbf{x}})(\mathbf{x}_j - \bar{\mathbf{x}})' + n(\bar{\mathbf{x}} - \boldsymbol{\mu})(\bar{\mathbf{x}} - \boldsymbol{\mu})'\end{aligned}$$

* See textbook.

We can now rewrite the likelihood as

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{np/2}|\boldsymbol{\Sigma}|^{n/2}} e^{-\mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}\left(\sum_{j=1}^{n}(\mathbf{x}_j-\bar{\mathbf{x}})(\mathbf{x}_j-\bar{\mathbf{x}})'+n(\bar{\mathbf{x}}-\boldsymbol{\mu})(\bar{\mathbf{x}}-\boldsymbol{\mu})'\right)\right]/2}$$

Sometimes it will also be convenient to write the term inside the exponent as

$$\begin{aligned}
&\mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}\left(\sum_{j=1}^{n}(\mathbf{x}_j-\bar{\mathbf{x}})(\mathbf{x}_j-\bar{\mathbf{x}})'+n(\bar{\mathbf{x}}-\boldsymbol{\mu})(\bar{\mathbf{x}}-\boldsymbol{\mu})'\right)\right] \\
&= \mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}\left(\sum_{j=1}^{n}(\mathbf{x}_j-\bar{\mathbf{x}})(\mathbf{x}_j-\bar{\mathbf{x}})'\right)\right] + n\mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu})(\bar{\mathbf{x}}-\boldsymbol{\mu})'\right] \\
&= \mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}\left(\sum_{j=1}^{n}(\mathbf{x}_j-\bar{\mathbf{x}})(\mathbf{x}_j-\bar{\mathbf{x}})'\right)\right] + n(\bar{\mathbf{x}}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu})
\end{aligned}$$

The MLEs $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$: The *maximum likelihood estimates* (MLEs) are

$$\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}} \text{ and } \hat{\boldsymbol{\Sigma}} = \mathbf{S}_n = \frac{(n-1)}{n}\mathbf{S}$$

Note that these are realizations of random variables. We call the random variables the *maximum likelihood estimators*.

To maximize $L(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with respect to $\boldsymbol{\mu}$, because of the way we rewrote the likelihood, we need to minimize the statistical distance $(\bar{\mathbf{x}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu})$. Since $\boldsymbol{\Sigma}^{-1}$ is positive definite, $(\bar{\mathbf{x}}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu}) > 0$ unless $\boldsymbol{\mu} = \bar{\mathbf{x}}$, so $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$.

To obtain $\hat{\boldsymbol{\Sigma}}$, note that we need to maximize $L(\hat{\boldsymbol{\mu}}, \boldsymbol{\Sigma}) = L(\bar{\mathbf{x}}, \boldsymbol{\Sigma})$, since the MLEs are defined as the parameter values that jointly maximize $L(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. So, we now need to find $\hat{\boldsymbol{\Sigma}}$ to maximize

$$L(\hat{\boldsymbol{\mu}}, \boldsymbol{\Sigma}) \propto \frac{1}{|\boldsymbol{\Sigma}|^{n/2}} e^{-\mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}\left(\sum_{j=1}^{n}(\mathbf{x}_j-\bar{\mathbf{x}})(\mathbf{x}_j-\bar{\mathbf{x}})'\right)\right]/2}$$

We use the following result; see the textbook for its proof. For a $p \times p$ positive definite matrix $\mathbf{B}$ and a scalar $b > 0$,

$$\frac{1}{|\boldsymbol{\Sigma}|^{b}} e^{-\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{B}\right)/2} \leq \frac{1}{|\mathbf{B}|^{b}} (2b)^{pb} e^{-bp}$$

for all $p \times p$ positive definite matrices $\boldsymbol{\Sigma}$, with equality holding only for $\boldsymbol{\Sigma} = (1/2b)\mathbf{B}$. With $b = n/2$ and $\mathbf{B} = \sum_{j=1}^{n}(\mathbf{x}_j - \bar{\mathbf{x}})(\mathbf{x}_j - \bar{\mathbf{x}})'$, we have $\hat{\boldsymbol{\Sigma}} = \mathbf{S}_n$.

Notes:

- The maximized likelihood is proportional to the inverse of the generalized variance

$$L\left(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}\right) = \frac{1}{(2\pi)^{np/2}} e^{-np/2} \frac{1}{|\hat{\boldsymbol{\Sigma}}|^{n/2}}$$
$$\propto (\text{generalized variance})^{-n/2}$$

  The smaller the generalized variance is, the more highly peaked the likelihood is at its center.

- According to the *invariance property* of MLEs, the MLE of $h(\boldsymbol{\theta})$ is $h(\hat{\boldsymbol{\theta}})$, where $\hat{\boldsymbol{\theta}}$ is the MLE of a parameter $\boldsymbol{\theta}$ and $h(\cdot)$ is a function. For example, the MLE of $\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ is $\hat{\boldsymbol{\mu}}\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}$, and the MLE of $\sqrt{\sigma_{ii}}$ is $\sqrt{\hat{\sigma}_{ii}}$.

- Since the joint density depends on the $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ only through the sample mean $\bar{\mathbf{x}}$ and $(n-1)\mathbf{S}$, we say $\bar{\mathbf{x}}$ and $\mathbf{S}$ are *sufficient statistics*. Thus, all of the information about $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in the data matrix $\mathbf{X}$ is contained in $\bar{\mathbf{x}}$ and $\mathbf{S}$.

Wishart Distribution: Let $\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_m \underset{\text{i.i.d.}}{\sim} N_p(\mathbf{0}, \mathbf{\Sigma})$ (the "i.i.d." means independent and identically distributed). Then the distribution of $\tilde{\mathbf{Z}} = \sum_{j=1}^{m} \mathbf{Z}_j \mathbf{Z}_j'$ is Wishart with m degrees of freedom. We say $\tilde{\mathbf{Z}} \sim W_m(\mathbf{\Sigma})$.

A couple of properties:

- If $\mathbf{A}_1 \sim W_{m_1}(\mathbf{\Sigma})$, $\mathbf{A}_2 \sim W_{m_2}(\mathbf{\Sigma})$, and $\mathbf{A}_1$ and $\mathbf{A}_2$ are independent, then $\mathbf{A}_1 + \mathbf{A}_2 \sim W_{m_1+m_2}(\mathbf{\Sigma})$.
- If $\mathbf{A} \sim W_m(\mathbf{\Sigma})$, then $\mathbf{CAC}' \sim W_m(\mathbf{C\Sigma C}')$.

Sampling Distributions: Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be a random sample from the $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution. Then

- $\bar{\mathbf{X}} \sim N_p(\boldsymbol{\mu}, (1/n)\boldsymbol{\Sigma})$.
- $(n-1)\mathbf{S} \sim W_{n-1}(\boldsymbol{\Sigma})$.
- $\bar{\mathbf{X}}$ and $\mathbf{S}$ are independent.

- $\bar{\mathbf{X}} \sim N_p(\boldsymbol{\mu}, (1/n)\boldsymbol{\Sigma})$.

This is analogous to the univariate case, in which we have $\bar{X} \sim N(\mu, \sigma^2/n)$.

- $(n-1)\mathbf{S} \sim W_{n-1}(\mathbf{\Sigma})$.

This is also analogous to the univariate case, in which we have that $(n-1)s^2$ is distributed as the sum of squared independent $N(0, \sigma^2)$ random variables, which is equal to σ^2 times a χ^2_{n-1} random variable.

Large Sample Size: Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be a random sample from a population with mean $\boldsymbol{\mu}$ and nonsingular (invertible) covariance $\boldsymbol{\Sigma}$. Then

$$\sqrt{n}\left(\bar{\mathbf{X}} - \boldsymbol{\mu}\right) \dot{\sim} N_p(\mathbf{0}, \boldsymbol{\Sigma})$$

and

$$n\left(\bar{\mathbf{X}} - \boldsymbol{\mu}\right)' \mathbf{S}^{-1} \left(\bar{\mathbf{X}} - \boldsymbol{\mu}\right) \dot{\sim} \chi^2_p$$

for $n - p$ large.

The preceding large-sample approximations follow from applications of the *law of large numbers* and the *central limit theorem*.

- The law of large numbers states that certain statistics get arbitrarily close to their corresponding parameters as $n \to \infty$. As an example, if $X_1, X_2, \ldots, X_n$ are independent with mean μ, then, for any $\epsilon > 0$,

$$P(\epsilon < \bar{X} - \mu < \epsilon) \to 1 \text{ as } n \to \infty$$

  and we say $\bar{X}$ converges in probability to μ. In the multivariate case, we have that $\bar{\mathbf{X}}$ converges in probability to $\boldsymbol{\mu}$ and $\mathbf{S}$ converges in probability to $\boldsymbol{\Sigma}$.

- The central limit theorem states large sample conditions under which the distribution of the sample mean is approximately normal, even if the population distribution is not normal. Specifically, let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be a random sample from a population with mean $\boldsymbol{\mu}$ and nonsingular covariance $\boldsymbol{\Sigma}$. Then

$$\sqrt{n}\left(\bar{\mathbf{X}} - \boldsymbol{\mu}\right) \dot{\sim} N_p(\mathbf{0}, \boldsymbol{\Sigma})$$

  for large n (relative to p).

Assessing Normality

Q-Q Plots: One way to visually assess whether data came from a *univariate* normal distribution is with a *quantile-quantile plot* (Q-Q plot), in which sample quantiles are plotted against their corresponding quantiles from the standard normal distribution. Suppose, for example, that $X_1, X_2, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$. Then, since the αth quantile of $N(\mu, \sigma^2)$ equals $\sigma q_\alpha + \mu$, where q_α is the αth quantile of $N(0, 1)$, the quantiles of $N(\mu, \sigma^2)$ and $N(0, 1)$ lie on a straight line.

To make a Q-Q plot with an observed sample $x_1, x_2, \ldots, x_n$:

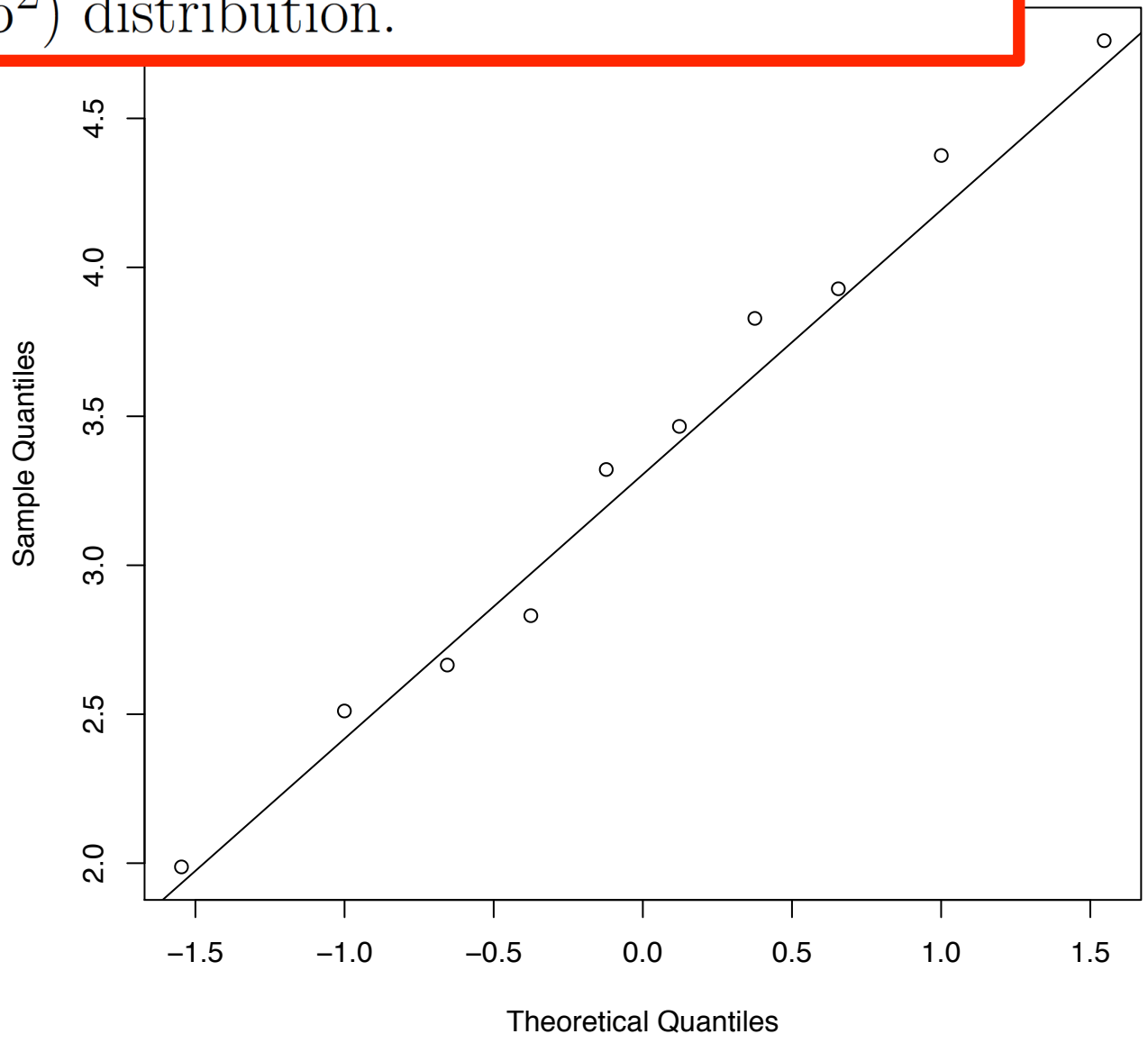
1. Let $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ be the values ordered from least to greatest. So, a proportion j/n of the sample is less than or equal to $x_{(j)}$.

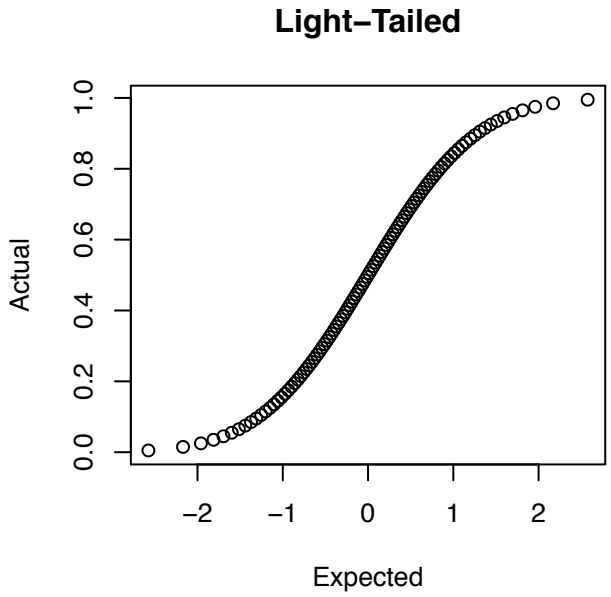
2. For comparison with the $x_{(j)}$, compute the standard normal quantiles $q_{(j)}$. Instead of the (j/n)th quantile, it is standard to use the $((j - 1/2)/n)$th quantile as a "continuity correction."

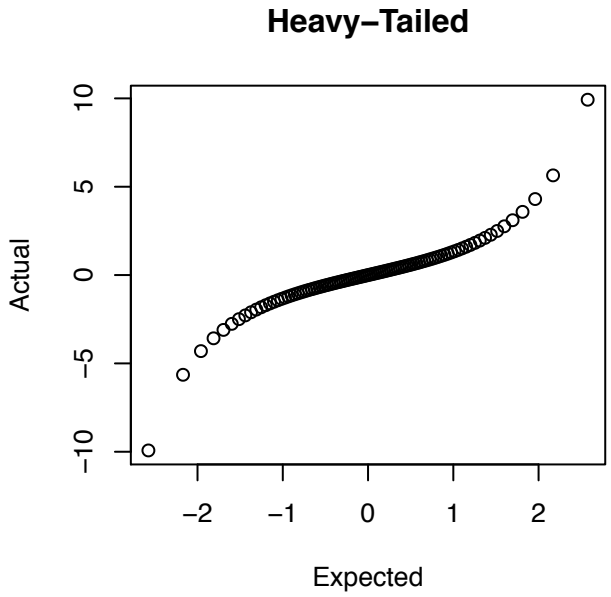
3. Plot the $q_{(j)}$ on the horizontal axis and the $x_{(j)}$ on the vertical axis.

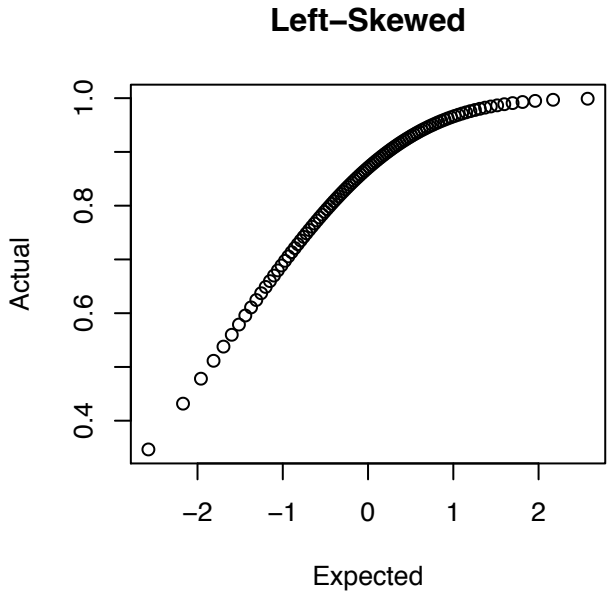
If the points roughly follow a line, the normality assumption is supported.

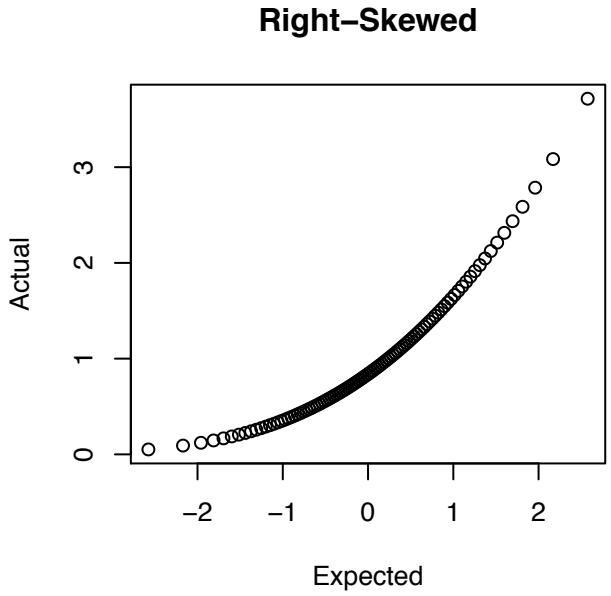
This is for a sample of size 10 drawn from the $N(3, 1.5^2)$ distribution.





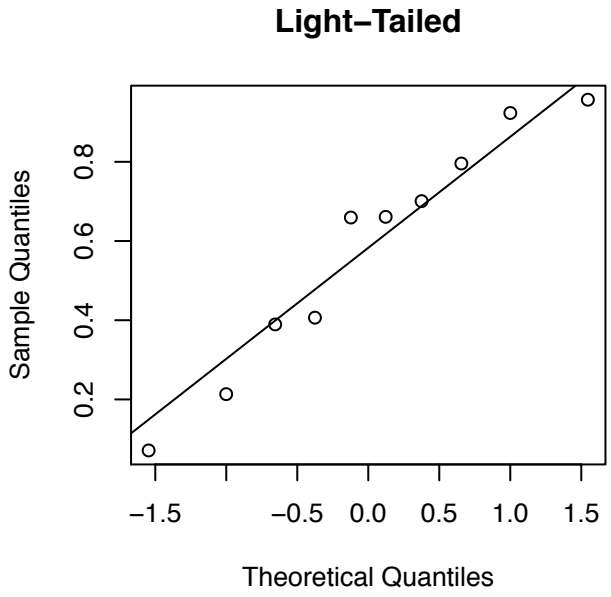


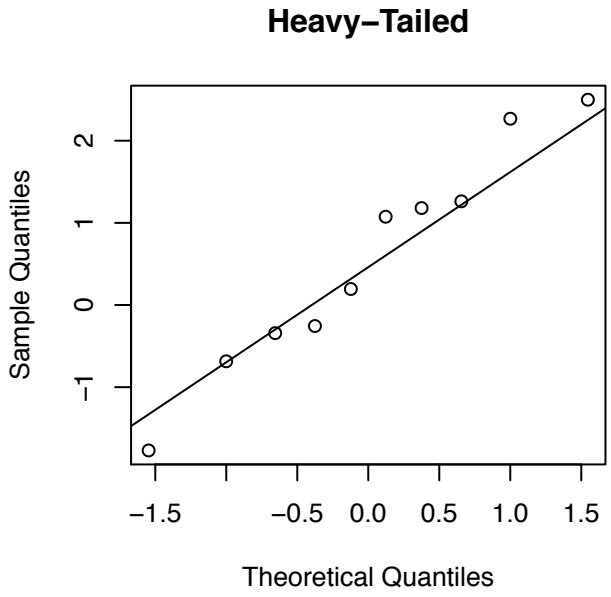




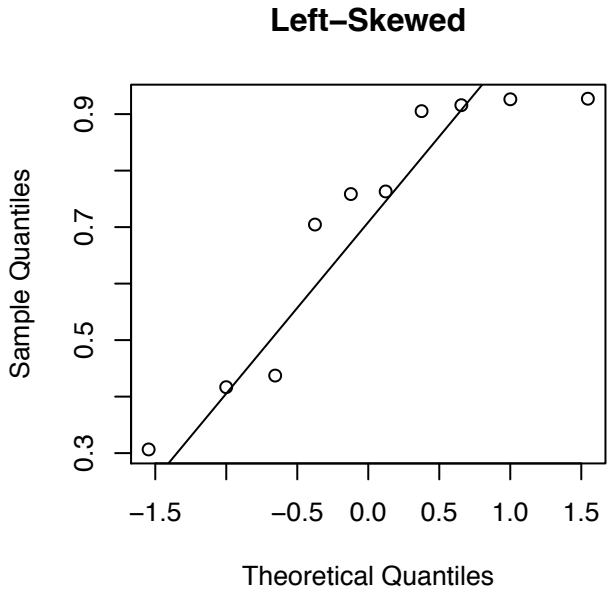
Data from non-normal distributions will show nonlinear patterns. The pattern will depend on the shape of the true distribution.

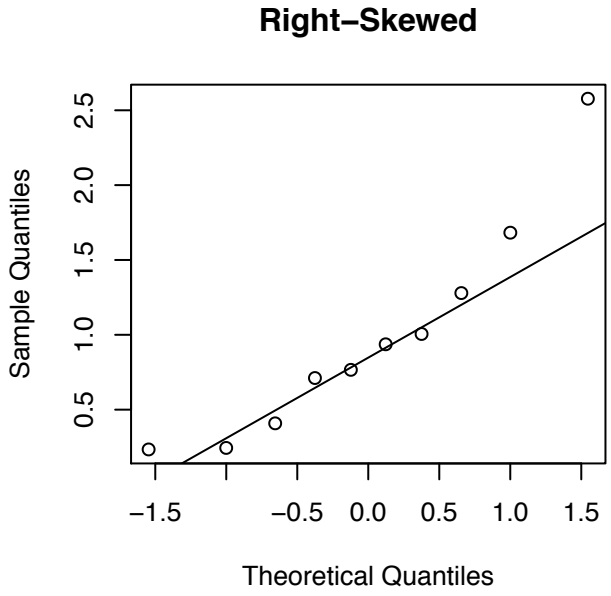
These four examples show theoretical quantiles. In practice (next slide), we only have sample quantiles. And the plots can be hard to interpret.





These are Q-Q plots made with samples of size 10 from the same distributions from the previous slide. The patterns are difficult to see.

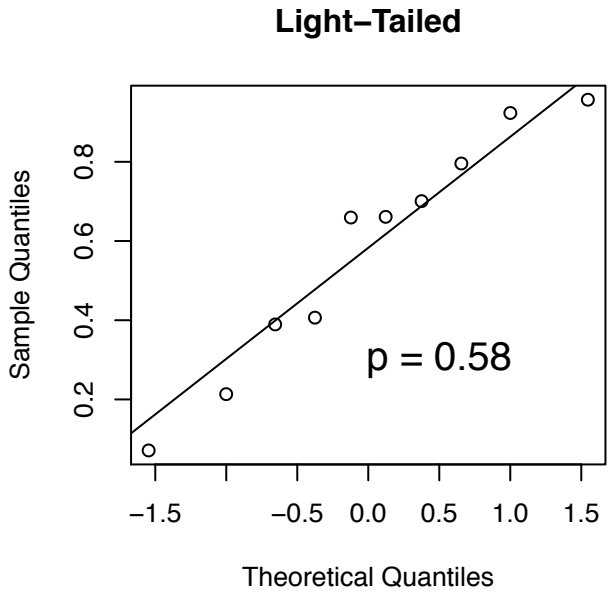


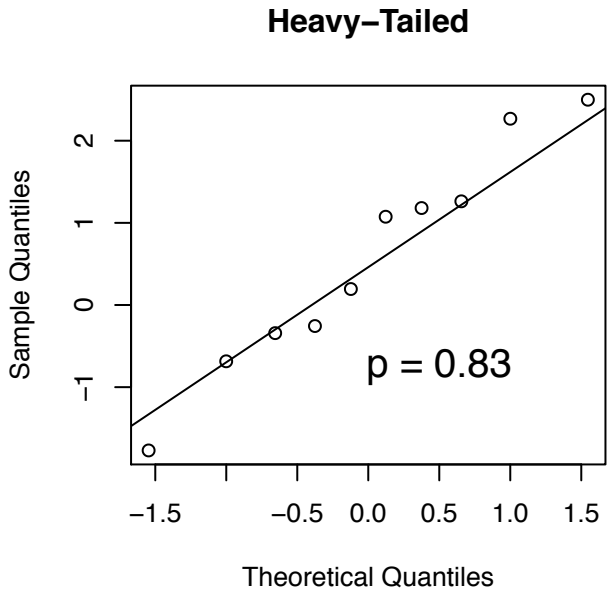


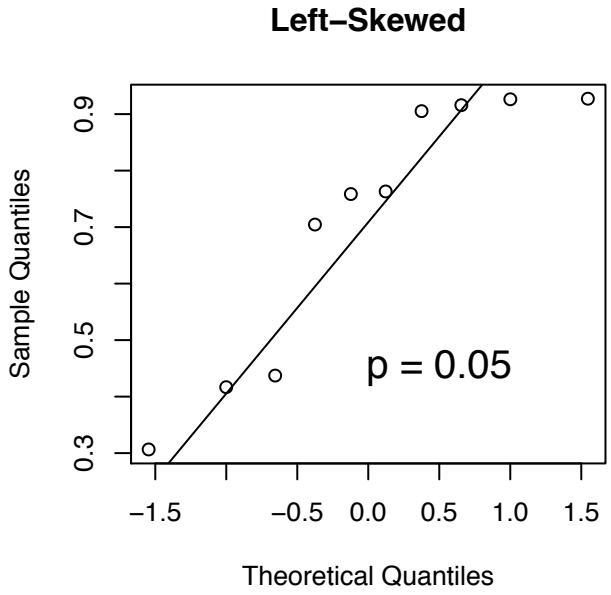
Univariate Tests: In addition to the use of Q-Q plots for visually assessing univariate normality, there are several formal tests.

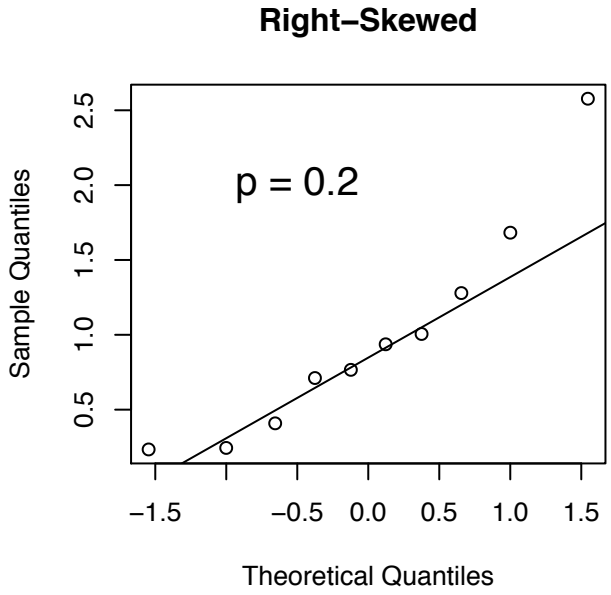
- *Shapiro-Wilk*: The test statistic is like a correlation between the sample and theoretical quantiles. The textbook refers to an alternative, which uses the usual correlation coefficient. See `shapiro.test` in `R`.
- *Anderson-Darling*: The test statistic is an integrated, weighted, squared difference between the sample ("empirical") and theoretical cumulative distribution functions (cdf's). See `ad.test` in `R` package `nortest`.
- *Kolmogorov-Smirnoff*: The test statistic is the largest difference between the sample and theoretical cdf's. See `ks.test` in `R`.

Shapiro-Wilk is probably the most widely-used test.

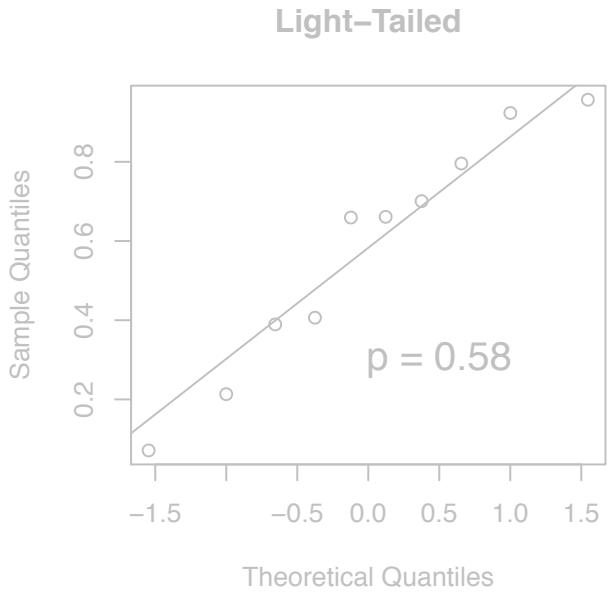


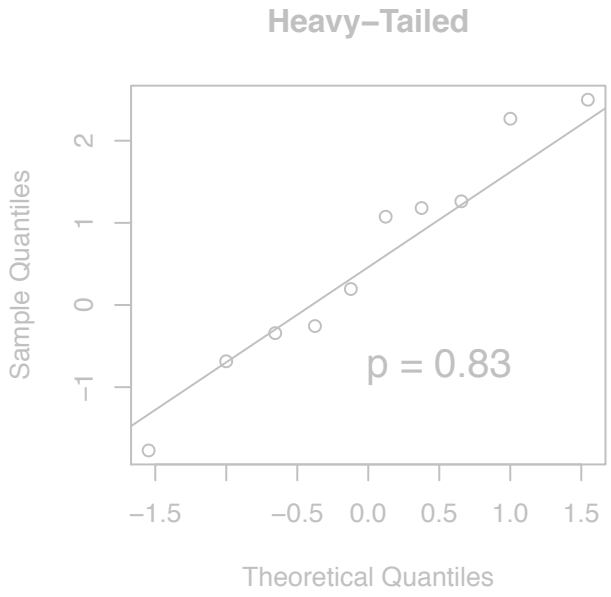




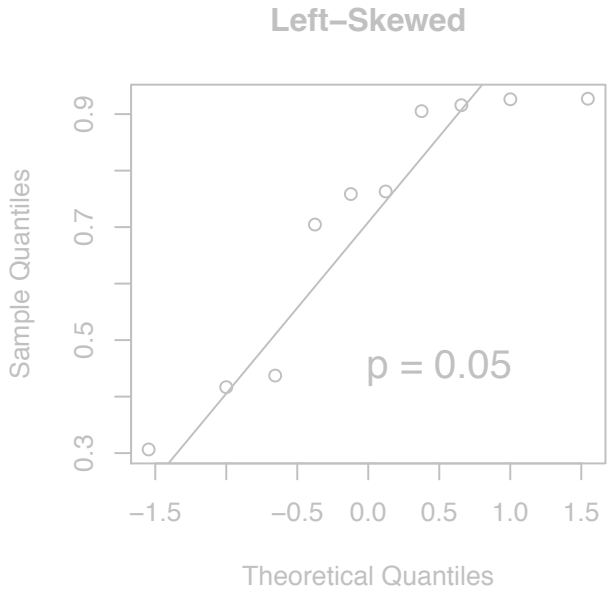


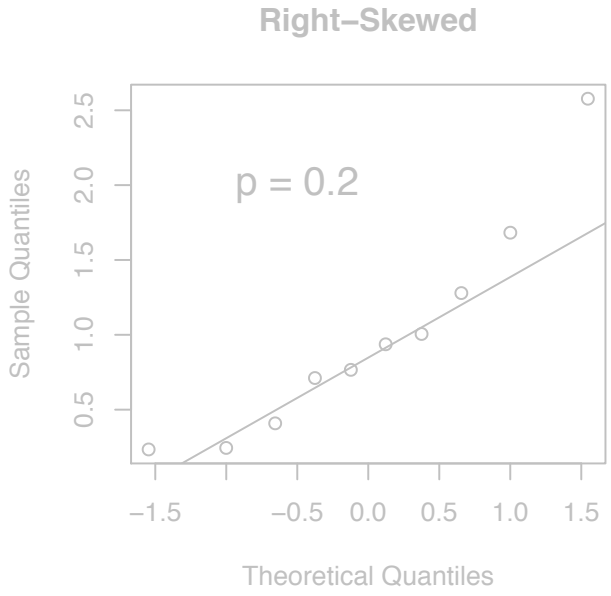
These are Shapiro Wilks p-values.





These are Shapiro Wilks p-values.





Assessing Multivariate Normality: The Q-Q plots and tests just considered are appropriate for assessing univariate normality. Assessing multivariate normality is more challenging. We present two techniques below. Assume that $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n \underset{\text{i.i.d.}}{\sim} N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and let $d_j^2 = (\mathbf{x}_j - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x}_j - \bar{\mathbf{x}})$, $j = 1, 2, \ldots, n$.

- Because the set of $\mathbf{x}$ for which

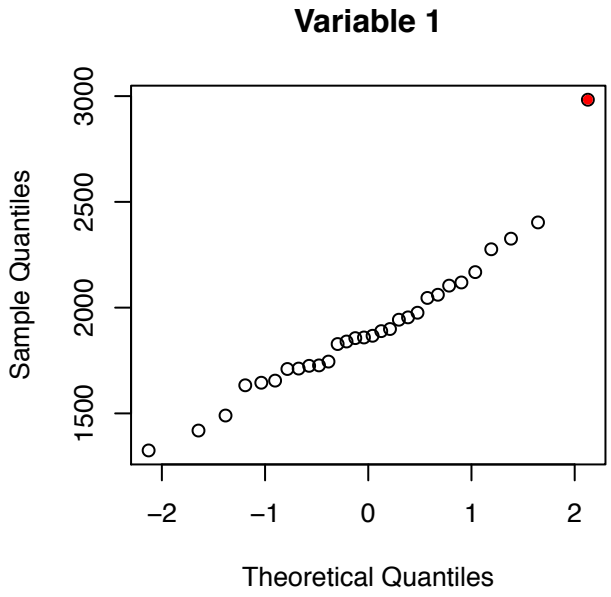
$$(\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \leq \chi_p^2(0.5)$$

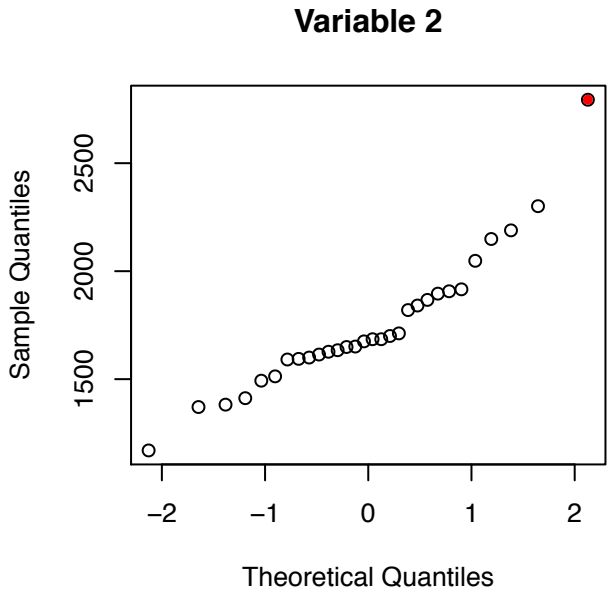
has probability 0.5, we expect *approximately* 50% of the d_j^2 to be $\leq \chi_p^2(0.5)$. If there is gross deviation from this expectation, we have evidence against multivariate normality.

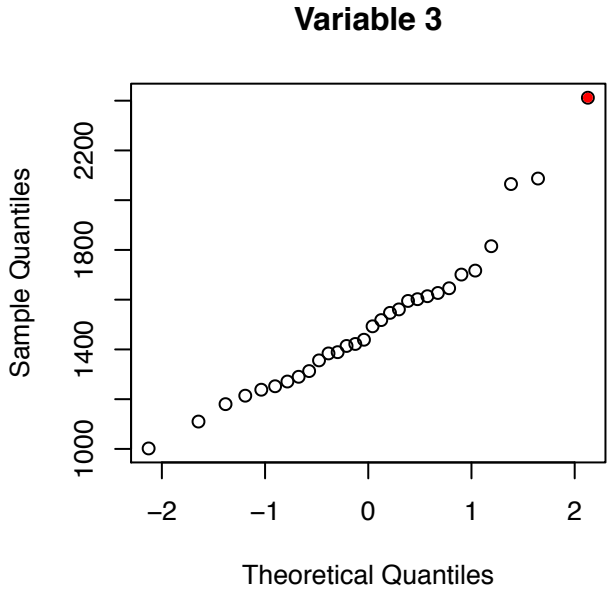
- If n and $n - p$ are both greater than about 25, then each of the d_j^2 are approximately χ_p^2 random variables. While they are not independent, we can still use them to make an approximate Q-Q plot. To do this, we plot the ordered d_j^2 against the corresponding quantiles of the χ_p^2 distribution. Because we compute quantiles from the χ_p^2 distribution itself, we expect the points to follow the line of equality if normality holds.

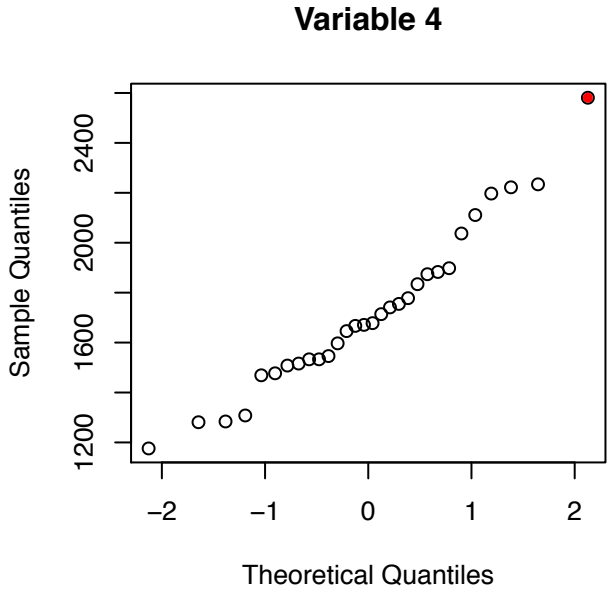
Example: For each of $n = 30$ boards, we have four measures of stiffness: (1) while being hit with a shock wave, (2) during vibration, and (3) - (4) are static tests. A subset of the data is shown below; the full data set is in Table 4.3 of the textbook.

board	x_1	x_2	x_3	x_4
1	1889	1651	1561	1778
2	2403	2048	2087	2197
3	2119	1700	1815	2222
4	1645	1627	1110	1533
5	1976	1916	1614	1883
6	1712	1712	1439	1546
7	1943	1685	1271	1671
8	2104	1820	1717	1874
9	2983	2794	2412	2581
10	1745	1600	1384	1508
⋮	⋮	⋮	⋮	⋮
30	1490	1382	1214	1284





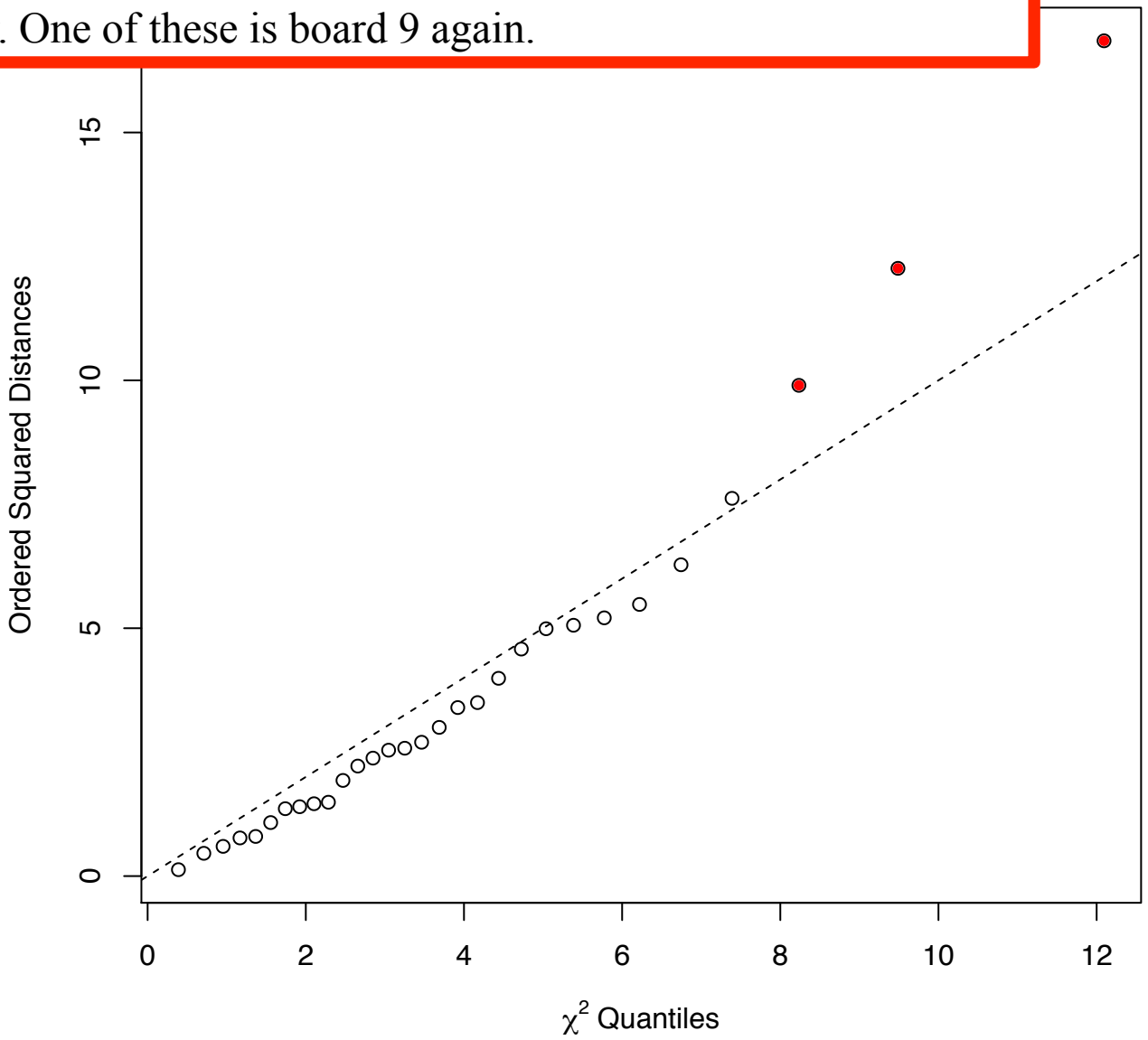




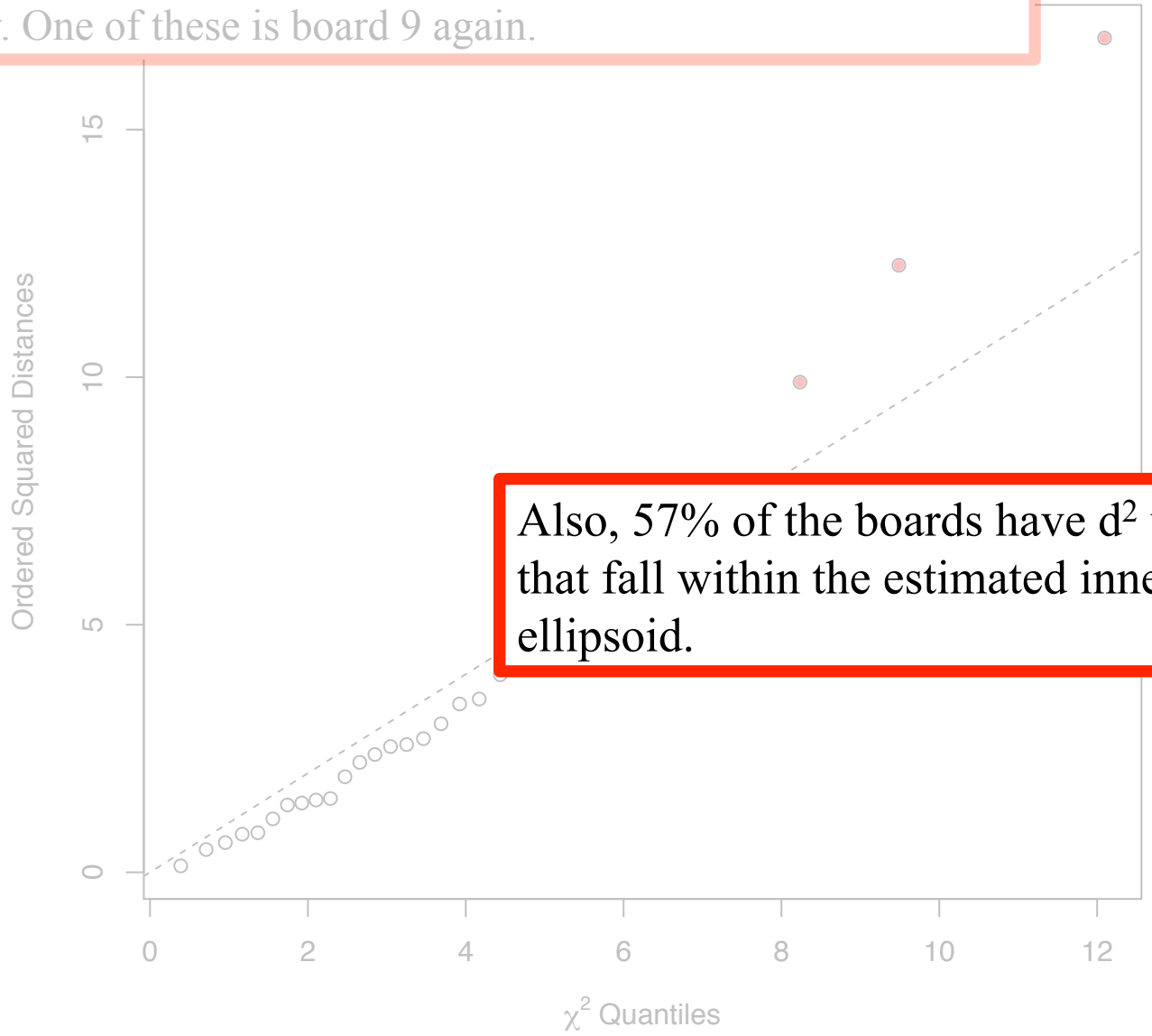
These are univariate Q-Q plots for the four variables. Normality appears reasonable, with the only concern being due to the red point, from board 9.

Remember that these pictures only speak to univariate normality, not to joint normality.

Chi-square plot for multivariate normality. The three boards with the largest statistical distance pull the curve away from the line of equality. One of these is board 9 again.

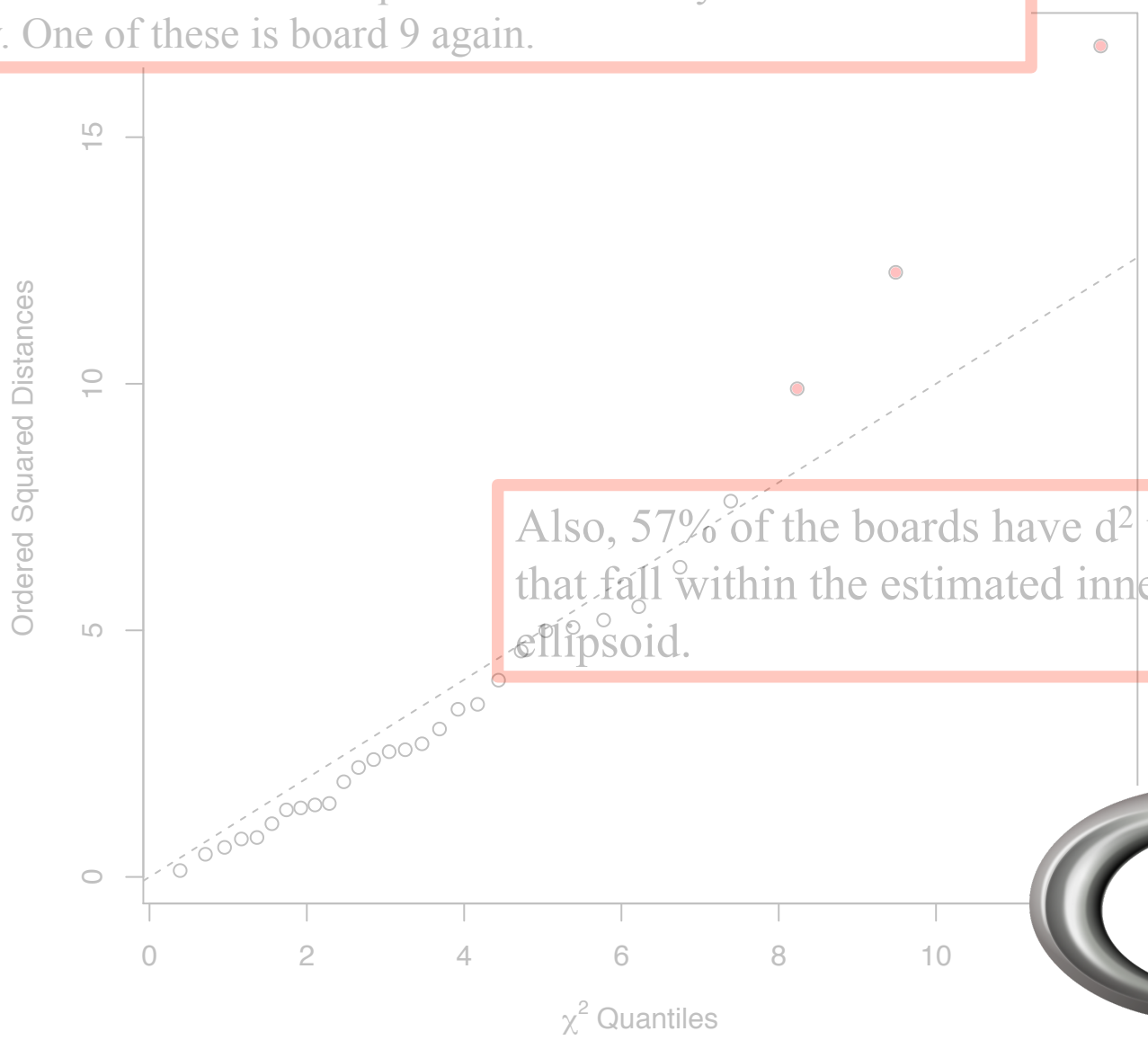


Chi-square plot for multivariate normality. The three boards with the largest statistical distance pull the curve away from the line of equality. One of these is board 9 again.



Also, 57% of the boards have d^2 values that fall within the estimated inner 50% ellipsoid.

Chi-square plot for multivariate normality. The three boards with the largest statistical distance pull the curve away from the line of equality. One of these is board 9 again.



Also, 57% of the boards have d^2 values that fall within the estimated inner 50% ellipsoid.

Outliers

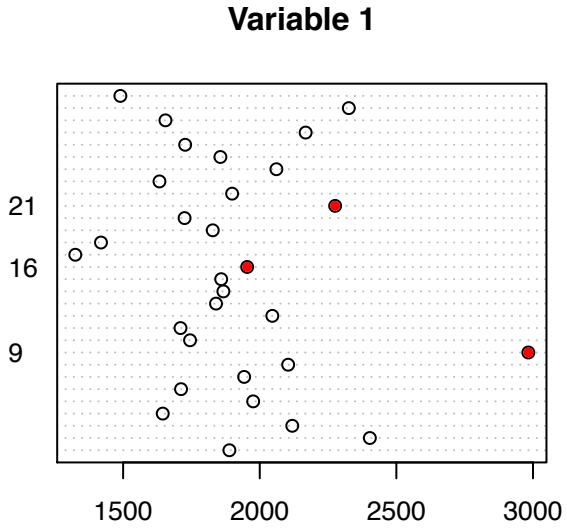
Outliers: An *outlier* is an observation that is far removed from the majority of the other observations. Determining how far removed an observation should be before being called an outlier is subjective. And the fact that an observation is an outlier does not necessarily mean it should be removed from the analysis.

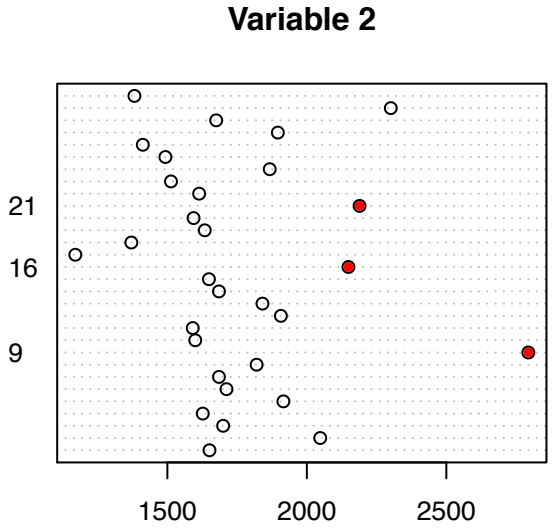
Strategies for detecting outliers:

- Make dot plots for individual variables and scatterplots for pairs of variables. Look for unusual points.
- Calculate the standardized values $z_{jk} = (x_{jk} - \bar{x}_k)/\sqrt{s_{kk}}$ for $j = 1, 2, \ldots, n$ and each column $k = 1, 2, \ldots, p$. Look for values that are more extreme than would be expected by chance.
- Calculate the squared statistical distances $d_j^2 = (\mathbf{x}_j - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_j - \bar{\mathbf{x}})$ and make a chi-square plot. Look for points that are unusually far from the origin.

Example Revisited: In the stiffness data, we already saw that boards 9, 16, and 21 looked unusual in the chi-square plot. This means that they may be outliers in a multivariate sense.

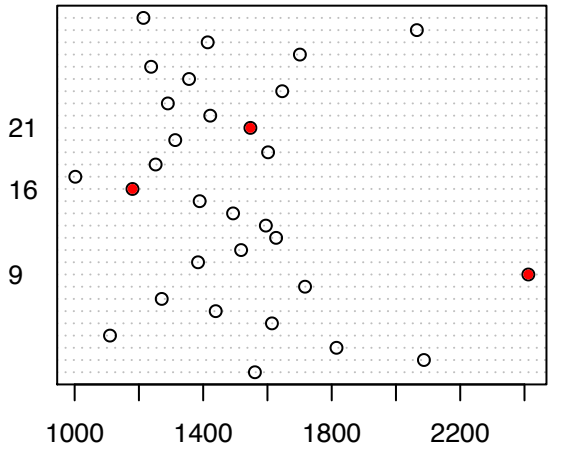
Considering each variable individually, we can examine the standardized values z_{jk}. With $n \times p = 120$ total standardized variables, assuming they are approximately normally distributed, we don't expect any values greater than about 2.6 in absolute value (because 2.6 is the 0.00833 / 2 quantile of the standard normal, and $0.00833 \times 120 = 1$). Rounding 2.6 up to 3 to reach a nice whole number, board 9 is the only one that is highlighted.

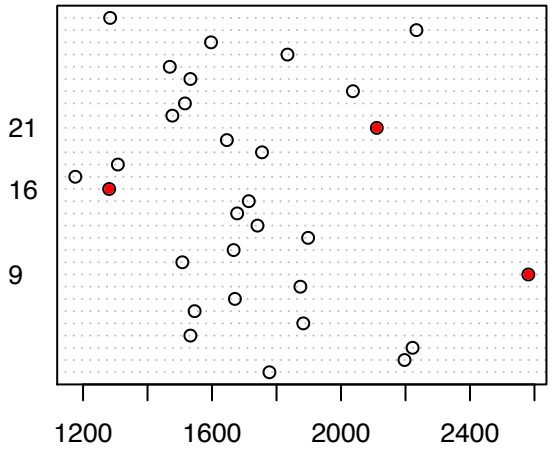


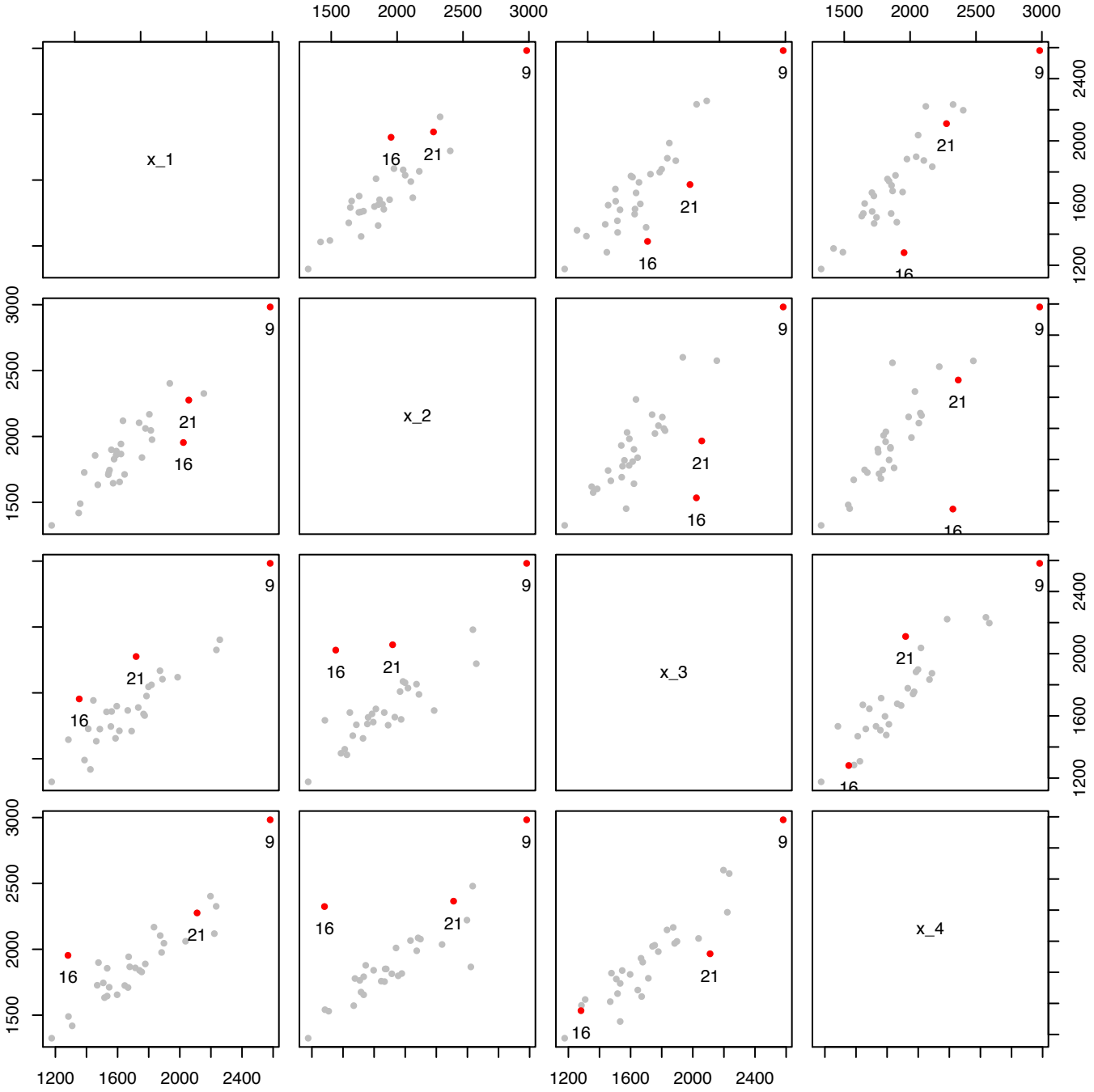


These are univariate dot plots. The values for boards 9, 16, and 21 are highlighted. These boards looked unusual in the chi-square plot from a few slides back.

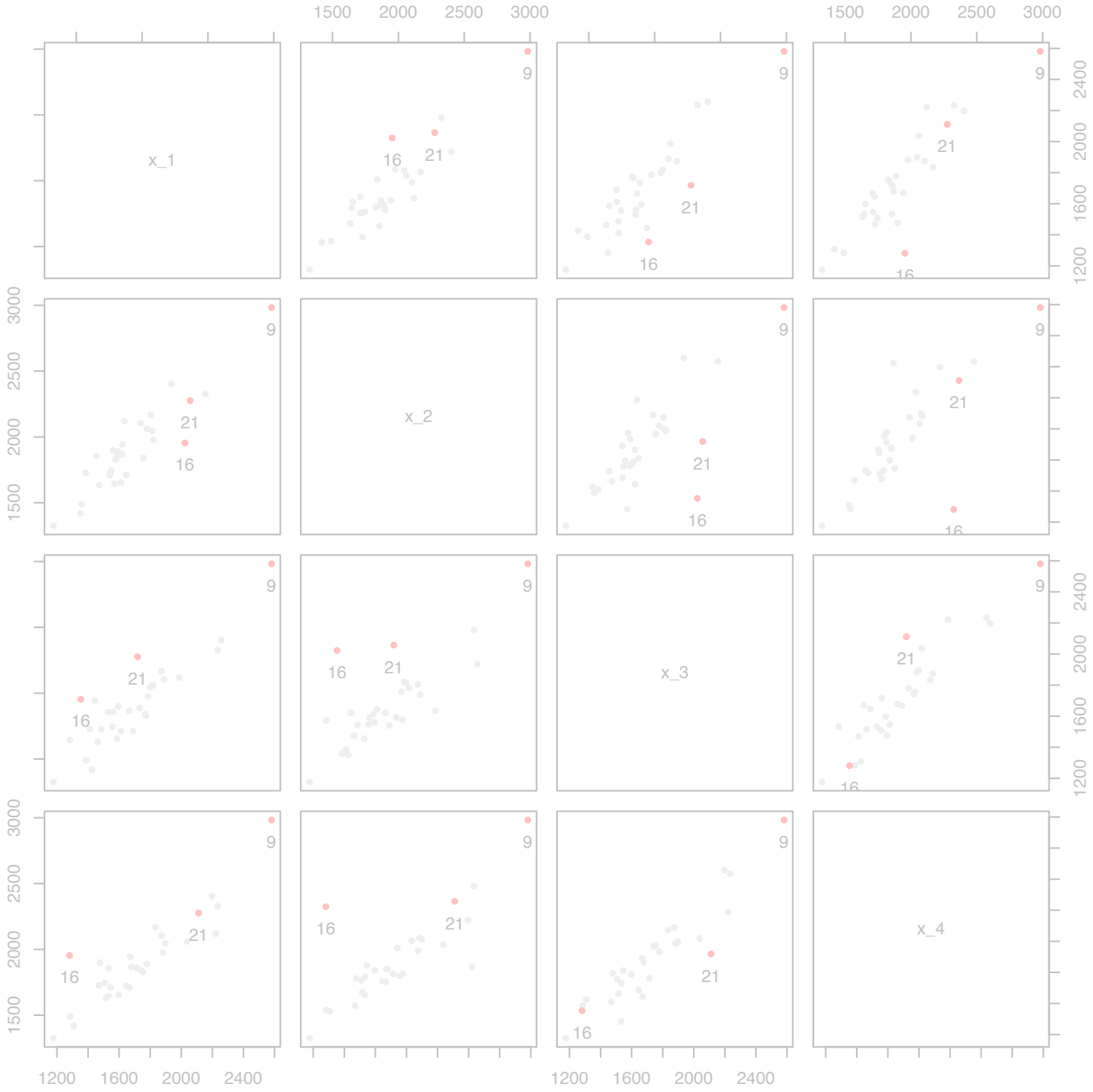
Considered one-by-one, only board 9 appears unusual.







Pairwise scatterplots. Boards 9 is always high for both variables. Board 16 stands out in some pictures but not all. Board 21 is mostly inside all of the scatters.



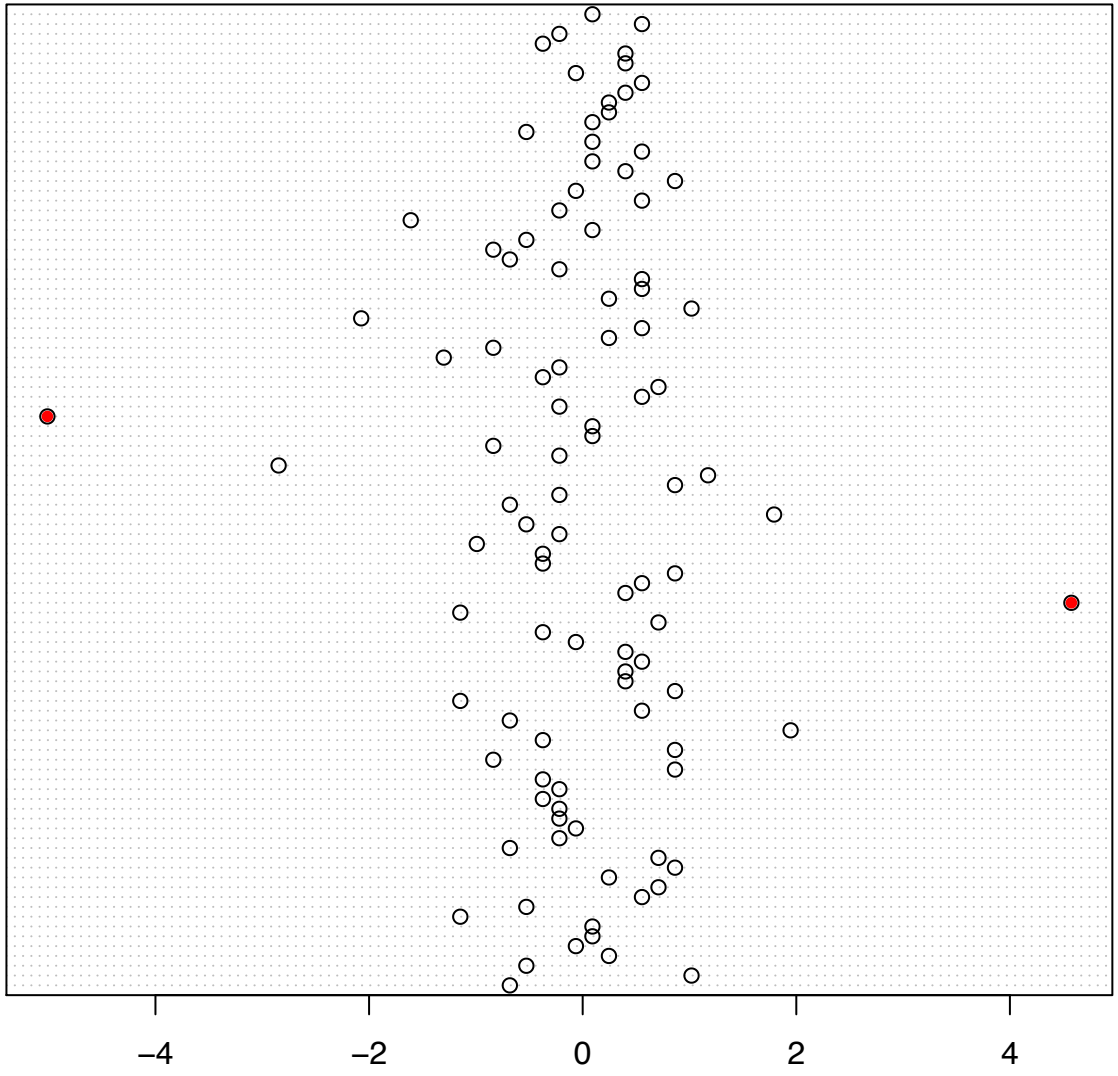
Pairwise scatterplots. Boards 9 is always high for both variables. Board 16 stands out in some pictures but not all. Board 21 is mostly inside all of the scatters.

Example: We have 100 measurements of weight in micrograms of 100 samples of a 10-gram chrome steel standard known as NB10. Each of the weights is actually of the form 9,999,xxx, so our numbers are just the last three from the actual microgram weights.

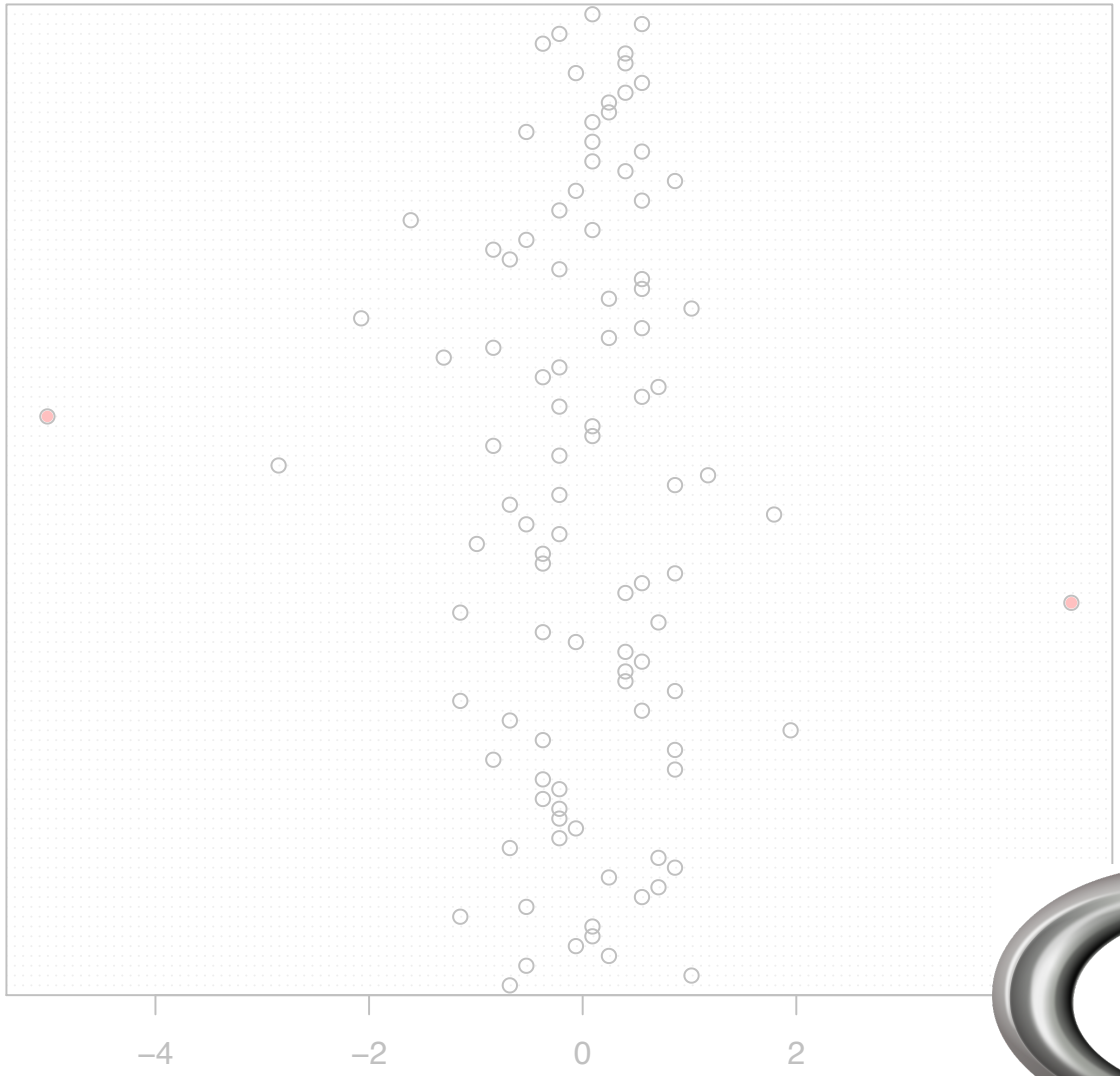
Two of the standardized weights exceed 3 in absolute value and therefore appear to be outliers. Should they be removed from the analysis? It turns out in this example that they should not be removed. The point of collecting these data was to learn about the weighing process. The outliers indicate that there can be upsets in this process. In other words, the weights are not *homogenous*.

Data obtained from here.

Dot Plot of Standardized Weights



Dot Plot of Standardized Weights



Data Transformations

When our data do not appear to have come from a normal distribution, we might try transforming the data. A useful family of transformations for achieving near normality is the family of *power transformations*, which replace an observation x with x^λ for some choice of λ. If $\lambda < 1$, we shrink large values of x. If $\lambda > 1$, we increase large values of x. And $\lambda = 0$ is defined to correspond to the log transformation.

A few notes:

- Power transformations are only appropriate for positive variables, but we can always add a constant to our measurements to make them all positive.
- Transforming a variable impacts its interpretation. If two transformations are approximately the same in terms of their ability to make the distribution normal, choose the one with greater interpretability.
- The "best" transformation is not guaranteed to result in normality. Always check normality after transformation.

Box-Cox: Box and Cox proposed the slightly modified family of power transformations

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

which is continuous in λ. Then the Box-Cox solution for the choice of λ is the value that maximizes the expression

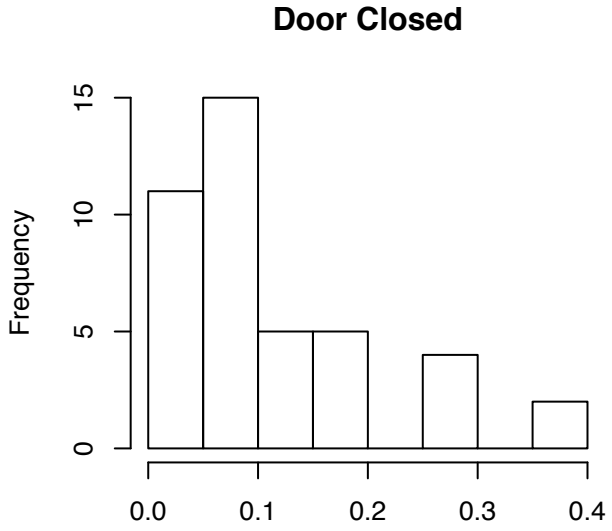
$$l(\lambda) = -\frac{n}{2}\ln\left[\frac{1}{n}\sum_{j=1}^{n}\left(x_j^{(\lambda)} - \overline{x^{(\lambda)}}\right)^2\right] + (\lambda - 1)\sum_{j=1}^{n}\ln x_j$$

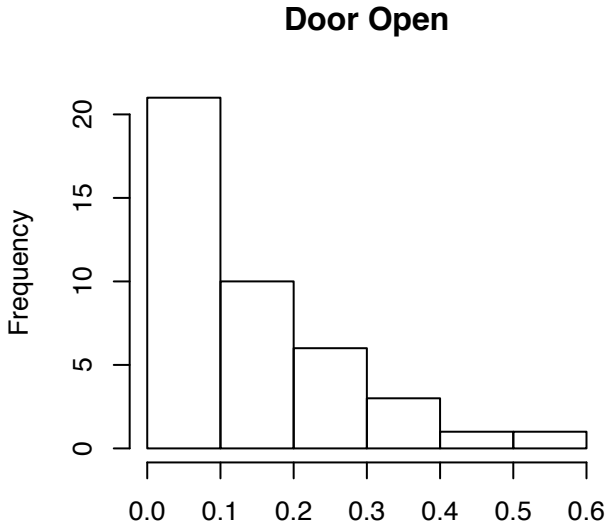
We choose λ numerically, evaluating $l(\lambda)$ over a range of possibilities.

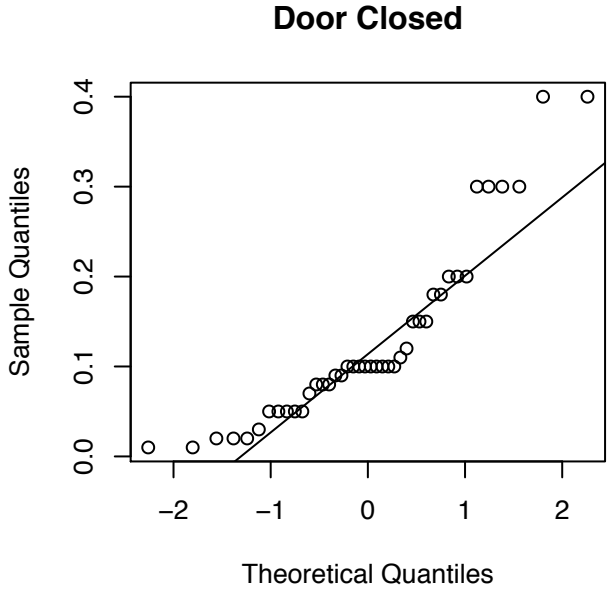
Example: For each of 42 microwave ovens, we have measurements of radiation emitted with the door both closed and open. Neither of the variables appear to be normally distributed individually (so they can't be jointly normally distributed), but separate Box-Cox transformations work well.

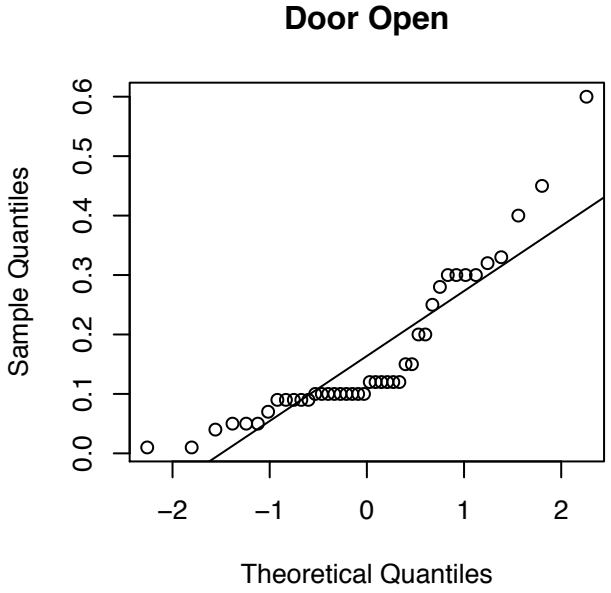
Oven	Closed	Open
1	0.15	0.30
2	0.09	0.09
3	0.18	0.30
4	0.10	0.10
5	0.05	0.10
⋮	⋮	⋮
42	0.05	0.12

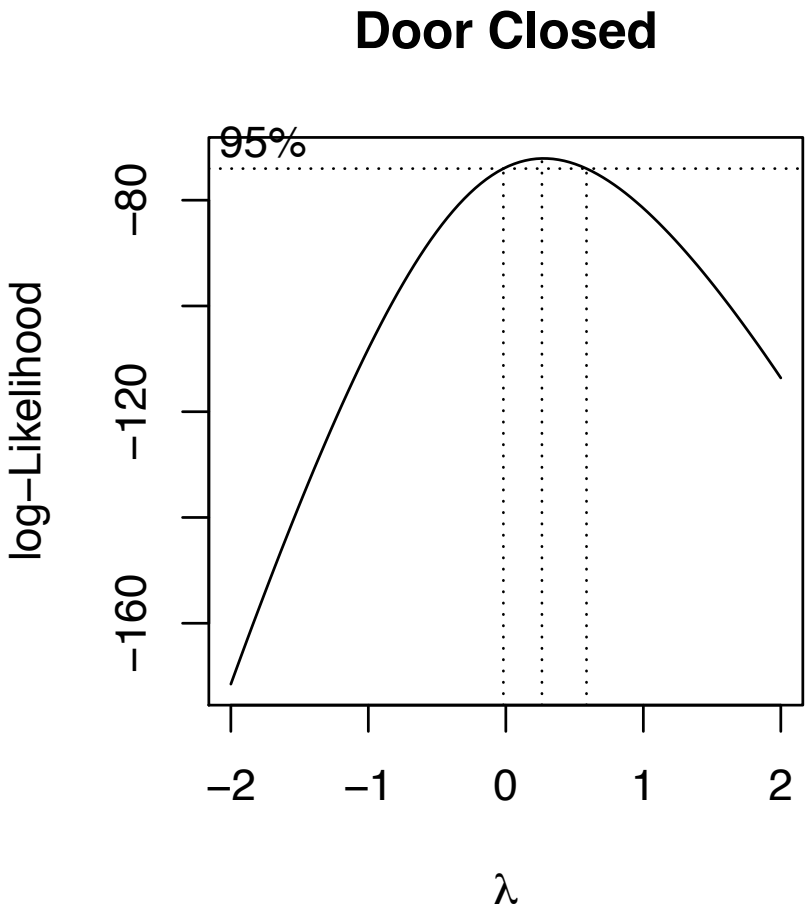
Data from Tables 4.1 and 4.5 in textbook.

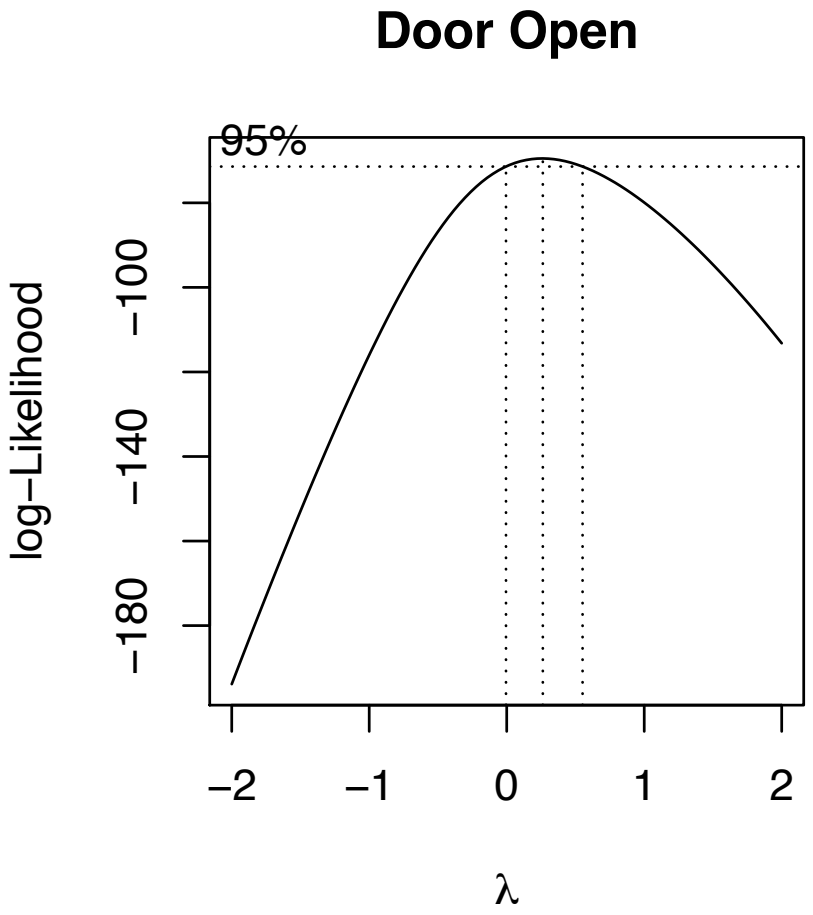

Door Closed
Frequency
0
5
10
15
0.0
0.1
0.2
0.3
0.4


Door Open
Frequency
0
5
10
15
20
0.0
0.1
0.2
0.3
0.4
0.5
0.6

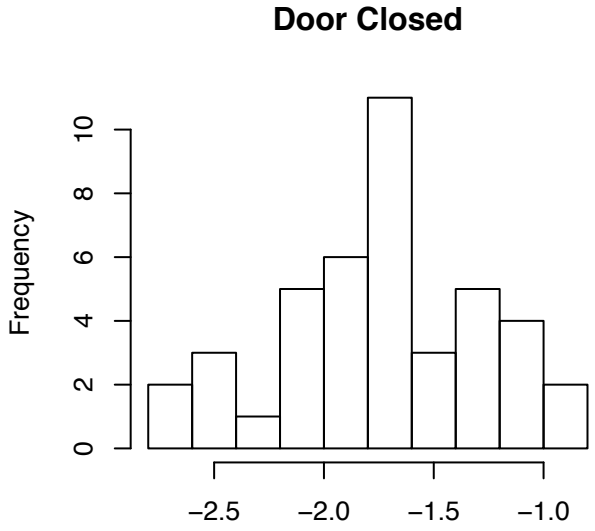

Door Closed
Sample Quantiles
0.0
0.1
0.2
0.3
0.4
−2
−1
0
1
2
Theoretical Quantiles

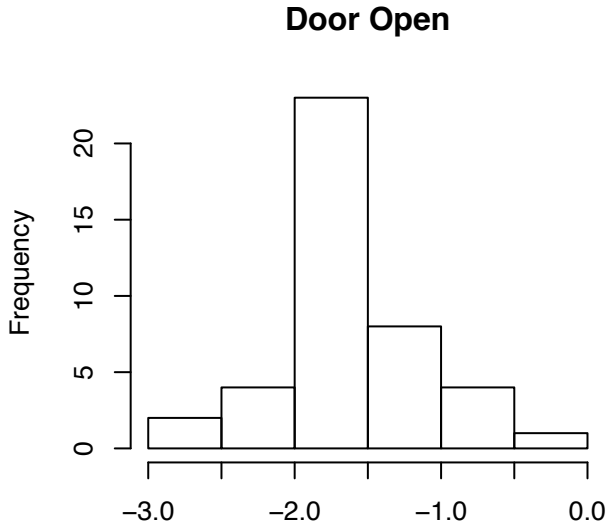

Door Open
Sample Quantiles
0.0
0.1
0.2
0.3
0.4
0.5
0.6
−2
−1
0
1
2
Theoretical Quantiles

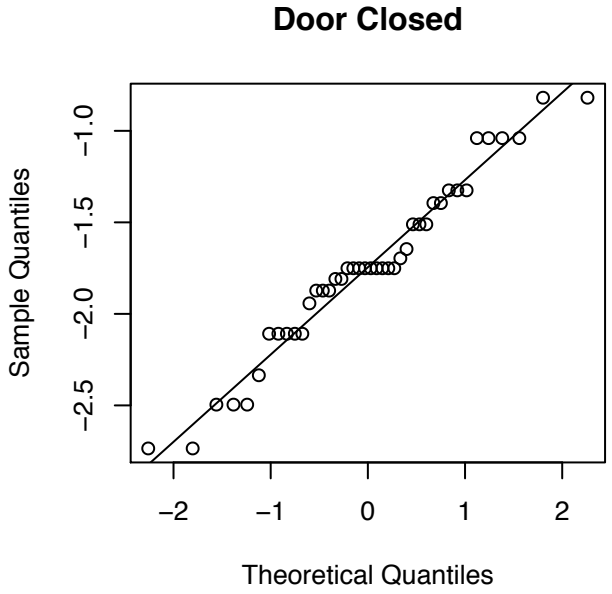


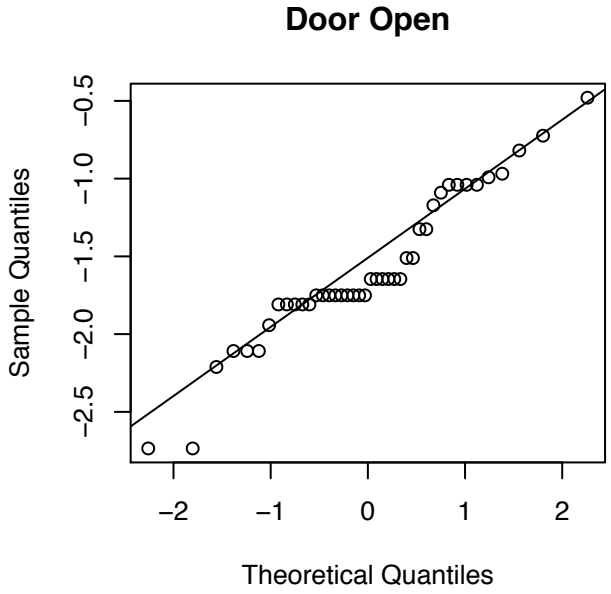


The optimal λ is about 0.25 for both variables.


Door Closed
Frequency
0
2
4
6
8
10
−2.5
−2.0
−1.5
−1.0


Door Open
Frequency
0
5
10
15
20
−3.0
−2.0
−1.0
0.0


Door Closed
Sample Quantiles
−2.5
−2.0
−1.5
−1.0
−2
−1
0
1
2
Theoretical Quantiles


Door Open
Sample Quantiles
−2.5
−2.0
−1.5
−1.0
−0.5
−2
−1
0
1
2
Theoretical Quantiles

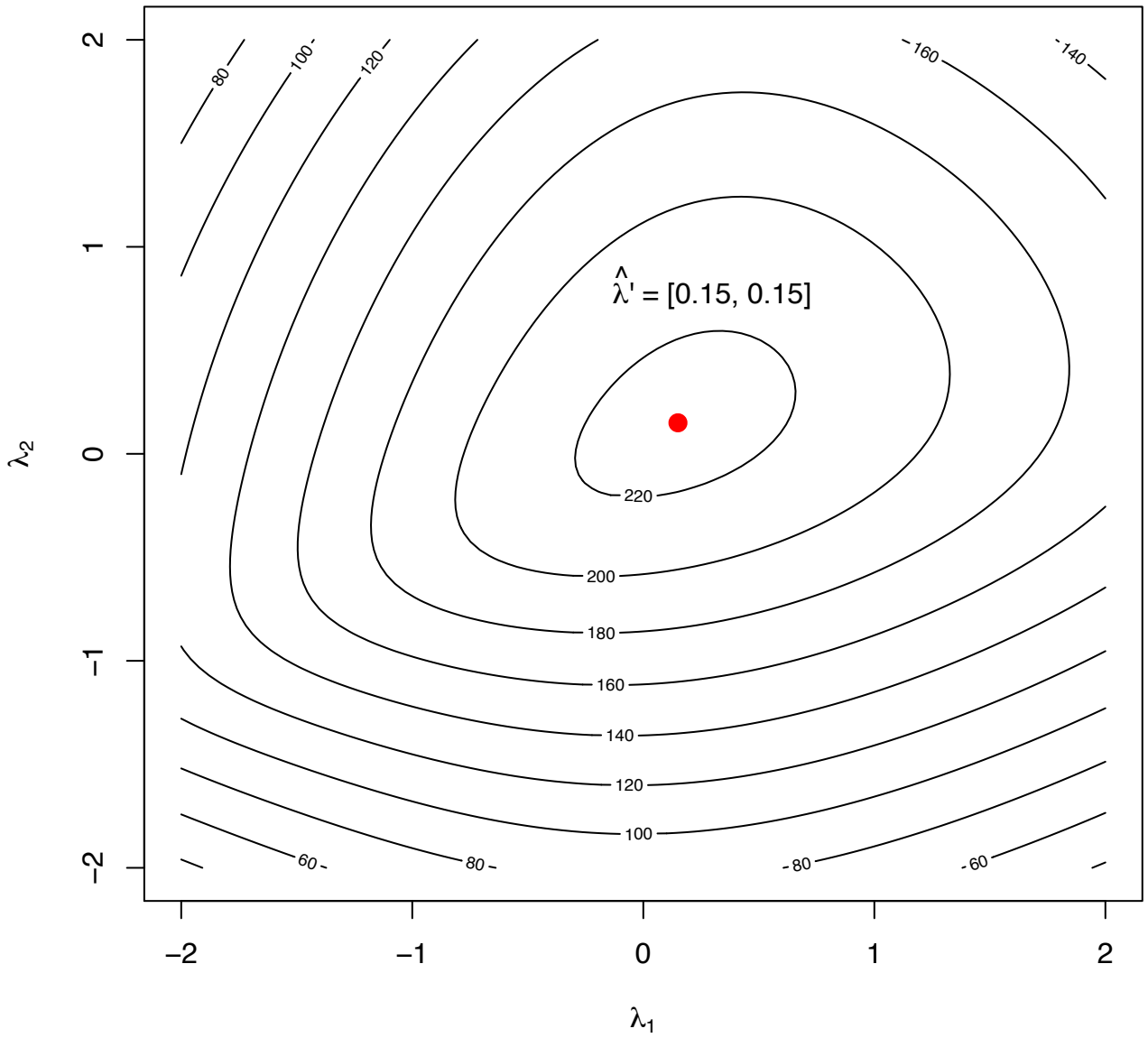
Multivariate Box-Cox: The Box-Cox procedure we just used is for univariate transformations. Since marginal normality doesn't guarantee joint normality, we might consider a multivariate extension of Box-Cox. We can define the new objective function

$$l(\boldsymbol{\lambda}) = -\frac{n}{2}\ln|\mathbf{S}(\boldsymbol{\lambda})| + \sum_{k=1}^{p}\left[(\lambda_k - 1)\sum_{j=1}^{n}\ln x_{jk}\right]$$

where $\mathbf{S}(\boldsymbol{\lambda})$ is the sample covariance matrix of the

$$\left(\mathbf{x}_j^{(\boldsymbol{\lambda})}\right)' = \left[x_{j1}^{(\lambda_1)}, x_{j2}^{(\lambda_2)}, \ldots, x_{jp}^{(\lambda_p)}\right]$$

and search for the *vector* of values $\boldsymbol{\lambda}$ that jointly maximize this function. This is obviously a more computationally intensive problem, and the textbook claims that the transformations suggested by the joint procedure often don't differ substantially from those suggested by individual applications of the univariate procedure.



The univariate and multivariate Box-Cox procedures arrive at similar transformations. We had $\lambda_1 = \lambda_2 = 0.25$ in the univariate case. Here, we have $\lambda_1 = \lambda_2 = 0.15$.